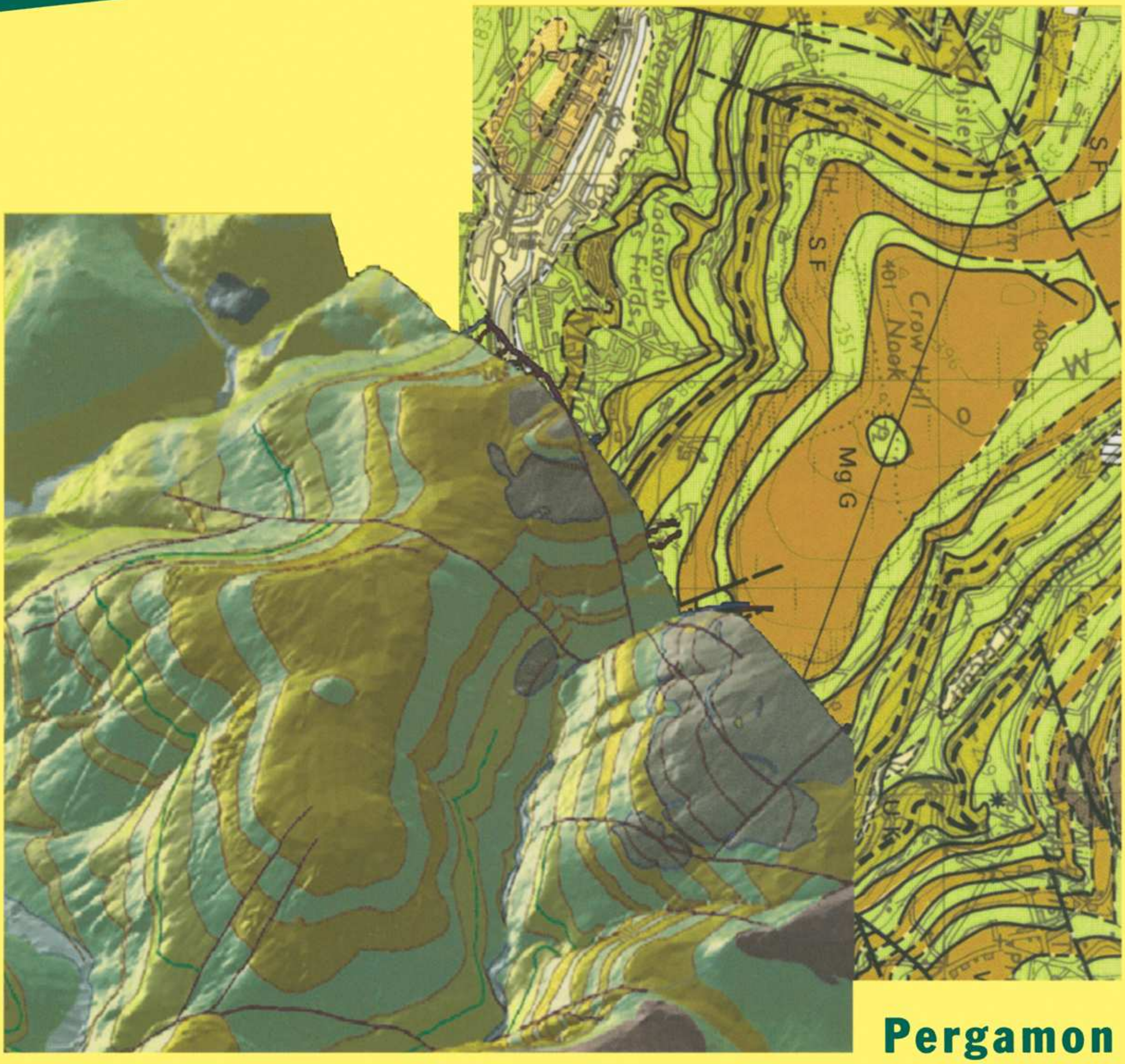
COMPUTER METHODS
IN THE GEOSCIENCES
VOLUME 17

T.V. Loudon

Geoscience after IT

A view of the present and future impact of information technology on geoscience



Pergamon

Geoscience after IT

A view of the present and future impact of information technology on geoscience

COMPUTER METHODS IN THE GEOSCIENCES

Daniel F. Merriam, Series Editor

Volumes in the Series published by Elsevier Science Ltd

Geological Problem Solving with Lotus 1–2–3 for Exploration and Mining Geology:
G.S. Koch Jr. (with program on diskette)
Exploration with a Computer: Geoscience Data Analysis Applications:
W.R. Green
Contouring: A Guide to the Analysis and Display of Spatial Data:
D.F. Watson (with program on diskette)
Management of Geological Data Bases:
J. Frizado (Editor)
Simulating Nearshore Environments:
P.A. Martinez and J.W. Harbaugh
Geographic Information Systems for Geoscientists: Modelling with GIS:
G.F. Bonham–Carter
Computing Risk for Oil Prospects: Principles and Programs:
J.W. Harbaugh, J.C. Davis and J. Wendebourg (with two diskettes)
Structural Geology and Personal Computers:
D. DePaor (Editor)
Simulating Oil Entrapment in Clastic Sequences:
J. Wendebourg and J.W. Harbaugh

**Volumes published by Van Nostrand Reinhold Co. Inc.:*

Computer Applications in Petroleum Geology: J.E. Robinson
Graphic Display in Two- and Three-Dimensional Markov
Computer Models in Geology: C. Lin and J.W. Harbaugh
Image Processing of Geological Data: A.G. Fabbri
Contouring Geologic Surfaces with a Computer: T.A. Jones, D.E. Hamilton, and C.R. Johnson
Exploration–Geochemical Data Analysis with the IBM PC: G.S. Koch Jr. (with program on diskette)
Geostatistics and Petroleum Geology: M.E. Hohn
Simulating Clastic Sedimentation: D.M. Tetzlaff and J.W. Harbaugh

*Orders to: Van Nostrand Reinhold Co. Inc., 7625 Empire Drive, Florence, KY 41042, USA.

Related Elsevier Publications

Journals

Computers & Geosciences

Full Details of all Elsevier publications available on request from your nearest Elsevier office.

Geoscience after IT

A view of the present and future impact of information technology on geoscience

T.V. Loudon
British Geological Survey
West Mains Road
Edinburgh EH9 3LA
UK

2000

PERGAMON

An imprint of Elsevier Science

Amsterdam - Lausanne - New York - Oxford - Shannon - Singapore - Tokyo

ELSEVIER SCIENCE Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB, UK

First edition 2000

Library of Congress Cataloging in Publication Data
A catalog record from the Library of Congress has been applied for.

British Library Cataloguing in Publication Data
A catalogue record from the British Library has been applied for.

ISBN: 0 08 043672 2
Previously published as a Special Issue of *Computers & Geosciences*, Vol. 26/3A, 2000.

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in The Netherlands

PERGAMON

Computers & Geosciences 26 (2000) v–viii

COMPUTERS & GEOSCIENCES

Table of contents

Motivation

Familiarization with IT

The emerging system

PERGAMON

Computers & Geosciences 26 (2000) ix

COMPUTERS & GEOSCIENCES

Editorial

Series Editor's Foreword

It (the personal pronoun) is IT (Information Technology — the nouns), a truism in the Information Age. This contribution is a look at the effect of information technology on the geosciences by one of the pioneers in the field of quantitative applications. Vic Loudon, with more than 30 years experience in the field and author of *Computer Methods in Geology* (1979), takes a look into the future from the present status of the subject from the vantage point of three perspectives — motive, implementation, and the emerging system.

The author notes that '...Information Technology deals with the tools for handling information, notably computers and networks. It affects how geoscientists investigate the real world, how they are organized, what they know and how they think. This book should help them [the readers] understand these changes and form a view on future trends. Benefits include more efficient and rigorous formulation and expression of ideas, and wider sharing and integration of knowledge'. The book covers all aspects of the subject and is the perfect introduction for the serious student or the dedicated professional. It contains just about everything you ever wanted to know about IT in an easily followed concise manner.

Rather than look at this volume as a summary of accomplishments to date and where we have been, it can be considered a look into the future and where we are going. Dr. Loudon considers the scientific and economic benefits from IT after introducing the subject. The second part is on familiarization of IT tools and includes a discussion of the system, project support, analyses, and managing an information base. The last part of the book looks into rethinking of the data, information, and knowledge (for the analyst) and reengineering of the system, design, user interface, and repositories (for the user).

Let the geoscientist read, digest, and react to this provocative and insightful book as we enter the 21st Century and a new era in the earth sciences.

D.F. Merriam
Geological Survey of Kansas

PII: S0098-3004(00)00051-0

This Page Intentionally Left Blank

Motivation

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A1–A3

COMPUTERS & GEOSCIENCES

Geoscience after IT Part A. Defining information technology, its significance in geoscience, and the aims of this publication☆

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Information technology deals with tools for handling information, notably computers and networks. It brings benefits such as more efficient and rigorous formation and expression of ideas, and wider sharing and integration of knowledge. This review should help practicing geoscientists and students to gain a broader understanding of these changes and form a view on future trends. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Information technology; Metainformation

1. Defining information technology

Geoscience after IT, published as a book (Loudon, 2000) and a special issue of Computers & Geosciences, offers a broad overview of the impact of information technology on the work of geoscientists, seen against the background of evolving global communication on the Internet.

Information technology (IT) refers to methods of handling information by automatic means, including computing, telecommunications and office systems. It deals with a wide range of mostly electronic devices for collecting, storing, manipulating, communicating and displaying information. Examples of IT devices are: computer software and hardware, including memory and disk storage; printers; the telephone, cable, broadcasting and computer networks; office copiers; facsimile (fax) machines; DVDs; video cameras; image scanners; televisions and monitors; data loggers and automated instruments in the field and laboratory; sensors on satellite cameras or downhole logging tools; digital surveying equipment.

IT applications seldom respect disciplinary boundaries. The focus here is **geoscience**, centered on geology but inevitably overlapping into such subjects as geophysics, geochemistry, economic geology, engineering geology, and soil science. I occasionally stray into related aspects of environmental science, surveying and geomorphology, but try to steer clear of such topics as hydrology, meteorology or oceanography, which may be parts of the Earth Sciences in the wide sense, but are well covered in their own specialized publications.

A primary task of geoscientists is to add to the base of recorded knowledge. Philosophers have made valiant attempts to say what knowledge is (see, for

☆ The views expressed by the author are not necessarily those of the British Geological Survey or any other organization. I thank those providing examples, but should point out that the mention of proprietary products does not imply a recommendation or endorsement of the product.

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00036-4

example Audi, 1998). Workers in computer expert systems and knowledge bases take a more pragmatic approach. Addis (1985) defines **knowledge** as "justified true belief", seen not as referring to a state of the brain, but as a shared human response to a complex environment. **Information** can be defined as something that adds to one's knowledge and understanding, and includes **data** (known facts) and their interpretation. The prefix *meta-* is sometimes used to refer to a higher logical type. Thus metalanguage deals with the nature, structure or behavior of language. Similarly, **metadata** is the name sometimes given to data about data, describing the data so that they can be understood beyond their originating project. The broader term **metainformation** refers to information about information. Definitions of knowledge, information and data seem to lead more rapidly than most to circularity. However, as the underlying concepts are familiar, these should serve our present purpose.

2. The significance of IT to geoscience

Modern IT offers opportunities for more effective handling of geoscience information in three main areas. The first is the obvious ability of computers to calculate, thus opening up possibilities of more rigorous analysis with quantitative and statistical methods, and more vivid graphical presentation with visualization techniques. A second area is the manipulation and management of information. This starts with the ability to move words around in a word processor or to move elements of a picture in a graphics system. It leads to the ability to capture data, store vast quantities of information in a database or document management system, and retrieve specific items on demand. A third area is hypertext linkage with rapid dissemination and display of multimedia information through worldwide telecommunications and the Internet.

IT influences the way in which scientists investigate the real world, how they are organized, how they communicate, what they know and how they think. They depend less of intermediaries like typists, cartographers, librarians and publishers for acquiring information and disseminating their findings. They can collaborate more widely, thanks to better control and flow of information. Individual workers and groups can enjoy greater autonomy within a defined, shared framework.

Taken together, the benefits from IT (see part B) include better science, cost savings, and speed and flexibility in data collection, storage, retrieval, communication and presentation.

3. This publication

3.1. Target readers

- Practicing geoscientists with a general interest in how modern information technology (IT) will affect their work and influence future directions in their science.
- Geoscientists, familiar with computer or IT applications in their own specialist field, who need a broader perspective on future trends.
- Students or educators specializing in IT applications in geoscience who require a top-down view of their subject.

3.2. Objectives

To provide an overview and rapid reference to assist readers to:

- understand the ways in which geoscientists can collect, record, analyze, explain, assemble and communicate information in the evolving geoscience information system;
- understand how IT affects methodology and enables hidden constraints imposed by traditional methods to be overcome;
- understand the theory underlying IT applications and know how to find examples and guidance for their implementation;
- form a view on future trends and thence develop a framework to influence new developments and operate effectively within them.

3.3. Structure and overview

One effect of IT can be to separate content from container. The same material can be held (as here) in an electronic archive and presented as a book, a special issue of a journal or a set of articles for browsing on screen or printing locally. I have tried to harmonize the results with established bibliographic conventions and terminology, and apologize for any remaining confusion.

Although I wrote this account for reading in sequence, there is probably enough repetition and cross-reference for you to refer to sections out of context. I hope you will have little difficulty in dipping into sections of interest from the table of contents, abstracts and keywords. Internal cross-references should help you to follow threads leading to similar topics.

The parts deal with the following topics.

Parts A and B: definitions and motivation.

Information technology deals with tools for handling information, notably computers and networks. Geoscience can benefit from IT through more efficient and rigorous formulation and expression of ideas, and wider sharing and integration of knowledge, Progress requires a broad systems view. This account should help geoscientists to understand the overall changes and form a view on future trends.

Parts C–H: familiarization with IT methods and the underlying theory.

Not all geoscientists are familiar with available methods of IT, although these influence all phases of a project and every time of information. This review looks for underlying principles, moving from individual to project to global requirements. It tracks the process of familiarization, from ubiquitous tasks like word processing through statistical analysis and computer visualization to the management of databases and repositories.

Parts I–M: the emerging system.

Earlier parts dwell on the benefits of IT and the nature of IT tools. For a clearer view of how geoscience and IT will interact, we need to reconsider our own methods of investigation: how we observe, remember and record, how we build knowledge from information, cope with changing ideas, and create a shared record in the geoscience information system. Our methods relate to the potential of IT: the flexibility of hypermedia, the developing standards for the global network of cross-referenced knowledge, and the particular value of well-organized structures of geoscience knowledge. They help us to understand the emerging geoscience information system, to define our requirements and to build on current initiatives and opportunities, which are outlined here.

Acknowledgements

I remember with gratitude the help of too many colleagues to list here. In particular, Dan Merriam and Graeme Bonham-Carter have made this publication possible, and Peter Henn and Lee Hodge of Elsevier have offered patient guidance. I am indebted to the British Geological Survey for access to their facilities and for moral and intellectual support. I am grateful to Dr Peter Allen for his encouragement over many years, and to John McInnes and Bob McIntosh for their constructive advice.

References

Addis, T.R., 1985. Designing Knowledge-Based Systems. Kogan Page Ltd, London 322 pp.

Audi, R., 1998. Epistemology: A Contemporary Introduction to the Theory of Knowledge. Routledge, London 340 pp.

Loudon, T.V., 2000. Geoscience after IT. Elsevier, Oxford.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A5–A13

COMPUTERS GEOSCIENCES

Geoscience after IT Part B. Benefits for geoscience from information technology, and an example from geological mapping of the need for a broad view

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Information technology can lead to more efficient, versatile and less costly ways of supplying and using information. The familiar paper journals and books of the geoscience literature are being supplemented, and some supplanted, by electronic versions offering new facilities. Geoscience repositories gain efficiency and flexibility in storage, management, access and presentation of data. Global standards help communications, sharing of facilities, integration of ideas, collaboration and delegation of decisions. An example from geological mapping illustrates how a broad view of computer methods leads, not just to better ways of delivering the same product, but to more fundamental improvements in expressing, sharing and generalizing the geologists' conceptual models. Familiarity with existing systems can blind us to their shortcomings: familiar methods may hide assumptions that are no longer relevant. The example suggests that maps, reports and supporting evidence can be linked by hypertext in a tightly connected model. Targeted distribution of appropriate, up-to-date information can replace the high-cost scattergun approach of conventional publication. This leads to a tentative identification of user needs. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Electronic publication; Global standards; Digital cartography; Conceptual model; User requirement

1. The geoscience literature

Greater efficiency often shows up as reduced cost. Some figures, quoted by a fact-finding mission of the European Commission (Goldfinger, 1996), therefore deserve some thought. They refer to the costs of a banking transaction, such as paying a bill or cashing a check, in the USA. Using a full service branch, a typical cost was $1.07; using telephone services $0.54; using Automated Teller Machine full service $0.27; using Personal Computer banking $0.015; and using the Internet World Wide Web $0.01. The authors point out that non-banking organizations, such as supermarket chains, can enter the banking field at low cost through the Web and can cherry-pick desirable customers from the traditional banks.

The costs are spectacularly reduced by removing the need to rely on manual records, the buildings to house them and staff to run them. Customers can control the transactions through an automated process from their own desktops. No routine support is needed from the

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

PII: S0098-3004(00)00037-6

supplier. Although some customers may pine for the marble halls and the human touch, the inconvenience of going to the bank and lining up for service is avoided. It is not difficult to draw analogies with obtaining geological information through the Internet, as opposed to purchasing traditional publications or visiting libraries or archives. Customers bear the small cost of printing the information on their own desktop printers, but have the opportunity to review it first on screen and benefit from the rapid delivery.

There has been wide discussion of the consequences of information technology for scholarly publication. Useful entry points are the work of Varian (1994) on the future of electronic journals, and the bibliography maintained by Bailey (1996). Odlyzko (1994, 1996) quotes some figures which give an idea of the scale of the costs. The number of publications in earth sciences has for some time been doubling every 8 years. Commercial publishers see a new journal as successful with as few as 300 paid subscriptions. Harvard University spends some $58 million a year on all its libraries. Subscription costs are about one third of total library costs, much of the remaining cost being associated with storing and managing books, serials and catalogs that are duplicated in numerous other libraries. To publish a scientific article in mathematics, Odlyzko estimates the research costs at $20,000, preparing and reviewing the paper another $5000, and publication costs $4000. The average number of readers for each article is about 20 (possibly less in geoscience). He suggests that if library users were asked to cover the publication costs by putting $200 into a meter in order to read one article in detail, or $20 to skim the contents, readership might fall. Instead, the costs are almost entirely concealed from the user.

High-energy theoretical physics is one area where new delivery systems are in place. Within a year of Ginsparg (1996) introducing a system, most communication of pre-prints in that field moved from paper to electronic communication. The main benefits are convenience, immediate worldwide availability, and the opportunity to comment and discuss.

Earth Interactions (Hepner et al., 1998), the on-line journal of the American Geophysical Union, the American Meteorological Society and the Association of American Geographers, offered a different rationale. It was launched for scientific rather than economic reasons. "The problem with you publishers is you think you add value. Well, you don't. You force me to reduce the information in my scientific papers so that they will fit on a flat printed page." The journal articles are refereed rapidly by e-mail and published on the Web. They offer such features as hyperlinks, animation, virtual reality, links to large datasets in external archives, "live mathematics", interactive 3-D display, forward references (that is, to later citations of the paper), linked comments and replies, and corrigenda. It is intended to become financially self-supporting and the editors believe that: "the availability of authenticated peer-reviewed scientific articles for future scientists require an income stream from the users. Furthermore, the marketplace will help to sort out the useful from the chaff". After an initial subsidized trial period, the editors expect to obtain revenue from author and subscription charges, similar to their print journals. Their preparation costs (as opposed to publication and library costs) are at least as high as for a print journal.

Odlyzko (1996) suggests that the major displacement of conventional publication will occur between 2000 and 2010. When it happens, he expects the change to be abrupt. The kudos of publication in electronic journals must then match that of their printed counterparts. Computer-mediated communication is likely to replace much print publication because of the lower publication and library costs, and because of the improved representation of scientific findings, increased flexibility and ease of use. IT promises quicker access to more appropriate and up-to-date information at lower cost. Alongside the development of new methods that take full advantage of these opportunities, the **backlog** or **legacy** information (existing documents from earlier technology) must remain an integral part of the changing information system. As described in part L, section 3, much printed literature is also available as an electronic copy.

2. Managing a knowledge base

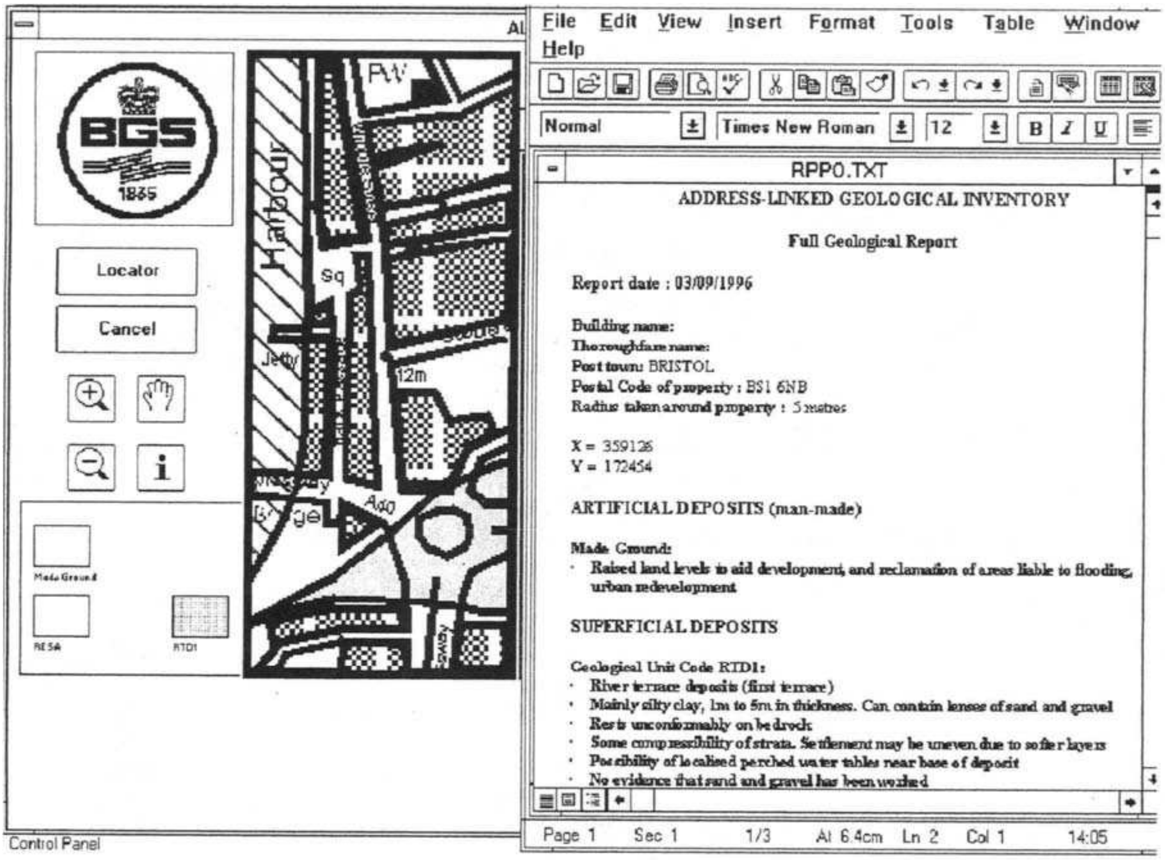
Technical journals are concerned with publishing self-contained articles on a wide range of topics. Commercial and state organizations, such as oil companies or geological surveys, have a different objective — systematically collecting and maintaining geoscience information to support specific activities. This calls for a different approach to IT. The British Geological Survey (BGS) is an example of a medium-sized survey. "The BGS is the UK's foremost supplier of geoscience solutions and is active in areas such as land-use planning, waste disposal, hydrocarbons exploration, civil engineering, minerals extraction, contaminated land, seismic and geohazard evaluation and understanding climate change" (BGS, 1998). A primary concern of BGS is therefore the management of a comprehensive knowledge base for a well-defined geographical and subject area.

BGS currently prepares and publishes a comprehensive set of geoscience maps and reports for the UK and surrounding seas, largely the results of its own surveys. Extensive archives of supporting information are held for consultation, including much information of variable quality from external sources. The archives take many forms, such as data files, field notes, borehole records, logs, charts, photographs, videos, sketches, thin sections, fossil specimens, satellite and aircraft imagery, references to external source of information, including papers, sources of expert advice, and so on. In addition to the publications, many of the archived records are publicly accessible for inspection on site.

The integrated collection of linked information can be handled more efficiently with IT. Low-cost storage in computer databases makes it possible to archive and index information, including some results of field survey, in a readily accessible form. The indexes support off-site searching of the archive. Compound electronic documents combine text and map information, and link it to images illustrating detail and pointers to additional evidence. Flexible retrieval and presentation can provide output customized for a wide range of users (see J 1.8), as in Fig. 1. A well-structured database with full metadata (H 3) is a basis for publication on demand. Knowledge, however, still remains largely in the heads of experts. They must understand the users' needs, as well as the content and structure of the information, to interactively control selection and output. As customers are best placed to understand their own requirements, the system should in time offer them direct access.

3. Sharing information

IT should simplify the sharing of information. Yet a computer user may need to learn new techniques for



ADDRESS-LINKED GEOLOGICAL INVENTORY

Full Geological Report

Report date : 03/09/1996

Building name:
Thoroughfare name:
Post town: BRISTOL
Postal Code of property : BS1 6NB
Radius taken around property : 5 metres

X = 359126
Y = 172454

ARTIFICIAL DEPOSITS (man-made)

Made Ground:
- Raised land levels to aid development, and reclamation of areas liable to flooding, urban redevelopment

SUPERFICIAL DEPOSITS

Geological Unit Code RTD1:
- River terrace deposits (first terrace)
- Mainly silty clay, 1m to 5m in thickness. Can contain lenses of sand and gravel
- Rests unconformably on bedrock
- Some compressibility of strata. Settlement may be uneven due to softer layers
- Possibility of localised perched water tables near base of deposit
- No evidence that sand and gravel has been worked

Fig. 1. Printed compound electronic document. Geological information relevant to the potential purchaser of a property is assembled in the BGS Address-Linked Geological Inventory. It is retrieved from various databases and GIS. The material is then edited, and provided to the customer on screen or on paper. A small section of an online report is shown here. British Geological Survey ©NERC. All rights reserved. Base map reproduced by kind permission of Ordnance Survey ©Crown Copyright NC/99/225. More at: http://www.bgs.ac.uk/bgs/w3/see/SERVICES.HTM.

every new application or change to the system. Even within an organization, different groups may tend to work independently, selecting their own computing tools and their own structure and format for storing data. Users may spend more time transforming data than solving geoscience problems. Sharing data between organizations adds further complexity.

A cross-section of oil companies addressed this issue by creating the Petrotechnical Open Software Corporation (POSC) as a not-for-profit corporation in 1990. "The standards and open systems environment to be facilitated by POSC represent a maturing of the industry that frees companies from worrying about the integration of their computer systems and lets them concentrate on areas of added value" (POSC, 1993). Although specifically addressing the requirements of the exploration and production business, it adopts existing standards where possible, developed further as necessary. It makes decisions through an open process supported by technical arguments, not commercial or special interests. Its work is made available to all, and is relevant to a wide area of geoscience. In addition to POSC, there are many other activities and groups promoting standards in related areas. The short-term costs of standardization cannot always be justified by putative long-term gains, and many standards have been superseded before being widely adopted. Nevertheless, the implementation of standards and better links through the internet are gradually overcoming the artificial barriers to communication.

The oil industry is establishing shared computer repositories where any of the subscribing companies can access the data. Because they are run by specialists, the repositories provide better and more secure facilities. Because they are collaborative ventures, they reduce duplication of effort in systems development, data collection and storage. Because the data meet agreed standards, they can readily be retrieved and analyzed by the subscribers. The quoted savings are immense.

There are other gains. The standards create a larger single market and so justify higher investments in developing applications software. The consistency of data collected to uniform standards pays dividends in such areas as quantitative analysis, visualization and database management (see parts F, G, H). Standard procedures and content also simplify project planning. Effort can be put into genuinely new investigations rather than reinventing and documenting old ideas. Global standards (L 6) simplify exchange of data across boundaries of discipline, organization and place.

IT means that scientists can themselves prepare documents, such as letters, memorandums and reports. This includes keyboard entry, preparing and inserting diagrams, selecting content and layout by inspection on screen, and reusing earlier work in new contexts (C). A computer template prepared by a graphic designer can ensure a uniform house style. In an academic community, lecture notes, student appraisals, and examples can be accessed more widely and more readily.

Project management is also helped by computer communication. Larger groups can collaborate effectively through rapid dissemination of planning documents, schedules and progress reports. Fewer layers of management are needed, because the information is available to all (M 3.1). Potentially, improved sharing of information offers more freedom of action to individuals, with more intelligence at their fingertips. Many of the benefits of IT are missed, however, if they are sought in too narrow a context, as the following example of geological mapping illustrates.

4. The need for a broad view

An **information system** is a means of recording ideas and sharing information. Geoscientists, for excellent reasons, tend to take their information system for granted, and may consequently give little thought to a basic need for improvement. Modern information technology, appropriately applied, makes the system more effective and efficient. There are many examples of computer applications that make a valuable contribution to part of the geoscience information system. To gain the full benefits, however, it is necessary to look at the system as a whole. Analysis of a system generally starts with a specification of the user requirement, but this can prove hard to tie down. For example, a study that I shall now describe began with a small, familiar part of the system and ended by pointing to some unrecognized requirements.

It was my privilege, many years ago, to study some fine examples of conventional geological maps. The maps are informative, attractive, accurate. The organization which produced them strives to respond to customer demands. My objective was to learn by comparison how the poor daubs then coming off the computer printer might be improved. With modern technology, the aim must be not just an imitation, but a better product. To detect areas of possible improvement, it may help to consider how information was transformed during the process of mapping.

The information available to the geologists as they strode across the landscape, hammers at the ready, is very different to that which reached the final published map on which they signed their names. The geology, in its infinite diversity, has been reduced to areas with uniform colors corresponding to a small set of mappable units. This categorization of **objects**, that is the things or entities of interest in the current context, is a basic part (taxonomy) of the scientific method (J 2.1).

Objects with similar attributes are grouped into named **classes** (grass, sheep, rocks), thus enabling one to make general statements about them and codify one's expectations about their properties and behavior. Not only can we talk about them, but in this case, can also show the distribution of classes of geological objects on a map. A dual statement is being made: the rocks designated on the map by a specific color have been identified as belonging to a particular formation; the formation comprises in part the rocks at the locations shown by the appropriate color on the map. We need to consider next, however, how adequately the map reflects the ideas in the minds of the surveyors.

4.1. Extending the language

We generally know more than we are able to express and share with others. Where technology allows us to express ideas in new ways, it can improve our ability to understand and share knowledge. The depiction of geology on the map is subject to cartographic constraints. Thicknesses of lines are chosen to be legible, boundaries are moved apart to be distinct, lines smoothed to avoid visual clutter, and so on. These are secondary to the interpretation in the field, which must surely have involved aspects, such as consideration of three-dimensional processes, which cannot be shown on the map. The geologists can therefore be said to have developed a **conceptual model** — a formalized mental image giving a simplified view of relevant aspects of the real world. The full conceptual model exists only in the minds of the authors, and must be further simplified for representation on the map.

The completed map is a permanent, shareable, public record of the authors' ideas. However, it necessarily imposes **physical constraints** on the representation of the conceptual model. The technical limitations of pen and paper are significant. Lines of even thickness and areas of uniform color are easily drawn, but artistic genius is needed to accurately depict our imperfect view of diversity, ambiguity and uncertainty. The limitations of our skills and tools force us to reduce the complexity of nature to a few mappable units.

Our mental images of objects are strongly influenced by their representation. Make a careful schematic drawing of a fossil, and it may be easier to remember the drawing than details of the original specimen. Draw firm boundaries on a map and they affect your view of the geology. The rock bodies are three-dimensional, and can be fully understood only in terms of the processes by which they originated and developed through geological time. The map is two-dimensional, supplemented by cross-sections, indications of the geometry such as orientation measurements (strike and dip), intersections with the topography, and possibly contours on a subsurface horizon. We have a view of the vertical relationships along the lines of cross-sections, with a rather hazier view in between. Ink marks on static, two-dimensional paper cannot represent satisfactorily a complex sequence of three-dimensional units and their spatial relationships, far less their origin and structural history.

A map at a uniform scale cannot accommodate the variation of information density on the ground. For example, the geological information for a map sheet may be limited to a few good coastal exposures with little solid geology exposed inland. A map which fully reflected this would have a thin zone of illegible clutter relieving the blank monotony of the rest of the sheet. If the main objective of surveyors in the field is to produce a map, therefore, they may give limited attention to the detail of good exposures, knowing that there is no room to show the results on the map. Again, the physical limitations of the medium influence the conceptual model, and thus the investigational procedure.

The ink marks on the map depict formalized symbols, such as stratigraphic codes, which do not imitate the appearance of the original objects, and patterns, such as formation boundaries, which are miniaturized versions of patterns on the ground, or, rather, in the geologists' conceptual model. The process of moving from observation in the field to representation on a map involves **generalization**, that is showing the salient features, possibly in a simplified form, and removing unnecessary detail (see Buttenfield and McMaster, 1991). When a smaller-scale map is produced for the same area, the original map is again generalized. The latter process can be readily studied with the aid of an enlarging photocopier. Fig. 2 indicates differences between features drawn at the scale of the original survey at 1:10,000 and as shown on the published map at 1:50,000. This may also throw light on the generalization during field mapping and the aspects of the real world that are conserved during that process.

The dike swarms or the coal seams in Fig. 2 are obviously exaggerated in thickness for legibility, and thus are not a true scaled reduction. Their exact numbers (actual or observed) are not shown, although variations in numbers may be reflected in some way, and it is possible that the spacing or relative spacing is also indicated. Their orientation is probably represented, and in a few cases their continuity and even variation in thickness, but not their length. An intricate pattern has been carefully displayed. Its exact meaning, however, is not immediately obvious.

Generalization resembles statistical **sampling** (F 3), in which a small number of items are selected to represent a larger whole. Generalization also reduces a large amount of information to a more manageable quantity that throws light on the overall situation. Here, the requirement is to be able to draw conclusions about the geology from the map. Statisticians insist that in order to arrive at statistically valid conclusions an appropriate sampling scheme must be followed and explained to the user (see Davis, 1973). The map in some ways resembles a sample, but what was the sampling procedure? Take, for example, a symbol showing the orientation of bedding in Fig. 2. It may be representative of the orientation within either a particular area, or within an outcrop, or a horizon, or a pattern of folding. It may be a random sample, a typical value, a particularly significant value, or selected haphazardly. It may refer to an area the size of a field notebook, or it may not. Certainly it is not a measurement from which one could confidently draw quantitative conclusions. Lines on the map show formation boundaries and the positions of faults. But it is seldom clear where the geologists observed their presence, or inferred it from landscape features, and where they were required simply in order to complete the geol-

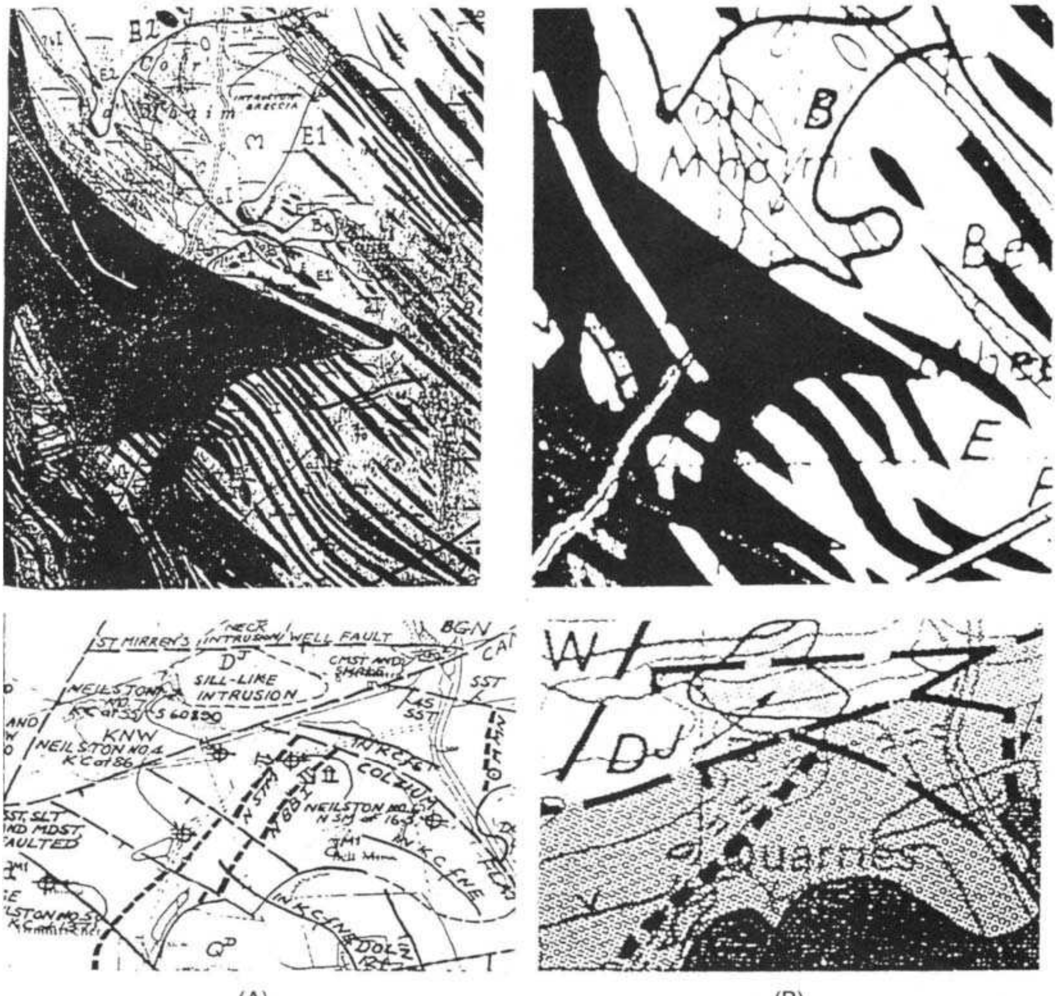


Fig. 2. Map generalization. (A) Fragments of maps at survey scale of 1:10,000; (B) maps for the same areas generalized for publication at 1:50,000, enlarged here for comparison. The amount of information is progressively reduced during observation, recording, abstracting and reading. This generalization process can be observed in action during scale reduction of a map. In the upper example (BGS Sheet 44W, Eastern Mull) geometrical properties of the dike swarm include average orientation, lenticularity, variation in density, dimensions and spacing. Some of these at least were affected by generalization. In the lower example of scale change (BGS Sheet 31W, Airdrie) minor faults were removed, coal seams selected, and the orientation of the coal seam adjusted to remove the effect of minor faulting. Presumably similar faults and coal seams exist that were too small to show at survey scale. British Geological Survey ©NERC. All rights reserved. Base maps reproduced by kind permission of Ordnance Survey ©Crown Copyright NC/99/225. More examples in: BGS Technical Report WO/94/3, electronic version at http://geolib.bgs.ac.uk, search by name (Loudon).

ogists' reconstruction. Despite the large scientific investment which the map represents, its content needs cautious interpretation. Nevertheless, despite the absence of any recognizable sampling scheme, it is possible to learn much about the geology from a geological map.

Perhaps the map is not an attempt at a precise geometrical depiction, but rather is telling a story which geologists have been trained to understand. It is not difficult to imagine geologists in the field developing their own conceptual model, and using a pre-existing map to check it against earlier observations and align their ideas with those of their predecessors. Geologists who could themselves have done the earlier work, know that particular marks are made under specific circumstances. By an intuitive process they could put themselves in the authors' shoes, project themselves into the authors' minds and visualize the conceptual model that lay behind the depiction on the map.

Intuition, that is apprehending something without any intervening reasoning process, it vital to science. You could not have read this far without it. Human beings are skilled in intuition, computers are not. But that is not the end of the story. What geologists show on a map is an inextricable mixture of hard fact and interpretation. With computer support, the language of the map could be extended so that aspects of the geometry, for example, could be rigorously sampled in the field, and the three-dimensional structure could be recorded and tested for consistency. But this is pointless if the rigor is lost in the final portrayal. These issues are explored in more detail in part G. Meantime, the point is that there is a hidden requirement for users to free their ideas from the constraints of scale, dimensionality and cartographic representation imposed by the paper map; to develop multiple conceptual models more freely; and to express them more fully and rigorously in the shareable record. Modern information technology can extend the means of expression and communication, but must be used with caution and awareness of the likely consequences.

4.2. Connectivity and integration

Devising a solution to the narrow user requirement just outlined could have damaging and dangerous side-effects, like taking aspirin for the pain of a stomach ulcer. The geological map is a small part of recorded geoscience information. Changes to one part can have knock-on effects elsewhere. For example, to obtain an account of the geology, even of a mapped area, the user is likely to turn to a written report. Both map and report are expected to share the same conceptual model, and refer to the same objects and object classes (H 5). One cannot be modified without affecting the other.

The map and the report are separate documents, possibly prepared at different times. Small maps are likely to be included in the report, as well as diagrams providing graphical information that might more naturally be part of the map. Many maps, on the other hand, contain long text descriptions that could fit equally well into a report. There is thus no sharp distinction of content between map and text material. They are separated because the two widely different formats could not readily be printed as a single document. Cross-reference from report to map is by the tedious mechanism of grid coordinates, and references from map to report are likely to be confined to a brief bibliography referring to the map sheet as a whole. For obvious reasons, no conventional document contains references to items published later than itself. The connections between documents could clearly be improved.

To grasp the full significance of a geological interpretation, the user may have to visit the area and retrace the investigation. After all, in any science there should be the option of checking conclusions by reexamining the evidence. Even where field notes can be examined, however, there is little guidance to the precise reasoning behind the conclusions shown on the map or recorded in the report. Although providing an invaluable context, published maps and reports may lack the specific information which is required for a detailed study, and give little indication of where that detail can be found. Think, for example, of the civil engineer looking for records to assess the foundations of a large building. If borehole records exist, they may have been used as supporting evidence in making the map, but neither explicitly cited nor evaluated. Information technology should be able to offer better solutions to supporting, retracing and sharing the investigators' ideas.

A surprisingly large part of most scientific papers is a reworking of earlier published material, recast to explain or support the author's viewpoint, but involving a degree of repetition which might be unnecessary if the original sources were more accessible. A somewhat broader solution would therefore take advantage of the greater connectivity that GIS and hypermedia (E 4) can offer, and thus the ability to integrate information from many sources. Material from the map, the report, diagrams, the database, computer applications, video and still photography, external comments, references to previous work and access to expert opinion could all be incorporated in a fully connected hyperdocument, using simple and familiar tools for access from the desktop. It would be unwise, however, to embark on such a project without considering its long-term development and the means of disseminating the information.

4.3. Deliver and print

The economics of the offset-lithography printing process affect both text publications and maps. Preparing the reproduction material is complex and requires scarce skills, possibly resulting in long delays. The costs of setting up a print run are comparatively high. Subsequently, each additional copy within the print run costs little more than the paper and ink. Some thousands of high-quality copies may therefore be printed in a batch. Identical copies are bound, dispatched, documented and stored throughout the world, in public, private and personal libraries. Interlibrary exchange schemes are organized to mail copies if they are not available locally. All this is to reach the handful of users who may be interested. The printed product is a permanent snapshot of the author's ideas at a particular time. Revision is costly and therefore infrequent. The information is likely in consequence to be out of date.

The geological map is complex and to the expert eye is full of information. But there is tension between the scientists' desire to record all their insights and the demands of the market place. Information to attract as many readers as possible may be included to justify the wide circulation, at the expense of clutter and complexity.

The end-user of the map may be an expert in another field with limited training in geology. Aspects of the geology may be important to the land-use planner zoning residential areas, the construction engineer planning a new highway route, the insurance agent concerned with geological risk to housing, the lawyer with a claim for ground-water pollution, the teacher explaining landforms, the company director financing mining development. All have their own requirements for specific geological information, which may or may not be available from the general map. They need a simple presentation of the information including relevant detail but free of clutter. One approach to meeting such needs is to prepare many thematic maps for the same area, meeting a range of potential requirements. A result, however, is greatly increased publication costs. A solution is to print on demand extracts selected from a **geographic information system** (**GIS**), a computer-based system for handling map information (see Bonham-Carter, 1994). Parallel arguments could be made about text reports.

Because many maps are published to a standard set of scales, the geological map can be overlain on a light table to correlate it spatially with other maps showing for example topography, soils, or land use. Unfortunately, it may be difficult to find the maps, their sheet boundaries may not match, there may be many small discrepancies due to different series being revised at different times, and the map underneath is never easy to read. GIS offers a solution in principle, seldom achieved at present because of the lack of availability of digital maps.

The geological map has been taken as an example, but is a small part of the recorded information on geoscience. Text publications share many of the same deficiencies. The language in which the reports are expressed forces a particular pattern of thought, such as categorizing the diversity of nature in predetermined molds, which may not always be the best option. A report of any kind is expensive and laborious to produce. It therefore tends to present a rather tidy and self-contained account of its topic, omitting unsuccessful lines of investigation and details that may be informative but do not contribute to the main theme.

None of this implies incompetence, but rather that ways of working are influenced, perhaps controlled, by the available tools. Around these tools a major industry of intermediaries, such as publishers, printers, booksellers and librarians, has grown and cannot change overnight. Geoscientists are the beneficiaries of a huge legacy of information, recorded and greatly influenced by the technology of the time. Now, technology is moving on and the information industry is regrouping. It is feasible to hold information of many types under the control of the originators or their proxies, and to select and deliver it electronically worldwide when required, for local editing and printing of both images and text under the control of the user. In the words of the Xerox Corporation, *print-and-deliver* is giving way to *deliver-and-print*.

5. Towards a user requirement

There are incompatibilities and conflicts between the old and new. Ways of thinking and ways of working that have been deeply ingrained over generations, may no longer be appropriate. We must consider not just the representation of existing data, but also the more effective representation of reality. We must bear in mind that new methods may bring risks of misunderstanding, which hidden features of conventional systems were designed to circumvent. The benefits we can expect from IT, and our objectives in using it, can be crystallized as a user requirement, a concept described in more detail in K 3. An apparently narrow user requirement, as in B 4.1, must be placed within its wider context. To the frustration of IT support staff, the user requirement tends to evolve as new methods are explored, and may be clear only after the work is complete.

In general terms, the user requirement identified so far is to share information more effectively and efficiently by:

- more complete and rigorous representation of conceptual models and their supporting evidence;
- reduction of repetition by better links;
- direct worldwide access to information that meets global standards;
- more appropriate control of information by originators and users, with less reliance on intermediaries;
- easier access to more rigorous analytical methods and visualization techniques;
- a move from fixed paper documents, to a shared, dynamic knowledge base, from which users can selectively retrieve and print information.

At this stage, I hope you agree at least on the need for geoscientists to arrive at an informed view of how we can best work with new information technology: informed by experience of the various tools that IT places at our disposal (parts C–H), and by insight into how we think and work within the geoscience information system (parts I–M).

References

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modelling with GIS. Pergamon, Oxford 398 pp.

Buttenfield, B.B., McMaster, R.B. (Eds.), 1991. Map Generalization: Making Rules for Knowledge Representation (Symposium Papers). Wiley, New York 245 pp.

Davis, J.C., 1973. Statistics and Data Analysis in Geology: with Fortran Programs. Wiley, New York 550 pp.

POSC, 1993. Petrotechnical Open Software Corporation, Software Integration Platform Specification. Epicentre Data Model, version 1, vol. 1, Prentice-Hall, Englewood Cliffs, New Jersey.

Internet references

Bailey, C.W. Jr, 1996. Scholarly electronic publishing bibliography. Houston: University of Houston Libraries, 1996–99. http://info.lib.uh.edu/sepb/sepb.html.

BGS, 1998. British Geological Survey home page. http://www.bgs.ac.uk/.

Ginsparg, P., 1996. Winners and losers in the global research village. In: Invited contribution for conference on electronic publishing in science held at UNESCO HQ, Paris, 12–13 February http://xxx.lanl.gov/blurb/pg96unesco.html.

Goldfinger, C., 1996. Electronic money in the United States: current status, prospects and major issues. http://www.ispo.cec.be/infosoc/eleccom/elecmoney.html.

Hepner, G.F., Sandwell, D.T., Manton, M., 1998. Earth Interactions Journal. http://earthinteractions.org/.

Odlyzko, A.M., 1994. Tragic loss or good riddance? The impending demise of traditional scholarly journals. http://www.iicm.edu/jucs_0_0/tragic_loss_or_good/html/paper.html.

Odlyzko, A.M., 1996. On the road to electronic publishing. Euromath Bulletin 2 (1June), 49–60 http://www.research.att.com/~amo/doc/tragic.loss.update.

Varian, H.R., 1994. Recent research paper of Hal R. Varian. http://www.sims.berkeley.edu/~hal/people/hal/papers.html.

Familiarization with IT

PERGAMON

Computers & Geosciences 26 (2000) A15–A19

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part C. Familiarization with IT applications to support the individual geoscientist

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Familiarization with IT proceeds best by first learning the basic skills, using them, and considering the consequences; thus developing a mindset to take advantage of future opportunities. A desktop computer is a good starting point. Word processing can be seen as a route from tangled thought to immaculate presentation. Spreadsheets can help to build, analyze and plot datasets. Data can be collected with forms and spreadsheets, with on-line instruments in the laboratory or in the field, or by scanning and maybe OCR. Standards are needed to communicate between systems. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Desktop computer; Word processing; Spreadsheets; Data capture

1. The route to IT familiarization

Not all geoscientists are familiar with available methods of IT, although these influence all phases of a project and every type of information. This review looks for underlying principles, looking at individual support in this part, moving on to project, then global requirements (parts D–H). It tracks the process of familiarization, from ubiquitous tasks like word processing through statistical analysis, spatial analysis and computer visualization to the management of databases and repositories.

You learn about computers by using them, just as you get to know a town by living in it. A guidebook can help by pointing out features you might otherwise miss, and explaining the background to improve your understanding and give an overview to tie it all together. I see no reason to repeat sections of computing manuals here, when you can readily find the real thing, often as online help and demonstrations of software on your computer. Instead, I offer suggestions about what to do and where to find things — a means to an end, not an end in itself.

This guidebook also has a theme: that information technology is changing the way we conduct geoscience, and to control that change we need to understand, not just the technology but also the way our science works. As I am unsure of what you already know, I start with basic concepts, and hope you will skip ahead if they offer nothing new. A search of the World Wide Web will provide details of many of the topics mentioned.

Many geoscientists have a good basic knowledge of computing, and add to it as required. The advantage of learning by experience is that the information may

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00038-8

be more relevant and so more easily remembered. The disadvantage is that broader issues may be forgotten and better methods neglected in favor of the familiar, but periodic review of similar applications elsewhere helps to overcome bad habits. Formal instruction through books and courses should lead to a wider appreciation of the possibilities, and to learning the best approach rather than one that was stumbled on by accident.

Those who are less familiar with the basics of IT or its geoscience applications have different priorities. They need to gain the basic skills of operating with keyboard and mouse. They must learn how simple tasks are handled on the computer before they are in a position to learn about more complex applications, where these skills are taken for granted. A step-by-step approach where skills are learned and immediately put to good use is strongly recommended. Word processing is one skill which is at once useful and also leads to familiarity with the keyboard, mouse and screen, and the idiosyncrasies of the graphical user interface (GUI). Handling electronic mail (e-mail), on-line public access catalogs (OPACs), and surfing the Web are then less daunting.

2. Desktop hardware

A good starting point for learning about computing applications is your own desktop computer, where skills can be learned by experiment, making mistakes in the privacy of your own office. Here is a place to learn of the user interface, word processing, spreadsheets, database and communication features. Desktop computers also support many small geoscience applications that are helpful in individual investigations.

The computer screen on the desktop is a window into information technology. It may be part of a personal computer or a workstation or a simple network computer with little computing capacity of its own, but with the ability to download instructions from a remote server. It may be static or portable. It may be for your exclusive use or it may be one of many machines which get their individuality from downloaded software. Thus you may be able to go to one of many different computers, enter your identifier and password and operate it as your own machine.

If you would be seen as technically aware, you would know the make and model of your desktop machine and its main server, if any. You would know the name of its operating system and the company which produced it. You would know whether the computer is free-standing, that is, unconnected to others, or networked. If the latter, you would know what type of network it is connected to and by what means, how extensive the network is, and the bandwidth of the connection. The protocols for graphics on screen, the memory size, the processor speed, the disk capacity, and any peripheral attachments including the printing facilities (type, manufacturer, resolution color availability, paper sizes), are additional points of interest. Some computing terms are reviewed in part E. The details for your own machine can usually be discovered by searching through the menus on the machine, by looking at the manuals, or by asking a well-informed colleague.

Most of use have little say in selecting a computer at work. If we do, the important points are likely to be how it fits in with existing equipment and planned developments. At home, other considerations come into play, but compatibility with one's employers' plans may still be crucial. For mainstream, general-purpose computers, you should be able, as when renting an automobile, to get in and drive, assuming, of course, some training and experience. If you plan to select and buy your own personal computer, there are many periodicals that offer advice. The vendors are likely to be biased, and it is well to precede any discussions with them with a little background research and to have in mind some figures for the details mentioned in the last paragraph. There is no point in offering further advice on purchase here, as it would be long out-of-date before publication. Remember, however, that a computer soon becomes obsolete, and the cost of purchase may have to be written off over as little as three years. Remember also that the main costs will lie in training and upkeep rather than the initial purchase.

Faced with an unfamiliar machine, progress can be made with a combination of experimentation, reading books and manuals, on-line help systems and demonstrations, advice from colleagues and vendors, and from training courses if they are readily available. Skills should be learned with a purpose in view, not studied and forgotten. An obvious starting point in computer use is word processing.

3. Word processors

Word processing skills are not difficult to acquire, and are a vital part of most geoscientists' work. Furthermore, the skills are fundamental to most other computer applications, and few of the present readers should be without them. A very small number of word processor packages dominate the market, all with similar features, and all suited to a geoscientist's needs. A choice between them might depend on the availability of local support, cost, their integration with other relevant packages, and perhaps the ease of handling mathematical or chemical formulas. They should be seen as part of the overall computing facility, not as a self-contained item.

There is no point in offering advice on their usage here, as this is best obtained from their on-line tutorials or manuals. However, we need to review their function, to see how it contributes to the information system. Think first of the earlier alternatives, which have shaped the information system as we know it. A familiar analogy is writing with pen and paper, starting at the top of the page and writing line by line until the page is full. Corrections can be made by striking out unwanted text and adding more wherever space is available. As the hand-written document can become rather illegible, the next step is to pass it to a professional typist who, by reading and reentering all the material, can prepare an equally inflexible document on a typewriter. Corrections can be marked by pen as before, at the expense of legibility, and retyping is inevitable for a document of any significance. Publication again requires reentry from a keyboard, with all the consequent possibilities for mistakes.

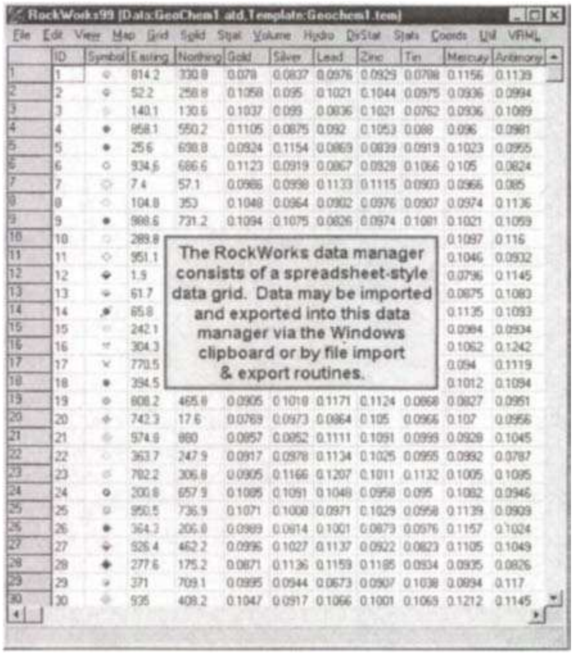
Word processing changes these procedures in several ways. The output from the computer printer is more legible than most handwriting. More fundamentally, ink is not put on paper until required. It is not fixed during the recording process. The record on the computer can be viewed on screen and altered at will. The repeated keyboard entry of the manual system is replaced by altering only what must be changed. Updated versions of a document, such as a list of references, can be prepared without reentering old material. The document need not be prepared in a set sequence, but can be built up from any starting points, and the contents rearranged whenever required. This is a major advantage for those whose thoughts seldom follow a straight line.

Content is separated from form of presentation, which can be altered separately. The computer record can generate a draft, a final document, or camera-ready copy for offset-litho printing. Parts of the text can be extracted for use in other contexts. With descriptions of, say, fossils or borehole records, items can be subdivided or expanded and automatically renumbered, without affecting the rest of the document. If the document is appropriately designed, tables of contents and indexes can be prepared and brought up to date automatically to match the current page numbering.

Typing skills, including high keyboard accuracy, good spelling, and the ability to visualize and plan the layout of the final product before it is typed, are less necessary. Scientists can record their own information, helped by tools such as a spelling checker, and adjust the layout as the document develops. This may be easier than preparing a hand-written draft for later typing. Although clear handwriting can be read mechanically, simple keyboard skills are not difficult to acquire and are likely to provide a better solution. Pen and paper remain more robust, versatile and cheaper in some circumstances. Despite the availability of rugged hand-held notebook computers, for example, it may be easier to use a conventional notebook in the field, and later decide what should be transferred to a computer record.

Word processing is a step on the way to the preparation of compound documents, which may incorporate data tables and diagrams as well as text (L 3). You can include placeholders to indicate where external information, perhaps selected from a database, is to be inserted later. You can then print circular letters or reports where the placeholders are replaced by information specific to the recipient, such as their name and address or paragraphs matching their interest profile.

The digital record can be accessed remotely, searched for keywords, and combined with images. Links can be inserted to other documents and to points in the same document. These need not refer to text, but can link to any electronic information, such as an image or data. They can be included in a compound document produced by a word processor, but require special facilities for manipulation and editing. Such concepts are greatly extended in the World Wide Web (E 4).



	ID	Symbol	Easting	Northing	Gold	Silver	Lead	Zinc	Tin	Mercury	Antimony
1	1	○	814.2	330.8	0.078	0.0837	0.0976	0.0929	0.0798	0.1156	0.1139
2	2	○	52.2	258.8	0.1058	0.095	0.1021	0.1044	0.0975	0.0936	0.0994
3	3	○	140.1	130.6	0.1037	0.099	0.0836	0.1021	0.0762	0.0936	0.1089
4	4	●	858.1	550.2	0.1105	0.0875	0.092	0.1053	0.088	0.096	0.0981
5	5	●	25.6	638.8	0.0924	0.1154	0.0869	0.0839	0.0919	0.1023	0.0955
6	6	○	934.6	686.6	0.1123	0.0919	0.0867	0.0928	0.1066	0.105	0.0824
7	7	○	7.4	57.1	0.0986	0.0998	0.1133	0.1115	0.0903	0.0966	0.085
8	8	○	104.8	353	0.1048	0.0964	0.0902	0.0976	0.0907	0.0974	0.1136
9	9	●	988.6	731.2	0.1094	0.1075	0.0826	0.0974	0.1081	0.1021	0.1059
10	10	○	289.8							0.1097	0.116
11	11	○	951.1							0.1046	0.0932
12	12	◆	1.5							0.0796	0.1145
13	13	◆	61.7							0.0875	0.1080
14	14	●	65.8							0.1135	0.1093
15	15	○	242.1							0.0984	0.0934
16	16	✕	304.3							0.1062	0.1242
17	17	✕	770.5							0.094	0.1119
18	18	●	394.5							0.1012	0.1094
19	19	○	808.2	465.8	0.0905	0.1018	0.1171	0.1124	0.0868	0.0827	0.0951
20	20	◆	742.3	17.6	0.0769	0.0973	0.0864	0.105	0.0966	0.107	0.0956
21	21	○	974.8	880	0.0957	0.0852	0.1111	0.1091	0.0999	0.0928	0.1045
22	22	○	363.7	247.9	0.0917	0.0978	0.1134	0.1025	0.0955	0.0992	0.0787
23	23	○	702.2	306.8	0.0905	0.1166	0.1207	0.1011	0.1132	0.1005	0.1095
24	24	●	200.8	657.9	0.1085	0.1091	0.1048	0.0958	0.095	0.1082	0.0946
25	25	○	950.5	736.9	0.1071	0.1008	0.0971	0.1029	0.0958	0.1139	0.0909
26	26	●	364.2	206.8	0.0989	0.0814	0.1001	0.0873	0.0976	0.1157	0.1024
27	27	◆	926.4	462.2	0.0996	0.1027	0.1137	0.0922	0.0823	0.1105	0.1049
28	28	◆	277.6	175.2	0.0871	0.1136	0.1159	0.1185	0.0934	0.0935	0.0826
29	29	●	371	709.1	0.0995	0.0944	0.0673	0.0907	0.1038	0.0994	0.117
30	30	○	935	408.2	0.1047	0.0917	0.1066	0.1001	0.1069	0.1212	0.1145

The RockWorks data manager consists of a spreadsheet-style data grid. Data may be imported and exported into this data manager via the Windows clipboard or by file import & export routines.

Fig. 1. Table or "flat file" of geological data. The array of data is held in a spreadsheet where it can be edited, manipulated or transferred to other programs for databasing, analysis or display. This spreadsheet is reproduced by permission of Rockware. More at http://www.rockware.com/.

4. Spreadsheets and business graphics

Most desktop computers offer systems for preparing and handling spreadsheets, which are of value in many areas of geoscience. Because each step of the calculation is clearly visible, spreadsheets are of particular interest in learning, teaching and exploring new ways of analyzing data. The spreadsheet is a table, or array, of numbers arranged as rows and columns (see Fig. 1). The array can be large, but because the program offers simple procedures for entering and adjusting data, spreadsheets are also useful for small illustrative tables to insert in a report. New rows and columns can be created by adding new data or by inserting a formula stating how the new entries are to be calculated from the existing data. Because the program stores the formulas, the entire spreadsheet can be recalculated automatically when entries are added or amended.

Spreadsheet systems generally include full documentation and on-line demonstrations, a good way to appreciate their characteristics and learn how to use them. Their application may lie in administrative tasks, such as keeping track of expenditure and staff time of various projects, week by week. Their geoscience applications, where many data are collected as tables, are surprisingly varied. A number of properties or characteristics are used as column headings and their values are recorded in a row for each item. In this way, the data are collected consistently. Many computations (see F 3, Fig. 1), including statistical and geophysical calculations, can be set out as sequences of formulas relating successive columns in a spreadsheet. This leads to rapid programming, since no complex coding is required. As intermediate steps in the computation are held as separate columns, they can be inspected to check on unexpected results, or to get a feel for the influence of different factors on the final result.

Simple business graphics are widely available for desktop computers. If data have been collected in the form of a spreadsheet, it is not difficult to experiment with their display. Simple pie charts, barcharts and x-y plots are a straightforward means of displaying the distribution of values and the relationships between variables, and are suitable for including in word processor documents. They are as relevant to science as to business, and you should certainly be aware of their existence.

5. Capturing data and images

Spreadsheets and database systems provide means of data entry and editing, including the possibility of creating your own form to enter and check data on screen (D Fig. 3). Field or other observations can be manually recorded on printed forms, designed to ensure that all the necessary data are recorded systematically. They make it easier to transfer the data to the computer at a later date, particularly if they match a form on screen. It is also possible, though not necessarily cost-effective, to take rugged and portable computers into the field, where they can store records as the observations are made (Briner et al., 1999). This simplifies review of the information and may help with initial analysis. The course of the investigation can then proceed on the basis of what has already been discovered.

Some surveying instruments and positioning systems can plot locations directly to a computer map. Up-to-date accounts of the satellite-based Global Positioning System (**GPS**) and trials of surveying instruments can be found on the Web (Graham, 1997), and a complete overview in Hofmann-Wellenhof et. al. (1997). Expensive data-collection instruments, such as geochemical equipment in the laboratory or some geophysical equipment in the field, are usually linked to a computer or to a digital recording device. A computer has the advantage that some processing of the data can be done at the time of capture, and it may be possible to adjust or control the instrument by feedback reflecting the incoming data.

Images can be generated from graphic programs, as mentioned in the last section. Existing images, such as diagrams, maps or photographs can be **captured**, that is recorded for use in the computer, with a scanner. The **scanner** captures the image as a raster, a set of colored or monochrome dots on an evenly-spaced grid, typically at a resolution of 300 dots per inch or 15 dots per mm, and perhaps four times as many for a high-quality image. Images use a considerable amount of storage space, from 1 to 24 or more bits for each dot, depending on color resolution. Once captured, image editing and enhancement is possible with appropriate software. Color and density can be modified, the resolution can be changed, the size and shape of the image can be adjusted, or part of the image can be selected by cropping. Images can be combined, for example by overlaying small pie charts of lithology ratios on a scanned map.

Scanning is an important means of capturing data. It can capture images of text on a printed page. **Optical character recognition** (**OCR**) systems can convert from the image to a word-processor representation of the characters. As this is a moderately expensive and error-prone process, it would not normally be used if the text had already been keyed in and was available in computer-readable form. It would probably, however, cost less than rekeying the document.

6. Information delivery and presentation

A surprisingly important ingredient in the success of computers is the widespread availability of good-quality printers. It is surprising because the rapid availability of information on the computer screen might seem to make the paper copy less necessary. However, paper remains a most convenient medium for reading and studying documents of any significant length. The ability to receive documents from a distant source and print them locally does, however, mark a significant change. Documents can be maintained by the originator, and obtained, edited and printed by the reader. The consequences (M 2.1) amount to a fundamental change in the procedures of publication.

Hypertext documents (E 4) can similarly be delivered and viewed on the screen. In print, of course, it is possible to arrange the document only as a single sequence, and the network structure is lost. Multimedia insertions, such as audio, video or computer programs, cannot be transferred to paper. Electronic documents can, however, be cited in a paper document (see IFLA (1998) and ISO (1999) for style guides). It is worth remembering also that presentations to an audience can be made with a suitable projector, and software such as Microsoft PowerPoint. The images and video from the computer screen can be projected to a large screen and sound to a loudspeaker system.

The ability to pass information between systems depends on shared standards. At a basic level, most systems can accept ASCII characters and thus a character string can be passed between them. Different versions of the same word processing system can generally exchange more detailed information, including adjustment of lines to left, right or center, font, point size, and italic, bold or underlined. By saving the document in rich text format (**RTF**), this information may also be exchangeable between different types of word processor. Similarly, images and multimedia can be transmitted in various standard forms using hypertext transmission protocol (**http**). The appearance of a page can be represented in the Postscript language (E 6), including both images and text, However, this is not appropriate if the text is to be edited by the recipient, as it might require redesign of the entire page layout. Portable data format (**PDF**) is a possible compromise, allowing full access to the text in order, for example, to search for a keyword, while preserving the appearance of the page. Where the recipient is more concerned with content than appearance, particularly where text searching and editing is required, a simple format, such as RTF, is more effective and efficient.

Some familiarity with the general use of desktop computers is a good starting point for anyone intending to make serious use of computer methods. They should then be able to create simple and compound documents on the computer, and understand the graphical user interface. It is argued later (L 6.3) that the ability to integrate information types will have important consequences for the process of publication. The ability to manipulate and analyze quantitative data will continue to transform the ways in which scientists express and exchange ideas. Most of the features of IT which are set to have a huge impact on geoscience can be seen in embryo in systems on the humble desktop machine.

References

Briner, A.P., Kronenberg, H., Mazurek, M., Horn, H., Engi, M., Peters, T., 1999. FieldBook and GeoDatabase: tools for field data acquisition and analysis. Computers & Geosciences 25 (10), 1101–1111.

Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J., 1997. Global Positioning System: Theory and Practice. Springer-Verlag, New York 389 pp.

Internet references

Graham, L.A., 1997. Land, sea, air: GPS/GIS field mapping solutions for terrestrial, aquatic and aerial settings. GIS World, January 1998. http://www.geoplace.com/gw/1997/0197/0197feat.asp.

IFLA, 1998. Citation guides for electronic documents (Style guides and resources on the Internet). International Federation of Library Associations and Institutions, The Hague, Netherlands. http://www.ifla.org/I/training/citation/citing.htm.

ISO, 1999. ISO 690-2, Bibliographic references to electronic documents. Excerpts from International Standard ISO 690-2. http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A21–A29

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part D. Familiarization with IT applications to support the workgroup

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Once geoscientists have acquired basic computing skills, the next step in IT familiarization is generally to use IT methods to collaborate within a project. The project is managed to achieve the objectives of a workgroup. Computers facilitate communication with e-mail, discussion groups and intranet links. There may be a need to formalize: standards, metadata and investigational design for all contributors to share compatible results; procedures to monitor and control the project; and document and database design to deliver a uniform product. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Project support; Workgroup communications; Workgroup documents; Metadata; Standards

1. Project and workgroup

Few of us, even as students, work alone. After acquiring basic computing skills, the next step is to look at the techniques for supporting a project and enabling a workgroup to collaborate more effectively. Communication and preliminary planning are obviously important, as are databases and quantitative models. This leads on to the wider scene, where we recognize the global scope of geoscience and the pervasive influence of information technology.

We began (part C, section 2) by looking at the desktop computer. Without external distractions, you can develop basic skills there, such as using a keyboard, a graphical user interface and basic tools for preparing text, diagrams and data files. Most geoscientists, however, must inevitably relate their own specialist expertise to the knowledge of others within a **workgroup** — a number of individuals brought together to work on a defined task. This creates additional requirements for information technology to assist in communication and coordination.

Tasks are normally handled as projects. A **project** is a managed activity with a set of objectives and a time scale, normally with identified requirements for resources of staff-time, equipment, services and information. The objective may be as small as the identification of a fossil, or as large as the production of a geological map of the world. A large project can be divided into subprojects. A very large project, say the geological surveying of the United States, might be regarded as a service activity with no final completion date. It would subsume many projects, concerned perhaps with completion of specific reports or map sheets.

A project is defined within a business context rather

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00039-X

than a purely scientific one. Consequently, the geoscience aspects would be designed and conducted differently for, say, a project estimating sand and gravel resources compared with one looking for oil and gas. **Business** is defined rather broadly as activities to meet the objectives of the organization. A graduate research study, say, would reflect the "business" of the university in terms of research and education. Knowledge of the business context in which the project is undertaken is essential for others to evaluate the results obtained and their significance in other contexts. For example, core samples from oil exploration might not be representative of the area because they were selected as potential hydrocarbon-bearing rocks, possibly obtained from anticlines selected by seismic survey.

2. Communicating in the workgroup

Communication between participants in a project can be improved by IT, and the closer coordination can make work more productive. Where it is impossible or undesirable for all participants to be accommodated at the same location, IT can offer good communication links over long distances (E 4). Tasks within a project can be divided among a range of experts, all with their own computing needs and solutions. They must combine their results and share their resources. Information must be transferred between machines and must be usable when it arrives. The benefits of working together, as well as the ability to make the results of the project available to others, therefore depends on compatibility and consistency within the project and comprehensibility to the outside world (H 2).

IT offers a variety of communication methods, from the familiar telephone and fax to e-mail, file transfer, teleconferencing, distributed computing and project management (D 5). Communication can be between two individuals, one individual addressing a group, or discussion within a group (one-to-one, one-to-many or many-to-many). The message can be actively directed at specific recipients (push), or can be made available, appropriately labeled, to be collected by anyone interested (pull). The exchange of information can be ephemeral or intended to provide a lasting record. It may consist of speech, written text, images, data files or computer programs, and be required for reading as a paper document, for display on a computer screen, for storage on the computer or on hard copy, or for computer analysis and manipulation. There may be a need for interaction, possibly requiring instant response, or possibly a considered response at a later date. It may be necessary or desirable to limit access to the information. Most of these needs are quite general, but the solutions depend on the available technology.

The attributes of conventional methods of communication are familiar. Standards and protocols are adopted, such as the behavior expected at an interview, the formality and sequence of a letter, the procedures, agenda, minutes and control of a meeting to ensure orderly discussion. Similar requirements arise in electronic communication, but may be resolved differently.

Electronic mail (E 4) communicates by means of messages entered at a keyboard. It automatically transmits the sender's name and address, and a copy can be stored on the sender's or recipient's own file, or on both. The recipient can read the message on the screen, print it, forward it to others with or without added comments, and can reply without re-entering the sender's address. Each message can be sent to an individual, members of a group, addresses selected from a database, or a complete mailing list. Mailing lists can be typed in, acquired from other sources or built up from user requests. Computer-readable documents, including data files and computer programs, can be sent by e-mail. Like the telephone, response can be immediate, or, like ordinary mail, can be timed to meet the convenience of the sender. The inevitable delays in handling paper documents, however, are partly overcome. Unlike fax, the information must be entered from a keyboard, which may make it more difficult to create a document, but is likely to make it easier to edit.

Although e-mail was designed primarily for sending text, options are available to attach images, voice and video in multi-media systems. Other special-purpose systems include **voicemail** for storing and forwarding spoken messages, and **teleconferencing** in which a group can hold a discussion through videophones without assembling physically at one location. **Usenet** provides discussion forums that are not limited to one place, time or discipline. They can be deliberately restricted to project members, to a nominated list of participants, or can be open to all. Entries can be selected, edited and controlled by the discussion leader, or circulated as received. Authors and their affiliations can be included at any level of detail, or entries can be anonymous. Those with an interest in the subject can be asked to respond, can be invited to join the forum, or the existence of the forum can be publicized, with or without an invitation to register.

Techniques like the Usenet, newsgroups and the World Wide Web (E 4) are directed primarily at global rather than local communication. Nevertheless, their procedures and protocol can be used in a restricted setting, with the advantage of compatibility with the wider world to which they may sooner or later be linked. The successful concept of the **intranet** is based on similar reasoning, offering local connections with the same software and characteristics as the Internet. These tools may assist in the preparation of documents and databases as a shared activity, with many contributors working on them together.

3. Sharing information, metadata

The ability to understand the work of others, including the language they speak, depends on a **shared coding scheme**, that is, expressions of ideas and information which mean the same to the sender and the recipient. This implies a shared background understanding. Outsiders may not be able to appreciate fully the results from a project because necessary background information is not available to them. It must be explained at an appropriate level, explaining enough for features specific to the project to be taken into account. For example, a petroleum geologist needs background information to evaluate core descriptions prepared by others during development of an oilfield. Information stating when the wells were drilled, the way the core was obtained, when it was described, under what conditions, by whom, with what ends in view, would all help in the evaluation of the description.

These aspects, however, would not be part of the data (the description of the cores). Rather they are **metadata** — data about the data — which may be helpful in their interpretation (A 1). The fundamental feature of metadata is this ability to carry information at a higher level than the data, and so assist in their understanding. Metadata are often recorded formally, as on the title page of a book (author, title, publisher, date of publication, etc.), or on the legend of a map. Also, the header of a downhole log records date, time and place, and is likely to include other information about the type of logging tool and the characteristics of the drilling fluid, all of which help in the interpretation of the log itself. There is a hierarchy of metadata. Information from the headings of downhole logs might be assembled as data in a database, with higher-level metadata referring to the wells for which the suites of logs were obtained.

The metadata that enable scientists to understand the work of others are not always explicit. Understanding often depends on the expert who can infer from past experience the significance and reliability of diverse sources of possibly conflicting information. In some cases, this expertise depends on knowledge of the techniques, procedures, personalities and local background, much of which (if only to avoid libel suits) would never be recorded. One effect of IT is to make information more widely available and thus to separate it from local background knowledge. In these circumstances, metadata that set out the constraints and limitations of data are increasingly important.

Standards may be thought of in this context as specifications or definitions intended to be generally followed, established by agreement, custom or authority, to ensure interchangeability, quality and reliability for least cost. They may be widely adopted for methods, vocabulary, instrumentation and the like. Other things being equal, a standardized approach has great benefits. Because projects have their own unique objectives stemming from their business setting, they cannot all be conducted efficiently in the same way. Nevertheless, standards that are appropriate, available, credible and relevant should clearly be used. The metadata should state which standards were followed, for these change with time. They should describe any deviations from these standards, and procedures specific to the project. If standards are unavailable or inappropriate, datasets should be described in detail, together with details of the project in which the data were collected. Much of this will be part of the project report and not specifically identified as metadata.

Another view of metadata is taken by librarians, museum curators and archivists who are concerned with formal resource description. They tend to see metadata as offering a brief description that can be used to catalog information (H 2). Yet another view is taken by the database analyst who uses metadata to bring together information about specific topics for subsequent analysis (L 5). There are widely diverse requirements for metadata, and many solutions are adopted.

4. Designing an investigation

There is an obvious need to plan any project. In some cases, the preliminary planning may be only a broad outline that expands and develops during the investigation (J 1.6). In other cases, a project based on a well-defined model may be planned in precise detail before work starts. Some geophysical studies are like this. In the project design, it is important to be aware of relevant standards, and to use them where appropriate. Documentation should describe all datasets and aspects of the projects that could assist in their interpretation.

One task of most projects is to record the salient information, selected from the vast amount that could be observed. This process of abstraction starts with the initial observations and is directed towards explaining and throwing light on topics that bear on the objectives. As described later (J 1), the outcome may be narrative and spatial descriptions and explanations, which the scientist develops through directed observation, and may involve quantitative models, which are likely to have a statistical component.

The statistical approach is mentioned here because some of the insights and vocabulary are widely relevant and because it calls for rigorous design of the investigation. The objectives of the project define the subject of interest and, hence, the **population**, or the total set of observations that might, in principle, be obtained about the subject of interest. The procedures for making measurements and observations (**operational definitions**) should be on record, to make it easier to verify the results (see Krumbein and Graybill, 1965). The objectives, the hypotheses under consideration, and past experience determine the procedures in an investigation. The procedures for deciding when and where measurements are made can be defined as a sampling scheme.

It is obviously impossible to make all possible observations, and so we seek a representative portion (a **sample**) from which we can draw conclusions about the properties of the whole population, and about the degree of uncertainty which is inevitable in making such inferences. A rigorous sampling scheme is essential to make a valid statistical interpretation of a set of measurements (see, for example, Griffiths, 1967; Davis, 1973). The measurements are expected to throw light on something specific (the **target population**). For example, if the purpose is to determine the overall content of uranium within a black shale unit and to map its regional variation, then the investigation should be designed with that in mind. Remembering that it is not possible to sample those parts of the shale unit that have been eroded away, and that it may not be practicable to sample those that are not exposed, the **available** (potentially accessible) population is greatly reduced. There may be some outcrops that are inaccessible in practice, reducing the available population further. Some will be easier to get at than others, and some samples can, therefore, be obtained at lower cost than others. A variable sampling density, if well designed, can be allowed for in subsequent analysis.

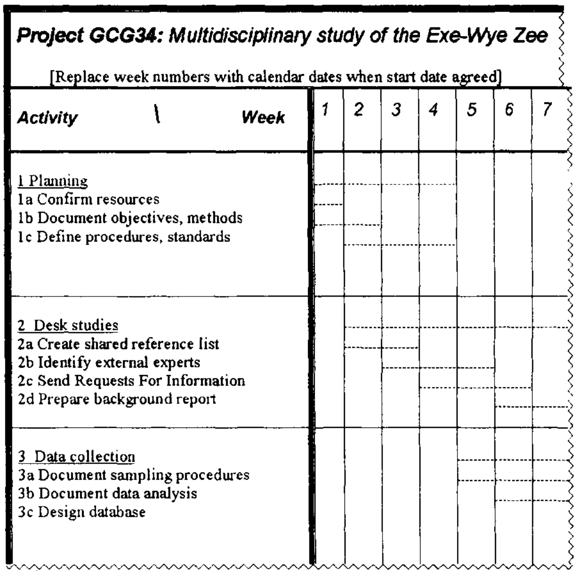
The procedures will inevitably change as the project proceeds and more is learned. The modifications should be recorded. In all projects it is helpful to ask from time to time whether the procedures of investigation introduce an unintended bias. If statistical arguments are used, do the sampling procedures give a representative, random sample? Is the sample representative of the population of interest, and the sampling density sufficient to support the conclusions? The sampling procedure should not be unnecessarily complicated, but must be devised to avoid misleading results. The procedures should be fully documented so that all participants can follow them and others can repeat the procedures to verify your results.

Another design issue that is troublesome in most projects is whether the objectives can be met with the available resources, such as time, manpower, information and equipment. This is considered next.

5. Project management

It is easier to estimate the resources required for small tasks than for a complex project. By breaking down the planned project into a series of steps, you can estimate the demands of each task separately. Records of similar completed projects may help, if you can find them. Project management software can then calculate the total resource requirements. Simple presentations, such as the **Gantt chart** (Fig. 1), help to monitor progress, and ensure that participants know what must be done in what order and on what timescale. Some systems allow individuals to maintain their own records, and combine them as a central record of progress for the project as a whole.

For a very large project or closely linked set of projects, **critical path analysis** (CPA) techniques may be useful. They are concerned with subdividing the project into a number of tasks and estimating the effort to complete each. Time dependencies between tasks are



Project GCG34: Multidisciplinary study of the Exe-Wye Zee

[Replace week numbers with calendar dates when start date agreed]

Activity \ Week	1	2	3	4	5	6	7
1 Planning 1a Confirm resources 1b Document objectives, methods 1c Define procedures, standards							
2 Desk studies 2a Create shared reference list 2b Identify external experts 2c Send Requests For Information 2d Prepare background report							
3 Data collection 3a Document sampling procedures 3b Document data analysis 3c Design database							

Fig. 1. Gantt chart for recording progress within a project. The work of the project is subdivided into tasks, making it easier to plan and monitor progress. Horizontal bars show the duration and time relationships of the tasks. "Milestones", such as dates for authorization of procedures, seminars, or delivery of reports, can be shown as symbols on the bar. Individual contributions may be identified separately and linked to the contributor's other commitments. The final documents can be extensive, and only a fragment is shown here.

identified. For instance, samples must be collected before they can be analyzed. The tasks can be placed in sequence, and by comparing estimates of the effort needed with the resources available, completion dates for each task can be estimated. The critical path through the network of tasks links those that determine the shortest overall time to complete the project. To meet that completion date, the latest time for completion of any task, whether on the critical path or not, can be calculated and the consequences of any delay can be assessed.

Despite project size increasing (thanks to IT support) few geoscience projects are so complex and time-sensitive that they warrant critical path analysis. However, the technique is of interest to geoscientists for another reason. It is a means of recording events of many distinct kinds that occur in succession within the same time frame. The time taken by a process leading from one event to another may be quantifiable, or in some cases may be unknown. Before and after relationships can be incorporated where they are known, whether or not they refer to the same kind of event. The software for critical path analysis can thus be applied to geological activities, processes and events, as well as human ones. It can be applied to stratigraphic relationships, linking events, such as the start and end of deposition of formations, tectonic events, fossil ranges, and the like. The mathematical framework is called **graph theory** and is concerned with networks, their properties and their visualization.

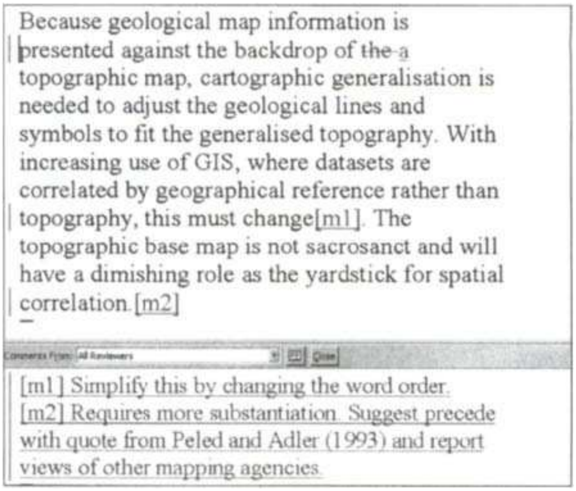
Management of a project can be assisted by IT. It provides an effective means for members of the workgroup to keep in touch. Assuming that they are connected by a network, such as an intranet, bulletins can be posted centrally, or can be distributed by e-mail, to give all concerned immediate access to developments within the project. Software is available for maintaining individual **diaries** and records of time spent on various aspects of the project. All participants can update their section of a shared **movement sheet** so that they can be contacted while out of the office, and meetings can be arranged at a time suitable for all. The computer system is being used as a device to improve social interaction. But, while the project as a whole may gain, individuals may lose some freedom of movement. There is, therefore, a risk of subversion by unwilling participants. An enforcer, such as the project leader's secretary, may be needed to ensure timely and accurate input.

6. Project documents

When a workgroup, rather than an individual, prepares project documents, word processing enables each of several authors to contribute sections, and the principal author to fit them together and edit as necessary. The editing process is made easier by facilities such as **red-lining**, which enables the editor to **markup** the document, to indicate suggested changes and add annotations without obscuring the original wording (Fig. 2).

Some sections of a project document, notably the references and table of contents, may have contributions from all the authors, and all may take part in the editing process. With a large and complex document, it may be helpful to set up a formal structure for the sections of the document, to clarify where responsibilities lie and simplify the task of building a single coherent document. The **SGML** approach to structuring a document (E 6) is particularly relevant if it is to be archived. The SGML markup language separates content from presentation (specified by a style sheet), and thus makes it possible to present aspects of the work in different ways for different audiences. It also subdivides the document into identifiable sections, which can be cataloged and retrieved separately, and possibly reused in other contexts. The production of a hypermedia document might be considered, with the possibility of linking in maps, photographs, models, datasets and video records. The potential advantages are described in L 6.3 where the drawbacks of publishing in this form are also mentioned.

IT can also provide the means for integrating the data collected for various aspects of a project. A simple database can be built up using the tools available on the desktop computer. On-screen forms can be designed (Fig. 3) for entering data on any topic. Links between topics can also be created. For example, the



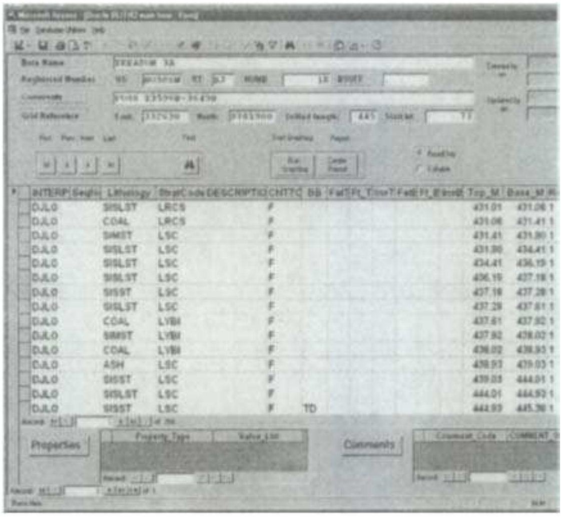
Because geological map information is presented against the backdrop of ~~the~~ a topographic map, cartographic generalisation is needed to adjust the geological lines and symbols to fit the generalised topography. With increasing use of GIS, where datasets are correlated by geographical reference rather than topography, this must change[m1]. The topographic base map is not sacrosanct and will have a dimishing role as the yardstick for spatial correlation.[m2]

[m1] Simplify this by changing the word order.
[m2] Requires more substantiation. Suggest precede with quote from Peled and Adler (1993) and report views of other mapping agencies.

Fig. 2. "Red-lining" in editing a document. This enables reviewers to annotate and suggest changes without obscuring the original wording.

results of a geochemical analysis could be related to the specimen from which the sample was analyzed, and this linked in turn to a description of a thin section cut from the same specimen. The linkage, however, must be part of the design of the investigation. The organization of the database (H 3) stems from the way in which the data are collected. The relationships between data items can be depicted on a diagram (Fig. 4) which shows types of data (entities), attributes and relationships. These **entity-relationship diagrams** can be quite complex, and can include data from many sources. The layout of a database can be generated from such diagrams using computer-aided support environment (CASE) tools. At the level of a project, however, a simple diagram should be sufficient, perhaps prepared on a computer drafting system to avoid the tedium of correcting hand-drawn diagrams.

Data are frequently stored in the form of **tables**, and can thus be treated as a spreadsheet. This structure can also be referred to as a flat file or a two-dimensional array. Quantitative data held as a table is known in algebra as a matrix (F 4). The obvious reason for this arrangement is that each column can hold records of measurements of a particular variable, and each row can hold the information for a different item. This assumes that the same variables are measured for each item. The variation within each variable can be studied separately, including perhaps their spatial distribution, but it is also possible to see how the different variables are interrelated (F 5). A number of different tables are required to cope with different topics, such as petrographic descriptions, lithological descriptions, geochemical analyses and so on. There can, however, be links between them. For example, the analyses may come from the same borehole, from the same location, or even from the same specimen. Care is, therefore, needed to ensure that the data records are not only consistent within tables, but also between tables (H 3).



INTERP	Seqn	Lithology	StratCode	DESCRIPTIO	CHT?C	BB	Top_M	Base_M	
DJLO		SISLST	LRCS		F		431.01	431.08	1
DJLO		COAL	LRCS		F		431.08	431.41	1
DJLO		SIMST	LSC		F		431.41	431.90	1
DJLO		SISLST	LSC		F		431.90	434.41	1
DJLO		SISLST	LSC		F		434.41	436.19	1
DJLO		SISLST	LSC		F		436.19	437.18	1
DJLO		SISST	LSC		F		437.18	437.28	1
DJLO		SISLST	LSC		F		437.28	437.61	1
DJLO		COAL	LYBI		F		437.61	437.92	1
DJLO		SIMST	LYBI		F		437.92	438.02	1
DJLO		COAL	LYBI		F		438.02	438.93	1
DJLO		ASH	LSC		F		438.93	439.03	1
DJLO		SISST	LSC		F		439.03	444.01	1
DJLO		SISLST	LSC		F		444.01	444.93	1
DJLO		SISST	LSC		F	TD	444.93	445.38	1

Fig. 3. Form on screen for entering data. The form, part of which is shown here, carries information at two hierarchical levels — the borehole and individual beds. It is a convenient format for displaying the data, and for authorized users to enter or edit information using standard codes. British Geological Survey ©NERC. All rights reserved.

The purpose of creating project documents is to make them available to others within or outside the project, and to organize the information for further analysis. IT opens up new possibilities of analysis, and formalizes structures for holding the data. These are considered later.

7. IT applications in the cycle of project activities

It is essential to think of IT support for projects from a number of viewpoints, and adopt the one best suited to the task in hand. It may help at this stage to look at the cycle of activities involved in a geoscience project, and indicate the types of IT support available for each activity. This may serve (at the risk of repetition) to remind the user of the available applications and how they fit together — a shop window of IT techniques where the potential user can browse and decide where to look further. They are arranged as an idealized set of activities (M 1) for carrying out a geoscience project (I 8.1), such as a gravity survey, preparing a soil map, or identifying a batch of fossil specimens.

The first activity might be to clarify the objectives of the project, determine the resources available, and plan its execution. The next activity could be to find existing, relevant information. Then data might be collected in the field or the laboratory. The data would be classified, analyzed and explained, perhaps by means of a computer model. The results would be presented, with visualization where appropriate. They would be discussed with others and their broad implications taken into account. They could then be reviewed, revised, edited as necessary, and published or otherwise made available to the intended audience.

Real life, of course, is not like that. Activities overlap and some may not be recognizable at all. The cycle of activities as a whole, and subcycles within it, may be repeated many times before the project is over. Initial results, for instance, may lead to revising the plan and calling for more resources. The scheme, which follows, is an idealized model with some features at least in common with a real project.

7.1. Planning, analysis and project management

An exploratory project by one individual may need little formal planning, ad hoc decisions being taken as the project proceeds. But a project using scarce resources and embedded in a larger investigation may require appropriate results on a tight time scale, and, therefore, need careful planning and rigid control. Computer support is likely to include word processing, spreadsheets and business graphics.

Communication can be assisted and formalized with programs specifically designed for project management. Methods include:

- Electronic mail and word processing for communication and preparation of project documents (D 2, C 3).
- Diaries, movement sheets and time planners to allocate staff resources, plan meetings and monitor progress (D 5).
- Spread-sheets to support costing and allocation of resources, such as staff time and equipment, and to monitor usage and costs (C 4).
- Gantt charts or critical path analysis to schedule tasks, identify milestones, monitor progress and adjust priorities (D 5).

7.2. Desk studies, literature search, archive search

The preliminary desk study assembles relevant existing material from available sources. Consider whether the value of old information justifies the cost of retrieving it, or whether collecting new information might be more cost-effective:

- On-line Public Access Catalogs (OPACs) can help with searching for references in your local library, or if need be, in major libraries throughout the world (H 2).
- Citation indexes can extend the search forwards in time from known sources (H 2).
- Searching the World Wide Web may yield useful information (E 4).
- Other workers in the same topic area may respond to e-mail or Usenet inquiries (E 4, D 2).
- Computer indexes to archives and repositories may be searchable remotely (H 3).

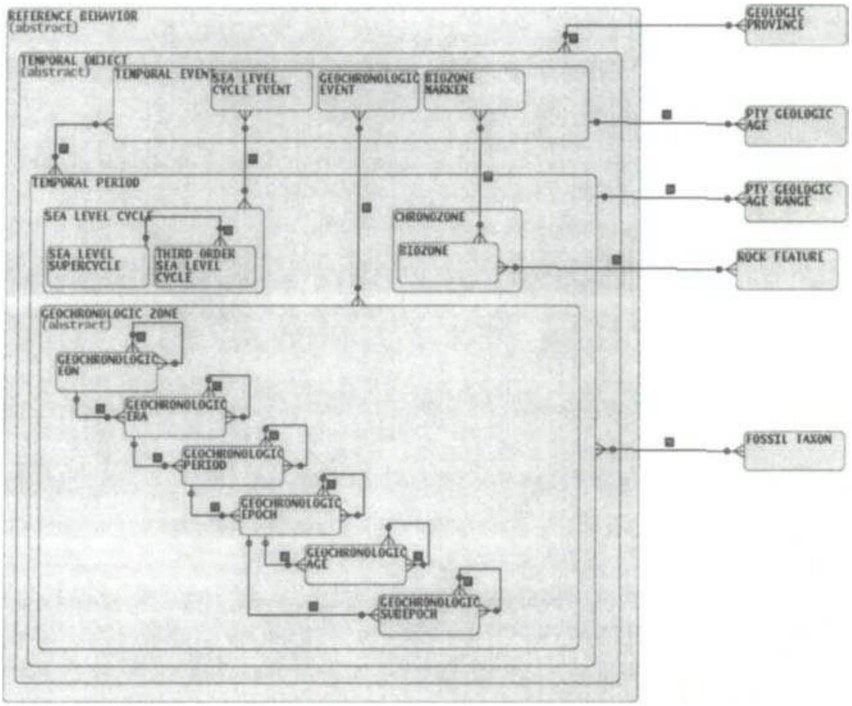


Fig. 4. Entity-relationship model. The entities are shown as boxes, in this example referring to geochronology. Aspects of their relationships are shown by the lines that link them. On pointing to the small box that appears beside each line, verb phrases appear that define the relationships. For example, each temporal period (top left) may be bound by one or more temporal event; each temporal event may bound one or more temporal period. Reproduced by permission of the Petrotechnical Open Software Corporation. More on the "Epicentre model" at http://www.posc.org/

Much of this information will have been prepared by librarians and should be in a suitable form for adding to your own lists of references (H 2).

7.3. Field and laboratory data collection

Many instruments in the laboratory or in the field, including much geophysical and oceanographic equipment, will automatically deliver digital, computer-compatible records. Data collection methods may even be adjusted automatically to conform to a responsive computer model. A quick comparison of cost, accuracy and time-saving will show whether this is worthwhile. Some thought should be given to the ultimate use of the data, and to the interfaces that enable it to reach the point where it is needed (such as a database) in an appropriate form. The computer does not necessarily make this easier. A number of alternative routes to acquiring data might be considered:

- Rigorously organized data, say for the collection of stream sediments for geochemical analysis, or for description of shallow boreholes, might be collected with preprinted forms or with prompt sheets (C 5), and the data later digitized manually or mechanically.
- The same procedure can be followed, but entering data directly to a computer or data recorder in the field (C 5).
- Electronic theodolites and range finders and Global Positioning System equipment (C 5) can assist field mapping and locating instrument stations.
- The data may be recorded for later entry, or the map can be plotted, edited and stored electronically in the field (C 5).
- Points, lines and symbols can be drawn in the field over a conventional base map, an air photograph or a satellite image. They can later be scanned or manually digitized, and the distortions corrected by computer (G 1, G 2).

7.4. Explanation, classification, modeling

Although the intuition and expertise of the human brain are essential in developing explanations, IT can assist the process by assembling, codifying, manipulating, analyzing and presenting the supporting information:

- Descriptive statistics, such as the averages of measured values, can readily be calculated, and X–Y plots drawn to look at possible correlations (F 3).
- Multivariate statistics can be computed which may throw light on complex relationships that would not otherwise be revealed (F 5).
- Numerical taxonomy offers procedures, such as cluster analysis, for classifying large numbers of items on the basis of their measured properties (F 5).
- Explanations may involve computer models (F 3, J 2.3), particularly in areas like geophysics or engineering geology, where the underlying relationships can be related to the laws of physics.
- Data analysis of the entities investigated in geoscience, and their relationships, may lead to diagrams, drawn and edited on the computer, which help to explain the structure of the information, as well as encouraging a more consistent approach (H 3).
- Explanations that would conventionally be presented as a written report can be given greater depth with hypermedia. Through access to additional background, such as video demonstrations, the reader can link the explanation to the supporting evidence (J 1.5).
- Techniques derived from studies of machine intelligence can formalize aspects of geoscientists' thought processes (L 5). These can be built into expert systems that can then apply the reasoning to other geoscience information.

7.5. Visualization, presentation

Geoscience is concerned with spatial processes and their interaction with geological objects through geological time and space. The importance of maps, cross-sections and block diagrams is, therefore, not surprising. Computer cartography plays a large part in the production of the maps. They can be regarded as an aspect of computer visualization, a subject that explores the application of graphical methods to the understanding of data.

- Bar charts, pie charts, and x–y plots (C 4) can help the user to grasp the relative frequencies and correlation of variables.
- Digital cartography and spatial models can show the pattern of variables in space, their spatial relationships and spatial correlation (G 1, G 2).
- Geographic information systems (L 4) and visualization systems (E 5, G 7) provide a more flexible means of displaying two- and three-dimensional relationships than the conventional approach of examining and overlaying maps.
- Hypertext and hypermedia systems make it possible to combine and cross-refer between text, images, models and map information in a more flexible manner (L 6).
- Portable display systems make it possible to present live demonstrations of multimedia to a large audience, through a suitable projector (C 6).

7.6. Reconciling information and aligning ideas

A significant part of the time spent on an investigation may be devoted to resolving conflicts between differing views, possibly within a project, possibly between ideas arising from different projects. The IT contribution to these debates is to provide discussion forums, with faster response and greater convenience, accuracy and global reach than conventional methods:

- E-mail (E 4), the Usenet (D 2), and the World Wide Web (E 4) are means of communication which meet different needs.
- The process of seeking the views of others, as in tendering for new facilities or in setting standards, can be formalized as Requests for Technology and Requests for Comment. A good example is the procedures followed by POSC (L 5).
- Digitized information from several sources, such as photographs, satellite imagery, and geological and geophysical maps, can be linked, adjusted to fit, compared and integrated on a computer screen (L 4).
- **Teleconferencing** allows a small group to see one another on screen and participate in the same discussion from different locations, or can make it possible for an individual to address a group of any size from another location. The advantage over a video recording is that the speaker can respond immediately to audience reaction and questions.
- IT encourages the separation of metadata and standards from other information (L 6.1), thus assembling key reference information where all can consult it. In this way, greater consistency of data should be achievable, and any disputes about nomenclature or standards can be placed immediately before the relevant authority or submitted to an appropriate forum.

7.7. Review, revision, editing

There are obvious advantages to scientific editors and publishers in receiving material by disk or e-mail. It can be forwarded without delay to referees for consideration or comment, without the cost and inconvenience of handling and mailing photocopies. Comments can be marked up as an integral part of the text (D 6), and alternative versions directly compared. The author can incorporate agreed changes without retyping the rest of the text. The publisher can pass the finished work directly to the plate-maker and avoid the process of rekeying with the inevitable errors and additional corrections that this must introduce. Most publishers must obtain or prepare computer-readable copy for their printer's phototypesetters, and may also wish to make an electronic version of the paper available to customers as an alternative to paper (M 2.1).

Documents prepared within a project, and multi-author papers generally, can benefit from IT methods, particularly if the authors are geographically dispersed. E-mail can be helpful in exchanging ideas rapidly, but it is also possible to create project-centered documents (D 6) which are accessible to all the authors, with agreed protocols for reading, writing or amendment. Similar multi-author procedures for describing, drawing, reviewing and amending diagrams and maps are possible using GIS software (L 4).

Documents, including maps (G 1), that are subject to rapid change and development, may not be published conventionally, but instead an electronic record can be archived on the computer and kept up to date. When a copy is required, the latest version can be made available. Changes can be logged and earlier versions recreated if need be. In-house documents, and those with limited circulation, may never be published conventionally, but can simply be stored on the computer for access on demand.

References

Davis, John C., 1973. Statistics and Data Analysis in Geology: with Fortran Programs. Wiley, New York 550 pp.

Griffiths, J.C., 1967. Scientific Methods in Analysis of Sediments. McGraw-Hill, New York 508 pp.

Krumbein, W.C., Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill, New York 475 pp.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A31–A39

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part E. Familiarization with IT background

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

The geoscientist who wishes to move beyond basic techniques and day-to-day IT applications must know something of the underlying concepts and vocabulary of IT. Communication is vital, linking your desktop to the world. Generic computing tools, widely used in geoscience, can handle documents, geographic information and database management. More basic tools, including long-established programming languages like Fortran, retain an important niche. Recent developments, such as Java and a range of markup languages, bring new flexibility and precision to the geoscience record. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Generic software; Computer communications; Programming languages; Markup languages

1. The need to look at the IT background

Geoscientists may have gone a long way to meeting their employers' immediate needs when they are familiar with the ways of working of a desktop computer and the software required for their projects. To move ahead, however, they must be positioned to meet future demands. This calls for a fuller understanding of some underlying concepts and some more advanced techniques that are now widely used in geoscience.

It is a big step to use a machine to help us organize our knowledge, and we should be aware of the ideas, largely from mathematics, which make this possible. One of the problems, and opportunities, of using computers, is that they manage and manipulate information in a different way from human beings. Some applications mimic earlier technology. Others, like quantitative modeling, are practicable only with a computer.

2. What computers do

Computers count. They can add two numbers together. They can compare two numbers and decide which is the larger. They can carry out simple instructions, such as: store the value of a number. They can store, retrieve and act upon a sequence of simple instructions, such as: obtain two numbers from specified locations, add them together, compare the result with a total calculated earlier, store the larger of the two totals. Because they can do such things, they can perform the full range of mathematical operations that reduce to a sequence of additions, such as subtraction, multiplication, exponentiation, converting to logarithmic or trigonometric functions.

Computers can be connected directly or through the telephone or other network, and data can be passed

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00040-6

from one to the other. Their striking characteristic, however, is not the complexity of the underlying ideas, but their extreme simplicity. Their power stems from an ability to perform very large numbers of simple operations, quickly, cheaply and accurately. They thus harness the power (and reflect the limitations) of mathematics without the need for laborious manual calculation.

Their importance in geoscience comes from the relevance of numbers and mathematics and from the ability to tie into advances in electronic engineering.

- Text characters can be coded as numbers. By ensuring that the numeric codes follow a widely accepted standard for all computers (the ASCII code), computer text can be exchanged. Because the codes get bigger numerically in alphabetic order, text can be arranged and selected alphabetically. A few more steps lead to the electronic library.
- Points in space can be coded as numbers, using coordinate geometry. They can be combined as complex images, such as photographs or satellite imagery, or as geometric objects in 2, 3 or more dimensions, such as the lines and surfaces depicted on a geological map. A few more steps lead to the computer-based spatial model.
- Numeric and graphical data, like a geochemical analysis or a downhole log, can be recorded, selectively retrieved, analyzed and graphically displayed. As standards are implemented, global integration of data can follow.
- Processes in geoscience can be modeled by mathematical operations represented by computer programs. Together with global data, they can be assembled as a more complete representation of aspects of knowledge.
- Text, imagery, spatial information, data, processes, telephone, video and audio can be linked in a hypermedia representation of the recorded geoscience knowledge base.

Electronic engineers and computer scientists have provided the tools. Progress in their application depends on experts in subject fields, such as geoscience.

3. The computing system

Some knowledge of IT methods and procedures is essential to understand the developing technology which now pervades the geoscience information life cycle. We can start with some general, obvious and basic definitions and concepts. Computing equipment, including processors, memory, disk storage, printers, display units and communication facilities, is sometimes referred to as **hardware**. This is distinct from the **software**, which includes the operating system, compilers and interpreters, and applications **programs**, which specify a sequence of computer operations to meet the needs of the end user. The program is run by the **processor**, which executes, or carries out, the instructions under the control of the **operating system**, making comparisons and performing elementary arithmetic operations as required. The necessary parts of the program, data, and final results are stored in **main memory**. Information that is too bulky to fit in main memory or will be required for a later session is held in **secondary memory**, such as disk storage.

We can, for convenience, think of the computing system as three subsystems: process, repository and interface. Data items are manipulated by **processes**, which follow a set of instructions supplied in the form of a computer program. The computer is designed to carry out a basic instruction set. This contains instructions for such tasks as moving an item of data from one location to another or performing simple mathematical operations on the data — add A to B, move A to B, compare A and B, and so on. Assembler language, which may be specific to the type of computer, is the means by which systems programmers, helped by a systems programming language such as C, can build up a program as a sequence of basic instructions to carry out a specific task. More complex instructions are written by applications programmers in a **high-level language**, such as Fortran or Basic. This is translated by a compiler or interpreter into the more basic instruction code with which the machine can operate. A **compiler** produces a coded version of the entire program, which can be run efficiently whenever required. An **interpreter** translates and runs the program line by line as it is entered, thus allowing greater flexibility for the programmer at the expense of more work for the computer. An application program is based on an **algorithm** — the set of rules to be followed to solve the problem.

Data, the records of observations and measurements, contain many individual values. Several values referring to different properties of one thing of interest are known as a **record**. Many records from many related items could be regarded as a **database**. At one time, a database was seen as an all-inclusive set of connected records for an organization. Inevitably, however, many distinct collections of data are put together for different purposes, and so we have a collection of databases known as a **repository**. There may be weak connections between the databases that were not realized or not taken into account when the data were collected. The repository might then be termed a **data warehouse** and special "data mining" programs devised to decipher the links between the various datasets. Geoscience data are often recorded as a table, sometimes known as a **flat file**, in which each vertical column refers to a particular property or variable, and each horizontal row contains the values for a specific item or **instance** (see part C, Fig. 1). A set of such tables, in which items are cross-referenced through key fields, and which are structured according to rules which reduce needless repetition of data, form a **relational database**, widely used in geoscience.

Developments such as **hypertext**, in which cross-references are embedded in a document, enable the reader to call up a related reference in the same or in another document by clicking on a highlighted word. **Hypermedia** extends this concept to include references to images (which may have clickable highlighted areas), video, audio, discussion groups and computer processes. Tabular, quantitative datasets and the associated relational database management systems no longer dominate computer information. This greater flexibility is supported by the **object-oriented** approach (H 5, J 2.4). A thing of interest is referred to as an object, which can be a data table, document, image, or any combination of hypermedia. The **object** is a self-contained entity and may include within it the processes or references to processes that are appropriate for manipulating the data it contains. Objects are placed within **classes**, which are structured as a hierarchy, and **inherit** attributes, and relationships to other objects, from classes at a higher level.

The **interfaces**, where parts of the system join, are often of interest. Of particular interest is the user interface, through which the user communicates with the computer and vice versa. For many tasks it is convenient for the user to type in sequences of instructions. Much communication, however, is now through a **graphical user interface** or **GUI**. This uses windows, icons, menus and pointers (**WIMP**). The **windows** are rectangular areas on the screen with a separate process (program) running in each. By **pointing and clicking** with the mouse the window can be moved, resized, hidden behind other windows or made visible by placing it on top, reduced to a small icon, enlarged back to full size, or closed to remove it completely from the screen. The actions of the computer can be initiated by typing instructions in the window, clicking on items in a **menu** (list of options) or on **icons** (small symbols that indicate pictorially what actions will result).

The interface between the repositories, where the objects are stored, and the user environment, in which they are assembled and processed, also deserves some attention. The users' application programs may be linked to the data through an **applications program interface** (**API**) which is compatible with both. If appropriate standards are followed, finding the required objects can be delegated to an **object request broker** (**ORB**), a program, which is part of the **middleware** (L 2) between client and server (E 4) and runs partly on each.

The purpose of the complexity is to enable operating systems to cope with the number and diversity of available sources, while providing the user with the ability to integrate at the desktop the numerous objects of interest from a multitude of sources (**distributed objects**) while retaining ease of use. Underlying the access to distributed objects is the ability of computers to communicate.

4. Communication

Scientists working on the same project have generally tended to be in close proximity, often in the same building. This facilitated discussion and sharing of information. Over these short distances, it is economically feasible to connect computers with high bandwidth coaxial cables or fiber-optic cables, thus giving rapid data transfer. The **local-area network** (**LAN**) built up in this way can be supported by powerful software. A wide range of computers and their operating systems are designed to be compatible with such software, which can support a large network of many hundred devices. A small office with only a handful of users can be networked with simpler systems at lower cost. As the network grows, the task of designing and maintaining systems becomes more complex, and an expert may be required to ensure that it is robust and works consistently.

Local area networks can be linked together through the worldwide network of networks — the **Internet** (D 2). Its **protocols** (the rules, definitions and conventions that govern a cooperating activity) can also be used on a local network, thus ensuring that the in-house network or **intranet** has the same characteristics, and can use the same software, as the Internet. For example, Web browsers designed for global communication can also be used locally. The cost of providing high capacity links over long distances is obviously much greater than that for links within a building. The Internet has been in existence for many years since it began as a research project of the US government. But it is only in the last few years that faster modems, better compression techniques and better software made it practicable to connect home or office computers through telephone lines, fiber-optic cables, microwave and satellite transmission. Telecommunications companies generally provide the physical links. **Internet service providers** (**ISP**) may contract to use some of this transmission capacity, and resell smaller amounts, together with appropriate software and services, to local businesses and individuals.

The emergence of third-generation mobile phone technology is freeing communication from physical connections (International Telecommunications Union, 1999). Broadband **wireless** links are made practicable by **cellular radio**. The area to be covered is divided into smaller patches called cells, each served by a low-power transmitter. The same bands of the radio spectrum can be used in different cells. A computer tracks all subscribers, handing them over from transmitter to transmitter as they cross each cell boundary. The wireless industry is likely to agree global standards in the early years of this century, and the overlap with the computer industry must increase. Geoscientists in the field, and remotely-controlled devices, will be fully linked to the information system.

Where one computer is supplied with information by another, the two computers are known as **client** and **server**. The server may be configured for this specific purpose and may supply several client computers with data and programs on request, possibly over a local area network. The server can be managed by specialist staff within the organization to ensure that secure, up-to-date information is available. A client computer, such as the one on your desk, may also access remote servers across a **wide area network**, to obtain information that is not available locally. The GUI (E 3) can develop into a network user interface. This also has a simple point-and-click procedure to select actions, but the actions are not confined to the local computer and windows can be connected to a remote server. This is achieved by means of a **Uniform Resource Locator** (URL), which is a form of address standardized within the Internet. It identifies the servers, of which a central list is maintained, and the file names, which are assigned locally. The URL also has a prefix indicating the protocol in which the contents will be transmitted, and a suffix indicating their format, as described later in this section.

Standards are essential to ensure that the communicated information is meaningful to the recipient. The Internet works because standard protocols (**TCP/IP**) are used throughout. The Internet Protocol (IP) defines the routing between computers. The Transmission Control Protocol (TCP) defines how data are wrapped in packets for IP to transmit. Other protocols, such as **NFS** (Network File System) and **HTTP** (Hypertext Transport Protocol) are compatible parts of the TCP/IP suite. Most modern operating systems provide links to these protocols. A computer can be linked to the Internet, or to an Internet Service Provider, through a **modem**, a device that, by modulating and demodulating the signal, allows computers to communicate over telephone lines. Where available, an **ISDN** link (Integrated Systems Digital Network) may offer higher speed at greater cost.

A local network can be linked to the Internet through a **router**, a computer dedicated to controlling the traffic between the network and the outside world. Security is always a problem with networked equipment, where interlopers prowl in search of passwords, credit card numbers and the like, in the hope of being able to obtain and possibly interfere with information to which they are not entitled. It may, therefore, be necessary to have password protection on all shared resources on the local network, as well as ensuring that password protection is adequately enforced on all machines connected to the Internet. The router may be connected to a separate computer, which has the task of maintaining security, providing a **firewall** between the local network and the outside world. Each device that can be accessed on the Internet has its own unique identification number (IP address) provided through the ISP or by the Internet Information Center. For most geoscientists, arrangements for networking are handled by the local computer communications manager, who is responsible for organizing and maintaining the local network.

TCP/IP is an example of an **open standard**, agreed by national and international standards organizations such as ANSI and ISO, and available for all manufacturers and suppliers to follow. There are also many ad hoc and **proprietary standards** that have been defined within a company, such as the Windows standards defined by Microsoft. The specifications of some proprietary systems, such as IBM's PC-DOS, have been put in the public domain. A consequence is the availability of compatible personal computers and software from many suppliers.

Personal computers can be self-contained, and if users are concerned only with their own computing, communication may be unnecessary. Even at this level, however, it may be advantageous to download data and programs from a central server rather than storing all that may be required on the local machine. Maintaining an adequate range of material in up-to-date versions can then be the responsibility of the systems manager. Workgroup computing requires a degree of interaction between the participants that demand good communication. Geoscience is a worldwide activity, however, and to take full advantage of the potential benefits, global communication is called for.

Fortunately, the means of communication are available. They take a number of forms (D 2). The most widely used means of communication, accessed by many tens of millions of users, is electronic mail (**e-mail**). The message is generally in straightforward text. The e-mail address of the intended recipient may be hard to find, as there is not always a reliable equivalent of the telephone directory. Large files or those with a more complex format, such as computer graphics or documents with a complicated layout, may be better sent by file transfer protocol (**ftp**). This involves establishing a two-way link before transmission, and the complexities are normally concealed from the user behind a simple **drag-and-drop** operation (using the mouse to move an icon from one point on the screen to another). Shared documents that are being worked on by a collaborating group might use a format suited to workgroup activities, such as MS-Notes. Discussion groups can follow **Usenet** protocols that can be found through Web search engines. Documents for the world at large can be prepared in hypertext markup language (**HTML**) (E 6) and made available through the World Wide Web.

The **World Wide Web** (WWW) consists of many millions of pages stored in standard formats on numerous servers throughout the world. It can be accessed through a Web **browser** — software that runs on desktop client computers, and allows users to make general searches, follow links, and display documents held on the Web. The Web pages are distributed across a wide range of servers and are connected through links that are embedded in the pages. The link appears to the user as a highlighted phrase in a text document or area on an image. Normally concealed from the reader, but embedded in the text at that point, is the address of a point in the Web pages in the form of a URL (Uniform Resource Locator). It looks something like this ⟨A HREF = "http://www.bgs.ac.uk/bgs/w3/free/reports.html"⟩ *text here* ⟨/A⟩

The first item (**tag**) enclosed within angle brackets indicates the start of an **anchor**, which is the link to another document or to a point in a document. The second set of angle brackets ⟨/A⟩ indicates the end of the anchor. On the reader's screen, the text within the anchor is highlighted (usually by printing in a different color) and underlined to indicate that it is a "hot-spot". Placing the cursor within the anchor changes the icon, typically to a pointing finger, and clicking activates the anchor. The HREF attribute contains a parameter within quote marks indicating the transfer protocol (here, http means hypertext transfer protocol), the name of the server (www.bgs.ac.uk), the path and name of the document (/bgs/w3/ ... indicates the directory and the file name) and the format (html means hypertext markup language). Optionally, it can move to a location marked by a flag in the original document. Clicking on the hot-spot causes the specified Web page to be retrieved from the server computer, and displayed on the screen at the flagged point.

The server name indicates the country name (USA if none is specified), preceded by the type of organization, such as com or org for a commercial organization, edu or ac for academic community, gov for government organization, and so on. This is preceded by an abbreviation for the name of the organization (bgs for British Geological Survey) and the name of the computer (here, www is the web server). This "domain name" identifies the specific server and is registered with the **domain name server** (DNS) which links the domain name to its unique IP address.

In addition to retrieving hypertext documents, as has just been described, anchors can point to other places in the document, or can access images. These are held in other formats such as .gif or .jpeg, rather than .html. This information is included in the anchor and is used by the browser to display the image correctly. Audio (.au) and video clips (.mpeg) can also be accessed from an anchor. The flexibility of this hypermedia system can be increased further by using the anchor to link to a computer program. This can then request information from the user through a simple form, and can perform operations such as searching a database and listing retrieved items on the screen.

Like many facilities accessed from the desktop, the Web contains its own documentation. The facilities it offers are rapidly expanding. Rather than attempting a description here, it is better to explore the documentation of your own installation. An up-to-date account of the range of facilities is available. There are also guides to authors, which describe the many types of tag that appear in angle brackets. They are normally hidden from the viewer, but control the appearance and structure of the page. Information can be obtained by following links from the ISP, the Web search engines or Web developers, such as the W3 consortium. Geoscientists can readily find their way to lists of relevant sites on the Web by using a search engine to find entries dealing with their own specialist subject. Alternatively, they can look at the Web pages of organizations such as university departments or geological surveys which provide links to related sources (Ingram, 1997; Butler, 1996). If you are new to the task, a demonstration from a local expert familiar with the system can be very helpful, but in the longer run there is no substitute for experience.

5. Generic software systems

Information comes in various easily-recognized types: text (the ordinary language used in most documents); spatial or graphical information (such as that found in maps and diagrams); structured data (like the tables of data in a database); and information like video or audio records that are less frequently found in this context.

Conventionally, information products have one predominant information type, as in the case of books and serials, maps, data files, video tapes. Major systems of computer software, mentioned in this section, also tend to focus on specific information types. These generic systems are designed to perform operations analogous to familiar actions with conventional products, such as: go to page 52, center the map on this latitude and longitude, select data where a specified variable lies within a given range. The metaphors make the integrated systems easier to use, and they now provide most of the general computing tools for geoscience.

The close links between information types and software systems suggest that they might give a good basis for organizing a course (or a book) on geoscience computing. This structure has not been followed here, partly because of the belief, expanded in J 1.8, that we should break away from these traditional divisions and explore ways to integrate all the information types that have a bearing on an investigation. Links among generic systems are being built into many of the more recent products, easing the task of integration.

Text documents are now generally prepared on a word processor. If they are subsequently published, they will be indexed in numerous computerized library catalogs, but only a few geoscience documents are at present archived as full digital records. For those that are, a markup language or a standard format (E 6, L 3) can ensure that their content can be organized and printed appropriately by computer. **Document management software** is available to manage and retrieve items from a repository of such documents.

Spatial information, which would normally be recorded on maps and cross-sections, can be managed and manipulated on the computer by a geographic information system (**GIS**). The GIS makes it possible to establish, manage, analyze and display a database of cartographic information. Contouring programs can interpolate three-dimensional data and display them as contour maps and cross-sections. Image-editing software can manipulate and adjust other images, such as photographs and satellite imagery. Computer aided design (**CAD**) and scanning software help to capture data and draw maps and diagrams. Visualization programs present datasets graphically, to make it easier to see the relationships between variables.

Structured data benefitted from computer methods at an early stage in the development of IT, as they could be handled relatively easily and cost-effectively with long-established programming languages, such as Fortran. The tabular layout is appropriate for much geoscience data as it enables like to be compared with like, and is well-suited to computer analysis. Relational databases fitted this layout well, extending it to keep track of complicated relationships. Relational database management systems (**RDBMS**) provided the means to separate **data management** (input, editing, deleting, updating, selecting, sorting and retrieving) from subsequent analysis and presentation. Statistical analysis and spreadsheet software make it possible to explore the properties and relationships of the data, and other quantitative models throw light on the underlying physical relationships.

Processes or computer programs are generally seen as distinct from the data, so that they can be reused with many datasets, while one dataset can be analyzed by many processes. This separation is not always appropriate, as some data are dependent on a particular process for their interpretation. For example, data points chosen to be representative of surfaces or lines on a map may recreate the original only if a specific process is applied to them. In an object-oriented system (H 5), objects are seen as linked data and processes, both of which, however, should remain reusable in other contexts.

Video and audio records have not been widely used for storing geoscience information. Now that they can be readily linked to hypermedia, however, there is considerable scope for their use in demonstrating, say, the appearance of a rock slice when rotated under crossed nicols, or a picture of a soil profile at the time of excavation. Specialist software is available for compressing these files to reduce their large size for storage or communication.

6. Programming languages

The importance of programming languages to the average user is diminishing. In most applications, user costs greatly outweigh machine costs. Building on the existing software repertoire is preferable to writing new programs from scratch. For most users, the well-established and commercially available generic systems, together with specific application programs, are sufficiently flexible to meet their needs, and it is more economical to buy than to build. Effort in selecting and understanding existing systems may be more rewarding than gaining skills in a programming language. In these circumstances, it is questionable whether it makes sense for a geoscientist to become a proficient programmer. The learning overhead is considerable, and practice is needed to remain fluent.

Most commercial systems deliberately hide the programming code from the user, and the task is to learn the idiosyncrasies of the system and the means of achieving the desired results. Until recently, software systems tended to be compartmentalized, often in a deliberate attempt to prevent the user's escape to rival systems through importing or exporting data. Programming skills made it easier to cross the interface. This is now less of a requirement as it is easier to find an exchange format supported by both systems.

However, good reasons remain for learning a computer language. For the applications programmer, a geoscience training supplemented by programming skills is a powerful combination. In areas like the development of quantitative models, the needs of the individual or the organization may be so specific that only home-made code will do, detailing the programmer's instructions step by step. In other cases, standard software may handle many of the tasks, but programming may be needed for specific additions. There is a large amount of existing code written within organizations or available from colleagues or the literature, for example, Press et al. (1992) and Universal Library (1999). You need programming skills to modify it for the task in hand, or to keep it up to date. Extensive libraries of high-quality subroutines are available for mathematical and statistical analyses, notably in Fortran. They can be included in your own programs. It can also be argued that programming skills provide a deeper understanding of how the computer works and thus of how methods can best be developed in future. A look through journals such as *Computers and Geosciences* (1997) suggests that extending the range of applications calls on an ability to program.

For most users, it is worth knowing something about computer languages in general, as they have much in common. A short course in one language could also give useful background. Languages you are likely to encounter include Fortran, Pascal, Basic, C, C+ + and Java. This section offers a very general introduction for the non-programmer.

The languages just mentioned are **procedural**, setting out line by line the sequence of procedures which the computer is instructed to follow, as opposed to stating the objectives and leaving the computer to select the method, as in SQL (mentioned later in this section). They deal with variables and resemble familiar algebraic formulas, such as $x = 1/2(y+z)^2$. In **Fortran** one might write $X = 0.5^*(Y+Z)^{**}2$. This, however, is not stating an equality. Rather it is indicating that the right-hand-side should be calculated, and stored in a variable called X. Perhaps = should be read as "becomes" rather than "equals". The meaning of the Fortran statement could be interpreted as follows: the names X, Y and Z refer to storage locations; if the names have already been used in the program, look up their locations, otherwise assign new locations for them; take the contents of Y and Z, apply the arithmetic operations indicated and store the result in X. The / denotes division, * multiplication, and ** raising to a power. Variables are usually given names which the programmer can remember more easily than X, Y and Z, thus: Distance = Time*Velocity. Data in sequence, as in time series or tables, are conveniently denoted by suffixes in algebra: y_{10} is the tenth measurement of y, y_i is the ith. Similarly, in Fortran, Time(5), Time(I), or Height(I, J) would represent the fifth and Ith measurements in a series called Time, and the entry in the Ith row and Jth column of a table of measurements (an **array**) called Height.

It is often necessary in a program to apply the same type of operation to each member of a series in turn. Rather than writing out each operation individually, it is written once, using **index** variables such as I and J rather than numbers. It is then placed within a **loop** which is an instruction to perform the operation, or set of operations, with stated values of indexes. In Fortran, it might look like:

```
DO I = 1, 15

    sequence of statements (operations)

END DO
```

The sequence of operations is performed from the beginning to the end, in this case 15 times. The variable I, which could also be the index of variables in the statements, takes the values 1, 2, 3 ... 15 in successive loops. To give the necessary flexibility, the programmer can cause control to jump to another point in the program under defined conditions. The command can be **conditional** on a variable having a particular value or a value within a certain range.

```
IF (Height(I, J) < 500.0) THEN
```

would indicate that control would pass to the next statement if the value of Height(I, J) is less than 500. Otherwise, control passes to a later statement that begins with the word ELSE.

It is thus possible, even without knowing much about a programming language, to get some idea of the calculations by looking at a program. Generally, one statement goes on one line, but & indicates that it continues on the next line. The ; separates short statements on the same line. Comments are generally inserted to explain the program to anyone reading the code. They are introduced by ! and continue to the end of the line. They are ignored by the compiler.

A surprisingly complex set of calculations can be built up from these simple basic building blocks. As the same set of operations can be useful in many different applications, they can be written as a self-contained **subroutine** or **procedure**, which is given a name and a means of indicating the variable on which it is to operate. Thus, Subroutine Sum(X, N) might be written to calculate the total of the first N values of the series called X. The subroutine can be invoked by a statement in another routine, such as Call Sum(Time, Number). Calling the subroutine is equivalent to repeating all the codes of the subroutine at that point.

Statement are also required to instruct the computer to acquire data from a particular source, or send it to a particular destination. It might, for example, request the user to enter information from the keyboard, or might read it from a disk, or send output to a screen or printer. The READ and WRITE or PRINT statement in Fortran indicate the variables holding the information, and where the data are to be acquired or delivered.

Fortran is a long-established programming language for scientific use, which has undergone substantial improvements over the years, and is still widely used in geoscience. The huge investment in existing programs and expertise mean that it is likely to remain in use for some time. It is a powerful language capable of representing complex tasks in numerical calculation. It is a reasonably tolerant language, allowing programmers to express the same idea in different ways, some inherited from earlier versions of the language. Programmers can consequently fall into bad habits which make their programs difficult for others (and themselves) to understand and to maintain or modify. For training purposes, therefore, a simpler language such as **Pascal** may be better because it takes a more stringent view of the way the sequences of commands (program **code**) are presented. It thus forces the user to acquire better programming habits. For less complex tasks, **Basic** in its various forms lacks the power of some other languages, but is simpler to learn and to run. Basic is interpreted, rather than compiled like Fortran (E 3), and it is therefore possible to spot mistakes as each statement is written, and the programmer can correct them before proceeding. Visual Basic is widely used to give programming flexibility in a desktop environment.

The WIMP graphical user interface (E 3) is currently the norm for the desktop computer, and is a more recent development than Fortran. The interface is handled at a deeper level in the computer software than the applications which were just mentioned, and special languages such as **Motif** have been written to help the programmer to organize the objects, such as the windows, cursors, icons, and menu-bars, which appear on the screen. More generally, languages such as **C** provide the basic facilities to access the systems functions of the computer. **C++** is its object-oriented counterpart. Programming at this level is a specialized activity. It implies a need to modify or extend the standard functions supplied by commercial systems, which may be as likely to confuse as to help the average user.

A number of other specialized languages deserve a brief mention. The success of the World Wide Web has encouraged some language developments. **Java** is designed to operate within a virtual Java environment. In effect it runs in its own operating system on the desktop client. The server supplies information to the client, including "**applets**" or small applications — processes or programs which operate on the information. The entire object, data and process, is thus supplied from the server. A simple, low-cost client can take full advantage of the server's power. Furthermore, the client can access a wide range of servers worldwide, receiving and combining applets from them all. The drawback is the heavy communications load and inefficiency in the handling of the data. **Perl** is another language which is widely used on the Web, for bringing to life information delivered by the server.

Markup languages place, within a document, symbols which can be read and operated on by appropriate systems. Thus a text report or document can be marked up to identify topics or the various sections, such as title, abstract, chapters, sections, paragraphs, references, or illustrations. The Standard General Markup Language (**SGML**) has been used in this role for some time (Seaman, 1999). The advantages of subdividing a document in this way are considered in D 6. Here, it should be mentioned that **HTML**, the hypertext markup language, is a subset of SGML which is used in many Web documents (E 4), and that **XML** (extensible markup language) has recently been developed as another simpler subset of SGML, with more powerful facilities than HTML. Markup languages can also be used to subdivide three-dimensional graphical objects using **VRML**, the virtual reality markup language.

Postscript is a page description language, describing the layout of text and images on a page, in a form that can be edited or modified. The Postscript files which it generates are widely used in medium to high-quality printing. **Acrobat** offers some of the features of HTML while preserving the page layout in a portable data format (**PDF**) (Kasdorf, 1998). **LISP** (LISt Processor) is another language used with text and graphics, where the information is stored as a consecutive sequence (string or list) of characters or of points on a line. It found an important niche in word on machine intelligence, and has been used in cartographic and work processing applications. Structured Query Language (**SQL**) has been widely adopted as a standard interface for querying relational databases (H 3). The advantage of this standard interface is that information can be spread across several databases, each with their own data management systems, and can still be processed by many clients. Communication is made possible by adhering to the SQL standards.

Computer languages can thus be seen as rigorously defined interfaces between the application and the operating system (Fortran, C), the GUI and the operating system (Motif), the client and the server (HTML, Java), the document and the printer (Postscript) and the database and the application (SQL). Numerous other languages, such as APL, Cobol and Ada have played their part in geoscience applications, but introduce no new ideas at this point. Special-purpose languages are available for some software products, enabling the user to modify or customize the products, without compromising the original code.

The computer can follow with speed and accuracy a set of rules expressed as the instructions for executing an algorithm. It lacks the capacity to understand the underlying reasons or to make decisions about unexpected results, tasks at which human beings are much more adept. The systems analyst and user must therefore decide what can better be done by machine and what should remain the task of the scientist. The best features of both can be combined in an interactive system (J 1.6) where the user can keep track of progress and guide the computer in its operations.

References

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in Fortran — the Art of Scientific Computing, 2nd ed. Cambridge University Press, Cambridge 963 pp.

Internet references

Butler, J.C., 1996. Another node on the Internet for those with interests in geosciences, mathematics and computing. http://www.uh.edu/~jbutler/anon/anon.html.

Computers & Geosciences, 1997. Computers & Geosciences Online. http://www.elsevier.nl/locate/compgeosci.

Ingram, P., 1997. The Virtual Earth: a tour of the World Wide Web for earth scientists. http://atlas.es.mq.edu.au/users/pingram/v_earth.htm.

International Telecommunications Union, 1999. IMT 2000: A vision of global access in the 21st century. http://www.itu.int/imt/.

Kasdorf, B., 1998. SGML and PDF — why we need both. Journal of Electronic Publishing 3 (4) http://www.press.umich.edu/jep/03-04/kasdorf.html.

Seaman, D., 1999. About Standard Generalized Markup Language (SGML). http://etext.lib.virginia.edu/sgml.html.

Universal Library, 1999. Numerical recipes on-line. Hosted by Carnegie Mellon University. http://www.ulib.org/webRoot/Books/Numerical_Recipes/.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A41–A50

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part F. Familiarization with quantitative analysis

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Numbers, measurement and calculation extend our view of the world. Statistical methods describe the properties of sets of quantitative data, and can test models (particularly the model that observed relationships arose by chance) and help us to draw conclusions. Links between spatial and quantitative methods, through coordinate geometry and matrix algebra, lead to graphical representations for visualizing and exploring relationships. Multivariate statistics tie into visualization to look at pattern among many properties. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Statistics; Matrix algebra; Visualization; Multivariate analysis

1. Background

Computing, in the sense of calculation, is a small part of IT applications in geoscience. But, even in traditional aspects of geology, the quantitative representation of spatial entities and models is important. To understand why, we need to look at some basic mathematical concepts, and see how numbers relate to properties and programs to physical processes. Readers with a mathematical background may prefer to skip at least the first part of this chapter, and those without may wish to skip details irrelevant to them.

2. Measurement and number

Real numbers form a continuous sequence in an exact order. Given any two numbers, say 1.415 and 1.416, you can find as many numbers as you need between them, for example, 1.4151 or 1.4150032. You can compare any two numbers to see if they are equal or if one is larger or smaller than the other. This leads to a **scale of measurement**. The continuous sequence of numbers is analogous to situations in the real world. For example, you can compare the thickness of any two beds of sandstone to see if they are the same or if one is thicker or thinner than the other. To avoid carrying sandstone around, you can measure their thicknesses by comparison with a standard tape marked in millimeters. Measurements enable comparisons between any two bed thicknesses. You can measure more than one property. For example, you could also measure the maximum grain size for the beds of known thickness and compare the two sets of numbers (bed thickness and grain diameter), pair by pair.

We can use numbers in several other ways that rely on different aspects of their properties and so must be handled differently (see Krumbein and Graybill, 1965; Davis, 1973, for a fuller account). We can assign numbers quite arbitrarily as object identifiers. An identification number, such as an **accession number** (numbers

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

PII: S0098-3004(00)00041-8

from a given range issued in sequence), need bear no relationship to the object's properties. Indeed it is easier to issue unique identifiers by separating identification from description. A program or a user who knows the identifier can find the object by consulting a numerical index. Integer numbers are appropriate for identifiers.

The **subject classification** on a book or library shelf, such as a UDC number, is rather less arbitrary. Similar numbers apply to related subjects, and the number hierarchy (hundreds, tens, units) reflects subject subdivisions (part H, section 2). Numbers with decimal fractions are convenient in book classification. One can insert additional subdivisions without limit, simply by adding more digits after the decimal point. By shelving books or arranging object identifiers in numerical order we bring together those on the same subject for convenient searching or browsing. A sequence of numbers can represent a strict order of categories. Typical **ordered categories** are Mohs' scale of mineral hardness and the Richter scale of earthquake intensity. The larger the number, the higher the value, but the steps between successive values are not equal.

Measurement compares a property of some object with a standard scale. Intervals are equal, although the zero value may be quite arbitrary. For example, most scales of temperature, unlike the Kelvin scale, place zero at a convenient, but arbitrary point. It would therefore be foolish to say that 20°C is twice as hot as 10°C. Nevertheless, we can reasonably say that the increase of temperature from 0 to 10° is half that from 0 to 20°. Other physical properties, such as length, have an obvious and unique zero value, and there is no difficulty in adding, subtracting, multiplying and dividing those quantitative measurements.

The number field can then lead to a useful model at a deeper level than categorization. Equations can mimic real physical relationships. For instance, physicists can write equations describing the relationships of temperature with the pressure and volume of a closed body of gas. Equations imply the ability to calculate and maybe predict. Aspects of physical systems have direct analogs in well-known arithmetic operations. This astonishing correspondence between the physical world and mathematics is the basis for **mathematical modeling** (F 3).

The quantitative approach introduces a new mode of thinking. Instead of seeing the subject of investigation as a set of discrete objects, such as formations and rock types, we view it as a continuum, with characteristics that we can measure and compare as they vary from place to place. Gravity and aeromagnetic surveys, satellite imagery, or regional geochemical studies of stream sediments are examples. If objects are seen as "things" represented by nouns, and processes resemble verbs, then quantitative measurements are more akin to adjectives, describing the properties or composition of the objects.

It is tempting to wonder how far this mode of thinking can extend. Could we, for instance, replace our rather arbitrary classifications of geological objects by a more quantitative view where we measure continual change. This is considered further in J 2.3, but classification is basic to science (J 2.1) and descriptions with adjectives and no nouns have little meaning. I argue later (L 6.3) that while, with IT support, the scope of quantitative studies will surely continue to expand, different modes of thought are complementary, each adding to the overall understanding. The more important role of IT may be to ensure that information of all kinds is readily available to the investigators. If this is correct, the scientist (or a multidisciplinary team) needs to understand and use an appropriate combination of methods and modes of thought.

Before collecting measurements, it makes sense to consider their intended applications. This is the next topic. For detail, see Griffiths (1967), Krumbein and Graybill (1965), Davis (1973) or Swan and Sandilands (1995).

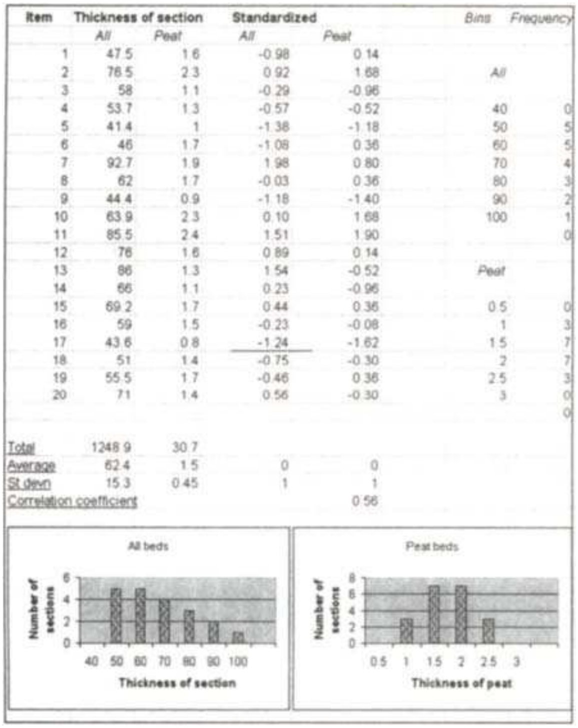
3. Descriptive statistics

We can manipulate numbers with simple operations of addition, subtraction, multiplication and division. They take us beyond individual comparisons to the properties of entire sets of measurements, and to general statements about relationships, say between grain size and bed thickness. **Statistics** (the branch of mathematics that deals with collecting, analyzing, interpreting and presenting numerical data) addresses these topics. One requirement is to characterize a set of measurements, like bed thickness, by fewer numbers that reflect the properties of the set as a whole. Important **statistics** (the measures or values calculated using the science of statistics) include the average value, also known as the **mean**. It is calculated by adding the measurements together and dividing the total by the number of measurements. We can measure the spread of values around the mean by the **variance** (the mean squared deviation from the mean) or by its square root — the **standard deviation**.

Statistics lead on from the description of a single **variable**, that is, a set of measurements of a single property, to explore the relationships between pairs of variables measured at the same point, such as bed thickness and grain size. An obvious approach would be to multiply each pair of measurements together and take their average (the mean cross-product). But the mean and standard deviation of each variable would greatly affect the result, and these have been measured already. Instead, we can **standardize** each variable by subtracting the mean from each value and dividing the result by the standard deviation. The mean cross-product of the transformed variables is known as the **correlation coefficient**, which has a value somewhere between +1 and −1. The extreme values are 1 if the bed thickness increases precisely as the grain size increases, and −1 if one decreases precisely as the other increases. The value is 0 if one variable shows no relationship to the other.

There are two general points here. One is that statistics measure different properties separately. Having calculated the mean, we remove its effects in calculating the next property, the standard deviation. We remove the effects of both in calculating the correlation coefficient. As a consequence, we can compare variation in a sequence of thick beds with that in a sequence of thin beds, and can judge whether the correlation of bed thickness and grain size is more pronounced in sandstone or in siltstone.

The other general point is that we are not dealing with sharply defined relationships. If we had measured the properties a few millimeters away, or made twice as many measurements, the results would have been different. If the processes of deposition had changed, with stronger currents, deeper water or different grain composition, the results would again differ. Statistical methods can measure the uncertainties of sampling and imperfect knowledge of the process. Their success depends on the skill with which the data are sampled and analyzed and on appreciation of the subject matter.



Item	Thickness of section		Standardized		Bins	Frequency
	All	*Peat*	*All*	*Peat*		
1	47.5	1.6	-0.98	0.14		
2	76.5	2.3	0.92	1.68	*All*	
3	58	1.1	-0.29	-0.96		
4	53.7	1.3	-0.57	-0.52	40	0
5	41.4	1	-1.38	-1.18	50	5
6	46	1.7	-1.08	0.36	60	5
7	92.7	1.9	1.98	0.80	70	4
8	62	1.7	-0.03	0.36	80	3
9	44.4	0.9	-1.18	-1.40	90	2
10	63.9	2.3	0.10	1.68	100	1
11	85.5	2.4	1.51	1.90		0
12	76	1.6	0.89	0.14		
13	86	1.3	1.54	-0.52	*Peat*	
14	66	1.1	0.23	-0.96		
15	69.2	1.7	0.44	0.36	0.5	0
16	59	1.5	-0.23	-0.08	1	3
17	43.6	0.8	-1.24	-1.62	1.5	7
18	51	1.4	-0.75	-0.30	2	7
19	55.5	1.7	-0.46	0.36	2.5	3
20	71	1.4	0.56	-0.30	3	0
						0
Total	1248.9	30.7				
Average	62.4	1.5	0	0		
St devn	15.3	0.45	1	1		
Correlation coefficient				0.56		

Fig. 1. Calculation of simple statistics with a spreadsheet. The total thickness of Pleistocene and Recent sediments were recorded at twenty boreholes, together with the thickness of peat in each. Simple statistics were calculated with a Microsoft Excel spreadsheet to examine the frequency distributions and their statistical correlation. See also Fig. 3.

Statistics are normally calculated by computer, particularly if the datasets are large. Good programs are readily available. The most flexible, although not the easiest to use, are subroutine libraries. The computer program normally calculates the values of statistical parameters using mathematical shortcuts. However, for teaching or exploratory purposes, spreadsheets and bar charts show intermediate steps and their effects on the individual items. For instance, they can show the original measurements converted to standard deviations from the mean (columns D and E of Fig. 1) and the user can examine the measurements in a local framework that may clarify relationships.

Some statistical programs help the user by providing an account of each method, a description of the algorithm, and examples of its use. The examples are unlikely to refer to geoscience, but it can be helpful to take an example as a template, and replace its variables and data with your own. An excellent range of textbooks is available on statistical methods and their applications. I have no plans to add to them, but do wish to point out the place of such techniques in geoscience investigations and to indicate some assumptions that constrain their application. The calculations of mean and standard deviation make no such assumptions. Their interpretation, however, raises many questions. The properties of the sets of beds constitute the **population** (D 4), as opposed to the actual measurements, which constitute a **sample** of the population. Sampling theory helps to clarify the link, so that conclusions about the population can be drawn from the sample, if appropriate sampling procedures (D 4) have been followed.

Circumstances determine the appropriateness of statistics. For example, an average thickness calculated from 49 siltstone beds, and one very much thicker conglomerate bed, would not be helpful. The result would alter greatly if we arbitrarily included another thick bed. A better procedure would be to study the thickness of conglomerate separately. Statistical measures make sense only for a clearly defined and coherent population. The **frequency distribution**, that is the pattern of relative frequencies of each measured value, can be examined on a bar chart or frequency plot (Fig. 1). Ideally, the frequencies are greatest in the center and fall off on either side to give the symmetrical bell-shaped frequency distribution of the so-called **normal distribution** (Fig. 2). A surprising number of actual distributions approximate to this, perhaps after a simple transformation, such as replacing the original values by their logarithms (F 5). It then makes sense to describe the distribution as a whole with a few numbers, such as mean and standard deviation. Otherwise, **robust statistics**, described in most modern statistics texts, offer a less complete means of description, but make fewer assumptions.

With some assumptions about the distribution and sampling scheme, it is possible, for example, to calculate the likely population mean and variance from the sample, and the probability of their lying within a particular range. A technique known as **analysis of variance** can show how mean values relate to sources of variation. For instance, if the ratio of Ca to Mg were determined in a number of samples, it could be of interest to see how it varied between formations, or between lithologies, or between analytical laboratories. With a carefully designed investigation, analysis of variance might be able to separate out the effects of each. Examination of the frequency distribution may, however, suggest a more complex situation, such as populations of different characteristics being sampled together. Descriptive statistics could then mislead by obscuring the real complexity.

Presumably measurements are made in order to draw some conclusions or to check some hypothesis. The conclusions must refer to something beyond the measurements themselves. Not "here are fifty beds that I have measured", but rather "these beds are noticeably thicker than their counterparts farther east, and the beds are thicker and the grain size coarser towards the base of the succession". Hypotheses about directions of sediment movement or deepening of the basin might in turn have prompted an interest in these findings. The hypotheses must be linked to the more general concepts in which they are embedded, and may lead to a mathematical model.

The analogy between the number field and the measurement of properties extends to the **mathematical model** — an analogy between mathematical operations (operating on the numbers) and physical processes (affecting the properties). Thus, adding together the thickness of beds in a vertical section is equivalent to finding their total thickness, a reflection, perhaps, of the total deposition of sediment at that point. Dividing the total by the number of beds to find the average is equivalent to recreating the original number of beds, but all of the same thickness. If nothing else, this may remind us that there is nothing magical about calculation. However, mathematical operations can mimic aspects of quite complex physical operations, often surprisingly well.

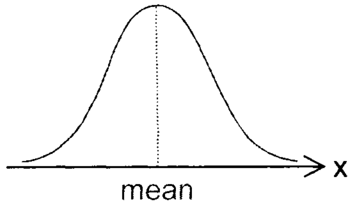


Fig. 2. Bell-shaped frequency curve of a normal distribution. The values of a variable that are deflected from their expected value (the mean) by many random events, might be expected to show this type of distribution.

If you develop one or more quantitative models before or during data collection, you can statistically compare the predictions of the models with the observed data to see if they conflict. This somewhat negative view is characteristic of scientific argument. If the observed values lie within the range of expected values, they give no indication that the model fails to match reality, and the model may, in consequence, be accepted. Acceptance is always tentative, for there may be other models that would also fit and would be more realistic in other ways. Quantitative models can help to investigate simple aspects of the process, such as: is it likely that the data reflect purely random events? Griffiths (1967) showed how salutary this approach can be. They can also be designed to throw light on the deep structures of the process (Wendebourg and Harbaugh, 1997). Science always seeks to disprove, and if data conflict with the predictions of the model, this could be taken as disproof of the model's validity.

The **random model**, in which events proceed on the basis of change alone, is widely used in statistics. It addresses the question of whether the results of analysis might merely be a consequence of random variation. If this can be ruled out, the argument for an alternative explanation is strengthened. A statistical model may have a **deterministic** element, giving precise values calculated from the mathematical model representing the physical process. It may also include a random element, reflecting the unpredicted, change events (although as pointed out in J 2.3, some non-linear deterministic systems are inherently unpredictable). The random element makes it possible to specify not only the most likely values of the outcome of the model, but also a range within which the values are likely to occur, taking random effects into account.

If the investigation is more than simply an initial gathering of ideas, and the data will be widely shared, then the investigator should describe the procedures and state how observations and measurements were made (D 4). Another scientist following the same sampling scheme and procedures would not expect to obtain identical data, but would expect the overall statistics to match within the calculated margin of error. An account of the sampling scheme can also help others to decide how far to rely on the results. By invoking statistical arguments, the process of geoscience investigation and reasoning can be made more rigorous and repeatable.

Quantitative methods have another role to play in geoscience. Measurements may be made in an exploratory way, without clear ideas about which of a range of hypotheses is being tested, but with hopes of unearthing a familiar pattern or detecting an interesting relationship (Tukey, 1977). Many geoscience investigations are concerned, not with the operation of physical systems, but with geological events and the environment in which they occurred. Sets of measurements may throw light on spatial patterns and interrelationships that guide the formulation of hypotheses. In geoscience, data are commonly displayed as a map or set of maps. Computer visualization studies (Cleveland, 1993; Gallagher, 1995) address the wider question of how to display data to detect pattern (G 2). They normally start with quantitative data and explore graphical means of conveying their significance to the human eye and brain. Field mapping faces the same problem of detecting pattern among a host of interrelated properties. It too can be seen as a form of spatial visualization. Statistical methods can then help in another way. They may offer concise summaries (like the mean) and reveal relationships that otherwise might not be noticed. They are unlikely to offer rigorous proof of the conclusions, but might point you towards a conclusion that you could test or support by other lines of argument more familiar in geoscience.

4. Matrix algebra and spatial data

The development of **coordinate geometry** by Descartes in the early 17th century gave the basis for spatial visualization of quantitative data, but also meant that spatial data could be brought within a quantitative framework. Position in space is measured by **coordinates** — distances from a zero point known as the **origin** along each of a set of **axes** at right angles. In consequence, spatial data can be manipulated, analyzed and managed on the computer. A range of computer techniques can be applied to information that would normally be recorded on maps and cross-sections. Those of us who think more readily in pictures than in numbers can make use of the correspondence between numbers and position, and between algebraic and geometric operations, as an easy way to gain an understanding of statistical methods. Many geoscientists may find it easier to visualize quantitative techniques as manipulation of points in variable space rather than as manipulations of numbers.

The link between computation and space, between algebra and geometry, is perhaps most obvious in matrix algebra, which enables a sequence of related operations to be written as one operation. Matrix algebra is an extensive and complex study in its own right, and is widely used in quantitative geoscience. Most computer users who are not programmers themselves can understand the results without understanding the details of the method. A few words of explanation here may make the process less mysterious to those who lack the mathematical background.

A table of quantitative values, such as those set out in a spreadsheet, can be regarded as a **matrix**. A single row of the matrix can be referred to as a row **vector**, and a single column as a column vector. The individual values or **elements** of the matrix are referred to in the spreadsheet by a letter indicating the column and a number indicating the row. In matrix algebra, the notation is slightly different, with the row and column both indicated by numbers. Letters are used, not to designate a specific column, but as placeholders that can be replaced by any number. Algebraic statements using the placeholder (or **index**) are thus quite general, and not tied to any particular numeric values. The matrix as a whole can be referred to by a name, in algebra usually a capital letter in bold type. The element has the same name, in lower case with the row and column numbers as suffixes. Thus, the element x_{ij} is in row i and column j of the matrix **X**. A typical notation in a programming language is X(i, j) where X is a name that could be several characters in length.

Matrices of the same size, that is, the same number of rows and columns, can be added. $\mathbf{Z}=\mathbf{X}+\mathbf{Y}$ means that each $z_{ij}=x_{ij}+y_{ij}$. Subtraction follows the same pattern. Multiplication, $\mathbf{Z}=\mathbf{X}\cdot\mathbf{Y}$, requires that the number of columns in **X** is equal to the number of rows in **Y**. The element z_{ij} is found by adding the products of the elements in the ith row of **X** with the corresponding elements in the jth column of **Y**:

$$z_{ij}=x_{i1}\cdot y_{1j}+x_{i2}\cdot y_{2j},\ \ldots,\ x_{in}\cdot y_{nj}$$

The ellipsis (...) indicates continuation of the preceding series in the same way, and n is the number of columns in **X**.

A problem frequently encountered in statistics and in some geophysical topics is that of solving a set of simultaneous equations, which could be written as $\mathbf{A}=\boldsymbol{x}$, where $\boldsymbol{x}$ is a column vector. In matrix notation, the general solution is $\mathbf{A}^{-1}$. Thus, matrix algebra is useful where each new value is dependent on several existing values in some systematic way. It provides a more compact and so more powerful notation than writing out each individual operation.

Returning to coordinate geometry, let us suppose that x, y and z are variables holding the latitude, longitude and elevation of a point in space. They can first be brought to the same units, say meters from an origin at mean sea level on the lower left corner of a map sheet. A set of points within the map sheet, say where samples had been collected, could be numbered with an index i, which takes the values 1, 2, 3, ..., n. As similar types of measurement are made for each sample, it is convenient to refer to them all in the same way. Each has an x, y and z value to indicate its location. To identify each sample, it can be given a suffix; thus the ith sample is located at x_i, y_i, z_i. The three values together, placed in brackets (x_i, y_i, z_i), form a vector, in the sense of a set of numbers referring to the properties of an object. In this case, because the elements of the vector are geometrical coordinates, (x_i, y_i, z_i) also denotes a vector in the geometric sense — a line from the origin (0, 0, 0) with length, orientation and direction.

As x, y and z, once they are measured in the same units, refer to similar things, it is convenient to refer to them with the same letter, say x_1, x_2, x_3 (or x_j, where $j = 1, 2, 3$). The values of x then have two suffixes, and the values can be arranged as a table with rows i, numbered from 1 to n and columns j, numbered from 1 to 3. In Fortran, the matrix is referred to by the more general term of an **array**. It is one-dimensional if it has one suffix, two-dimensional if it has two, and so on, whether or not this is its geometric dimension. The geometric operations that might be applied to these vectors are described in G 4. Their algebraic equivalents may involve changing each of the values of each row vector (x_{i1}, x_{i2}, x_{i3}) in a way that depends on all three of the current values. If the corresponding values after the operation are called y, then we can write:

$$y_{i1} = ax_{i1} + bx_{i2} + cx_{i3}$$

with similar equations for y_{i2} and y_{i3}, making three in all. Rather than referring to the constants, such as a, b and c, with separate letters, they can be seen as a matrix in their own right with three rows and three columns, say **T**. The transformation of the entire data matrix **X** to new values **Y** can then be written as $\mathbf{Y}=\mathbf{XT}$. Some important geometric operations have equivalents in matrix algebra that can be implemented on a computer system. As described in G 4, the transformation matrix **T** can represent familiar operations of moving objects about and changing their shape. It is a basic tool in creating computer maps and multidimensional spatial models.

5. Multivariate statistics

The link between numbers and space, between algebra and geometry, works in both directions. Spatial features can be represented by numbers; quantitative data points can be visualized as a cloud of dots. They can be manipulated in a space where the coordinate axes, at right angles to one another, are marked off in the units of measurement of the individual variables. The units refer to any measured property, such as bed thickness, grain size, gravity or uranium content. There is no limit to the number of axes, but we have trouble visualizing more than three at a time, as we appear to live in a three-dimensional world. Visualization may help us to understand the statistical relationships of a set of variables (see Cook, 1998). Statistics are concerned not just with single variables and comparison of different sets of measurements of the same variable, but also with the relationships between different properties of the same objects. This leads to techniques of **multivariate** analysis (a variate is a variable that exhibits some random variation).

Given a set of quantitative data, say, a collection of measurements of fossil shells, statistical methods are a guide to possible conclusions. Many different properties could be measured on each shell, such as length, breadth, thickness, length of hinge-line, and number of growth lines. We might wish to investigate how the properties are related to one another. We might also need some means of measuring the characteristics of the set of measurements as a whole, to compare them with another set from a different locality. The task has moved from comparing individual measurements to that of comparing aggregates of measurements, that is, a set of distinct items gathered together.

A starting point, however, is to look at the characteristics of each variate in terms of statistics, such as the mean and standard deviation (F 3). Each variate can then be **standardized** to zero mean and unit standard deviation. This frame of reference may make it easier to compare their relative variation. The cloud of standardized data points is centered on the origin and has equal spread or dispersion along each axis. Measures, such as skewness and kurtosis, based on the third and fourth powers of the deviations from the mean, can be calculated to assess the symmetry and shape of the frequency distribution of each variable (F 3). However, there is no substitute for their visual inspection with a bar chart, histogram or scatter diagram.

Some frequency distributions are quite unevenly distributed about the mean, such as the grain size of sediments or thickness of beds in a vertical section. A **log transformation**, which compresses the scale at the higher end, can bring the distribution to a more tractable form. Many other transformations are possible, and may be justified simply because the subsequent analysis is more straightforward. It is more satisfying if there is a physical justification for the transformation. For example, if an organism doubles in size each year, the distribution of size at random times will be logarithmic, reflecting a multiplicative, rather than an additive process. Replacing the original measurements by their logarithms converts the numbers to a simple straight-line distribution.

Statistical reasoning, as opposed to description, tends to assume that variates approximately follow the so-called **normal distribution** — the familiar bell-shaped curve of Fig. 2. Under a number of assumptions, it is possible to compare the actual sample with that expected from a random set of events, and determine the likelihood of this "**null hypothesis**" being incorrect. A number of excellent textbooks, including some for geoscientists (for example, Davis, 1973), give fuller information on these powerful methods.

The relationship between two variates can be measured by the correlation coefficient (F 3), or by a regression equation. The **regression** equation predicts the value of one variable (y) from that of another (x). The regression equation describes a line, using the formula $y=a+bx$, selecting the values of a and b to minimize the sums of squares of deviations between the measured and calculated values of y (see Fig. 3). The average squared deviation is a measure of the closeness of fit of the data to the line. The line that best fits the data could be regarded as a mathematical model of the relationship. As before, transformations that give the individual distributions a shape, like the normal distribution, are helpful. The null hypothesis of no correlation can be tested, given a number of assumptions, and only if it fails is there a need for a scientific explanation of the relationship.

The situation becomes more interesting where several variates are measured on the same items. You may be able to visualize the data as a cloud of dots in n dimensions, each dot representing one item, and each axis representing the standardized measurements for one variate. This is easiest with two or three variates, but the calculations are similar in higher dimensions. Each axis is regarded as independent of the others, which in geometry means that they are shown at right angles to one another. View the cloud in any two dimensions, and a correlation may be apparent that can be measured by the correlation coefficient. The correlation coefficients can be calculated for each pair of variates separately and shown as an $n \times n$ matrix, where n is the number of variates. There may be underlying processes that affect many of the variates together. Possibly there are several such processes affecting different variates in different ways. The matrix of correlation coefficients may throw light on the structure of the data, which in turn might suggest underlying causes.

One procedure for analyzing the correlation matrix is known as **principal component analysis**. It simply rotates the cloud of points in n dimensions (try visualizing it in three), until the largest variance is along one axis (the first principal axis), the greatest remaining variability along the second, and so on. The least variance is that around the final principal axis. The number of principal axes is the same as the original number of variates. They are still at right angles, but have been rotated together, as described, to a new orientation. When the points are referred to the new frame of reference (the principal axes), the new variates are known as the principal components. It is likely that most of the total variance will be accounted for by the first few components. If the remainder show no interesting pattern and appear merely to reflect ran-

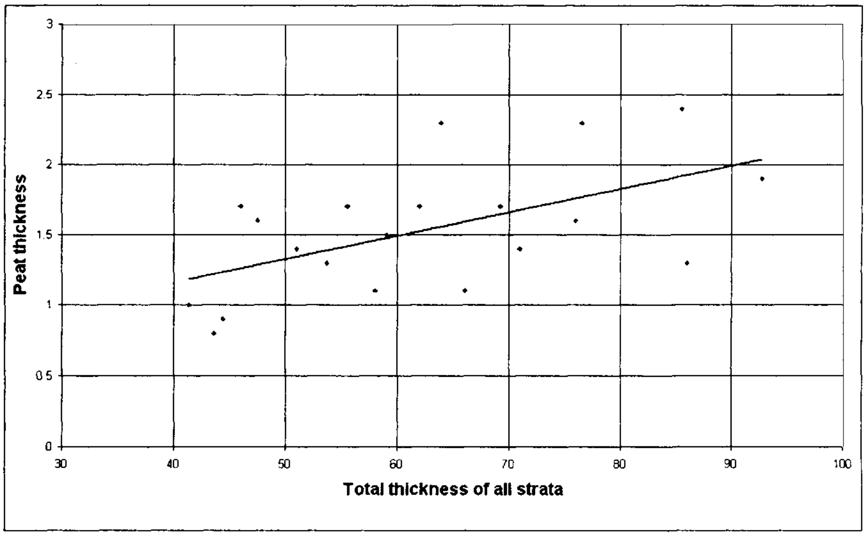


Fig. 3. Line of best fit between two variables. The data of Fig. 1 are plotted here to examine their correlation, and a regression line added. The chart was prepared from the spreadsheet with Microsoft Excel software.

dom effects, they can be disregarded. The result from the principal component analysis (PCA) program is thus a smaller set of new variates and a statement of how much of the variance each represents. The relative contribution of each of the original variates to each principal component is defined. The challenge for the geoscientist is to decide whether the principal components reflect underlying causal processes, or if not, how they might be accounted for. For example, the measurements of many aspects of the size and shape of fossil shells might be related to a few features of the environment like wave energy, nutrient availability, depth and clarity of water. This approach has been extended and elaborated as **factor analysis** (see Reyment and Jöreskog, 1993).

As well as the correlation coefficient between two variables, we noted the regression equation as an alternative way of looking at the relationship. This again is not limited to two variates. An equation $y=a+bx_1+cx_2+\cdots+gx_n$ representing a straight line in n dimensions, can be fitted to n variates x_1 to x_n, so that the value of the selected variable y can be predicted from all the x variates, minimizing the total sum of squares of differences between its measured values and the values predicted by the equation. Unlike PCA, which treats all variates alike, **multiple regression** focuses on one variate, with all the others seen as contributing to its variation.

As an aside, if the number of terms in a regression equation is greater than the number of data points, additional information is required to give a unique equation. Methods of **linear programming** achieve this by introducing an objective function that must be maximized to yield the greatest benefit to the system. This has applications in allocating raw materials to different products, as in models that allocate chemical elements to the mineral constituents of a crystallizing igneous rock. The more usual statistical case is overspecified, with many more data points than terms in the equation, and the least-squares criterion just mentioned, or a similar alternative, is used to arrive at a unique surface.

The form of the regression equation suggests how it is possible to fit curves, other than straight lines, to a relationship between two variables x and y. From the values of x, it is a straightforward task to calculate x^2, x^3, ..., x^n. We can then write an equation similar to that given above:

$$y=a+bx+cx^2+\cdots+gx^n$$

If we look at a graph of the powers of x (Fig. 4), we see that adding them in different proportions can generate quite complicated curves.

Many natural sequences are periodic, retracing a sequence again and again, like rotations of the Earth around the Sun. This type of sequence can be mimicked mathematically by a series in which, instead of an x value increasing along a line, we take an angle θ measured out by a radius rotating around a circle from 0 to 360° repeatedly. As it rotates, the sine of the angle θ changes from 1 to 0 to -1 and back again. A complex periodic curve (see Fig. 5) can be generated by taking, not powers of x, but sines of multiples of the angle θ:

$$y=a+b\ \sin\ \theta+c\ \sin\ 2\theta+\cdots+g\ \sin\ n\theta$$

The power series and the sine and cosine series have the mathematical property that successive terms have a smaller influence as the series continues. They are therefore suitable for approximating to an arbitrary curve. The slope and curvature at any point are readily calculated from the equations, and the form of the

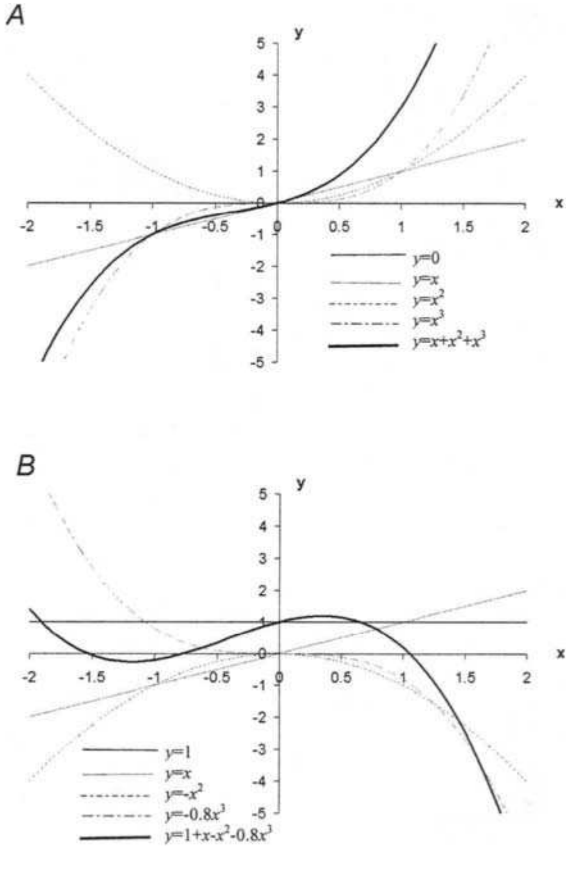


Fig. 4. Generation of polynomial curves. The basis functions for a cubic polynomial, and the combined curve from adding them together, are shown in A. In B, the coefficients are altered to give a different curve, which is still smooth, has the same number of inflection points, and heads for plus or minus infinity at each end.

power series is suited to statistical calculation. They are not appropriate, however, for extrapolating the relationship beyond the data points. The periodic curve repeats itself indefinitely, and the power series heads off to infinity.

Statistical tests (to which power series are well suited) can measure the "goodness of fit", that is, how well the curve fits the data, compared with expectations from a random relationship. The test is based on a number of assumptions, notably that a random sample has been obtained from an underlying distribution that is close to the normal, bell-shaped, curve. It follows that the sample is expected to be drawn from a single, homogeneous population. There are situations where we suspect that this is not the case. Our sample could have been drawn, without our knowing it, from populations that were formed in different ways by different processes, and they might have quite different properties. It is convenient, therefore, to have a means of searching for different groups within the dataset.

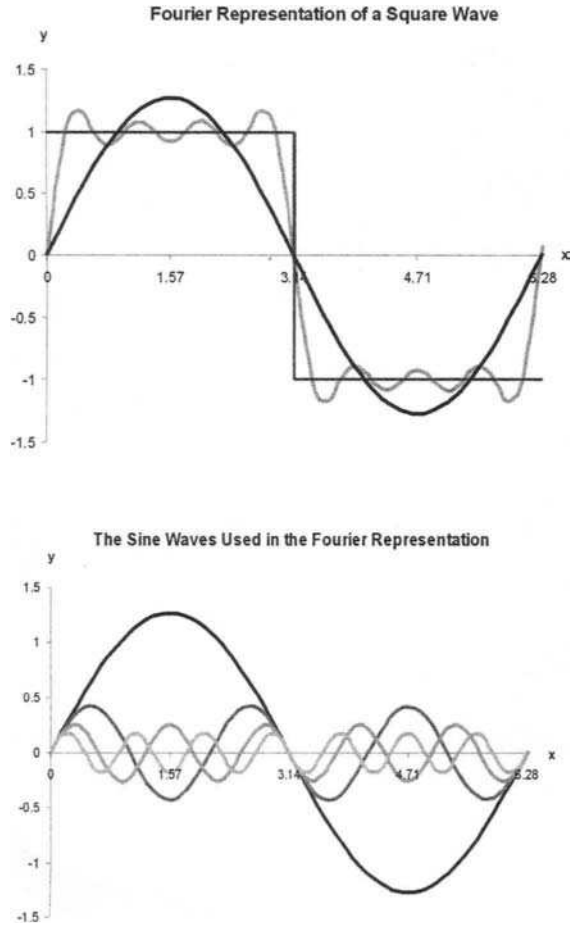


Fig. 5. Complex periodic curve. The upper diagram shows how sine waves can be combined to approximate even awkward shapes, such as a square wave. A single wave offers a first approximation, which can be improved by combining it with appropriately weighted harmonics, shown individually in the lower diagram.

Cluster analysis does this by looking for the most similar items in a dataset, combining them as one item, looking for the next most similar items, and so on until all the items in the dataset are combined. The similarity of two items, or its opposite, can be measured in various ways, such as the distance between them in standardized variate space. If the dataset is homogeneous, the clustering will proceed uniformly. If there are a number of natural groups or clusters, then the clustering is more likely to proceed with sudden breaks. Closely similar items are brought together, followed by a break before the next most similar items are found. This is a hierarchical process with clusters of clusters amalgamating as bigger clusters. Further examination of the characteristics of each cluster may suggest why they fall into groups (different species, different environments, different weathering, and so on). Cluster analysis can point to the existence of non-homogeneous populations, and lead to better analysis. If it is known before the investigation that several groups are present, **discriminant analysis** (see Davis, 1973) provides equations for assigning new items to appropriate groups. Techniques of this kind are used in **numerical taxonomy**, where measurements of sets of properties are the basis for classification. They may not be appropriate where objects are classified on the basis of an underlying qualitative model, as is often the case in geoscience (J 2.3).

Most multivariate techniques can simply be regarded as arithmetic transformations of a set of numbers, and do not necessarily require any underlying assumptions. The results may suggest ideas to a geoscientist who can then proceed to test them by other means. Calculating the descriptive statistics is then purely an exploratory exercise. Visualization, by displaying patterns through interactive graphics, follows this approach (see Cleveland, 1993). However, statistical tests of significance, and indeed any conclusions that depend on the numbers themselves, almost certainly imply some assumptions. Perhaps the most important and the most difficult requirement is to ensure that the items recorded (the sample) are truly representative of the population about which the conclusions are drawn. This applies, of course, not just to quantitative measurements, but to any observation of the natural world. There is, however, a danger that in the course of carrying out the complex manipulations of the data, original constraints and limitations are forgotten. The subject matter is all important.

References

Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, NJ 360 pp.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York 349 pp.

Davis, John C., 1973. Statistics and Data Analysis in Geology: with Fortran Programs. Wiley, New York 550 pp.

Gallagher, R.S. (Ed.), 1995. Computer Visualization, Techniques for Scientific and Engineering Analysis. CRC Press, Boca Raton 312 pp.

Griffiths, J.C., 1967. Scientific Method in Analysis of Sediments. McGraw-Hill, New York 508 pp.

Krumbein, W.C., Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill, New York 475 pp.

Reyment, R.A., Jöreskog, K.G. (Eds.), 1993. Applied Factor Analysis in the Natural Sciences. Cambridge University Press, New York 371 pp.

Swan, A.R.H., Sandilands, M., 1995. Introduction to Geological Data Analysis. Blackwell Science, Oxford 446 pp.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA 499 pp.

Wendebourg, J., Harbaugh, J.W., 1997. Simulating oil entrapment in clastic sequences. In: Computer Methods in the Geosciences, vol. 16. Pergamon, Oxford 199 pp.

PERGAMON

Computers & Geosciences 26 (2000) A51–A62

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part G. Familiarization with spatial analysis

T.V. Loudon*

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Spatial pattern is crucial to geoscience. It enables us to compare the observed properties of objects and explain their spatial relationships, correlation and distribution in terms of the geological circumstances and processes that formed them. Digital cartography and the spatial model provide computer support for looking at objects and properties in their spatial context. Transformations enable us to move and shape the objects for better visualization and analysis. Spatial statistics can quantify configuration for analysis of form and structure. The fractal model reminds us of the divergence between our models and reality. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Digital Cartography; Spatial relationships; Spatial models; Spatial transformations; Fractals

1. Digital cartography

A vector (x_{i1}, x_{i2}) can represent a point in two dimensions, and a vector (x_{i1}, x_{i2}, x_{i3}) can represent a point in three dimensions. A straight line joining two points is also a vector in the geometric sense (it has length, orientation and direction). A string of vectors, joined end to end, linking a set of points, can represent a curved line on a map. A set of lines joined end to end can delimit a closed area on the map, sometimes referred to as a **polygon**. The point, line and area are sometimes known in digital cartography as a **vertex**, **edge** and **face**.

Computer systems for drawing maps may be specifically developed for cartography, or could be computer-aided design systems, adapted for cartographic purposes. Areas can be filled with a selected color or ornament, and symbols and text positioned at points selected by the user. Fig. 1, for example, shows a diagrammatic geological map prepared on a desktop computer. Other cartographic representations show, for example, a vertical section of beds measured at an exposure or from a borehole (Fig. 2), or a fence diagram (Fig. 3) showing the variation in thickness of each of a sequence of beds across a number of measurement sites. Symbol maps of various kinds can show the spatial variation of one or more variables. Programs for a personal computer are available at reasonable cost to support a range of these tasks.

Existing maps can be digitized by clicking on selected points. A sequence of points can be digitized to record a line, and a sequence of lines to enclose an area. Because the short lines joining points are vectors, this is known as vector digitizing, and generates a **vector format**. Another way to capture existing images is with a scanner. This creates a **raster format**, in which the image is made up of a rectangular grid of small cells (picture elements, picture cells, or **pixels**). The

* Corresponding author.
E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

PII: S0098-3004(00)00042-X

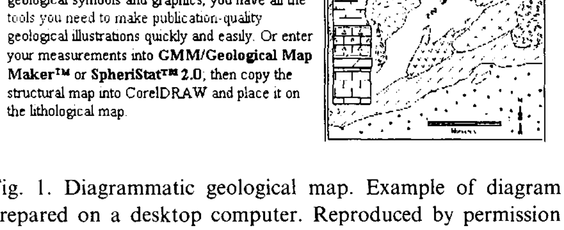
scanner assigns a value to each pixel to indicate the color at that point. The computer stores the values for the array of pixels. A minimum of one bit records each pixel in a monochrome image, but up to 24 bits or more may be used to store the proportions of red, green and blue in a pixel of a color image.

The raster format is an effective and efficient means of recording and storing complex images, such as air photographs, satellite imagery or "busy" topographic maps. Vector representation is more flexible and efficient where there are relatively few lines, as in a geological map. It has the advantage that individual lines and areas can be referred to and processed separately. Thus, one could select all the areas enclosed by the Cretaceous boundary, and find specimens within it by a **points-in-polygon** process.

Raster and vector methods are frequently used together. For example, a topographic map might be scanned and displayed as a raster image on the screen. Points and lines of geological significance could then be traced by moving a cursor on the screen and recorded in vector mode. This is a widely used procedure for digitizing geological maps, having the advantage of providing separate but linked representation of the topography and the geology in exact register. The two layers can be overlain when required and displayed together. Each is also available separately in its most useful format. Conversion from one format to the other is possible when required, and vector format is generally rasterized before it is displayed on the screen or printed. Many rasters, however, such as satellite images, could not be represented satisfactorily by vectors.

2. The spatial model

Ideas about how quantitative methods fit into broader schemes of investigation are taken up later (L 6.3). Meanwhile, we may note that most quantitative work in geoscience arises in geophysics and other subjects where instruments provide much of the raw data. Examples from geophysics include gravity, aeromagnetic or exploration seismic surveys, geomagnetism and global seismology. Other examples can be found in some geochemical and oceanographic surveys, aerial photography and satellite imagery. The initial data processing may be concerned more with correcting and clarifying the instrumental records than in exploring geological ideas. The mathematical models may be specific to the tools used in probing the properties of the Earth. The consistency of well-calibrated instruments, and the ability to correct for extraneous effects, are powerful features of these methods. Standardization makes it possible to conduct surveys over large regions or of global extent. Conclusions are not limited by local circumstances.



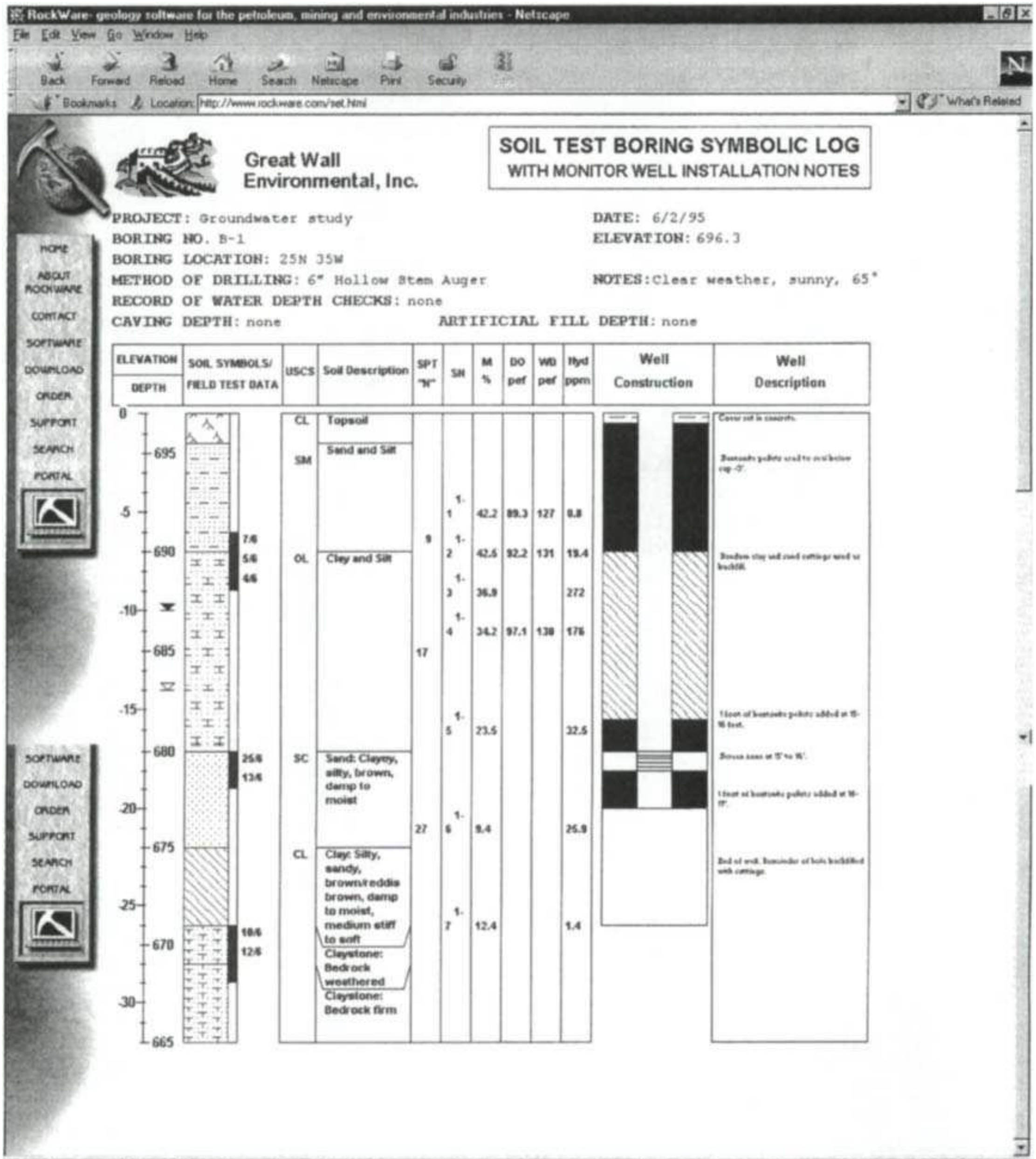
RockFill provides a variety of regular and pseudo-random patterns representing igneous, metamorphic and sedimentary rock types. When used with **GeoSymbol™**, our set of 190 geological symbols and graphics, you have all the tools you need to make publication-quality geological illustrations quickly and easily. Or enter your measurements into **GMM/Geological Map Maker™** or **SpheriStat™ 2.0**; then copy the structural map into CorelDRAW and place it on the lithological map.

Fig. 1. Diagrammatic geological map. Example of diagram prepared on a desktop computer. Reproduced by permission of Rockware. More at http://www.rockware.com/

The significance of the results is not likely to lie in a single measurement at one point, but rather in the regional spatial patterns, and their correlation with spatial patterns from other types of survey. Integration of the ideas from the various topics could proceed through examination and comparison of maps showing the final interpretations of the topic experts. However, **digital spatial models** (computer representations of objects and their properties in geographic space) offer more powerful methods of representing, analyzing, comparing and displaying the data. Spatial data, such as measurements of gravity, or formation elevations determined from downhole logs, may be recorded as part of a general database system, and still be part of the spatial model.

Traditional geological maps, showing the distribution of formations and other features at or near the ground surface, are now generally prepared by digital cartography (G 1). Contour maps, showing the variation of properties across an area, may also be drawn by computer (see Watson, 1992). The case for digital methods usually refers to long-term cost savings, flexibility, rapid production and ease of editing and updating. However, the longer-term benefits may depend on end products beyond the published map. The map, after all, is no more than an illustration of the underlying concepts (B 4.1). IT offers the prospect of expressing the latest conceptual model more comprehensively, and linking it to the quantitative surveys mentioned in F 5 (Förster and Merriam, 1996; Grunsky et al., 1996). The model can thus include three-dimensional rock bodies, their disposition and configuration, their composition and properties, their changes through geological time, the evidence and confidence in the conclusions (M 2.3). There is a prospect of freeing spatial information from the limitations of the paper map (B 4), and integrating it within the computer model (L 4). The ability to integrate concepts, and therefore the scope of the spatial model, is determined by the human interpreter, not by the capabilities of the software. Ultimately, therefore, the model reaches out beyond the computer with links to the scientists' unstated knowledge.

The computer spatial model, unconstrained by cartographic limitations, can express the conceptual model more fully. More of the ideas in the minds of the surveyors can be recorded and become the shared resource of the community. Maps can then be seen as ephemeral documents to assist visualization, illustrating selected two-dimensional projections of the model (see MacEachren, 1998 and the special issue of the International Journal of Geographical Information Science introduced by Kraak, 1999). The collection and archiving of spatial data are freed from the limitations of the map. Users would choose their own map content and display, after interacting with the model to clarify the availability of information and to select


RockWare- geology software for the petroleum, mining and environmental industries - Netscape
File Edit View Go Window Help
Back Forward Reload Home Search Netscape Print Security
Bookmarks Location: http://www.rockware.com/set.html
What's Related
HOME
ABOUT ROCKWARE
CONTACT
SOFTWARE
DOWNLOAD
ORDER
SUPPORT
SEARCH
PORTAL
0
695
5
690
10
685
15
680
20
675
25
670
30
665
7/6
5/6
4/6
25/6
13/6
10/6
12/6
Document: Done
Start
Ibetas - HyperTerminal
Microsoft Word - FIG10.DOC
RockWare- geology s...
10:25 AM

Great Wall Environmental, Inc.

SOIL TEST BORING SYMBOLIC LOG
WITH MONITOR WELL INSTALLATION NOTES

PROJECT: Groundwater study
BORING NO. B-1
BORING LOCATION: 25N 35W
METHOD OF DRILLING: 6" Hollow Stem Auger
RECORD OF WATER DEPTH CHECKS: none
CAVING DEPTH: none

DATE: 6/2/95
ELEVATION: 696.3
NOTES: Clear weather, sunny, 65°
ARTIFICIAL FILL DEPTH: none

ELEVATION / DEPTH | SOIL SYMBOLS/ FIELD TEST DATA | USCS | Soil Description | SPT "N" | SN | M % | DO pcf | WD pcf | Hyd ppm | Well Construction | Well Description

USCS	Soil Description
CL	Topsoil
SM	Sand and Silt
OL	Clay and Silt
SC	Sand: Clayey, silty, brown, damp to moist
CL	Clay: Silty, sandy, brown/reddis brown, damp to moist, medium stiff to soft
	Claystone: Bedrock weathered
	Claystone: Bedrock firm

SPT "N"	SN	M %	DO pcf	WD pcf	Hyd ppm
	1-1	42.2	89.3	127	0.8
9	1-2	42.5	92.2	131	19.4
	1-3	36.9			272
	1-4	34.2	97.1	138	176
17					
	1-5	23.5			32.5
27	1-6	9.4			25.9
	1-7	12.4			1.4

Cover set in concrete.

Bentonite pellets used to seal below cap -3'.

Random clay and sand cuttings used as backfill.

1 foot of bentonite pellets added at 15-16 feet.

Screen zone at 17' to 18'.

1 foot of bentonite pellets added at 18-19'.

End of well. Remainder of hole backfilled with cuttings.

Fig. 2. Log of vertical section. Example of diagram prepared on a desktop computer. Reproduced by permission of Rockware. More at http://www.rockware.com/

from it to meet their requirements. Comparison and correlation with other models, such as topography, land-use, or patterns of mineral deficiency in livestock, could be based on the models themselves, not on their cartographic representation. These future prospects are discussed in L 4 and M 2. The objective at this point is to look at the methods of handling data in the context of a spatial model.

A geological map enables the user to find (in the map key) a list of the things of interest (objects), such as stratigraphic units, displayed on the map. The map shows the location of the objects, related to topography and to a geographic grid or latitude and longitude. It gives an indication of the pattern and form of the objects. Contour maps may show, for example, chemical composition or engineering properties, and their spatial variation. The digital spatial model should combine these tasks and perform them more effectively. Like the map key, the metadata (D 3) associated with the model should identify the objects, perhaps grouping them within object classes or topics that might correspond to projects or to types of map, such as geology, geophysics or topography.

The location of data must obviously be recorded in the model. The configurations of objects and their distribution patterns and correlations may be investigated numerically, but are more likely to be examined by eye. The data must be displayable on a computer screen or printed as a map, and this involves transforming their three-dimensional locations to appropriate positions on the display. The transformation must preserve significant relationships among the data. We therefore look next at what these relationships are, how the transformations can be carried out, and which quantitative methods might be appropriate for the spatial data.

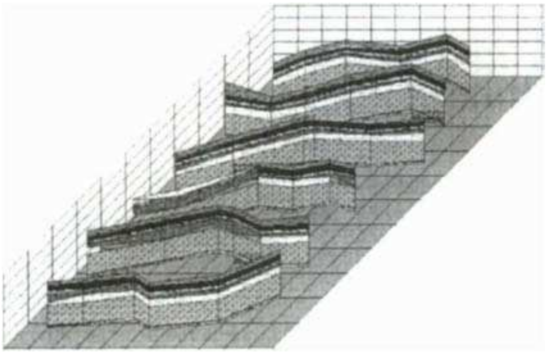


Fig. 3. Fence diagram. Example of diagram prepared on a desktop computer. Reproduced by permission of Rockware. More at http://www.rockware.com/

3. Spatial relationships

Spatial relationships can be important to the interpretation of a geological map, such as a formation boundary veeing upstream or a fossil being collected from a particular rock unit. They fall into two main categories — topological and geometric. **Topological relationships** are those that are unchanged by rubber-sheet deformations. If you imagine a map drawn on a thin elastic sheet, the sheet could be distorted by stretching and the relative position of points altered. Straight lines could become curved, angles between them altered, distances changed. Topological relationships are **invariant** (unaltered) during these distortions. Examples are: a point lying on a line; points, lines or areas coinciding or contained inside an area; lines touching, branching or intersecting; lines bounding an area; and their opposites, such as a point lying outside an area.

Geometric relationships, on the other hand, require a consistent **metric**, that is, distance and direction must be measured consistently throughout the space. Geometric relationships include: above, near, adjacent to, farther from, parallel to, converging with, at right angles to, interdigitated with, larger than, and their opposites and approximations. Unlike topological relationships, geometric relationships can generally be quantified, as in: twice as large, 234 m from, converging at 20°.

Some relationships may be obvious in the field, such as an outcrop being on the north side of a river below a road bridge. The relationship might be shown with some difficulty on a map, but is liable to be lost if the map is generalized for scale change. A printed geological map may be locked into a specific base map by overprinting. The spatial model seeks the flexibility of allowing each topic and each project to be managed separately, preferably meeting standards (L 4) that allow easy interchange of data. As this allows datasets to change independently, important relationships between them should be recorded explicitly. If need be, the short section of river and road bridge might be digitized and made available permanently within the geological model. It can then be compared with the same fragment of river on the topographic model during display, and any necessary adjustments made.

Spatial information can be transformed in various ways, altering the geometric relationships between objects. The transformations are an important part of computer graphics and essential for manipulating and visualizing the spatial model. They must be used with care to avoid accidentally distorting patterns or altering spatial relationships. For example, different map projections alter the size, shape and location of areas and cause difficulty in overlaying maps. Spatial transformations are also used in recreating surfaces from point data and in multivariate statistics. Understanding their effects is a useful background.

4. Spatial transformations

Spatial transformations of a geoscience model generally alter geometry rather than topology. However, it is possible that, say, two separate lines would coincide on the display when the scale is reduced. On a printed map, a cartographic draftsman might move the lines apart, to preserve the relationship at the expense of accurate positioning. On an interactive system, the ability to zoom in and clarify the relationships makes this unnecessary, and the true positions of the lines could be preserved.

For display purposes, a simple set of geometric transformations is available (Foley, 1994). **Translation** is bodily movement of an object relative to the origin. **Rotation** involves the object being turned about an axis. They are both **rigid-body** transformations that do not alter the size or shape of an object. **Stretching** changes the length of an object along an axis, thus changing its shape and altering distances and angles, except in the special case of **enlargement** where scaling is the same in all directions, and only the size and distances are altered. **Projection** involves reducing the number of dimensions, as when a three-dimensional body is projected on a two-dimensional screen. The linear transformations just mentioned are known as **affine** transformations. **Perspective** change gives a more lifelike appearance to the projection of an object by including perspective effects, such as apparent size diminishing with distance and parallel lines converging.

These transformations, which are illustrated in Fig. 4, are carried out on the computer by matrix multiplication (F 4). For display purposes, a sequence of transformations may be called for, such as rotate about the x-axis by 30°, dilate to twice the length along the new y-axis, then rotate back by −30° about the x-axis. This can be represented by a sequence of matrices, which could be applied in turn. It is more efficient, however, to multiply the transformation matrices together to obtain a composite matrix of the same size representing the entire sequence of transformations. The composite matrix is then applied to each of the original data points.

The transformation matrix is a square matrix, with two rows and two columns (2 x 2) for transformations in two dimensions, 3 x 3 for three dimensions and so on. The transformations, and the matrix multiplications, must be carried out in the correct order. Translation involves addition, rather than multiplication, but can be included in the multiplication by adding an extra dimension. The three-dimensional data are transformed in four dimensions. The extra dimension is also needed for perspective transformations.

The main application of these spatial transformations is for graphical display and manipulation, such as projection of three-dimensional objects onto two-dimensional paper or a computer screen. They also play a part in structural geology, process modeling,

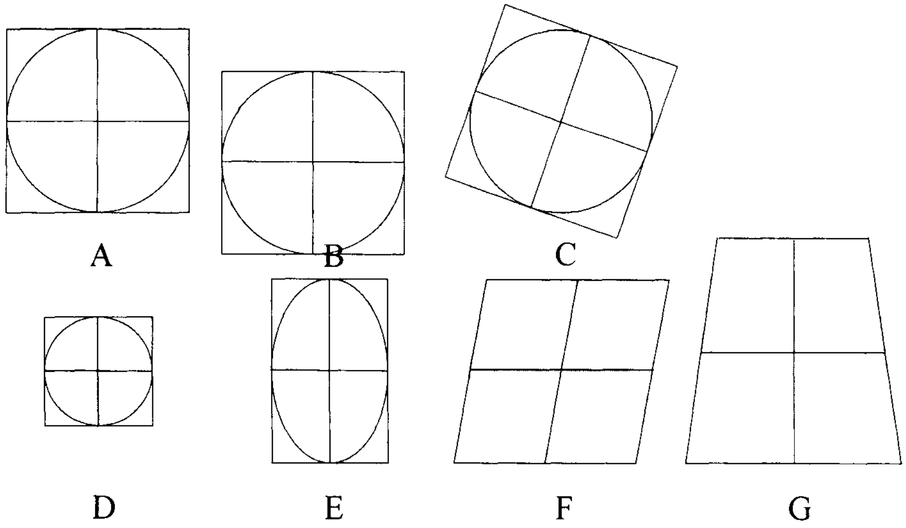


Fig. 4. Some geometric transformation. The figure at A is translated in B and rotated in C. The scale is changed in D, and the figure is stretched in E, with only the horizontal scale being altered. The figure is sheared in F. There are obvious three-dimensional equivalents. More complex transformations include perspective projections (G) in which the change of scale varies across the figure.

and such areas as multivariate statistics, surface interpolation and shape analysis.

An image can be subdivided into small patches. Each patch can be transformed separately, while forcing continuity across patches by modifying the transformations to ensure that the patches meet smoothly at the edges, by methods comparable to blending functions (G 5). The process is sometimes known as rubber sheeting. **Distorted** images such as oblique air photographs can be corrected. For example, points could be selected on the distorted photograph as landmarks for which exact coordinates could be found on a map. The distorted image could then be corrected by joining adjacent landmarks by lines and transforming the patches between them to conform to the map coordinates.

Spatial transformations also arise in **map projections**, which offer solutions to the problem of projecting an irregular oblate spheroid (the surface of the Earth) onto two-dimensional paper. The transformations are generally more complex than those just described. Different projections maintain different aspects of the original geometry. Areas, angles, distances and continuity cannot all be preserved when part of the surface of a sphere is distorted for representation in two dimensions. Aspects that can be neglected on the screen are important on printed maps. For instance, the user is likely to expect roads to meet across sheet boundaries without sharp breaks in direction.

Continuity from area to area raises problems as even the coordinate system, based on latitude and longitude, has local variations. The Earth is irregular, and mathematical approximations must be used for such concepts as the **datum** (the origin of the local coordinate system including mean sea level) and the **geoid** (a mathematical surface approximating, at least locally, to the shape of the earth). Different geoids and datums have been used at different times and in different places. Map coordinates and projections are a subject in their own right (see, for example, Snyder, 1987). However, if you combine spatial data from different sources, you should at least be aware of this potential source of error.

5. Spatial statistics and surface fitting

The statistical methods described in F 5 apply to properties measured at points in space. In geoscience, spatial patterns and spatial correlations are often crucial. Statistical methods describe the observed patterns quantitatively, guide interpolation between sampled points, and measure spatial correlations. The patterns are likely to reflect underlying geological processes that have influenced many measurable variables. Possible aims are to put individual points in context by fitting them to a surface, then to compare and correlate surfaces from the same area to learn more of the environment in which they developed, the formative processes and their effects. Specific tasks might be to estimate the elevation of a formation at a drilling location, the amount of ore in a mineral deposit, or the amount of folding across a cross-section.

As always, the validity of conclusions depends on appropriate sampling techniques (D 4). The sample should be random, in the sense that each item in the population has an equal chance of being included in the sample. You might try to obtain a representative sample of the lithologies in a rock body by selecting collecting points as near as possible to, say, random points on a map grid, or every 25 paces along a traverse. The sample would be unrepresentative, however, if hard sandstones formed outcrops, with interbedded soft shales covered by soil and vegetation. If you were studying the relationship between, say, petrography and geochemistry, then spatial randomization could be misleading. It might be better to ensure that every petrographic composition of interest had an equal chance of selection. We should continually review our raw material and ask how appropriate our models are, and whether it is possible to improve their correspondence to reality.

A surface is most readily analyzed if the data points are evenly spaced on a rectangular grid. But most data are positioned for reasons unconnected with surface analysis. Uneven clusters of points are found in oilfields, or soil samples from a housing development. Sampling points may lie along lines, such as geophysical traverses, geotechnical measurements along a new highway, or geochemical analyses of samples along streams. But even if the sampling pattern rules out statistical conclusions, quantitative methods may still be of value.

Time series, sequences of measurements at successive times, have been widely studied in statistics. In geoscience, time series arise in geophysics, hydrology, oceanography and other subjects. Similar methods can be applied to sequences along a line in space, and some can be extended to two or more dimensions. Regression techniques are one way to relate the variation of a property to its position in time or space. For example, the grain size (p) of samples could be related to distance (x) above the base of a vertical section, by the equation:

$$p = a + bx$$

The straight line which this represents is unlikely to show the variation adequately, and more terms from a power series or a trigonometric series can be added (see F 5):

$$p = a + bx + cx^2 + dx^3 + \cdots + mx^n$$

or:

$$p = a + b \sin x + c \sin 2x + d \sin 3x + \cdots + m \sin nx$$

In two dimensions (x and y) the equations represent **surfaces** and are slightly more complicated. For example, they might refer to the top of a formation based on its elevation (p) as measured at wells:

$$p = a + bx + cy + dx^2 + exy + fy^2 + gx^3 + \cdots + my^n$$

or:

$$p = a + b \sin x + c \sin y + d \sin 2x + e \sin x \sin y + f \sin 2x + \cdots + m \sin ny$$

The regression equations are calculated as in F 5 to minimize the sum of squares of deviations between the measured and calculated values of p. An excellent mathematical introduction is provided by Lancaster and Salkauskas (1986).

Let us take as an example the elevations of the top of the Cretaceous as known from fifty wells over the flank of a depositional basin. One possibility would be to fit a regression equation of fifty terms. As this equals the number of wells, it would fit the values perfectly, and the sum of squares of deviations would be zero. The position of contours on the quantitative surface could be calculated and drawn as a contour map. There is little reason to suppose, however, that the mathematical surface would match the geology. There is no obvious reason why geological processes would produce forms shaped like polynomial curves (F, Fig. 4), and the result will be greatly influenced by the spatial distribution of data points, which were almost certainly not positioned with this type of analysis in mind.

A better approach would be to fit a simpler third-order regression surface to the data, that is, an equation with terms up to the third power of x and y. The result would be a smooth surface that was less sensitive to the point distribution. It shows slow systematic change and is known as a **trend surface**. The **residuals**, that is the deviations between the data points and the fitted surface, can be displayed and mapped separately. Variations at different scales have been separated. The trend might be the result of regional tilting, subsidence and folding, the residuals the result of depositional and erosional features. By mapping the residuals separately, features might be noticed which were not apparent on the original map. Attempts at a rigorous mathematical justification for such conclusions tend to be rather unconvincing, but are unnecessary if the results can be justified by geological reasoning.

In order to arrive at a unique surface, the least-squares criterion can be applied (F 5), minimizing the sum of squares of deviations of data points from the fitted surface. An alternative is to fit **spline surfaces** (Lancaster and Salkauskas, 1986) that minimize the tension, or more strictly, the strain energy, in the surface (thinking of the surface as a flexible sheet). This model can give a good fit to some types of geological surface, and might, for example, mimic the shape of a folded surface.

For geophysicists, the use of periodic functions needs no introduction. The seismologist, for example, deals with shock waves generated by earthquake or explosion, and their modification by the rocks through which they pass. Waves, by definition, have a repetitive form and this may readily be described by fitting a function of sines and cosines (see F 5), a process known as **harmonic** or **Fourier** analysis. Rather than characterizing the wave by values at successive moments of time, a static view of the **power spectrum**, or relative importance of components of different wavelength, can be calculated. The values in the spectrum relate to the constants, a, b, c, ..., in the sine-wave equation. The power spectrum incorporates a great deal of information about the form of the wave, and makes it possible to compare the amplitude of different wavelengths (spectral analysis) as a separate issue from their time of arrival or position in space.

Fourier analysis has benefitted from the existence of many accurate time series collected digitally, and from the development of the Fast Fourier Transform (FFT), an efficient algorithm that reduces the computing load. Although the main applications are in seismic work, geomagnetism and gravity studies also study periodicity with Fourier transforms. Methods such as the fast Fourier transform can be applied to large digital datasets such as satellite imagery and downhole logs. With imagery, they can give information about the texture and its variation that may be related to changing geology.

For a geologist, the value of spectral analysis is less clear. The notion of treating distance from the origin as an angle may seem bizarre, but poses no mathematical problem. More fundamental is the relevance to the geological model. The periodicity of deposition of sediment, for example, might be related to geological time, with distance from the base of the section an inadequate substitute. Random events during deposition are likely to throw subsequent periodicity out of phase with that below. Similar comments could be made about variation in two dimensions, on a geological surface.

One way around this difficulty is to look at the relationship of each point to the adjacent points, as this is less affected by phase shifts. A mathematical function can be fitted **globally**, that is to the entire sequence or surface of interest, or **piecewise**, that is to a small number of adjacent measurements at a time (see Watson, 1992; Houlding, 1994). Piecewise fitting involves treating small parts of the section, or small patches of the surface, separately, and aggregating the results. The patches can be selected in many ways, preferably in some coherent and unambiguous manner. One way of creating surface patches is with Delauney triangles (see Bonham-Carter, 1994), a unique set of evenly shaped triangular facets with a data point at each apex.

A polynomial function can be fitted separately to each triangle. The simplest would be a flat plane. The planar facets would meet at the edges of the triangles, but the abrupt breaks in slope would be distracting to the eye, and would have no geological significance (Section 6). A cubic function could therefore be fitted to each patch, and a **blending function** (see Watson, 1992, Foley, 1994) applied at each vertex to ensure continuity from one patch to the next (Fig. 5). This is one approach used in computer contouring and in computer graphics and visualization. The **finite element methods** used by engineers for representing complex shapes adopt a similar approach, also including four-sided shapes (Lancaster and Salkauskas, 1986; Buchanan, 1995). Functions known as **wavelets** (Graps, 1995; Strang, 1994) can similarly be fitted to small patches. They lead to an interesting analysis by measuring the fit of wavelets to the surface while altering their position, amplitude and scale. There are occasions when local and regional variations are both of interest, as in trend-surface analysis, where local fitting methods can be applied to the residuals from the global trend. Piecewise fitting provides more detail where more is known, thus partly overcoming the sampling problem.

Many algorithms work best with data spaced on a regular grid. Therefore, one method may be used to interpolate from the raw data to grid points, and a different method to analyze and visualize the surface based on the grid points. The drawback is that one artifact is then being applied to another, obscuring the link between model and reality.

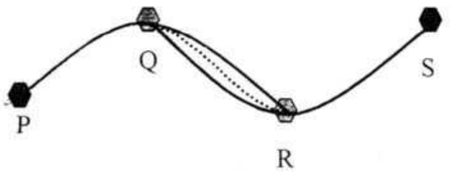


Fig. 5. Blending functions. Separate curves are fitted through the points P, Q, R and Q, R, S. They can be blended into one another where they overlap by the dotted curve, which is a weighted combination of the other two. The weight that the blending function gives to the curve PQR varies gradually from 1 at Q to 0 at R, while the weight given to QRS varies from 0 at Q to 1 at R. Blending functions can also smooth transitions between surface patches.

Other approaches are mentioned in G 7. Principal component analysis or PCA (F 5) can be applied to spatial data, including satellite images. Where quantitative information is available for each dot on the image (pixel) from each of a number of bands of the electromagnetic spectrum, PCA can distribute this information across a smaller number of principal components that can be mapped separately. They may be more readily interpreted than the original. PCA can also be applied to measurements of a suite of downhole logs, but is less likely to be successful. Since each of the logging tools, such as the microlog and the laterolog, are measuring properties of different volumes of rock, the results are not comparing like with like. It would not be clear what each of the principal components could be referring to, and the geologist's insights, which stem from an understanding of the strata and the tool, would be lost.

6. The fractal model

A computer program can define a sequence of mathematical operations. It enables the scientist to express a conceptual model of a geological process. This is particularly satisfying where theoretical ideas of the process match the fitted function. Not infrequently, however, it turns out that the conceptual model and reality differ in important ways. For example, we may visualize geological boundaries as smooth surfaces, even when we know from field observation that this is not so. Aided and abetted by the cartographer, we simplify by smoothing. It is inevitable that our records reflect our mental images, but smooth models can mislead and we should be aware of their divergence from the real world (J 2.3).

Mandelbrot (1982) takes as an example the question: How long is the coastline of Great Britain? Faced with the question, many of us visualize a map, maybe even wondering what scale is appropriate. Few think of the coast itself, with its headlands, beaches, boulders and breaking waves. We know that the answer must depend on a conceptual model, and we simplify reality in order to think about it. The more detailed our model, the longer the coastline seems to become. The length of a line joining the main headlands might give a first approximation (see Fig. 9). However, the line becomes longer without limit as we extend it to give the outline of each bay; the detail of each small irregularity; the outline of each sand grain; the microscopic and atomic structure of the interface. The question has no single number for an answer, but leads to an interesting discussion of measurement techniques. A precisely defined conceptual model can define the distance. There is no uniquely appropriate model, however, and none that matches reality in every respect.

Mandelbrot pointed out that continuity, which is an essential feature of most of our models, is not typical of natural phenomena, which tend to be discontinuous at all scales. **Continuity** refers to the characteristic of a mathematical function of having the possibility of creating a very small zone about a point, within which the value of the function does not significantly change. On a geological surface, for example, continuity may refer to elevation, slope (rate of change of elevation), curvature (rate of change of slope), and so on (Fig. 6).

In the natural world, we can seldom demonstrate mathematical continuity at any level. It must therefore be a feature of our perception. When we look at the distant outline of a mountain ridge, the small-scale variation is blurred, and we see an apparently smooth edge against the sky. As we examine the edge, the changes in direction, breaks in slope and even changing curvature catch our eye. In preparing maps and models, we therefore avoid such breaks if we know they have no geological significance. For the same reason, flat triangular facets joining data points are generally an unsatisfactory representation of a geological surface. The discontinuities of slope between facets draw the eye to the sampling pattern and obscure the underlying geology.

The approach taken by Mandelbrot is to study a number of mathematical functions that he terms fractals (Fig. 7). As well as lack of continuity, they exhibit another feature of interest to geoscience. They show **self-similarity**, that is, the pattern created by the function looks the same if we enlarge a small part of the original (Fig. 8). The concept is familiar to the geologist. Microfolds can mimic regional structure, and trickles of water on a mud bank may form a delta like a scaled-down Mississippi. He extends this to self-affinity, where vertical exaggeration or other affine transformations (G 4) alter the pattern. Manipulations of the fractals can produce images uncannily like real islands, mountains and clouds. The functions that generate them, however, are not obviously related to geological processes.

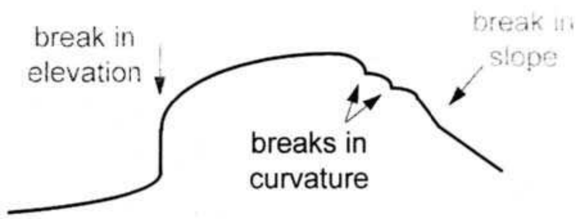


Fig. 6. Continuity of elevation, slope and curvature. Large and sudden changes in elevation, such as cliff faces, are obvious when looking at a landscape. Elevations on either side of the break differ, with no gradual transition. Large breaks in slope and curvature (the first and second derivatives of elevation) may also be apparent. Close inspection reveals similar breaks at all levels of detail.

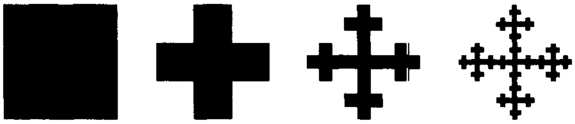


Fig. 7. Example of fractal function. Starting on the left of the diagram, a simple square is the initial object (initiator). The object is copied and attached to the north, east, south and west sides to form a new object, here reduced in scale. The same simple process (generator) is repeated, going from left to right, to give ever more complex objects. Reproduced by permission of David G. Green. More at http://life.csu.edu.au/complex/tutorials/tutorial3.html

In studying fractals, we can visualize the length of a coastline being measured by a hypothetical set of dividers. We measure the length of the coast by stepping the giant pair of dividers around it. The apparent total length increases as the span of the dividers diminishes (Fig. 9). We can plot the resulting calculated length against the length of the measuring device. The plot throws some light on the intricacy of the convolutions at various resolutions. The overall slope of the plotted curve has a bearing on a property (the fractal dimension) of the configuration of the coastline as a whole. Accounts of the numerous applications of fractals in geoscience can be found in journals such as *Computers and Geoscience* (Agterberg and Cheng, 1999) and textbooks such as those by Turcotte (1992) and Barton and La Pointe (1995).

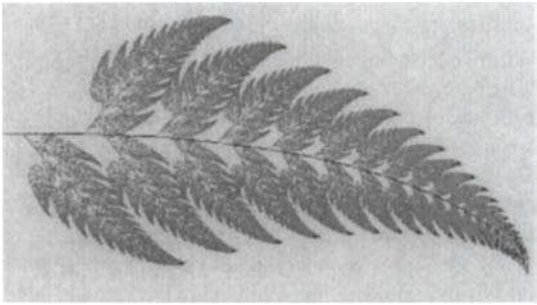


Fig. 8. Self-similarity of fractal. Enlarging one of the segments of this fractal fern would create another image similar to this frond, and so on. Many fractal functions show this property of self-similarity over a range of scales, as do the results of many geological processes. Reproduced by permission of David G. Green. More at http://life.csu.edu.au/complex/tutorials/tutorial3.html

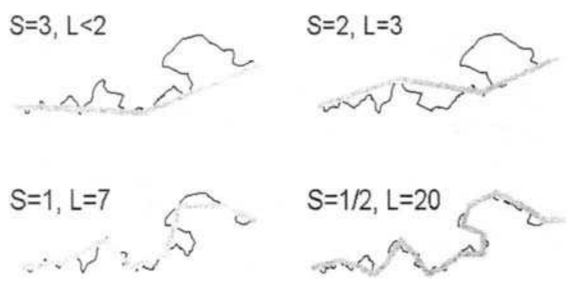


Fig. 9. Coastline and yardstick length. As the length S of the yardstick (or the span of a pair of dividers) decreases, the apparent length ($S \times L$) of the coastline increases. Reproduced by permission of David G. Green. More at http://life.csu.edu.au/complex/tutorials/tutorial3.html

7. Spatial configuration

We have looked in Section 5 at the **disposition** of a property, at ways of describing the distribution in space of its values. An obvious next step is to look at the **configuration**, that is, the spatial form and structure of the values. An important insight that geologists bring to spatial data is their expectation of how knowledge of nearby values might influence prediction. The distribution of commercially exploitable minerals in the Earth's crust, for example, is obviously uneven. Experience and concepts of formative processes guide geological expectations of their distribution. The study of **geostatistics** (for an introduction, see Isaaks and Srivastava, 1989) attempts to formalize and quantify this knowledge, and then to use it with available data to predict unknown values. Its roots lie in mining geology, where it offers techniques, for example, for estimating ore reserves from core sample data. The methods apply to estimating values at points, and also to estimating large volumes from small volumes, such as ore reserves from core samples or comparing readings from a laterolog and a microlog.

An obvious approach to describing the configuration is to compare values at different distances apart. One can then draw a graph of distance versus the similarity of values. The results are not tied to one location, but make a general statement about spatial correlation within the dataset. The power spectrum (Section 5) takes this approach, as does the correlogram, which shows correlation coefficients between data points at a given distance apart, plotted against that separation distance.

Geostatistics also considers the amount of change to be expected between values at various distances apart. For each separation distance, one can sample the difference in values, calculate the variance and plot it against the separation distance in a so-called **semi-variogram** (Fig. 10). Subsequently, one can estimate the value of an unsampled point on the basis of surrounding data points weighted according to the semi-variogram. This process is referred to as **kriging**. It can take direction as well as distance into account, as ore bodies, say, may be elongated, with greater variation in one direction than another. Not surprisingly, nearby points predict unknown values better than those farther away. Thus, tightly clustered data points are all to some extent providing the same predictive information. Kriging therefore gives them less weight than more isolated data points.

There is a major advantage in looking at the general configuration separately from the individual values. The general summary is not totally dependent on the sample. The sample gives an indication of the likely form of the semi-variogram, but conclusions about the population require geological inference. One can select a general curve that is compatible with the sample but takes experience of other similar situations into account.

Structural geologists and geomorphologists may prefer to think of change between values at various distances apart in terms of slopes rather than differences in height. Slopes are easier to measure, are invariant under translation, are not limited to a single surface, and relate more obviously to their conceptual model. For computer processing, slopes in two, three or more dimensions can be conveniently represented by **direction cosines** (Fig. 11) for manipulation and analysis. The subject of **differential geometry** (Kreyszig, 1991) is concerned with the intrinsic geometric properties of surfaces, such as principal axes and lines of greatest

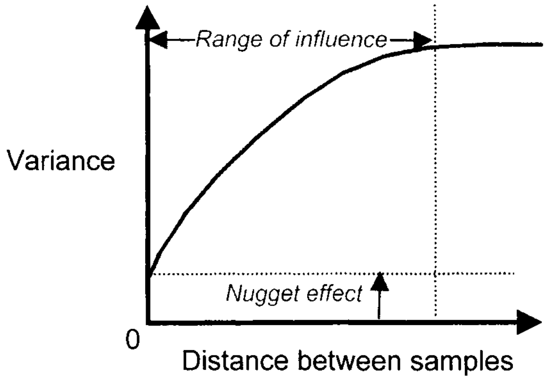


Fig. 10. Variogram. A geological property varies from place to place. Its values are likely to be more similar at nearby points than at those further apart, at least over a limited range of influence. This pattern can be quantified in a variogram, which plots the geographical distance between samples against the mean squared difference of their values. Samples taken at the same point may not give identical results, and rapid change can occur over short distances (the nugget effect).

and least curvature. It offers a mathematically rigorous account of geometric features that resemble geological features such as fold axes. It is relevant to considerations of shape and form as opposed to size and position.

These various methods can take into account detailed information from past and present studies. They do not, however, supplant the geologist's insight into the geological setting, the processes and complex relationships. Already some spatial modeling programs enable geologists to input additional information about, say, the position of faults. To benefit fully from the capabilities of both the computer and the human brain, interactive methods must be the way ahead. This will require the development of computer processes that accept interactive control by users who can formulate their background information in appropriate terms. In turn, this will need greater appreciation by geologists of the underlying mathematical concepts.

There are many instances where it is difficult for geologists to convey their ideas with a conventional contour map. A buried landscape might be sculpted with a pattern of river valleys. One cannot place the valleys accurately on the map if elevations are known only at widely spaced boreholes. There are two choices. One could contour a smooth surface showing the most likely elevation of the surface based on known data, but giving no indication of the detailed form of the surface. Alternatively, one could draw an illustrative surface, matching the data points and between them showing the nature of the surface and hence its likely origin, but with the features in arbitrary and possibly misleading positions. In computer graphics, the process of superimposing texture on a surface to make it look more realistic is known as **rendering**. A computer system should offer the flexibility of separating likely position from likely form and shape.

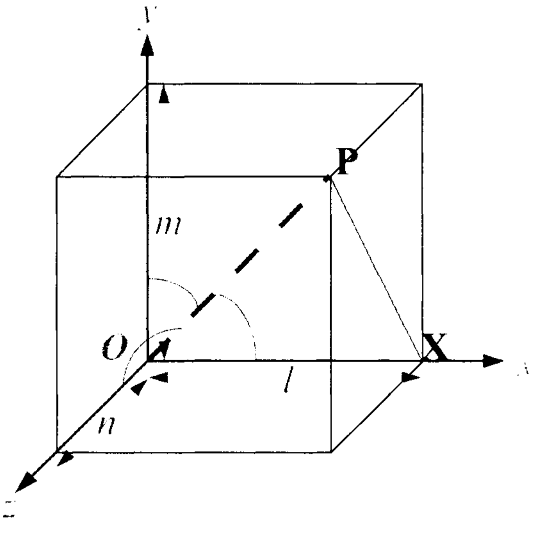


Fig. 11. Direction cosines. The line from the origin O to the point P is of unit length. It represents the orientation of a line parallel to OP or the pole to a plane perpendicular to OP. The cosines of the angles which OP makes with the x-, y- and z-axes give the direction cosines l, m and n. Using Pythagoras' theorem, it can be shown that $l^2+m^2+n^2=OP^2=1$. The direction cosines can be treated as lengths when applying geometrical transformations, but rescaling will then be needed to ensure that their sum of squares is again one.

Another issue arises with positioning, say, the edge of a steep-sided carbonate reef or a vertical fault. One might know the overall geometry, but not the position of the feature, which might lie anywhere between two widely spaced wells. Hand contouring could lead to arbitrary positioning of the feature. Computer contouring could lead to a smooth slope between the wells. But this is not realistic for a vertical feature, like betting that a coin will land on its edge because heads and tails are equally probable. The solution requires a clearer perception of probabilities. Probabilities cause problems in hand contouring but are well suited to computer calculation. Many contouring programs can provide probability envelopes, in effect indicating the most likely value of a surface, and bands on either side where it is decreasingly likely to occur. Near the steep feature and between the two widely spaced wells, there are two likely positions of the surface, separated by an improbable zone. The problem is not calculating the positions but in providing a comprehensible display. An exact definition of what the contours represent, such as most likely position or most likely shape, and different displays showing different aspects, can make the situation clearer.

Surfaces that repeat as with a thrust fault, or roll over as with a recumbent fold, are again difficult to contour manually. The computer contouring procedures described earlier are also unable to handle them. The computer graphics solution uses parametric coordinates s and t (Rogers and Adams, 1976), which are drawn on the surface itself, and related to the conventional coordinates (x, y and z) by polynomial equations. Visualization of the results may require a block diagram rather than a conventional contour map.

It is clearly necessary to have a wide choice of methods available for handling spatial data. There is no single method appropriate to all circumstances. Geostatistics and spectral analysis inevitably provide different solutions to the same problem. To find the best solution, the geologist must combine wide background knowledge with some understanding of the computer techniques.

In order to provide adequate flexibility, the computer system must offer a wide range of processes that one can apply to the data. The data, and the general supporting information, must be managed within a flexible framework that makes it readily available as and when it is required for any process. Information management is the topic of part H.

References

Agterberg, F.P., Cheng, Q. (Eds.), 1999. Fractals and Multifractals (special issue). Computers and Geosciences 25 (9), 947–1099.

Barton, C.C., La Pointe, P.R. (Eds.), 1995. Fractals in the Earth Sciences. Plenum Press, New York 265 pp.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modelling with GIS. Pergamon, Oxford 398 pp.

Buchanan, G.R., 1995. Schaum's Outline of Theory and Problems of Finite Element Analysis (Schaum's Outline Series). McGraw-Hill, New York 264 pp.

Foley, J.D., 1994. Introduction to Computer Graphics. Addison-Wesley, Reading, MA 559 pp.

Förster, A., Merriam, D.F. (Eds.), 1996. Geologic Modeling and Mapping. Plenum, New York 334 pp.

Grunsky, E.C., Cheng, Q., Agterberg, F.P., 1996. Applications of spatial factor analysis to multivariate geochemical data. In: Förster, A., Merriam, D.F. (Eds.), Geologic Modeling and Mapping. Plenum, New York 334 pp.

Houlding, S.W., 1994. 3d Geoscience Modeling: Computer Techniques for Geological Characterization. Springer-Verlag, New York 309 pp.

Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, Oxford 561 pp.

Kraak, M.-J., 1999. Visualization for exploration of spatial data. International Journal of Geographical Information Science 13 (4), 285–288.

Kreyszig, E., 1991. Differential Geometry. Dover, New York 352 pp.

Lancaster, P., Salkauskas, K., 1986. Curve and Surface Fitting. Academic Press, London 280 pp.

Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. Freeman, San Francisco 460 pp.

Rogers, D.F., Adams, J.A., 1976. Mathematical Elements for Computer Graphics. McGraw-Hill, New York 239 pp.

Snyder, J.P., 1987. Map Projection — a Working Manual. United States Geological Survey Professional Paper 1395. Government Printing Office, Washington.

Strang, G., 1994. Wavelets. American Scientist 82, 250–255.

Turcotte, D.L., 1992. Fractals and Chaos in Geology and Geophysics. Cambridge University Press, Cambridge 221 pp.

Watson, D.F., 1992. Contouring: a guide to the analysis and display of spatial data. Computer Methods in the Geosciences, vol. 10, Pergamon, Oxford, 321 pp.

Internet references

Graps, A., 1995. Amara's wavelet page. http://www.amara.-com/current/wavelet.html.

MacEachren, A.M., 1998. Visualization — cartography for the 21st century. International Cartographic Association Commission on Visualization conference, May, Warsaw, Poland. http://www.geog.psu.edu/ica/icavis/poland1.html.

PERGAMON

Computers & Geosciences 26 (2000) A63–A73

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part H. Familiarization with managing the information base

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

The geoscience record stores information for later reuse. The management of bibliographic, cartographic and quantitative information have different backgrounds. All involve: deciding what to keep; structuring the record so that information can be found when needed; maintaining search tools, indexes and abstracts; defining the content by reference to metadata. The current approaches to managing the literature, spatial information and quantitative data may be subsumed in a more comprehensive object-oriented model of the information base. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Information management; Disposal; Search; Metadata; Object-oriented methods

1. The framework

1.1. The requirement

The availability of quantities of data for analysis and display created a need to organize and store this information. Users could then revisit results and explore other ways of analyzing the data. The discovery of interesting relationships within one dataset might lead to investigation of similar relationships elsewhere, through access to a wide variety of related datasets. A database could combine data from many sources, and the user could select subsets for retrieval. A clearly defined interface ensured that retrieved data could be accepted by the programs for analysis. The programs could be reused with a variety of data, and datasets could be reused for analysis by a variety of programs (I 2.3). Management of the database came to be regarded as a task in its own right, and tools such as relational database management systems were developed. They have been successfully (and often unsuccessfully) applied to geoscience data.

Databases are appropriate not only for quantitative data, but also for indexes to other types of information. For example, cores, samples and specimens can be cataloged, and the records stored and retrieved from a database. The names of wells, boreholes or outcrops from which they were obtained can be recorded, with information about locations and depths, dates, investigators and the like. The result is structured, tabular information that can be linked to other datasets by means of key fields, such as well name.

When spatial information, traditionally shown on maps, became available to the computer, similar requirements arose for spatial data management. Geo-

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00043-1

graphic information systems were extended to include data management, with new options such as finding specimens from an area shown as Upper Cretaceous (points in polygon), or finding outcrops within the Upper Cretaceous where the lithology is oolitic limestone (polygon overlay).

Librarians and archivists have a long-standing interest in managing information. Indeed, the separation of their work from that of the geoscientist might be seen as an early distinction between information management and analysis. Their use of IT to organize and manage their collections has contributed many decades of experience in classifying, storing and retrieving the documents that record geoscience knowledge.

The different types of data management have tended to remain separate, with scientists looking after their own databases, cartographers and image analysts managing spatial data, and librarians managing published documents. Although they deal with different information types, however, the activities are similar for all. The information is likely to be:

- recorded and edited;
- acquired in a collection and in due course disposed of;
- assigned an identifier;
- annotated with source and date;
- structured, marked up, linked to other information;
- classified, described and indexed;
- assessed and evaluated;
- stored;
- retrieved;
- copyrighted to establish and maintain intellectual property rights;
- supplied on request to users who are entitled to access it;
- updated.

With published documents such as papers, books and maps, the three main players are the author, publisher and reader. Editors act as intermediaries between the author and publisher, and librarians and booksellers as intermediaries between the publisher and reader. The counterparts of author, publisher and reader in the more general situation are the contributor, manager and user of information. The manager is responsible for most of the tasks just listed, with contributors responsible for the content they supply, and users for their own selection and retrieval (M 1).

1.2. Acquisition, context and disposal

There are fundamental questions of what information is worth holding, for whom, and for how long. The answers should determine whether it is stored, how it is stored and how it is made available. With or without IT, storing information is not necessarily a useful activity. Each scientific study is project-oriented, taking place within a specific framework of business needs and scientific theory, hypotheses and models. Inevitably, much information has little value outside the project, and can be disposed of when the project is complete.

The value of data is determined by its significance to a model. Data items which control or elucidate aspects of the model have greater value than those which merely confirm what is already known. Mapping a thousand square miles of exposed sandstone might add little to the model of the area. Discovering one microscopic fossil in the same formation might be of intense interest, throwing new light on the model, both locally and globally. The model is paramount.

To be useful to others, information must be placed in a context in which it can be understood. It must have links to integrate it with the main body of scientific knowledge. Communication between author and reader depends on mutual understanding, and this in turn depends on standardization and quality control imposed by managers and editors. Its success, or otherwise, can be seen in the ease with which the reader accepts the conventions and language of a published paper or map.

The final products from a project are generally documents. They are the main level at which information is communicated. They provide the context within which the raw and processed data and other evidence can be understood. There is no scientific link between, say, a single measurement of gravity at a point and the proportion of tin in a stream sample at the same point. The connection is through the products — the gravity map and the geochemical map — and an interpretation of the patterns of distribution of the variables against the background of the underlying geological model, such as the possibility of a buried granite body. The availability of the final products is central to the ability to integrate different sources of geoscience information. Standards, for example for a geological map, may apply to the published end product rather than to the internal details of a project.

The long-term structure of recorded geoscience knowledge is based on publications. The formal literature must be organized in such a way that the user can find relevant information. It is the responsibility of editors, with the help of referees, to ensure that a coherent set of documents is produced, and the responsibility of librarians to ensure that the records are secure and accessible. Each publication is a major work that can be carefully assessed, cataloged and stored. Systems, including a legal framework, are in place to safeguard copyright and to hold and disseminate publications in perpetuity. Thus, although most libraries dispose of material they no longer require, even documents with obsolete ideas and disproved concepts are not totally banished from the record (I 6).

Each project is exploring unknown scientific territory. It is unrealistic to suppose that totally predetermined standards can be followed, as this would assume that all eventualities could be foreseen. Data collected for each project are partly specific to that project. The project objectives influence the design features of the investigation, including the underlying model, sampling scheme, sampling density, types of data, data collection methods, operational definitions, instrumentation and measurement procedures. The data can be fully interpreted and evaluated only in the context of the project design. In these circumstances, supporting data must be linked through the project documents, rather than directly within a comprehensive database (D 4). This suggests that informal records should be retrievable by tying them to the formal literature. Archive specimens or data could be referred to in the published paper, thus enabling readers to locate that additional information. But this is not always practicable.

Except for some long-term archives, project data are seldom seen as part of the formal literature because they are ephemeral and subject to change. They are likely to be maintained and possibly made available by the originating individual or organization, but may be altered as ideas change, and disposed of when the owner loses interest. They are thus not part of the permanent record, even if some at least ought to be. If the user knows the reasons for the study and something of the manner in which it was carried out, data can be abstracted from it that are of value in other contexts. For example, a local geological model might be based on fortuitously exposed rock outcrops, but could also make opportunistic selective reuse of data collected for other purposes. Those data might be derived from other projects employing different models, such as oil exploration, mining and quarrying, underground water production, and site investigation. Direct reference to the data and project descriptions through one coherent database rather than through the literature could make the links clearer and simplify the interfaces for analysis and display.

The conventional literature is also inappropriate for maintaining some large datasets. Properties such as gravity can be measured consistently by instruments calibrated between projects and following a predetermined sampling pattern. Such measurements over a large area have obvious value in regional studies. Similar comments could be made about, say, aeromagnetic, seismic, downhole logging, geotechnical and geochemical surveys. Even borehole descriptions that follow detailed guidelines can be consistent and comparable over wide areas (NLfB, 1999). In effect, these become regional projects focused on narrow aspects of geoscience. They may co-exist with projects with different scope and aims. For example, a site investigation might follow external standards in describing the borehole records because, although they added a small overhead cost, this could be justified by the value added to the data.

IT raises the possibility of smaller but more frequent updates to the knowledge base, new criteria for acquisition and disposal, and a shared context that integrates formal and informal documents. Possible consequences are discussed in M 2.1.

1.3. Search strategies

A crucial aspect of managing information of any kind is the ability to find appropriate material when it is needed. If the user knows the name of an object, it should be possible to discover where to find it by looking it up in a catalog or index. If not, various other search techniques are possible, based on a description of the topics of interest. The search can be modified by

- extending the area of interest
  - ○ use broader terms in a bibliographic search
  - ○ zoom out in a geographic search
  - ○ extend the range in a database search
- restricting the area of interest
  - ○ use narrower terms in a bibliographic search
  - ○ zoom in, in a geographic search
  - ○ narrow the limits in a database search
- moving to related concepts
  - ○ use related terms in a bibliographic search
  - ○ pan to adjacent areas in a geographic search
  - ○ select related criteria from the data model in a database search.

Ideally, a computer system would allow the user to combine bibliographic, geographic and database criteria, using the technique most appropriate to each stage of the search. Various commercial systems, including GIS, are moving in this direction.

The human brain is adept at searching, particularly through well-structured material. The scientists' background knowledge enables them to detect clues to what is relevant, even where this is not expressed directly either in the material or in their own explicit search criteria. Users can assess the relevance of documents or images by browsing through them, focusing on and following up items of interest. Abstracts and index maps may save the trouble of looking through the full document. The structure of objects may be shown by a paper's table of content, a map explanation or a database entity-relationship diagram, and this may narrow the area of search.

Control numbers, like the familiar UDC of library shelves (H 2), indicate subject areas and organize ideas in hierarchical classifications. Because similar numbers refer to similar topics, browsing among adjacent objects, for example on library shelves, can be profitable. Classification of documents implies that catalogers have examined their content, and assigned categories accordingly. As this is a worldwide activity, the categories should follow a global standard, and classification should follow standard cataloging rules (H 2). A computer index of titles or abstracts can be searched for a keyword or combinations of keywords (H 2), as can the original document if the full text is held on a computer.

Documents can also be found from a general written account of the subject by following references to more specialized papers, or by looking at a small-scale map to find areas to examine at a larger scale. References and hypermedia links may lead to other relevant material. It helps if the strength of the links can be estimated, in terms of current access rates, numbers of previous users making the connections, or their evaluation of the links' significance. Examples can be seen in Web documents, such as electronic bookshops (Amazon.com, 1996).

Spatial data represented on a map or in a GIS can be searched by geographic locations using the grid of geographic coordinates, by looking on the map face for the color codes and symbols shown in the explanation, or by relating the geoscience data to features on the underlying topographic base map or to other map overlays. Because the location of data on the map reflects their position on the ground, the map (or GIS) can be browsed for items within an area, close to a feature, coinciding with a point, and so on.

Classification of items in a database (H 3) can follow similar procedures to those used by the librarian. A data model can show the relationships of concepts. Terms can be defined and standardized in data dictionaries. Data searches can therefore be narrowed to appropriate variables, including spatial coordinates, then to a selected range of their values. If data are to be widely shared, the data models and dictionaries must follow widely accepted standards. This has been achieved in limited areas, such as global studies of oceanography, seismology or geomagnetism. At present, most geoscience databases are restricted to the organization that created them. However, the work of such organizations as POSC (1999) suggests that a more general framework for geoscience is emerging.

Computer systems lack the scientists' background knowledge, but (being machines) can efficiently perform mechanical searches for specific keywords or numeric values, and can follow recorded links. The past experience of other workers can be recorded, such as: many past users looking for references to "carbonates" found that some references to "limestone" were also relevant. As always, the computer and human brain should pull together, making best use of the abilities of each. A well-structured search might alternate between the computer extending the search on the basis of links and structure, and the scientist narrowing or redirecting the search on the basis of background knowledge of the subject and the requirements.

Things can be found more easily if they are carefully organized. We now look in turn at how librarians, database managers and cartographers set about this task. There are obvious benefits in combining the best features of all their approaches, and it is not surprising that their methods are tending to converge. In H 5, we consider how distributed objects can assist convergence.

2. Documents

The number of relevant published documents in most fields is large enough for librarians to benefit from computer support in acquiring, storing and cataloging them. As can be seen in electronic journals such as D-Lib (D-Lib, 1995), many librarians are enthusiastic pioneers of IT methods.

A library requires a description of all the bibliographic items in the collection, such as books, serials, maps, videotapes or computer files. The objectives are to know what is there, and where it is, to arrange the material sensibly on shelves, to help users to find the information they require, and to monitor its use. The items must be uniquely identified, and should be retrievable by author, title, or subject. A catalog is therefore required containing at least this information. There are obvious advantages in adopting standard cataloging procedures. An initial aim was to help libraries to exchange information on their holdings, and obtain comprehensive lists of new publications as a guide for acquisitions. Obtaining entries from a central source, such as the Library of Congress, can reduce the considerable cost of cataloging. Sets of national and multinational standards have evolved (see Mulvany, 1994), many based on the International Standard Bibliographical Description (**ISBD**).

For referencing or cataloging purposes, each item must be identified uniquely. The International Standard Book Number (**ISBN**) and Serial Number (**ISSN**) address this need. The ISBN and ISSN indicate by numeric codes the language of publication, the publisher's imprint, and a number assigned by the publisher for the edition of the book, or issue of the serial. This control number is likely to appear on the cover as a machine-readable bar code. Other control numbers may be assigned by the library, particularly for works that have no standard numbers.

The Anglo-American Cataloguing Rules (**AACR2**) are widely followed in most English-speaking countries and have been translated into many other languages. They help to ensure that each catalog entry has a similar style. The **Dublin Core** (DCMI, 1998; L 3) has comparable objectives for simpler systems. Subject matter can be classified and assigned numeric codes, such as the long-established **Dewey Decimal Classification**. Developed from it is the Universal Decimal Classification (**UDC**), a more general classification covering the whole field of knowledge. The classification is hierarchical, going from the general to the particular. Thus, 54 indicates chemistry, crystallography and mineralogy, 549 mineralogy, 549.32 sulfides of metals, 549.324/.326 disulfides of iron and related sulfides, and 549.324.31 pyrite, melnikovite. Where books are arranged on library shelves by such a numbering system, those on similar topics should be together, making it easier to browse through the subjects of interest.

An obvious snag with this arrangement is that a document may deal with more than one topic. For example, it might deal with the geophysical as well as the mineralogical consequences of pyrite deposits. The UDC code can accommodate several subjects, separating the codes by devices such as colons, that indicate the relationships between the subjects. Multi-dimensional shelving to reflect this would be inconvenient. With a computer catalog, of course, there is no difficulty in handling classification by multiple criteria. Users can access the catalog remotely from the desktop, through a simple user interface. Many libraries provide on-line public-access catalogs (**OPACs**) of this kind (L 3), which are freely accessible and easy to use. Lists of OPACs can be found on the World Wide Web (NISS, 1999).

A list of references at the end of a paper points to earlier, related work. With computer support, the process can be reversed to point to papers written later which refer to the target paper. A **citation index** produced in this way, such as the Science Citation Index, enables you to start from a key reference in your subject area, and locate later works which deal with the same topic (Garfield, 1983; Institute for Scientific Information, 1999). These forward references can also be used to analyze the structure of cross-reference in the literature, and to throw light on the range and number of references to a paper and the caliber of the journals in which they appear. This is one tool used to evaluate the contributions of individuals or organizations to the scientific literature. Access to a citation index can be expensive, and the coverage of the literature is inevitably incomplete.

Rather than classifying a paper with a long string of UDC codes it may be easier to record a list of keywords reflecting the main subjects. A **thesaurus**, such as GeoRef (Shimomura, 1989), is normally available to indicate which terms have been used for cataloging. For searching, it may suggest synonyms ("see also"), broader, narrower and related terms, if the first choice is not appropriate. Lexicons may also be available with definitions of the terms. Geoscience catalogs of this kind are commercially available, on-line or on disk. They may be available as a library service on a local computer network.

Retrieval by combinations of keywords can be effective, but is not classification in the strict sense, which depends on analysis of idea content. The UDC is "a universal classification in that an attempt is made to include in it every field of knowledge, not as a patchwork of isolated, self-sufficient groupings, but as an integrated pattern of correlated subjects" (British Standards Institution, 1963, p. 6). The notion of providing a map of human knowledge reappears in the concept of ontology as used by workers in machine intelligence (L 5), the pattern of linkages in hypermedia (E 4), and in the object-class hierarchies and entity-relationship diagrams used in database work (H 5, H 3). UDC provides perhaps the simplest notation, with a view to mapping fields of knowledge onto the linear sequence of the library shelf. Remarkably, this simple approach has proved to be of considerable long-term value.

The storage and exchange of bibliographical records on computer systems require decisions about the format in which they are held. The **MARC** format (MAchine Readable Cataloguing) meets this need. Various incompatible versions arose, leading to the development of **UNIMARC**, which facilitates the exchange of records created in any MARC format (Library of Congress, 1999). It specifies a wide range of fields that can be identified by a standard tag or by their position in the record. Librarians have for some time been using the ANSI **Z39.50** protocol for communicating between their computer systems in order to access bibliographic catalogs (Biblio Tech Review, 1999). With its help, they, or end-users, can access several on-line public-access catalogs (OPACs) in numerous libraries worldwide (NISS, 1999) from a single search. Extended services include ordering or borrowing a document, collecting fees and even updating the database. New and more general exchange formats, based on SGML (E 6), wait in the wings.

From the point of view of the librarian, as opposed to the user, computer systems can also help with their housekeeping tasks. These include stock control, keeping track of loans, acquisitions, disposals, and exchanges. Bar codes attached to each document identify it when it is borrowed or returned, and similar codes on borrowers' cards mean that the transaction can be largely automated. The computer records should help librarians to manage their collection, and to combine cataloging activities in a virtual union catalog. Library software integrates the housekeeping tasks with the user services mentioned earlier. The integration of library tasks with broader information systems is discussed in part M.

Librarians are traditionally concerned with published documents. As these are complete and unchanging, editing and digitizing would normally be completed before publication (D 6) and are not seen as part of the management process. However, there is a growing need to manage electronic documents. These may be continually modified to reflect the most recent views and to maintain links to datasets and spatial data. They must be continually modified to conform to current formats and updated systems. One short-term possibility is to store the electronic documents as part of a database. In the long term, information management may merge the library and database functions. Meanwhile, libraries must evolve to meet new requirements, including the growing number of electronic publications, the requirement to digitize and markup existing publications, and the need to keep track of numerous versions of a document. The possible consequences are discussed in L 3.

3. Database

Data are collected within a project to meet its objectives and methods. The dataset may nevertheless be retained as a **persistent object**, which continues to exist beyond the project. Its continuing value was considered in H 1.2, including reuse by the investigators, or others, to follow the reasoning behind the project's conclusions, to confirm the results, to add to the data, and possibly to verify the data by repeating at least some of the observations. The persistent object may be evaluated, authorized, and published or deposited in an archive for long-term availability. The stored object is likely to be a computer file, which is readily exchanged and designed to interface with programs for analysis. The programs may be included as part of the same object or may be persistent objects in their own right. An account of the data or programs is likely to be published in a conventional journal, and the same editorial team may evaluate them and provide archiving facilities.

Successful sharing of detailed information depends on a common purpose. Data are more consistent when they are collected in a similar way by similar instruments. The extent of integration may depend on the supporting information technology. Long ago, paper records about the geology of an oil field, for example, were seldom exchanged far beyond the original operators. When service companies started to offer better instrumentation (for instance in seismic exploration, downhole logging and core analysis) standardization became easier, and data were more widely exchanged. As electronic methods of data storage and analysis developed, global standards were introduced, for example, POSC (1997), and data management began to be seen as a task that need not be part of the core activity of an oil company but could be outsourced to a shared facility. Wider standardization between broad fields of activity, say, between oil exploration and geological survey, is more difficult, with fewer obvious gains on either side. Overlapping standards are nevertheless developing, often driven by IT solutions that are adopted in a number of different fields.

Close dependence of data on the project limits their value in other contexts. Ideally, datasets would not be limited to a specific project, but would play a wider part in geoscience knowledge as a whole. Data might then be continually revised as more is learned. The effect of changes in one area might be propagated through to all related areas. The concept of a database was devised with such integration in mind. It began with the hope of bringing together all the data from an organization, making them more generally useful outside their original context, and avoiding repeated collection of the same information for different purposes.

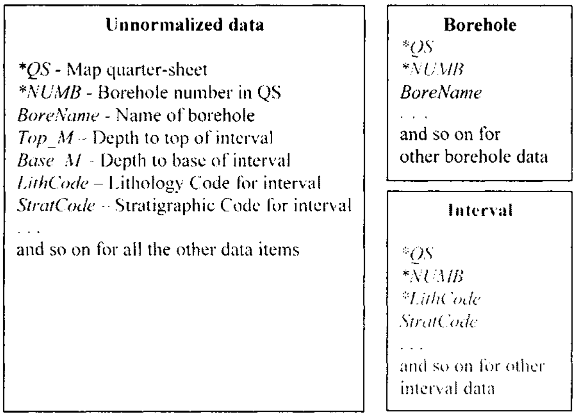
A **database** provides a structure in which a wide variety of data from many projects can be recorded and kept up to date by a database management system, while maintaining internal consistency and providing a uniform interface to the outside world. The trade-off is that greater generality can bring additional, often unacceptable, overheads to an individual project. They include the need to analyze and preplan the activity, and to follow standards that may create additional work and reduce flexibility with no obvious benefit to the project. Individual contributions, and hence specific responsibility and credit, can sometimes be lost within an integrated database. Nevertheless, the gains even from limited integration can be substantial, and large organizations and scientific consortiums have made good use of database techniques. The implications are relevant to most geoscientists.

Relational databases provide a widely used structure for geoscience data. Relational database management systems provide the means of managing them. Data are stored in tables of a particular kind known as relations. The columns may be referred to as variables or domains and the rows as tuples. The relational design aims to reduce **redundancy**, that is, repetition of information. The reason is that redundancy causes problems when information is changed. The same changes must be made wherever the information is repeated, otherwise inconsistencies arise. If information is held once only, only one item need be altered. All references to that information will automatically access the revised item.

In a relational database, the relations are designed (by a process known as normalization) to avoid repetition. For example, if a number of beds from the same borehole are described, the data about the borehole are held once, and each bed description refers to them rather than repeating them. The reference is by means of a **key field** — a column in the table that contains the identifier for the borehole data (see Fig. 1). As well as reducing redundancy, this structure has the advantage that each relation contains uniform sets of data, with similar variables for each record.

The relational database is appropriate only for well-structured data, that is, items such as categories or numerical data that fall naturally into a tabular arrangement. Some flexibility can be obtained, however, by regarding other data, such as a string of text, an image, or a string of points representing a line on a map, as a **binary large object** (BLOB). It can be held separately, in its own format, and referenced when required from a key field in a relation.

The benefits of a relational scheme come from a structure that can accommodate simple datasets, and can also be scaled up to encompass large and complex datasets. It is a robust structure that can be up-dated, corrected, re-organized and extended as required. It can be linked to simple procedures for editing data through forms on screen. Data can be retrieved through a standard language (E 6), **SQL (Structured Query Language)**. This provides a simple means of specifying which items have to be retrieved. The user specifies whether a variable or each of a combination of variables is equal to, greater than or less than specified values (numerically or alphabetically). SQL also allows the user to indicate which variables are to be returned and in what format. It thus provides a convenient and flexible interface between most relational database management systems and programs for analysis of the retrieved data.



Unnormalized data

*QS - Map quarter-sheet
*NUMB - Borehole number in QS
BoreName - Name of borehole
Top_M - Depth to top of interval
Base_M - Depth to base of interval
LithCode – Lithology Code for interval
StratCode – Stratigraphic Code for interval
. . .
and so on for all the other data items

Borehole

*QS
*NUMB
BoreName
. . .
and so on for other borehole data

Interval

*QS
*NUMB
*LithCode
StratCode
. . .
and so on for other interval data

Fig. 1. Organizing lithological descriptions in a relational database. On the left is a list of data for each bed, with the data for the borehole repeated each time. On the right, borehole information is placed in one table, and bed descriptions in another, thus reducing the need for repetition. The starred items, QS and NUMB, are together a unique identifier for the borehole. They form a "composite key". It alone is repeated in each bed (or interval) to link its description to the appropriate borehole. LithCode is a "foreign key" which links the compact, standard lithology code to a dictionary that provides the full descriptive terms. This structure is reflected in the form shown in part D, Fig. 3.

Retrieval requires some knowledge of the structure and contents of the database. The database may be used informally as a working tool, where all concerned are familiar with the contents. An outsider, however, needs access to metadata describing the content and layout. For a large relational database, the design of the relations and their links must be carefully considered. The completed design is referred to as a data model or schema. Systems analysis and data analysis may precede the implementation of the database. The results of the **data analysis** can be expressed in **entity-relationship diagrams** (see L, Fig. 5). The data are regarded as a set of entities, such as wells, cores, lithologic descriptions, stratigraphic classifications, and so on. The diagrams show the relationship between the entities. For example, samples might *be_part_of* cores, and the cores might *contain* samples. The diagrams may be supplemented by **data dictionaries** that list the recorded variables, and contain definitions of the variable and entity names. By examining the diagrams, users should be able to see what information is available and how it is related to the items of primary interest. From the data dictionaries they should be able to establish the exact sense in which terms are used.

Preliminary planning (D 4) is required to achieve a shared understanding of all the data, and avoid redundancy and ambiguity. There are various levels at which this can be attempted: within a project, an organization, an activity (such as oil exploration), or for the science as a whole. At each level there is a trade-off between local and general objectives. The complexity of the analysis obviously differs for each. A small project may call for no more than an informal record of metadata. At the other extreme, for a large organization or a general synthesis within an area of science, skilled specialists may be needed to conduct the analysis. **CASE** tools (computer-aided support environment) can provide computer support, from constructing entity-relationship diagrams to structuring the database. It has been said that data analysis aims to ensure that there is a common understanding of all the data held in a database, with no duplication, ambiguity or redundancy. These attributes could not apply to geoscience literature, where ambiguity and redundancy are essential. In principle, analysis is not restricted to a database, but could include separate text and spatial information. The need for a more comprehensive view leads to object-oriented methods (H 5).

4. Spatial data

Some vendors of Geographic Information Systems like to demonstrate the ability of their product to zoom in from a map of the whole country to an enlarged area, panning across the map to center on the point of interest. As the detail increases, a town takes shape, then individual streets and their names appear. Moving to a larger scale, buildings are individually identified. A new theme is selected, perhaps utilities — gas, water, sewers, electricity, telephone and television cables — showing their depth below street level and their exact position superimposed on a street plan. Or perhaps the interior plan of a building is displayed, zooming in to individual offices to show the position of the furniture, the wattage of a light bulb and the date it was last renewed, or perhaps a photograph and CV of the occupant with a number on which to click to establish a videophone link.

The demonstration is misleading, of course, and usually followed by a confession that with any other combination of areas and topics, the screen would be an embarrassing black. Unfortunately, most spatial data are collected in diverse projects to incompatible standards and can neither be shared nor integrated. Until global solutions (L 4) are widely adopted, each organization, or worse, each application, may have to adopt its own standards, and plan, where appropriate, for future migration to global standards.

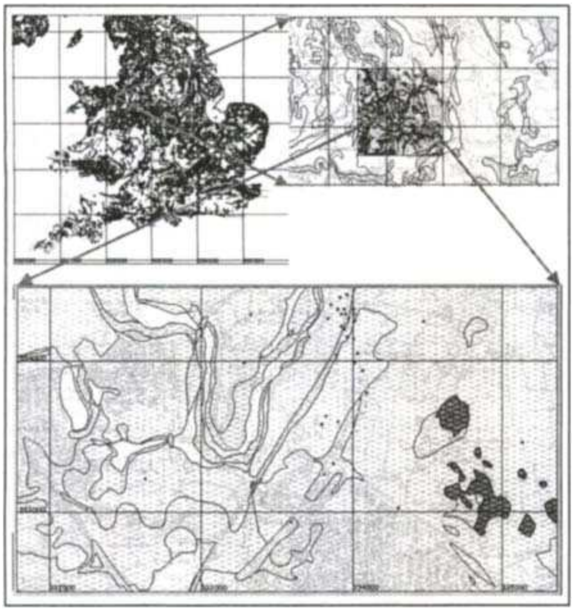


Fig. 2. Spatial search. Within a single topic, such as Drift geology, GIS enables the user to zoom into an area of interest, seeing how it relates to the broader picture. These extracts are from the BGS Geoscience Data Index. British Geological Survey ©NERC. All rights reserved. Base maps reproduced by kind permission of the Ordnance Survey ©Crown Copyright NC/99/225.

What it suggests, however, is the power of spatial search allied to visualization. Systems of this kind have been demonstrated in geoscience and are partially implemented within a few organizations (M 2.3). They offer the ability to see what data are available for which topics and where they are located. The user can visualize spatial data in their correct relative positions (see Fig. 2), and, maintaining spatial relationships, examine the pattern of their distribution and the spatial correlation with patterns of other types of data. The user might select a sequence of operations, such as the following. Look at the surface geology, superimpose contours on a subsurface horizon, zoom out to view the regional setting, zoom in to see which fossil species were found at an outcrop, and examine some thin sections on screen. Look (Fig. 3) at the gravity map and magnetic anomalies, examine well records, core descriptions and downhole logs, see the 3 d seismic reconstruction (J, Fig. 1) slice by slice, look at enhanced photographs of the landscape and processed satellite imagery (as in J, Fig. 2). Pan southwestward to look at geochemical stream sediment analyses downstream, using quantitative tools, such as SQL, to select only analyses with appropriate concentrations of defined elements.

The search techniques just described are specific to spatial information. Because we are well able to **visualize** space, it can be a powerful metaphor for other data. For example, an organization chart can show a hierarchy of employees, with related departments placed side by side to represent their organizational closeness. It could be more effective than an office plan for locating and communicating with appropriate staff. A stratigraphic table presents a sequence of formations laid out in space to correspond to their sequence in time. Entity-relationship diagrams display a spatial image of related concepts. Cross-plots of quantitative values (F, Fig. 3) correspond to a map in space. Techniques for spatial analysis are therefore not confined to data in geographical space, but also apply to a variety of other data represented in metaphorical space.

The tools for managing spatial data are available in many geographic information systems (GIS). They access an underlying database in which most of the data are spatially indexed. They must be able to provide rapid access to spatial objects. For this, the structures in which the data are held are crucial. The spatial coherence of the objects and phenomena, that is, the fact that they lie within a limited, continuous area, determines the form of a suitable structure. Arranging the data first by the x-coordinate, then by the y-coordinate, would not be appropriate, as it would split the coherent object into successive strips

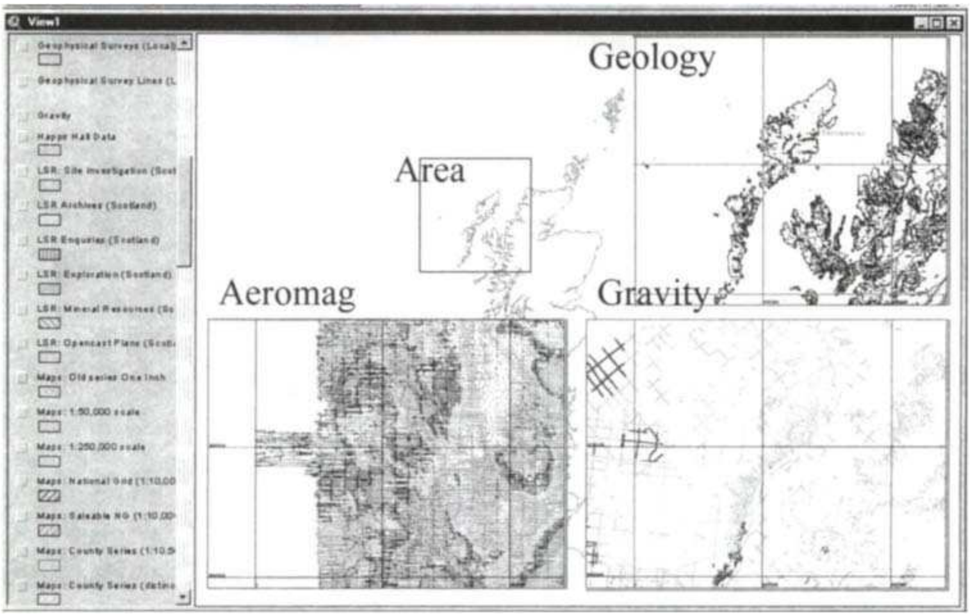

View1
Geophysical Surveys (Local)
Geophysical Survey Lines (L
Gravity
Mapper Hall Data
LSR: Site investigation (Scot
LSR Archives (Scotland)
LSR Enquiries (Scotland)
LSR: Exploration (Scotland)
LSR: Mineral Resources (Sc
LSR: Opencast Plans (Scoti
Maps: Old series One Inch
Maps: 1:50,000 scale
Maps: 1:250,000 scale
Maps: National Grid (1:10,00
Maps: Saleable NG (1:10,00
Maps: County Series (1:10,5
Maps: County Series (defini
Geology
Area
Aeromag
Gravity

Fig. 3. Aspects of geoscience as topics on a map. From a GIS-based system, the user can select and compare many properties for the same area. These extracts are from the BGS Geoscience Index. British Geological Survey ©NERC. All rights reserved. Base maps reproduced by kind permission of the Ordnance Survey ©Crown Copyright NC/99/225.

mingled with pieces of unrelated objects that happened to be as far east. Instead, a **quadtree** structure divides space into squares, which in turn are divided into smaller squares and so on. The larger squares can be referenced to show the approximate location of relevant data, and successive layers in the tree of smaller squares give progressively more detailed views of the object. Provided the quadtree structures match, they give an efficient route to studying the spatial correlation among different objects (Mark et al., 1989). Even pixels in raster format (G 1) can have a quadtree index. **Octrees** are the three-dimensional equivalent, and a means of representing, modeling and correlating objects in three-dimensional space (see Jones, 1989). The three-dimensional equivalent of a pixel is known as a **voxel.**

Spatial objects, which might include sets of points, areas and volumes dealing with a particular topic, must be identifiable. Users must be able to refer to the object or object class, and to do so must be aware of its existence. The system must therefore display metadata in a well-organized form which is easy to search. They may be provided as an index, list or menu. To take full advantage of the spatial aspects of the system, it should also be capable of displaying a summary of the object's spatial distribution as a small-scale map. Users should be able to select objects by pointing to them on a display, or defining the search area. The area may be a circle, rectangle, swathe of specified width along a line, or an irregular polygon. All of these should be selectable by moving the screen cursor. The area of search might alternatively be a polygon or polygons representing another object or object class. Objects might be retrieved depending on whether they are wholly within the search area or only partly within it, and might be retrieved in their entirety or only that part which lies within the search area. By this means, point data lying within a formation could be retrieved, or the parts of a formation coinciding with a particular lithology.

At all times, the display should provide a means of visualizing the disposition and configuration of the objects of interest. It must be possible to zoom in to see detail and to zoom out to see the context. This implies, however, that the spatial information must be available at various levels of detail, and generalization is unlikely to be a completely automatic process. The problems with providing comprehensive access to spatial data thus seem to arise, not from deficiencies in IT, but from the absence of standard methods for providing spatial data and the absence of incentives to make them available in an appropriate form. Solutions have been proposed (see L 4).

5. Object-oriented methods

By involving computer systems in the management and manipulation of information, we introduce machines into an intuitive process. In order to clarify the role of the machine, we need to take an introspective view of the process, thinking explicitly about the structure of our information, thought processes and objectives. Formal analysis may not be required, but it can help to have some knowledge of current methodologies for analysis, such as the entity-relationship modeling mentioned in H 3. Object-oriented methods (see also J 2.4) offer a more comprehensive and flexible approach to including the full range of information types and to providing a natural means of handling distributed objects and relationships. They are well suited to geoscience with its complex objects in various versions and its long and complex transactions (Worboys et al., 1990).

Object-oriented (O-O) methods address the way we represent our ideas about the real world and how, by abstracting and formalizing our knowledge, we can implement them on a computer system. Starting from our perception of the real world, the methods proceed through analysis and design to programming and database (J 2.4). They offer an integrated view of large and complex problems and can place it in a systematic engineering discipline. This is the realm of the specialized consultant, and at first sight has little relevance for the circumscribed problems of the geoscientist. Nevertheless, the insights justify some acquaintance with the techniques. Some major studies, notably POSC (1999), take an object-oriented view of geoscience information, and large organizations are moving towards a similar framework. Geoscientists should be aware of these developments and may even be able to align their ideas with them.

Object-oriented analysis and design do not necessarily lead to O-O programming or database, although this may prove desirable in the long run. O-O programs are appropriate for developing the graphical user interface, but less so for numerical calculation. At the time of writing, RDBMS are more robust and better supported than OODBMS.

Analysis is seen as the practice of studying a problem domain. It leads to a consistent set of diagrams and protocols constituting an abstract system, which can convincingly be defended as an adequate understanding of the problem. This leads in turn to a complete, consistent and feasible statement of what is needed from the computer system. **Design** then takes this specification of externally observable behavior and adds details needed for actual computer system implementation. These include details of human management, task management and data management. Careful design ensures that objects can be reused for other purposes, and that the system as a whole can readily be altered and extended.

Some authors, such as Henderson (1993), make a distinction (not followed here) between entities, which they define as existing in the real world, and objects, which are their counterparts in the computer implementation. The object is a record with attributes, each of which has a value, together defining the **state** of the object. A **method** alters the state of an object or causes the object to send **messages** — the means of communicating with another object. The interface is limited to message passing. This **encapsulation** hides the structure and implementation details of the object from other objects. It ensures a simple interface that shows only the external aspects of the object, which are accessible to other objects. It reflects **abstraction**, the principle of ignoring those aspects of a subject which are not relevant to the current purpose, in order to concentrate more fully on those that are.

The objects correspond to entities in the real world whose states and relationships we wish to track. As the entities change, there are parallel changes in the objects. The objects are grouped into **classes**, descriptions of one or more objects (an individual object is sometimes referred to as an **instance**) with common features. Instances **inherit** the features of the class they belong to, and possibly those of a higher level superclass (see, for example, Cattell, 1991; Graham, 1994; Blaha and Premerlani, 1998).

The influential **Object Management Group (OMG)** is committed to consistent development of these methods. Their documents can be found on the Web (OMG, 1997; Netscape Communications Corporation, 1997). Particularly relevant is their work on global exchange of objects through a standard interface — the common object request broker architecture **CORBA**. As objects are not restricted to particular information types, and **distributed objects** can be held on any server for access by any client, O-O methods seem to offer a flexible basis for integrating a great deal of geoscience work. They offer the prospect of harnessing the power of hypermedia to link diverse information types and objects distributed among many repositories, through a uniform user interface.

Parts A–H dwell on the benefits of IT and the nature of IT tools. The extent to which the benefits can be achieved depends on future developments. For a clearer view of how geoscience and IT will interact, we now need to reconsider our own methods of investigation: how we observe, remember and record, how we build knowledge from information and cope with changing ideas. These methods relate to the strengths and weaknesses of older systems as well as the potential of IT — the flexibility of hypermedia, the developing standards for the global network of cross-referenced knowledge, and the particular value of well-organized structures of geoscience knowledge. They are outlined in parts I–K and should help us to understand the emerging geoscience information system, and to build on initiatives and opportunities such as those reviewed in parts L–M.

References

Blaha, M., Premerlani, W., 1998. Object-oriented Modeling and Design for Database Applications. Prentice-Hall, Upper Saddle River, NJ 484 pp.

British Standards Institution, 1963. Guide to the Universal Decimal Classification (UDC). British Standards Institution, London 128 pp.

Cattell, R.G.G., 1991. Object Data Management: Object-oriented and Extended Relational Database Systems. Addison-Wesley, Reading, MA 318 pp.

Garfield, E., 1983. Citation Indexing: its Theory and Application in Science, Technology, and Humanities. Wiley, New York 274 pp.

Graham, I., 1994. Object Oriented Methods, 2nd ed. Addison-Wesley, Wokingham 473 pp.

Henderson, P., 1993. Object-oriented Specification and Design with C++. McGraw-Hill, Maidenhead, Berks 263 pp.

Jones, C.B., 1989. Data structures for three-dimensional spatial information systems in geology. International Journal of Geographical Information Systems 3 (1), 15–31.

Mark, D.M., Lauzon, J.P., Cebrian, J.A., 1989. A review of quadtree-based strategies for interfacing coverage data with Digital Elevation Models in grid form. International Journal of Geographical Information Systems 3 (1), 3–14.

Mulvany, N.C., 1994. Indexing Books. University of Chicago Press, Chicago 320 pp.

Shimomura, R.H. (Ed.), 1989. GeoRef Thesaurus and Guide to Indexing, 6th ed. American Geological Institute, Falls Church, VA.

Worboys, M.F., Hearnshaw, H.M., Maguire, D.J., 1990. Object-oriented data modelling for spatial databases. International Journal of Geographical Information Systems 4 (4), 369–383.

Internet references

Amazon.com, 1996. Welcome to Amazon.com. http://www.amazon.com/.

Biblio Tech Review, 1999. Information technology for libraries. Z39.50 — Part 1 — an overview. http://www.gadgetserver.com/bibliotech/html/z39_50.html.

DCMI, 1998. Dublin Core metadata initiative, home page. http://purl.oclc.org/dc/.

D-Lib, 1995. D-Lib Magazine. The magazine of digital library research. Corporation for National Research Initiatives, Reston, Virginia. http://www.dlib.org.

Institute for Scientific Information, 1999. Home page with information on ISI citation databases. http://www.isinet.-com/.

Library of Congress, 1999. The Library of Congress standards. http://lcweb.loc.gov/loc/standards/.

NISS, 1999. Library OPACs in HE [Higher Education in UK]. http://www.niss.ac.uk/lis/opacs.html.

NLfB, 1999. Die Bohrdatenbank von Niedersachsen (in German). http://www.bgr.de/z6/index.html.

Netscape Communications Corporation, 1997. White paper — CORBA: catching the next wave. http://developer.netscape.com/docs/wpapers/corba/index.html.

OMG, 1997. The OMG (Object Management Group, Inc.) home page. http://www.omg.org/.

POSC, 1997. POSC Specifications — Epicentre 2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/Epicentre.2_2/SpecViewer.html.

POSC, 1999. POSC Specifications — Epicentre 2.2, upgrade to version 2.2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/.

The emerging system

PERGAMON

Computers & Geosciences 26 (2000) A75–A85

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part I. A view of the conventional geoscience information system

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

We need a strategy to cope with fundamental changes in our ways of working, based on a clear view of what we do and why. A systems view is essential to relate each part to the whole. This model of the geoscience information system should take into account: the need to separate data and process to enable reuse; the modes of thought of the geoscientist and how memory orders our thoughts; the shared geoscience record (knowledge base) and its interface with users; how ideas are linked as a network; how knowledge is organized; how a general overview can be maintained and linked to detail; how projects relate business objectives to the knowledge base and incentives keep the system alive. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Systems view; Thought modes; Ideas network; Business aspects

1. A scheme of ideas

"It is a profoundly erroneous truism", wrote Whitehead (1911), "that we should cultivate the habit of thinking about what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them". But elsewhere (as quoted by Laszlo, 1972) he wrote: "in creative thought, common sense is a bad master. Its sole criterion for judgment is that the new ideas should look like the old ones ..." and "... the true method of philosophical construction is to frame a scheme of ideas, the best that one can, and unflinchingly to explore the interpretation of experience in terms of that scheme ... all constructive thought, on the various topics of scientific interest, is dominated by some such scheme, unacknowledged, but no less influential in guiding the imagination. The importance of philosophy lies in its sustained effort to make such schemes explicit, and thereby capable of criticism and improvement" (Whitehead, 1929).

Kuhn (see part K Section 1.2) resolves these apparent contradictions by distinguishing between "normal" science and revolutions in science. Normal science thrives on routine incremental additions to the body of knowledge, where the paradigm can be taken for granted. But revolutionary change calls for validation of ideas against basic principles. Our present concern is with changes to the supporting information technology that pervades all of geoscience. To grasp their significance, we need a broad view.

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

PII: S0098-3004(00)00044-3

1.2. The need for a top-down view

Specialization helped the advance of science by overcoming the limited capacity of individuals to store and process information. By working in groups, each gains knowledge in depth of a specialized topic, complementing the knowledge of others. Difficulties arise, however, in communicating between specialist fields. Trapped in bubbles of specialization, we spin in our own eddies and lose track of the mainstream.

In day to day work a **bottom-up approach** is inevitable, concentrating on the detail and then fitting the pieces together. But rather than trying to assemble the jigsaw puzzle by building outwards from the pieces in your hand, it may be better to study the picture on the box lid. The system description should follow a **top-down approach**, seeking always the broad picture without which the detail might be misplaced. We therefore look at the behavior of the system as a whole and the structure and relationships of its specialized components. The objectives are to understand and control the massive changes stemming from the current advances in IT.

1.3. A glimpse of a broad panorama

We have looked in earlier sections at many aspects of IT. It offers many benefits. Information technology can help to deliver information efficiently in an appropriate form, where and when required. It can help to represent geoscience knowledge more comprehensively, linking ideas in a more fully connected network. It can help to manipulate the information more effectively with techniques to model, search, visualize, generalize and reconcile ideas.

We look next at the process of acquiring and recording knowledge. Geoscientists maintain in their minds a general model of the workings of geoscience and quite specific information about their own area of specialization. Each scientist has a unique view of the world, but in total, their knowledge overlaps to a large extent. The thought processes of the individual are modified by interactions within workgroups, and developed by training and study. They give rise to an explicit ordering of ideas as recorded knowledge.

We shall look further at the role of IT tools in supporting the thought processes of geoscientists. We usually take our thought processes for granted. However, by considering them explicitly, we may make better use of the tools and improve the information system. To understand the processes, we can draw on the work of brain specialists, on philosophers who study thought processes, and introspection of our own procedures. Change to one part can have unexpected consequences elsewhere (B 4), so we must study the system and its interrelated parts as a whole.

This leads to a view of future systems. IT networks connect many information sources, containing narrative text, spatial data and interpretations, structured databases, computer models, references to material and links to experts. This is the emerging global **hypermedia knowledge repository**, also known as **cyberspace**. It needs structures that guide contributors and users on where to put things and where to find them. Local structures are built and maintained by geoscience information communities and supported by shared metadata. Smaller projects can relate to these and remain in tune with global developments. The system must handle incentives (cash or kudos), because driving forces keep the system alive.

2. Systems

A **system** can be defined as a collection or set of interrelated and interacting objects or entities, including their relationships and behavior, which can usefully be studied as a whole. Early writers in this field, such as Beer (1967) and Laszlo (1972), stress the wide applicability of the systems approach. They note the importance of the **gestalt** principle — that the organized whole is more than the sum of its parts. More recent writers (such as Addis, 1985; Van Lehn, 1991) have applied systems insights in specific fields, notably computer system design, and data and systems analysis. The object-oriented approach to design and analysis (H 5, J 2.4) owes much to this background.

The **information system** is where geoscience and IT meet. It deals with recorded knowledge, and its associated tools, activities and procedures. It is concerned, therefore, with how and why we collect information, process and record it, structure it, analyze it, draw conclusions, and make the results available to others. It is concerned with the information industry — not just with the work of computing specialists, such as analysts, programmers, designers, database managers, systems and network managers, but also with the more traditional work of scientists, surveyors, authors, editors, referees, publishers, librarians, archivists, booksellers, reviewers and readers. It is concerned not just with the tools of modern IT but equally with products of the older technologies.

2.1. Designing change

We might think of the conventional information system as being like a set of books, maps and reports arranged on a shelf and cataloged in a card index. With full IT support, we need a new model. It may resemble more an interwoven fabric of objects and processes in cyberspace. Such major changes mean that we must look at the architecture of the geoscience information system. **Architecture**, in this context, is defined as the structure of components, their interrelationships, and the principles and guidelines governing their design and evolution over time. We will look at the metadata of dictionaries and models that define terms and relationships; the hierarchies of object classes; the systems by which the information is managed and manipulated; and the frameworks in which information is organized to tell a coherent story.

An **information system strategy** deals with the change from existing to new ways of working. It might apply to an individual, workgroup, project or company. Regardless of the scope, it addresses three questions (CCTA, 1989):

- Where are we now?
- Where do we want to be?
- How do we get there?

The IS strategy would normally apply at corporate level, and lead to an implementation plan that concentrated on actions to introduce new IT developments. These might involve specification of new software and the design of data models using CASE tools (H 3). But our immediate concern is different. We are looking at the science as a whole, as a broad background for more detailed studies.

2.2. Subsystems, interfaces, models and metadata

For descriptive purposes, the system can be broken down into **subsystems**. We can think of each of these as a system of smaller extent, selected to include objects and processes that naturally belong together. Complexity of behavior should be incorporated within the subsystems as far as possible, thus simplifying the interfaces between them. The **interface** is that shared boundary between systems or parts of a system, or the means of interaction between them that makes joint operation possible. An interface device, for example, provides compatibility by enabling one item of equipment to communicate with another. An example may clarify the value of such an approach.

It causes few external problems if we replace an established self-contained procedure by a new one. Indeed, this can be a good way to gain initial familiarity with computing. For example, a geologist might decide to experiment with a computer technique for drawing graphic logs. Nothing beyond that single task need be affected. If, however, the routine preparation of graphic logs is to be automated, the data must be available to the computer, and a database is a possibility. Standards for recording data exist (L 4) and maybe the data could be incorporated in a shared archive, available also for drawing cross-sections and map making. Rather than storing the paper logs, the images might be recreated when required, in forms appropriate to specific users, perhaps geophysicists, geochemists or soil scientists, each with their own needs.

The self-contained task has become open-ended, hinting at broad new possibilities. Meanwhile cartographers in the same organization might be automating their map data on different principles. Solutions to small computer tasks can grow haphazardly in this way (and as described in B 4) from many different points, and because they tend eventually to overlap and conflict, the outcome is generally unsatisfactory. If a good **system model**, or description of the system, is available, then standard interfaces between the subsystems can be defined. Provided the standard interfaces are maintained, operations in one specialized subsystem can be adjusted without affecting others.

In describing the system, however, we should bear in mind another issue (K 1.2). Kent (1978, page 93) comments: "A model is a basic system of constructs used in describing reality. It reflects a person's deepest assumptions regarding the elementary essence of things. It may be called a 'world view'. It provides the building blocks, the vocabulary that pervades all of a person's descriptions ... A model is more than a passive medium for recording our view of reality. It shapes that view, and limits our perceptions. If a mind is committed to a certain model, then it will perform amazing feats of distortion to see things structured that way, and it will simply be blind to the things which do not fit that structure". It is unwise therefore to adopt a specific world view without serious thought, or to take a fixed view of a changing world.

There are many ways of looking at systems, and many possible models of varying extent. Hence, there is a need for a higher-level description of information that enables it to be understood outside the context in which it was collected. This is the **metadata**, the data about data mentioned in A 1. Metadata might contain definitions of relevant objects and refer to a model (known as a **data model**) indicating the relationships between them. Individual models are likely to refer to specific subsystems, dealing with topics such as geophysics, stratigraphy, spatial aspects, or business aspects. Sections in L 5, L 6.1 look at some attempts to relate these to one another and thus define geoscience metadata as a whole. The initial step, however, is an analysis of the geoscience information system.

2.3. Scope and components of the system

System boundaries are somewhat arbitrary, and so we must define the **scope** of the system, that is, its extent, what it consists of and how it works. We need to identify the components of the geoscience information system and their interactions, the participants and their roles, activities and driving forces. For present purposes, it is important to identify aspects of the system that are essential for carrying out its task, rather than those which reflect the limitations or historical development of technology.

Fig. 1 depicts the real world at the base and recorded information at the top, with scientists as individuals and in groups moving to and fro between them. Recorded information is shown as a triangle. At the base is the detailed raw data, collected by observation and measurement. An important scientific activity is reworking those data by generalization, interpretation, explanation and indexing. These operations progressively reduce the information to more concise forms, thus ensuring that they can be more widely shared. In doing so, they alter the original data, creating new information. The triangle in Fig. 1 narrows as the volume reduces, as a reminder of these essential processes of **abstraction**. The process transfers ideas from observation to explanation, from data to metadata.

The real world as studied by geoscientists presumably exists independently of those who study it. It can therefore be treated as a system external to the information system. Scientists, who are thought by some to exist independently of their work, are again external systems. They interface with the real world through their investigations, and with the information system through their work in studying, improving and extending the information base. The **business environment**, and the **background theory** used to explain the data, much of it from other disciplines such as physics, chemistry, biology, mathematics and engineering, can also be regarded as separate systems. An attempt to show their relationships diagrammatically (Fig. 2) makes it clear that the geoscientists, not the information system, occupy the central position, and the interactions between the major systems are mediated by the scientists. The lasting record of the scientists' work is seen in **information repositories** — the vast body of recorded information in books, serials, reports and maps, as well as archived data, cores, samples and specimens. In addition, much unrecorded knowledge is held in the minds of individuals.

Analysis of a system involves identifying components that can be studied separately in more detail, without losing sight of their interfaces and relationships. It soon becomes clear that many subdivisions, based on different criteria, are both possible and desirable. One possible basis for subdivision starts by separating the data (in the repositories) from the operations or **processes** that are applied to manage and manipulate them. The process of, say, depicting a number of features on a map, or of contouring a set of elevation values, is similar regardless of the area or dataset to which it is applied. When the processes are carried out manually, the distinction is between the data and the set of procedures and techniques. Where computer methods are used, the distinction is between the data and the program. An advantage of separating process and data is the prospect of **reuse**. A geologist who has learned the techniques of contouring can reuse them repeatedly in many different situations. Computer software can be prepared and maintained once, but used many times by many users. The same dataset can be analyzed by many different procedures. These benefits depend on clear, consistent interfaces. The means of linking separate data and processes are developed in the object-oriented approach (H 5). Following Kent (1978), this leads to a three-part division of the system: repositories, processes and their interface to the outside world.

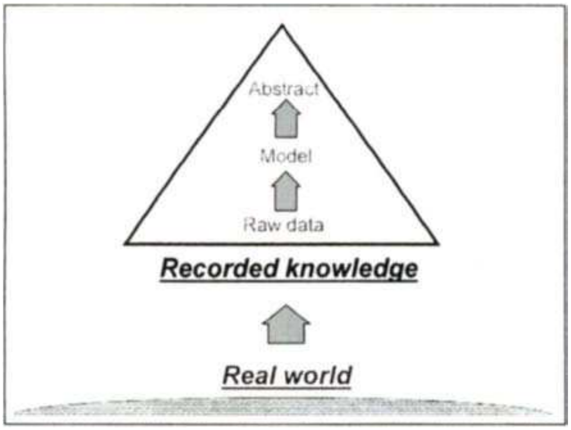


Fig. 1. Scientists collect and record information. From observations and measurements of objects in the real world, information is assembled, assessed and added to the store of recorded knowledge.

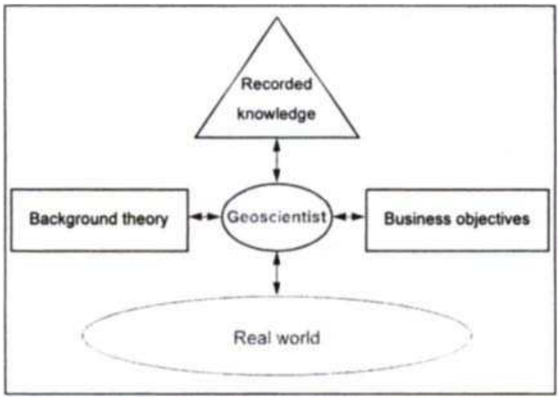


Fig. 2. Linking information to background theory and objectives. Geoscientists link the repositories of recorded geoscience knowledge and knowledge of other fields, business requirements, and their interactions with the real world.

The **user interface** (which can have more specific titles like machine interface, human-computer interface or HCI) is concerned with communication between the information system and the user (L 2). It includes output from the computer to screen or printer, and input to the computer from scanner, keyboard, mouse, or other device. With our broad definition of the information system, we must include traditional devices, such as pen and paper, typewriters and typesetters.

The user interface must give access to such functions as:

- input;
- data management;
- data manipulation;
- modeling;
- output.

It should handle these in a consistent manner. Furthermore, as mentioned in parts C–H, it must deal consistently with activities involving varying degrees of collaboration:

- personal computing, individual thought processes and concepts developed and inherited from training and study;
- interactions within a workgroup, preparation of project-centered documents and explicit metadata;
- developing the main corpus of knowledge at the corporate and global level to advance the science as a whole, maintaining the canon, and sharing knowledge with known and unknown colleagues of past, present and future.

3. A student looks at the real world

Moving on from general systems aspects, we focus first on the conventional (non-IT) geoscience information system and how it handles recorded knowledge. As we work through from this to IT-based systems, it may help to have in mind a geoscience investigation in which you were personally involved. My own choice of an example is close to my home.

Near the center of Edinburgh, Scotland, there is a tiny mountain, not 250 m high, with the unusual name of Arthur's Seat. It is the remnant of a Carboniferous volcano. Generations of geology students have been brought here to meet the realities on which the science is based. From a viewpoint on Salisbury Crags, they can look across the grassy slopes of the Queen's Park, and see the smoke-blackened tenement houses of the Royal Mile, like spikes on a long tail stemming from a smaller volcanic crag to which the Castle clings. Beneath the slopes are sandstones, known from boreholes which once provided water to the breweries beside the Royal Mile. South of the Royal Mile is part of the ancient University. Here, intellectual battles between Neptunists and Plutonists once raged (Wyllie, 1999). The eventual paradigm shift, deciding in favor of igneous rocks cooling from a molten magma, rather than precipitating from a primeval ocean, presumably resulted in wholesale reconsideration of earlier observations.

From the viewpoint, the students can be led along a track at the base of a cliff of crystalline rock, examining the color and texture of freshly broken samples. They can observe the contact between this crystalline material and the underlying bedded sandstone, with its water-worn grains set in a silty matrix; can see the hardened, discolored sediment where it was baked in contact with the igneous rock, and observe the diminution of crystal size where the igneous rock cooled rapidly at the lower contact. Observations during a short walk along the base of the igneous body suggest that it lies parallel to the bedding of the sandstone. Then, vertigo permitting, a few observations at the top of the cliff reveal similar contact relationships in reverse. Having now identified the igneous intrusion as a sill, it can be positioned on a map, with measurements by compass and clinometer of its regional slope and that of the surrounding sediment.

The strange hollow at the base of the Crags unrelated to any apparent river erosion, and the crag and tail formation of the Castle Rock, can be explained to the student as glacial features, with an invitation to imagine the huge glaciers of the Ice Age grinding overhead on their way from the distant Highlands to the sea, gouging away the landscape, impeded here and there by harder volcanic rock. An ice-striated rock face is conveniently to hand. Now gather round and look at this map showing orientation and type of glacial features across central Scotland and observe our position in the regional picture.

The volcano itself is a little more difficult. The outcrops on Arthur's Seat show the now-familiar crystalline rock. The landforms suggest sheet-like lava flows, all with a regional dip to the east. Microscopic examination of thin sections back in the laboratory confirms that successive sheets have distinct petrographic characteristics which can be conveniently recorded on data sheets. Some careful mapping in that rough terrain and a coherent story emerges of a sequence of events (deposition, intrusion, volcanic activity, tilting, erosion, glaciation) consistent with the evidence and with what is known from the surroundings and elsewhere.

From the summit looking eastwards one can see the estuary of the River Forth opening out into the North Sea. A separate excursion will take us 60 km east to the coast at Siccar Point, noting on the way the Old Red Sandstone of Devonian age, exposed along the coastal cliffs as thick near-horizontal beds. From the car park, we will stumble across the beach on rough vertical ridges of Lower Paleozoic gray shale and sandstone, etched by waves and partly obscured by sand and boulders brought in by last winter's storms. We come to a point where the vertical strata are plastered over by boulders and sand, as elsewhere but now consolidated. Following the consolidated beds we realize that we are looking at the local base of the Old Red Sandstone. The more imaginative students will perceive, as Hutton did on this very spot in 1788, that, had they been old enough, they could have stumbled across a rather similar beach in Devonian times. His visit is vividly recorded by Playfair (1805). As for the now-vertical Silurian strata, they too must once have been deposited as horizontal beds. Across Playfair's giddy abyss of time, the present is seen as the key to the past.

From the cliff-tops we can look out across the North Sea, now criss-crossed by seismic tracks, the underlying strata penetrated by thousands of wells as it developed into a significant source of oil and gas. Its three-dimensional subsurface geology is known in great detail thanks to modern instrumentation, with digital records that can be displayed on the computer screen.

After these experiences, the students may be able to draw some tentative conclusions about how geoscientists work and think, some more obvious than others. Points relevant to IT will be discussed as they arise. First, we need to review the process of investigation just described.

4. How memory orders our thoughts

We observe much, remember little, and record less. We direct our attention towards information which clarifies the developing geological picture. The more relevant the material, the longer we remember it. We extract ideas from the geological picture to modify or reinforce our background knowledge. For example, observations of heaps of loose rock fragments at the base of the cliff might enter **short-term memory** (which enables us to remember small amounts of information very accurately for a brief period). But having avoided tripping over them, we soon forget them, just as we remember in detail the flow of words from the professor just long enough to interpret their significance. **Episodic memory** holds information much longer (though not always reliably) and allows us, in our mind's eye, to recreate sequences of past experiences and events. Although the episodes are typically an autobiographical view of our own experiences, we may similarly build in memory an internal narrative of processes, states and events in the geological past (the conceptual model of B 4.1). Thus the observations which enable us to build a picture of the intrusion of the sill might be passed on from short-term to episodic memory by tying them to events in its geological history. These can be linked in turn to the broader history of the volcano, the region, and more general aspects of the geoscience model. Recognizing that the rocks involved in the intrusive episode apparently followed well-established principles relating, say, crystal size to rate of cooling, confirms the value of ideas such as the present being the key to the past. This reinforces concepts in **semantic memory**, which deals with acquiring and using general knowledge, and with background understanding of what is true and what is significant.

This suggests that our brains have a built-in ability to abstract and generalize, building summaries that are remembered when the detailed observations are long forgotten. In activities involving short-term memory, thought may provide an instant response. At the other extreme, activities involving episodic and semantic memory may give rise to deep cogitation and reflection, with a repeated review of the options, reaching conclusions only after a long period. The first may lead to entries in a field notebook, the second to publication of a considered opinion.

Psychologists and neurophysiologists have identified several levels of memory and have been able by experimental investigations to map them to distinct regions of the brain (Pinker, 1997). As well as those already mentioned, there are other levels relevant to our present purposes. **Procedural memory** involves learning motor and cognitive skills, sometimes executed subconsciously, as in trimming a hand specimen with a hammer or sketching in a notebook. **Spatial memory** refers to spatial pattern and the relative location of objects in space. These concepts have been carried across to studies in machine intelligence (in Brachman and Levesque, 1985) and to data analysis (Kent, 1978).

The different levels of memory are not wholly distinct, however. Recent research (see McCrone, 1997, 1999), which studies the working of the conscious brain with non-invasive techniques, emphasizes the intricate interconnections of the brain. It concludes that events influence, and their perception is influenced by, the state of the brain as a whole. The brain's responses to input are dynamic, non-linear and thus largely unpredictable, quite unlike those of the digital computer. Specialized regions of the brain do perform specific functions. But the brain was not designed component by component. It evolved as a whole in response to its environment over past generations and developed in response to events in its own lifetime. This fits the common-sense view that we should regard the computer not as an extension to our brains, but as a shared tool for managing and manipulating stored knowledge more effectively for the benefit of the human users.

It is clear that we must record information and knowledge if we wish to share them widely. We must store records to make them available in the future to others and to refresh our own uncertain memories. In these records, we can detect narrative text recording stories from episodic memory, data files and field notes that communicate with short-term memory, maps and diagrams that match our spatial memory, textbooks that feed our semantic memory. The immediate reasons for considering **conventional systems** (that is, those not designed with modern IT in mind) are to see where improvements are desirable, what knock-on effects they might cause, and how our legacy of recorded information can be carried forward into a new environment. The description, therefore, follows the system subdivisions mentioned in I 2.3 of interface, repository and process, to which business aspects are added.

5. Interfaces in a conventional system

5.1. Access

In conventional systems, users can interface directly with the knowledge base. This includes the formal and informal literature, in-house records and archives, computer files and the vast pool of knowledge held in scientists' minds. Access to recorded information may involve intermediaries, such as librarians, curators, record clerks or booksellers. These powerful figures are equivalent to the middleware (L 2) of a software system and influence what is visible to the user.

Conventional publications in geoscience assume that readers are familiar with current thinking. Using them effectively depends on the user knowing, or finding out, which books and serials deal with relevant topics at an appropriate level of generality and complexity. Unpublished knowledge may be passed on formally through managers and supervisors, and informally through a network of colleagues, by discussion, presentation and demonstration.

The interface is also involved when users contribute to the knowledge base. Informally, information can be passed on without delay. A typical published scientific paper, however, is prepared and edited with great care over a period of some months or years, and a similar time may elapse between submission of the completed paper and its final publication. Many copies of the paper are distributed and printed, although only a fraction of one per cent of the published copies may be read in any detail. Each contribution to the published record is identifiable, permanent and unchanging.

5.2. Connectivity

Knowledge has its source in the highly connected networks of the human mind, with many links that cannot be carried through to written records. Selections of the same material are therefore repeated in different forms for different readers, such as geophysicists, geologists, or engineers. Some connections, however, can be recorded. They include references to related papers for background, corroboration or detail; or to specific points within a paper by page number or quotation; to points on a map by geographical reference. Amendments and comments may be published later, perhaps unknown to readers.

Much information is repeated from earlier sources, probably reorganized and with added comments relating it to the specific project in hand. Each paper is like a patchwork of pieces drawn from many sources, cut to shape, augmented, and sewn together to produce something new and largely self-contained.

The reworking means that the author can repeat the story with changes, introducing a personal opinion. This gives room for local dissent within the world view, and thus for evolution of ideas. In the process, the provenance of ideas may be blurred, and the precise changes and their implications may be apparent to the reader only on making a detailed comparison with the earlier work. The information is organized and arranged to follow long-established conventions enforced by editors and publishers, but, within the house-style of the journal, each paper has a different background and distinct viewpoint. The apparent simplicity thus conceals the vast labor, perhaps the greater part of the project, spent on searching and evaluating earlier work and either reconciling it with new ideas and observations (K 1.4) or ignoring or contradicting it.

6. Conventional repositories

When an author records information for use by the geoscience community, there is a need:

- to ensure that it meets standards which enable users to understand it;
- to ensure that it is connected with other information on which it depends;
- to store it safely in suitable repositories;
- to identify it for reference purposes;
- to catalog it for subsequent identification and retrieval.

International standards define the codes and procedures for identifying each contribution, and cataloging rules which provide metadata for retrieval by author or topic (H 2). The system must be able to attract and accept contributions, evaluate them, ensure that intellectual property rights are upheld, and reward all concerned (I 8.2).

Unpublished work may be restricted to a local file. Data records are generally held by the originators, or the commissioning group, or in archives established by geological societies or other organizations. International sharing of data on, say, global seismology, geomagnetism or aspects of oceanography, relies on networks, with collaborating groups exchanging computer files to agreed standards (L 5). Cores, samples and specimens are stored in museums or in company or specialist archives, such as state-funded core stores.

6.1. Repetition

A striking feature of the geoscience knowledge base is the extraordinary amount of **redundancy**, that is, repetition of the same information. Scientists all undergo lengthy training to develop a shared understanding. The courses necessarily cover much of the same ground, and many of the differences may be due to the fact that they are taught, not by the finest teacher of that topic, but by the one who happens to be in post at that particular place. Understanding is instilled through numerous examples and expositions, and it may be left to students to arrive at their own general conclusions or world views. Whether this is an important part of their training or merely reflects the bottom-up view of the instructors no doubt depends on the circumstances. The wide availability of textbooks does much to ensure that the best ideas are available for sharing. In due course, computer-aided instruction and distance learning (Butler, 1996) will no doubt also improve standards by making the best teaching methods more widely available.

Published information is stored in hundreds, if not thousands, of identical copies distributed through many general, specialist and personal libraries, and may be independently cataloged in many of them. Repetition is also a feature of the detail, such as descriptions of vertical sections or well samples, where ditto marks are ubiquitous. Much basic information is laboriously copied from informal reports to include in published papers which repeat part of what has already appeared in other publications. The material used for teaching is likely to have been reworked many times, perhaps from an original research report to part of a regional assessment to an account of a new method to a paragraph in a textbook. Text accounts overlap with information shown on a map and vice versa. The reader who consults several sources is likely to encounter the same information in many subtly different forms.

Redundancy is inevitable, and much of it is probably desirable. Scientists, for example, could not communicate without shared understanding based on similar training. Measurements may be repeated to check their validity. Information is clarified and reinforced by repetition, which is why you may reread the same passage several times. The reader will be influenced by frequently encountering similar ideas, and repetition may therefore be a means of indicating what is important. Undesirable features of redundancy are that it increases costs and makes change difficult to control. A significant part of copying information is unnecessary and unhelpful and might well be reduced by appropriate information technology. Hypertext can link an item to any relevant context. Explicit evaluation could avoid the need to repeat for emphasis. A long-term goal may be to create an IT knowledge base in which the same recorded information can be filtered and presented in different ways for many purposes from, say, teaching to risk assessment.

6.2. Organization of content

The literature deals with entities of interest, which might be referred to as objects (J 2.4), although they are seldom identified as such and their definitions may differ subtly between authors. The object classes, such as stratigraphic units, localities, fossil or rock types, may be linked implicitly or explicitly to published definitions, or may be defined within the documents which use them. Lexicons and dictionaries provide metadata to standardize the vocabulary. Textbooks and monographs record standard procedures, methods and classifications. The relationships among objects are unlikely to be formalized in a data model, but instead are developed in a text narrative.

A geological document is concerned with more than one narrative thread of events. It weaves together a set of stories concerning various aspects of the geology (J 1.1). An account of the reasons for undertaking the investigation and of the methods and instrumentation might be woven in with separate threads giving conclusions about the geological history of the area, aspects of the stratigraphy, petrography, structural and economic geology, and so on, all related to one another in a single coherent document.

Text narrative is generally concerned with describing observations, sequences of actions during investigation, and the processes of explanation, such as a chain of cause and effect. Diagrams and photographs provide some local spatial context, but the main spatial framework may be handled as a map, possibly separated from the text. Spatial information locates observations and interpretations and describes form and shape, spatial patterns and relationships, and movements in space and time. In geoscience, the narrative accounts are also likely to refer to spatial objects in a spatial environment. Narrative and spatial information, although analyzed in distinct areas of the brain, must therefore be closely linked. In their presentation, however, aspects may be separated because information types can be difficult to combine in one publication. Maps are most useful in large format, so that the user can see the overall pattern, can trace variation or search for detail, while aware of the context through peripheral vision. They are therefore published separately, and may be produced independently of text documents for the same area. Tacit knowledge is passed on interactively through training, discussion and demonstration. It may be reinforced by informal papers, but by definition cannot be part of the conventional literature.

7. Processes in the conventional system

7.1. Generalization

As specialists, we may wish to study parts of a paper in detail. For the rest we require only an abstract or summary of salient points. It is therefore helpful to have an explanatory title, a table of contents, abstract and index that provide a quick overview and an easy route to topics of interest. Also, it is only a small number of papers that interest any one reader, and they are easily missed. A significant part of the literature is concerned with reviewing and summarizing earlier work from a variety of viewpoints. This increases the amount of redundant information, but by reading the review articles we are less likely to miss significant new ideas, regardless of where they were first published. Similarly, maps are available at various scales, so that we can maintain a general overview, as well as finding detail for areas of particular interest.

A geological survey, for example, may provide illustrated written accounts at several levels of detail, from a regional guide covering a large area to an archived description of a microfossil. It may also provide maps at different scales with different levels of discrimination and resolution, from a postcard map of the entire country to a field sketch of a single outcrop. Although geological processes differ at different scales, detail and summary are inevitably interrelated. Abstraction and generalization apply to all these aspects and to all information types. By reducing the volume of information, they make it more widely available beyond its own specialized area.

The scientific literature sifts ideas from earlier papers by reference and quotation. As readers tend to concentrate on recent work, the effect is constantly to revive the more important ideas. They are implicitly placed, amended if need be, within the current paradigm. Other, probably less significant, ideas are allowed to sink into the sludge of old, unread papers. Some useful concepts, not recognized as relevant in their time, could also disappear without trace. However, the metadata (or metainformation) created by librarians, curators and catalogers may save them from oblivion. The metadata have a number of functions:

- to provide a structure or framework of topics to organize the contents of the information system;
- thereby, to make it easier to find specific information;
- to define terminology and describe methods and procedures, thus clarifying communication.

The classifications, definitions and the rules for applying them are generally agreed by committees of experts, defined by international standards, and applied by librarians or imposed by editors. The results are in dictionaries, lexicons, library catalogs, and the placement of books on library shelves (H 2). Specialist journals reflect, and may help to define, topic-based subsystems of the information system.

8. Business aspects

8.1. Projects

Contributions to the knowledge base generally stem from projects, and can be fully understood only in the light of their procedures. A **project** is a manageable activity which has objectives, resources, and structure. It may involve one individual, carrying out, say, a site investigation or an academic research study, or it may have a multimillion dollar budget and involve collaboration among many institutes. Each geoscience project has its own objectives, determined by its business setting, and therefore is unique in its methods and information content. Perhaps because of repetition, the conventional documents needed to provide background geoscience information for undertaking a project can be surprisingly few — a small number of maps and reports, perhaps access to a small part of an archive, and the use of a handful of well-known textbooks. The results of a project may be recorded in their turn as scientific papers. A paper generally refers to a single project, although several papers may deal with different aspects of the same project. Some details of how the investigation was designed, reflecting the project and its business setting, are generally necessary to interpret the results correctly, and are therefore described in the paper.

In designing an investigation, there is a trade-off between meeting specific project needs efficiently, and generating information that can be widely shared. Conventional procedures do not impose a solution, but do respond to market forces. For instance, a speculative seismic survey may be extended to cover a wider area, in the hope of selling the results to more oil companies. The scope of an academic study may be deliberately extended to reach a wider audience, or to make publication easier.

There is significant overlap in the basic information required in many projects. As a wide variety of investigations rely on the same basic information, it can be more efficient to collect it once in a single wide-ranging survey, than repeatedly whenever it is required. Organizations such as geological surveys and shared repositories are supported for this purpose, often with state funding. Information from a geological survey, a museum archive or in-house company records, is more highly structured, but less flexible, than the general scientific literature. It attempts to provide one coherent and consistent picture through the close collaboration of many workers. Individual topics, such as limestone resources or the findings of an aeromagnetic survey, may be the subject of separate reports, maps or series, or may be included as sections within accounts of specific areas, or both. All should tie together as aspects of a shared view — a cohesive but limited part of the literature. They constitute the results of a single, large project, and are the base data on which much else is erected (M 2.3).

8.2. Driving forces

The importance of business aspects permeates the information system (M 1, M 3). The development of specialized journals may not be unrelated to the identification by commercial publishers of two driving forces: the anxiety of authors to publish papers, even to the extent of paying for the privilege; and the need for librarians to purchase the result, with little sensitivity to the price. In contrast, some worthy computer projects have neglected the potential driving forces, and failed through not anticipating the passive hostility of those who were expected to contribute but not to benefit (Peuquet and Bacastow, 1991).

Authors are among the more highly motivated participants in the information system. Their career prospects are strongly correlated with the ability to produce a stream of valuable papers. The value of unpublished reports may be judged by a manager, with or without help. The manager, who is likely to have asked for the report for a specific purpose, is in a good position to judge the result. Publications are judged by their volume and by the prestige of the journals in which they appear, reflecting decisions by the editor and referees employed by that journal. The extent to which other authors cite the papers may also be taken into account, and with books, the comments of reviewers may carry some weight.

Likely motives of publishers, booksellers and editors are to earn an honest living and prosper, influenced or even dominated by enthusiasm for the science. The enthusiasm is likely to be shared by the referee and reviewer, possibly reinforced by a desire to remain in the network. Catalogers, curators and librarians are normally paid for their work, although this by no means rules out enthusiasm. They influence what is available and who sees it. Readers are presumably the most important participants, being the ultimate beneficiaries of the system's existence. Their motives are presumably to gain information for a variety of reasons, including curiosity, which reflect their business concerns. In the circumstances, their role is a surprisingly passive one. They seem to have little direct influence on the evaluation, operating instead through intermediaries or in their other roles such as authors or referees. The cost of the system is largely paid, not directly by the readers, but by government or institutional support for libraries. Publications costs may be recovered, but seldom the cost of the underlying research, which is paid for by other means.

In due course, these tasks will be handled differently. Any new structure must, however, incorporate the legacy of earlier information and work practices. The alterations and extensions which new technology is introducing need a solid foundation in the conventional knowledge base.

References

Addis, T.R., 1985. Designing Knowledge-based Systems. Kogan Page, London 322 pp.

Beer, S., 1967. Cybernetics and Management, 2nd ed. English Universities Press, London 240 pp.

Brachman, R.J., Levesque, H.J. (Eds.), 1985. Readings in Knowledge Representation. Kaufmann, Los Altos 571 pp.

CCTA, 1989. The Information Systems Guides. Wiley, Chichester.

Kent, W., 1978. Data and Reality. North-Holland, Amsterdam 211 pp.

Laszlo, E., 1972. The Systems View of the World. Braziller, New York 131 pp.

McCrone, J., 1997. Wild minds. New Scientist 156 (2112), 26–30.

Peuquet, D.J., Bacastow, T., 1991. Organizational issues in the development of Geographical Information Systems: a case study of U.S. Army topographic information automation. International Journal of Geographical Information Systems 5 (3), 303–319.

Pinker, S., 1997. How the Mind Works. Norton, New York 660 pp.

Playfair, J., 1805. Biographical account of the late Dr James Hutton, F.R.S.Edin. Transactions of the Royal Society of

Edinburgh, vol. V.-P.III. Reprinted 1997, In: Hutton, J., Black, J. RSE Scotland Foundation, Edinburgh, Scotland.

Van Lehn, K., 1991. Architecture for Intelligence — The 22nd Carnegie Mellon Symposium on Cognition, Lawrence Erlbaum Associates, Hillsdale, NJ.

Whitehead, A.N., 1911. An Introduction to Mathematics. Thornton Butterworth, London 256 pp.

Whitehead, A.N., 1929. Process and Reality. Cambridge University Press, NY 429 pp. Reprinted 1969.

Wyllie, P.J., 1999. Hot little crucibles are pressured to reveal and calibrate igneous processes. In: Craig, G.Y., Hull, J.H. (Eds.), James Hutton — Present and Future, Special Publications, vol. 150. Geological Society, London, pp. 37–57.

Internet references

Butler, J.C., 1996. Another node on the Internet for those with interests in geosciences, mathematics and computing. http://www.uh.edu/~jbutler/anon/anon.html.

McCrone, J., 1999. Going inside — the neuronaut's guide to the science of consciousness. http://www.btinternet.com/~neuronaut/index.html.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A87–A97

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part J. Human requirements that shape the evolving geoscience information system

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

The geoscience record is constrained by the limitations of human thought and of the technology for handling information. IT can lead us away from the tyranny of older technology, but to find the right path, we need to understand our own limitations. Language, images, data and mathematical models are tools for expressing and recording our ideas. Backed by intuition, they enable us to think in various modes, to build knowledge from information and create models as artificial views of a real world. Markup languages may accommodate more flexible and better connected records, and the object-oriented approach may help to match IT more closely to our thought processes. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Thought processes; Human communication; Geoscience markup language; Object-oriented analysis

1. Communication at the interface

1.1. Interwoven threads

A great deal is lost when we force our thoughts into the straitjacket of shared conventional records (part I, section 6). Imagine yourself leading a geological field excursion. You would certainly talk, producing strings of words — narrative descriptions, accounts of sequences of events, reasoning and explanations. You would weave the ideas together to tell a story, probably supplementing the narrative with gestures, pointing to features of interest and drawing diagrams, perhaps with a stick in the sand. You might look at detail with a hand lens, then stand back to see the wider picture. You might refer to recorded knowledge: "I will read you a brief account of the regional geology; you can see where we are on this map". You might pass on tacit knowledge by demonstration: "look at the outcrop here and you will see what I mean".

The spatial context of your observations and hypotheses gives them coherence, but the narrative is essential to tie the elements together. You might cope with leading a field excursion with a broken arm, but would have problems if you lost your voice. Different parts of the brain focus on different types of information, but given the opportunity, they work together for a clearer view of the big picture (or the big story).

Our experiences may be single sequences of events, put in words as a narrative thread. Repetitive events, like seeing a similar sequence of beds again and again, merge into a single strand, with only exceptions (the

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00045-5

fossiliferous bed) remembered separately. In our memories, the threads are woven together as a complex fabric. Our brains are constantly trying to make sense of our experience by drawing analogies, abstracting and summarizing. The result is ideas with form and structure. We have each built our own background understanding or world view that provides the frames (K 1.2) in semantic memory where we can accept and evaluate new ideas.

The ability to integrate information types, so important in the field, is hampered by the need to package recorded information in separate containers for narrative (books and reports), spatial information (maps), data (databases) and discourse (discussions). We therefore need to look more closely at what we are trying to do and how we might prefer to do it.

Three long-standing tools for sharing knowledge, as well as helping individuals to develop and organize their own thoughts, are language, image and demonstration. Two more recent tools are mathematical methods and computer software. These tools shape the way we think and what we think about, as well as the way we perceive the world. They correspond to different **information types** (text, spatial, tacit and structured information) and lead to different styles of thought, presentation and processing procedures. The earlier techniques evolved so long ago that we have lost sight of their origin, but in planning for new technology we must bear in mind their characteristics. The Encyclopedia Britannica (1973, *Language*) is as always helpful on such matters.

1.2. Language and narrative

Long ago and far away, our ancestors grunted and made noises, and then sequences of grunts, ascribed meaning to them, and evolved a living **language** — a sequence of tokens denoting objects, actions and agents. At a similarly primitive level, our ancestors looked around them, scratched maps on the sand, drew pictures on the cave walls. Spoken language could be used for discussion and command. It was a means of communicating to a group, for the speaker could broadcast the same message to all within earshot. Images were at first a secondary, more personal communication, directed to a select few, and requiring special skills even to draw crude sketches. But suppose early man could have broadcast not just to the ears, but also to the eyes of his colleagues, with the ability to capture, record and communicate images in full detail to one and all. Methods of communication would surely have evolved differently. That power, denied to our forebears, is now available to us. Does it make a difference? Has an optimal system evolved naturally, or are we constrained by historical accidents?

Leatherdale (1974) pointed out that sixteenth-century philosophers assumed that, in experience, we always encountered well-defined or discrete "things", and that they seemed to conceive of the business of language as the adroit matching of words to these separately given things of which they are the mark or sign. The more recent view is that language can be explained in terms of socially agreed understanding of words: not on any intrinsic connection between signs and the world, as that would imply absolute properties of language independent of human culture. Contexts need to be invented, and stories created, to make a character string meaningful (Laszlo, 1972).

In evolutionary terms, episodic memory (I 4) presumably developed as a means of learning from one's own experiences — a single thread of events winding through the continuum of space along the arrow of time. Language matches this pattern of thought with linear strings of words and sentences, referring to past, present and future. Narrative skills evolved to create **stories** as surrogates for experience, told, retold and remembered (inaccurately). The story that began as a surrogate experience may by constant reworking acquire the mythic quality of a folk tale. It **generalizes** by pulling out crucial, illuminating events, implying much more than it states by reacting with existing ideas in the listeners' minds and influencing their semantic memories.

From speech, written language evolved to reach a multitude of users separated in space and time. **External representation** of knowledge (outside the human mind), for example, by writing and drawing, goes far beyond the here-and-now of storytelling. Scientists, in building their knowledge base, are not limited to their own episodic memory. They can contribute to and access a vast repository of information. They can do so at a time and place of their choosing, examining a wide range of specialized sources, past and present, in summary and in detail.

Faced with the infinite complexity of the real world, a complete description of it is unthinkable. Instead, stories are told in innumerable ways, to illustrate a multitude of experiences from personal viewpoints. Yet, since they deal with the same reality, all refer in a sense to the same underlying story. Because the overall story is large and complex, scientists must specialize in subdisciplines. A major part of storytelling, or scientific writing, is therefore devoted to linking to earlier accounts, establishing common ground, clarifying, reconciling inconsistencies, and noting new information, discrepancies and ways to resolve them. **Discourse**, that is, the expression and interchange of ideas, is the means of clarifying and reconciling the accounts, by discussion within a workgroup, or through the slow process of conventional publication.

1.3. Spatial concepts

We can describe spatial pattern and relationships, rather inadequately, in ordinary language such as: "filling broad channels from a northerly direction which were later buried to a great depth". However, evolution provided us with specific abilities to handle and memorize spatial data (McCrone, 1999). Spatial skills bring evolutionary advantages in catching things, or not bumping into them, as in swinging from tree to tree, where (with luck) the main sensory input is by eye. We can communicate exact and accurate spatial information as **images** (representations of the semblance or likeness of an object), such as diagrams, sketches, maps, photographs and video clips, which make use of these skills. Note, however, that losing ambiguity is not always helpful. If we depict the channels described at the start of this paragraph on a map, they either join up in one direction or the other, or both, or not at all. This might bring unwanted connotations of tributaries, deltas, braided streams or whatever. The ambiguous statement may reflect genuine ignorance. The graphical representation can force an appearance of certainty that does not reflect the real situation.

Unlike narrative that places events in sequences of single threads, spatial thinking lets us build extensive structures, such as geological maps and cross-sections, that give a comprehensive view over space and maybe geologic time. We can zoom in to see the detail, zoom out to see the spatial context, pan around to see the situation elsewhere, and compare spatial patterns arising from different topics. Narrative text cannot offer these abilities, but can be intimately linked to a spatial representation.

We can represent spatial forms by combinations of geometrical objects, such as points, lines, areas, surfaces, fields and volumes, with well-defined mathematical properties. They can therefore be handled precisely on the computer. Within this rigorous framework, we can assemble spatial objects drawn from a wide range of topics, say, topography, borehole records, formation boundaries, lithologies, engineering geology, planning zones and proposed construction sites, and process them together. We can visualize the location, spatial patterns and relationships of sets of objects using interactive computer displays that take advantage of our spatial skills and the accuracy of human short-term memory. Our visual systems have evolved to process moving images, helping gannets to catch fish and motorists not to collide. Computers can exploit this talent, helping us to visualize changing spatial patterns, like the development of a sedimentary basin.

1.4. Structured data

Much more recently than language and images, another type of knowledge representation was invented, namely, numerical measurement and quantitative modeling (F 1). An advantage of numbers is their ability to define relative magnitudes as precisely as required. The eye is adept at comparing two magnitudes, for instance the relative sizes of two fossil specimens seen side by side. It is much less skilled at comparing objects seen in different places or at different times. **Measurement** against a standard scale uses the numerical series, the most exact order we can form. It provides the portable yardstick that makes it possible to assemble any number of side-by-side comparisons, as well as comparing the magnitudes as a group and examining subtle variations in space or time (also measured numerically).

Mathematical models (F 5) build on this ability of numbers to represent magnitude, and on the similarities between mathematical operations and physical processes. Sets of measurements and relationships between them can be summarized, properties can be sampled to represent the underlying situation, uncertainty can be measured, states can be compared and processes simulated. Practical applications in geoscience, including the quantitative description of spatial data and analysis of images, depend largely on computer support, and a computer program is a precise and convenient means of sharing the model.

Mathematical and spatial models (along with the data collected to investigate them) are both analogies (J 2.3) between the real world and the properties of numbers. Quantitative properties and relationships can be visualized as patterns in space. The map can also be seen as a means of visualization (MacEachren, 1998). We can thus bring quantitative and cartographic methods into a single numeric framework. Both analogies are at best fuzzy approximations to the real world, despite the exact mathematics of the internal reasoning.

The standards, rules and conventions of the more highly structured areas of the information repository add value by creating a coherent and easily accessible database, possibly derived from many independent sources. All are ultimately set within text narratives, which explain the objectives, the conventions, the reasoning and the conclusions. Detailed narrative threads may refer to quite specific spatial items or aspects of spatial and quantitative models. The reader should be able to view the relevant items highlighted against their spatial context, and be able to move freely between the narrative and the visualization.

1.5. Tacit knowledge

Perhaps the greatest amount of geological information is held, not in the written record, but in total in the minds of all geologists. The geologist who has surveyed an area develops a mental picture of the geology more complete than that shown on maps and described in papers. Much is **tacit knowledge** which is acquired through practice and cannot be articulated explicitly (Kuhn, 1962) — known, but not expressed. For example, you might instantly recognize a specimen which you could not identify from the most exhaustive description, just as you learned to ride a bicycle by demonstration (transferring knowledge held in procedural memory), not by written instruction. In a discussion, or a field excursion, much can be learned that could never be written down. The importance of tacit knowledge means that education, training and learning throughout a scientist's career need demonstration, discussion and directed experience — to communicate what we cannot express.

A reminder can recall forgotten memories. Hence, the menus on a computer screen. *Recognition* of a command is easier than *remembering* the exact wording of a computer instruction. A valuable feature of an information system may therefore be to "jog the memory", to present cues and analogies that can stimulate ideas in the pool of tacit knowledge. A second valuable feature can be the ability to access the tacit knowledge of others through discussion and inquiry. A third feature of the system could be the use of images and video demonstrations to illustrate, for example, the procedures used in collecting samples, or the precise points examined on an outcrop. These could clarify a narrative account, and would help others to repeat and test the observations.

1.6. Knowledge-based and rules-based investigation

Examples presented to students (I 3) are not typical of the procedures of experienced geologists surveying an unknown area. During an initial survey, a comprehensive set of observations is likely to be made and recorded, if only to avoid the cost of revisiting each outcrop. For a graduate research project, a local, self-contained problem might be sought, preferably with significance in a wider context. The abundance of such problems in geoscience makes it an attractive subject to study. On the other hand, the search for oil is more likely to employ techniques that are well established on a global scale, in tune with the uniform business objectives. The fact that the model is well defined before the investigation begins (being based on a clear user requirement) increases the scope for rules-based activity and so for automation.

This points to a distinction between exploratory investigation and systematic pre-planned investigation. The first is **knowledge-based**, starting from some preconception of the geological setting and developing and extending the explanation with each new observation. Evidence is constantly sought, by new observations or reworking data, to confirm or modify the current interpretation and choose the next step of the investigation in the context of growing background knowledge. A narrative account is built in episodic memory. The second is **rules-based**, where the pattern of the investigation is decided before work starts. In contrast to the exploratory search of the knowledge-based project, the rules-based project is analytical, studying known relationships by collecting and studying appropriate data, following a well-defined model. Standard procedures are specified, and instructions set out for following them. The resulting data are therefore consistent, and can be compared with one another and with data from other projects that followed similar rules. Short-term memories are recorded, and can be accurately recreated from the database.

Rules-based projects can be well suited to quantitative measurement and extensive instrumentation. The seismic investigations of the North Sea, and downhole logs from the subsequent drilling, provide examples of data collected by instruments according to pre-defined rules. Their consistency makes them particularly helpful in revealing a regional pattern. A rigid, pre-determined structure can also extend the reach of the designers of an investigation by delegating data collection to instruments or methodical human data gatherers. The plan is inflexible and cannot readily be adjusted in the light of the initial findings.

Knowledge-based projects have fewer precedents to guide the activity. They explore the unknown, and procedures must be modified as more is learned. The students who arrive at the outcrop not knowing what they are going to see are in this position. In fact, projects are likely to involve both rules-based and knowledge-based procedures. For example, a seismic survey may collect data according to a predetermined scheme, but the interpretation of those data is knowledge-based, evolving as ideas are tested and knowledge of the geology of the area grows. The student, carrying out an exploratory investigation, may nevertheless follow predetermined conventions when measuring strike and dip, and might even randomize the sampling procedure to aid subsequent analysis. Every rules-based activity is ultimately set in the knowledge-based framework of the science as a whole.

The distinction between knowledge-based and rules-based activities is important in the present context, because (work on artificial intelligence notwithstanding) knowledge remains largely the preserve of the human being. The machine, on the other hand, can be adept at following rules. Bear in mind, however, that automation can *support* free thinking and trial and error. An IT response to knowledge-based activities is to use flexible multimedia, creating fully connected and searchable documents. An IT response to rules-based activities, on the other hand, is a rigorously defined model and database supporting standardized applications. It makes full use of instruments for data collection and analysis.

The computer is well suited to accurate long-term storage of complex images and detailed tabular data, such as lists of fossils or results of geochemical analysis; but the brain (where accurate detail is restricted to short-term memory) is not. The brain *is* well adapted to handling, within a frame of existing background knowledge, the loose structure of descriptions, analogies and explanatory reasoning typical of a narrative account; but the computer is not.

IT should aim to harmonize knowledge-based human skills with rules-based computer modeling. As the approaches overlap and combine, the information system should optimize the abilities of both. An important part of the solution is **interactive computing** — a conversation between the user and the machine, in which the screen display is modified rapidly in response to instructions or decisions entered usually by keyboard or mouse. The ability of the computer to follow rules quickly and accurately is complemented by the ability of the scientist to use background knowledge to control the progress of the activity.

1.7. Modes of thought

The various information types (text, spatial, structured, tacit) are handled differently by the brain. Each supports a different style of thought, communication and IT system. We use them, separately or together, to model knowledge and information in various modes of thought and investigation, such as the following:

Narrative — one can pass on information, or develop a point of view, by telling a story. As Francis Bacon pointed out in 1652 (see Leatherdale, 1974), one can revisit the reasons for reaching a conclusion by going over in one's mind the events that led to it. Each part of the narrative depends on the story so far. By telling the tale to others, they too may follow the line of reasoning.

Temporal — geological explanations trace the course of past events, relating observations to a conceptual time sequence of past conditions (states) and processes.

Spatial — geoscience maps and images provide the means to locate observations and link them to spatial pattern and spatial relationships. To understand the pattern of ice flow, or sequence and extent of lava flows, the students (I 3) were led naturally to a map. The meaning of the observations depends on their spatial context.

Demonstration — narrative description is more powerful if augmented by actually demonstrating what happened, in the field or laboratory (the **ostensive** approach). For a full picture, it may be desirable to retrace the work, and so share the experience, of an earlier investigator.

Quantitative — the benefits of precise measurement have wide application and are immediately obvious in, for example, hydrocarbons exploration, where detailed prediction of the form and properties of the strata is required, leading to estimates of the location and magnitude of the oil and gas reserves.

Statistical — statistical theory provides a rigorous basis for marshaling complex evidence for testing hypotheses, and estimating probabilities and uncertainty, by computation from appropriately sampled observations and measurements.

Process-response — the concept that physical, chemical and biological processes operated in the past as they do now, is the basis for much geoscientific thinking. A coherent picture of past processes should be internally consistent and should predict responses (consequences) comparable to those of present-day systems.

Experimental — some processes and responses can be explored by experiment, that is, under circumstances that the scientist can control, leading to a more exact understanding of the relationships.

Trial and error — where the course that an investigation will take is not clear, a **heuristic** approach may be adopted, trying out a range of possibilities, following those which seem most successful, and modifying them as more is learned.

The information type and mode of thought play a large part in determining whether IT methods are relevant and which methods are appropriate. Conversely, technology influences the ways we think and the combinations of modes of investigation.

1.8. The need for a Geoscience Markup Language

This chapter is concerned with where we want to go, not how we get there. However, the requirement may be clearer if we have a mechanism in mind. Conventional methods of recording information have deficiencies. The poor connectivity enforced by earlier technology results in high redundancy and inflexibility. Processes to manage, manipulate and explain information are frozen along with their representation. Change is cumbersome, because minor corrections in the literature are easily overlooked, and when ideas change, the full knock-on effects on other work are seldom obvious.

We are looking for a mechanism that can offer better facilities for new investigations, while incorporating legacy material. A markup language could be one approach to improving communication. It can represent conventional narrative text, but can also include tags. Unlike HTML, where the tags control presentation and links, XML (E 6) can also tag words or sections by topic (this is a fossil name, this is the section on structural geology). Presentation is handled separately through a style sheet and can be controlled by the reader.

As its name indicates, XML is extensible. The ability of users to define their own tags could result in a large and unwieldy language. Therefore, specific dialects of XML have grown up and been partly standardized within subject areas. Thus, Chemical Markup Language (CML) can tag chemical formulae, and display them, or models of the molecular structure, with a choice of conventions and notations. Other dialects have been developed for fields such as mathematics and music, to handle their specialist notations and offer flexibility in presentation.

A Geoscience Markup Language (GML) could also be a specialist subset of XML, and thus have the ability to tag words or sections (modules) of the text to reflect their topic. Such modules could be linked to others, within the document or elsewhere. We think of the conventional literature as subdivided into encyclopedias, books, serials, reports, notes, maps, and so on. The subdivisions are based on physical form and process of delivery. Instead, we could visualize a GML document as a collection of modules brought together for a particular purpose, with many of them reusable in other contexts. Authorship takes on a different meaning where modules are assembled from many sources, possibly offering alternative explanations of the same phenomena.

A markup language, however, also offers the possibility of linking text closely and selectively to modules from, say, metadata, data, software, models, demonstrations on video or sources of expert advice. Many of these would be accessed through a database management system or a GIS (E 5), although this need not be apparent to the casual user. A module could be displayed in different ways to meet individual needs (a table of data or a graph, a contour map or a perspective view). It would be an integral part of the document. But it would also offer the possibility of moving to and from a different environment, such as a GIS to explore the spatial context, or a database for comparison with analogous datasets from other investigations. Another user option could be to select the level of generalization, while retaining the ability to drill down to additional detail when required. Change to one module would be seen by all modules connected to it, and knock-on effects could be traced through the links.

The technology is largely in place, and starting to operate in some other subjects. Such languages immediately offer greater flexibility and expressive power to the author. Geoscientists have special needs for interworking with spatial, historic and stratigraphic information, and have their own vocabulary and procedures. They could therefore justify a separate dialect of XML. But there is a long and painful learning curve, and geoscientists will be involved in much trial and much error before a robust solution for everyday use can emerge.

With an appropriate markup language, linkage to the underlying context of assumptions, laws and hypotheses could be recorded and therefore the effects of changes in the underlying ideas could be clarified. To take this further, however, we need to consider the process of building knowledge from information, and the object-oriented approach that stems from this.

2. Processes and the repository

2.1. Explanation

Having looked at communication (J 1), the next question is how scientists explain their observations and make sense of streams of observational data. How do we build knowledge from information? The Encyclopedia Britannica (1973) is again helpful with its entry on *Scientific Method*, where it defines the pursuit of science as "the search for knowledge and understanding through formulation of the laws of nature". The **theoretical function** of science is that of providing explanations of natural phenomena by discovering relationships between these phenomena and other events. These relationships fall under general laws that enable us to make predictions as to what events to expect in particular circumstances, and sometimes, by controlling the circumstances, to ensure that these events will occur. The **practical function** of science, that of enabling us to adjust our lives to nature and, sometimes, nature to our lives, thus derives from its intellectual function — that of explaining phenomena by means of scientific laws.

A starting point for scientific explanation is **classification** (systematically assigning objects to categories based on their properties). Recognizing an object as a sediment or an intrusion has implications about its geological behavior. The words used to name objects are nouns. Adjectives, like red, angular or hard, describe their attributes, but are less useful for classification, saying little about how the objects behave. The students' visit to Salisbury Crags (I 3) began by distinguishing, identifying and naming the sedimentary and intrusive rocks, relating them to their behavior in the geological past. The extension of this activity to formal data analysis is described in H 3.

Scientific discovery involves finding **hypotheses** (suppositions made as a starting point for further investigation) that could be refuted by further experience (see Popper, 1996), but which nevertheless survive testing by observation or if possible by **experiment** (observations made in circumstances over which the scientist has control). A hypothesis in I 3 was that the rocks of the cliff face are part of an intrusive sill. This was tested by thinking about what additional observational evidence might be found and then looking for it. Although the original events are beyond the reach of experiment, a detailed model of aspects of the processes, say, the baking of the sandstone, might be tested experimentally given appropriate facilities to replicate the high pressures and temperatures involved. The purpose may be to find more laws about the behavior of the things we can observe or to incorporate the results in a broader explanatory theory.

Scientific laws enable us to organize our thinking into coherent systems, as well as to make predictions. The laws are at many different levels of generality, arranged in a hierarchical system in which laws at a low level are logical consequences of sets of laws at a higher level, and so on. The lowest-level laws are general propositions whose instances are directly observable facts, but higher-level laws may be theoretical concepts in a wider system explaining new phenomena. Explanations in geoscience may present the observed situation as a logical consequence of preceding events and of more fundamental regularities, such as the laws of physics, operating on initial conditions of a geographical or historical kind. If we view the process of successive explanation as the erection of a hierarchy of laws of increasing generality, there is no reason to prevent different hierarchies from being constructed in different ways to explain the same phenomena (Kent, 1978).

2.2. Analogy

"To explain the origin of hypotheses I have a hypothesis to present. It is that hypotheses are always suggested through analogy. Consequential relations of nature are infinite in variety and he who is acquainted with the largest number has the broadest basis for the analogic suggestion of hypotheses" (Gilbert, 1896, quoted by Leatherdale, 1974).

Analogy is the resemblance in some particulars between things otherwise unlike. Analogy can be a resemblance in an ensemble of qualities, or of properties or attributes. In metaphor, the mind sees and expresses an analogy. The metaphorical use of language in science arises when familiar vocabulary is extended to describe novel insights and interpretations. Thus, in coining the term electric current, ideas were carried across from the familiar current in a river. **Metaphor**, according to the Oxford English Dictionary, is "the figure of speech in which a name or descriptive term is transferred to some object different from, but analogous to, that to which it is properly applicable". Analogy in logic, according to the same source, is the process of reasoning from parallel cases, based on the assumption that if things have some similar attributes, other attributes will be similar. Most of the truly fruitful facts about nature, Leatherdale (1974) suggests, have been discovered by reasoning from analogy.

According to Leatherdale, explanation involves an inescapable use of analogy. This is partly because the unobserved part of the description in an explanation, being unobserved, cannot be directly described. It must be verbalized and conceptualized in terms of other experiences. Explanation works by analogy of content, as well as of structure. When the analogy is well marked in terms of content, or observable characteristics, we speak of a **model**. The model is an essential tool, in that it enables us to think about the unfamiliar in terms of the familiar.

Models enable us to construct and meaningfully describe the concepts of theories in the same way as metaphors enable us to think about or describe things or concepts not normally describable in a literal vocabulary. Because they function in this way, they give meaning to, and thus an explanation of, theories. This in turn enables them to connect theory with observation and experiment. Thus, in the belief that past processes obeyed the same physical and chemical laws as today, analogies are drawn in geology with present-day processes, as in comparing an unconformity to a present-day erosion surface, in seeing finer crystals as indicating more rapid cooling, or in explaining changes in sandstone petrography as baking against an igneous intrusion.

2.3. Model and reality

Geoscience investigations are usually concerned, not with creating a new model, but with refining an existing one. They build on earlier work and must be closely linked to past records. The model is concerned not only with what is there, but also how it came about — how the operation of physical, chemical and biological processes, in a sequence of events in geological time, brought about the observed consequences. The model influences the classification of objects. For instance, the geologist sees an important distinction between a granite and an overlying pebble conglomerate, despite their similar composition and appearance, because they formed in quite different circumstances.

As pointed out earlier (B 4), the neat and tidy classification of rocks shown on a geological map or reported in the literature is unrealistic. The overlap, ambiguity and uncertainty, so painfully apparent in the field, have been banished. Crisply bounded areas of uniformity have somehow replaced the fuzzy boundaries and mish-mash of intercalated variation. Goodchild (1992) suggested that: "We need better methods for dealing with the world as a set of overlapping continua, instead of forcing the world into the mould of rigidly bounded objects." Quantitative techniques (F) are a possible candidate for representing the geological "continuum", although Mandelbrot pointed out that continuity is conspicuous by its absence in natural phenomena (G 6). We now need to consider whether the categories are necessary, or an artifact imposed by inadequate technology.

The distinction between model and reality is an important one. The model must be tuned to human thought, while reflecting something useful about the real world. A continuous model, such as a contour map, may be an appropriate representation of discontinuous reality, as long as the discrepancy is not important in the context in which the model is being used. The danger comes when the limitations are forgotten and questions that cannot be properly answered by the model, such as the length of a coastline (G 6), are addressed within it. Separate models relying on different assumptions are required for different purposes. For example, a statistical model might regard a process as deterministic and predictable, together with a superimposed random element for which only the statistical properties (as opposed to individual instances) could be predicted. On the other hand, a dynamic nonlinear model might regard the process as deterministic (in the sense of following natural laws), but unpredictable because small variations in the initial conditions could lead to a large change in the outcome (Baker and Gollub, 1996).

There is another issue. The model must be one that the available data and technology can support. New IT solutions extend the range of models that are realistically available. Looking at the computer display illustrated in Fig. 1, for example, there is a clear possibility of modeling three-dimensional rock bodies in new ways. The image can show discontinuous areas. Zooming in to part of the image could cause the areas to fragment and reveal more detail, as discontinuous as before. The scope of the model in Fig. 1 is limited. It is a stunning image when seen in full motion on the screen, and no doubt serves its purpose well. But it is tied to just one set of properties, related to acoustic impedance within a body of rock.

Take another example. Satellite imagery records a

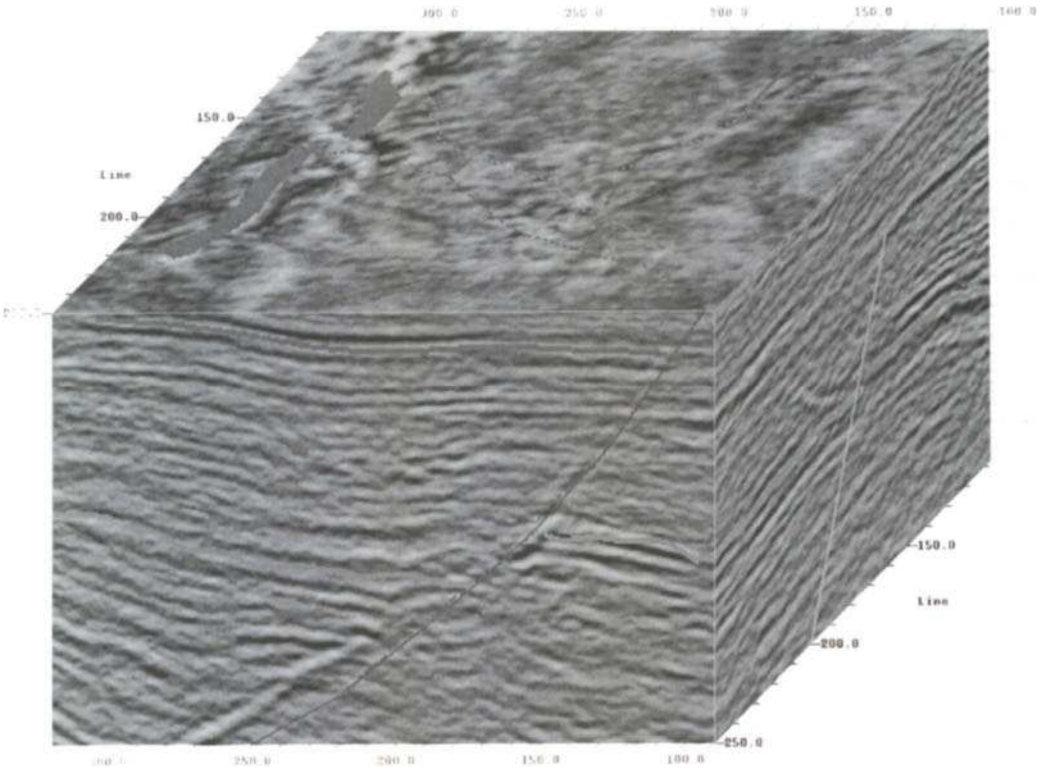


Fig. 1. Display of 3-D seismic data. Animation enables you to move through the data volume to follow structural and stratigraphic trends. Reproduced by permission of Landmark Graphics Corporation. More at http://www.lgc.com/

number of related properties, namely separate bands of the electromagnetic spectrum measured simultaneously. The properties can be analyzed quantitatively and, for example, classes based on discriminant analysis (F 5) can be displayed (Fig. 2). They again serve their purpose well, and are a useful reminder of the variation hidden in conventional cartography. But they complement and extend earlier methods, rather than displacing them. One reason is that three-dimensional seismic surveys and satellite imagery are unusual in their dense and regular sampling patterns that make detailed analysis possible. Relating them to other variables sampled on other patterns by other means calls for background knowledge and human interpretation. Limitations of access and measuring procedures mean that the data for most variables are inevitably inaccurate and incomplete.

Our aims cannot be solely descriptive, for geoscience is concerned with recreating a story from incomplete evidence. It is a story about objects, identified by nouns, given meaning by the models in which they participate, and their properties, described by adjectives of qualitative assessment or quantitative measurement. The creation of the main object classes, such as stratigraphic units, and the assignment of instances to these classes, seldom depend on subtle quantitative comparisons. They depend on drawing analogies and spotting crucial features. They depend on building up a pattern of behavior of objects within models and relating them to a place in the overall scheme of things (the current paradigm), more likely tied to fuzzy concepts than to measurable properties.

Even where an example or instance of an object class is described, such as the lithology of a core or well sample, a descriptive term (say, biohermal dolomite) may place it in a category that reflects an impression of many characteristics. We thus benefit from the ability of the human brain to recognize complex patterns specific to the context. This lends itself to narrative description, not to point-by-point quantitative comparison. From an initial broad appreciation of the situation, we observe and describe to extend our model. It is not difficult to think of examples where several conceptual process models are invoked. In the example at I 3, they dealt with sedimentary deposition, igneous activity, regional tilting and glacial erosion. The processes they refer to are worldwide. But we were looking at their consequences within a small area. Without conscious deliberation, we selected objects (the rock bodies) that took part, with the same definition, in each of the models. We naturally placed the objects and the processes at appropriate positions within the same framework of space and geological time. In M 2.3, we look at spatial and stratigraphic models which make that framework explicit for computer processing. Meantime, we note that objects seem to be chosen by a subtle process that depends on the intuition and background knowledge of the human mind.

At a general level, object classification and identification (H 5) are well suited to our thought processes, and may be assisted but not greatly changed by new technology. Quantitative methods, on the other hand, are well suited to computer processing. They are

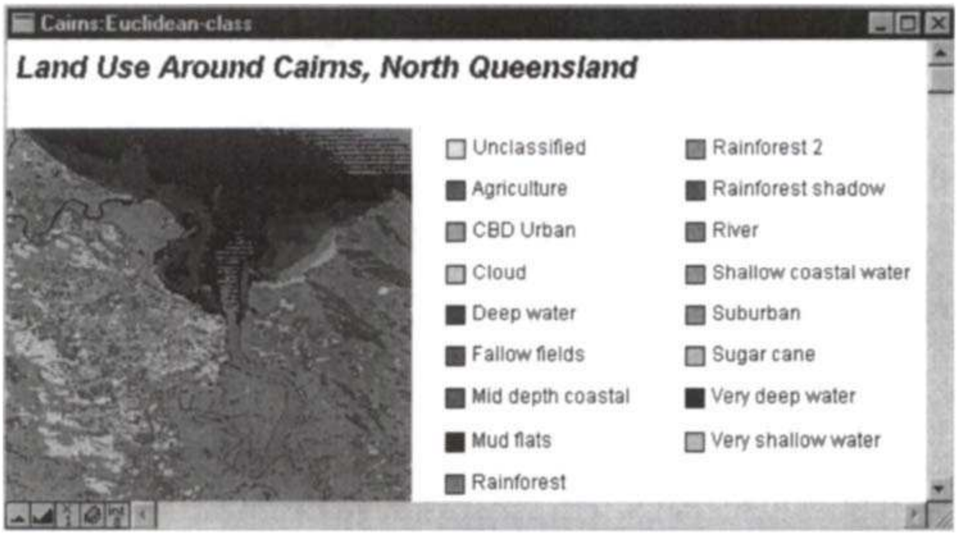


Fig. 2. Classification of land use from a satellite image. Example of satellite imagery classified by an iterative technique. The user indicates typical areas for each class, the computer extrapolates by quantitative analysis of the spectral bands and displays the color image, the user corrects and extends the classification, and so on. Published by permission of Rockware. More on http://www.rockware.com/catalog/pages/dimple.html

appropriate where there is a clear physical model that can be represented mathematically, as in seismic processing, gravity corrections, and so on. Their crucial contribution may lie, not in interpreting the geology, but in clarifying the geological significance of the records by removing extraneous effects (F). Where wider conclusions follow quantitative analysis of the raw data, as for example in seismic stratigraphy, they may result from non-quantitative reasoning. The new insights nevertheless depended on technology extending the reach of human thought.

Subtle variations in properties or composition (M 2.3) may be detected by quantitative analysis (F 5), as, for example, the identification of distinct lava flows from petrographic studies in the example of I 3. Statistical reasoning, based on randomly sampled measurements, can be a surprisingly powerful approach, even when based on the apparently weak concept of testing whether observed patterns were likely to have arisen by change.

IT offers the opportunity to build models that span the modes of human thought (J 1.7), combining their individual strengths. Quantitative reasoning, such as a computer process that simulates states and events, can be embedded in a narrative that explains its significance, limitations and context. Quantitative reasoning can be linked to cartographic and spatial thinking by computer visualization. Individual measurements gain meaning from the context of spatial pattern. Human insight, intuition and modes of thought remain supreme, but can be expressed in new ways. The wide range of models which IT supports can lead to better understanding, provided their properties and the limitations of their analogies are clearly appreciated.

2.4. The object-oriented approach

Our thought processes, constrained by technology, ultimately determine how we record and handle data. The object-oriented approach attempts to match those processes with IT procedures. According to Coad and Yourdon (1991), three methods of organization pervade our thinking about the real world:

- differentiation of experience into particular objects and their attributes;
- distinction between whole objects and their component parts;
- distinction between different classes of objects.

It is not difficult to think of examples from geoscience in terms of rock types, fossils, stratigraphic units, geological processes and so on. For instance, here is an outcrop (object), somewhat overgrown and deeply weathered (attributes), beside the river and under the bridge (spatial relationships). The outcrop (whole object) consists of beds of sandstone (component parts), containing grains of quartz and mica (components). It is interbedded (spatial relationship) with shales (different object class), and contains (spatial relationship) fossils (different object class).

Reality is a seamless web of infinite complexity, but the human mind can cope with only a limited amount of information at one time. Abstraction reduces the complexity by separating out a **model** dealing with a small number of things that are important to the purpose in hand. All words, language and data are abstractions and incomplete descriptions of the real world. There is, thus, no correct model of a situation, only adequate or inadequate ones. An **object model** describes the structure of objects in a system, their identity, relationships with other objects, attributes and operations. Common relationships are *being* (as in sandstone *is a* sedimentary rock), *having* (as in this sandstone *has* graded bedding) and *doing*. A **dynamic model** describes those aspects of a system concerned with time and the sequencing of operations — events that mark changes, states and organization, whereas a **functional model** captures what a system does, without regard for how or when it is done.

Language, images, quantitative modeling and demonstration all share the tendency to see the world in terms of objects, attributes and processes, from which may spring the noun, adjective and verb structure of our language (Leatherdale, 1974). Thus, communication in geoscience, by whatever means, concerns **processes** (which cause things to change) and **objects** (the things of interest), the object classes, and their **attributes** (properties, composition, relationships and behavior). An object should not be constrained by information type. One object, such as a borehole description, might comprise a text description and a geographic reference. It thus includes both text and spatial information types, which might be stored separately and accessed by different software. Many of the objects invoked in a narrative have second homes in other, possibly more structured environments. For example, a paper describing a fossil locality might include a list of species that could also appear in a paleontological register, and could be plotted on a map of fossil distribution, and linked to a stratigraphic table. The user must therefore interface with distributed objects, related to various topics, and represented by a mixture of information types.

Object classes, by definition, belong in a hierarchical sequence (H 5), **inheriting** attributes from classes farther up the hierarchy. Thus, *sandstone* may inherit properties from its superclass *sedimentary rocks*. A data model (I 2.2) can assemble object classes into topics, such as (examples in brackets): stratigraphic (formation), bibliographic (document), petrographic (thin section), paleontological (specimen of fossil). The topics are not mutually exclusive, so that a fossil description could be both a paleontological object and a bibliographic object. Within each topic, rules and standards can ensure that information is consistent and comparable. The fossil, as a paleontological object, is named according to the rules of fossil nomenclature, described according to paleontological conventions. The fossil description, as a bibliographic object, is cataloged according to international rules. A single object may thus be firmly embedded in at least two topic areas. We return later (L 6.1) to the application of object-oriented methods in analysis, design, programming and database work.

References

Baker, G.L., Gollub, J.P., 1996. Chaotic Dynamics: An Introduction. Cambridge University Press, Cambridge 256 pp.

Coad, P., Yourdon, E., 1991. Object-oriented Design. Yourdon Press, Englewood Cliffs, NJ 197 pp.

Encyclopedia Britannica, 1973. William Benton, Chicago.

Gilbert, G.K., 1896. The origin of hypotheses, illustrated by the discussion of a topographic problem. Science, N.S. 3, 1–13.

Goodchild, M.F., 1992. Geographical data modeling. Computers and Geosciences 18 (4), 401–408.

Kent, W., 1978. Data and Reality. North-Holland, Amsterdam 211 pp.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago 172 pp.

Laszlo, E., 1972. The Systems View of the World. Braziller, New York 131 pp.

Leatherdale, W.H., 1974. The Role of Analogy, Model and Metaphor in Science. North-Holland, Elsevier, Amsterdam 276 pp.

Popper, K.R., 1996. Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, London 431 pp.

Internet references

McCrone, J., 1999. Going inside — the neuronaut's guide to the science of consciousness. http://www.btinternet.com/~neuronaut/index.html.

MacEachren, A.M., 1998. Visualization — cartography for the 21st century. International Cartographic Association Commission on Visualization Conference, Warsaw, Poland, May. http://www.geog.psu.edu/ica/icavis/poland1.html.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A99–A108

COMPUTERS & GEOSCIENCES

Geoscience after IT Part K. Coping with changing ideas. Defining the user requirement for a future information system

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

The information system must deal with the diversity of ideas in geoscience and their changes through time. To communicate information, ideas must be aligned and molded to fit a shared view of the world. Change can be traumatic and may be deferred until obvious benefits force old ideas to give way to new, and even then individuals only partly reconcile their ideas. The mechanical records of IT must reflect the flexibility, overlap, ambiguity, inconsistency, conflict and evolution of human interpretation. These, and other needs considered in earlier parts of *Geoscience after IT*, are brought together as a statement of what we want from the system, set out as a user requirement. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Aligning ideas; Paradigms; Learning curve; User requirements

1. Change

The geoscience record is in constant flux. New ideas and new data are continually being added, and old ideas and data revised. Conventional methods struggle with limited success to maintain a record which is readily accessible and up to date. If we are to find better ways, we need to form a view on how change works. IT must cope, not just with the changes it creates, but also with the diversity of ideas in geoscience and their changes through time.

1.1. Flexibility and sharing knowledge

The information system must be flexible in order to respond to change. For example, words can retain their place in a growing science only because their meaning depends on the context. Think for instance of the word *fault*, and the ideas it might bring to the mind of field geologists using the concept to explain the outcropping of sediments of unexpected age, and possibly searching for landscape features to mark the fault as a line on the map. Their views of its characteristics and connotations differ from those of, say, the seismic interpreter, the seismologist locating an earthquake epicenter, a prospector looking for fault-related minerals, or the structural geologist studying the movement of continental plates. The same word used by a geologist a century ago would carry subtly different implications, embedded in the knowledge and thinking of the time. The computer engineer, for whom the word has a totally different meaning, could be forgiven for failing to see even a metaphorical connection. A

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00046-7

keyword search for documents about faults could be unhelpful. Nevertheless, the ambiguity associated with analogy (part J, section 2.2) gives room for growth and extension of ideas.

Information gains its meaning from its context. Geoscience information is gathered and made available to the information system from a variety of large and small projects (I 8.1). The projects are not devised within the information system, but are undertaken for reasons that stem from their **business setting**, which determines the objectives. The objective may simply be to satisfy curiosity. More likely the studies are directed to, for example: the search for oil and mineral wealth or help in its exploitation; collecting background information for protection of the environment; avoiding geological hazards or optimizing land use; or an attempt to understand more clearly the processes that formed the earth. The project objectives affect the sampling scheme, type of data, data collection method, and operational definitions. The data can be fully understood only through knowledge of the project and the approach used.

The development of a model of some sort precedes and is the subject of every investigation. Data from different projects may use the same terminology but different models, and thus be misleadingly similar but not fully compatible. This is one reason why the concept of a database as a pool of shared information (H 3) must be approached with care in geoscience. The important relationships among projects may be between models rather than between datasets. It is entirely possible that the models may be implicit rather than explicitly defined, which adds to the difficulties of data integration. Furthermore, a range of alternative models (multiple hypotheses) may be considered in parallel within a single investigation, as advocated by Chamberlin (1897). Yet from a multitude of independent projects there springs a coherent and integrated body of knowledge, as though coordinated by unseen hands. How does this happen, and how will it be affected by changing information technology?

The scientific process strongly encourages a shared view of the world. Indeed, a primary purpose of science is to relate a myriad of observations to a few scientific laws. Explanation is the means of integrating numerous concepts and results. Conformance with accepted procedures is encouraged by peer review, editors and referees, examination boards, textbooks, and standards organizations. Industry may encourage standards, for example to make more efficient use of information collected during hydrocarbon exploration. Government, with an interest in royalties and thus in the overall efficiency of the process, may reinforce this with legislation. On its own account, government may play a part by funding surveys of, say, topography, geology, soil science, hydrology, or oceanography. As long-term organizations, surveys tend to develop a uniform house style for investigation and presentation of their results.

A benefit of a standard approach is that it simplifies the exchange and integration of information. Object classes and their relationships, which can be defined formally in data analysis, provide a context into which new information can be fitted. Hypotheses are erected for further investigation and linked to the current hierarchy of scientific laws. **Standards** are created by assertion and negotiation; enforced or encouraged by custom, education, agreement, peer pressure and sometimes legislation. They all contribute to a shared frame of reference in which ideas are more readily exchanged, part of the map for scientific research (K 1.2). Establishing the relationships between models and harmonizing the underlying concepts is an important theme in the geoscience literature.

There is, however, a **trade-off**, that is, some benefits are gained at the expense of others. Collecting data to be widely useful imposes an additional cost on a project, possibly unnecessary for the immediate objectives. A standard approach limits flexibility, and can lead to an unduly narrow view. **Diversity** arises from divergent objectives, fragmentation of disciplines, rival or competitive organizations seeking a new niche, research into new possibilities, availability of better or cheaper non-standard methods, and attitudes such as preferring ownership to communication of information. Diversity is particularly associated with the early experimental phase of a new development. As ideas mature, and a general paradigm gains wide acceptance, the emphasis of the science and the attitude of the scientists change from innovative to methodical. Standards are valued more highly. Exploratory investigations, which are knowledge-based and proceed by trial and error, may be supplemented by systematic, pre-planned rules-based studies.

Diversity also arises from ideas changing with time. Philosophers remind us that we can expect all scientific information ultimately to be wrong. Information repositories, such as the scientific literature, contain much that we accept, if only because the scientific community has so far failed to disprove it. Other information we might regard as no longer entirely valid because it conflicts with more recent ideas or new data. However, there are many strands in a complex explanation and in the observations that support it. They involve ideas from many sources at varying levels of generality, put together in different ways to explain the same phenomena. The POSC Epicenter Model (POSC, 1997) relates observations to "**activities**", thus bringing distinct versions of data, possibly collected at different times with other instruments or objectives, into the same setting.

A study that we regard as based on unacceptable reasoning may contain information that has residual value for unforeseen use in a new context. For example, a borehole description with unbelievable stratigraphy might yield useful data on lithology. The use of analogy and metaphor introduces an element of ambiguity and flexibility to scientific reasoning (J 2.2). By permitting interpretation in several contexts, analogy offers the prospect of reworking old material and finding residual value in otherwise obsolete information. Its inevitable imprecision helps cross-fertilization where data or ideas are placed in a new context. It is not surprising, in these circumstances, that a large part of any project is devoted to the difficult tasks of finding, assessing and reinterpreting earlier studies, driven by the need to accommodate change.

1.2. Paradigms

We tend to see what we look for, and more strikingly fail to see what we do not look for. Minsky (1981), in his well-known work on machine intelligence and the human-computer interface, used the concept of **frames** to describe the intricate context in which ideas are embedded by the human mind. An idea communicated from one individual to another can be fully understood only if the recipient (man or machine) has an appropriate frame in place to receive it. In other words, the ability to grasp an idea depends on what you already know. Data dictionaries (H 3) reflect this concept by defining and placing in context the terms used to record data. Laszlo (1972) made a broader statement: "There is no theory without an underlying world view which directs the attention of the scientist. There is no experiment without a hypothesis and no science without some expectation as to the nature of its subject matter. The underlying hypotheses guide theory formulation and experimentation, and they are in turn specified by the experiments designed to test the theories."

Observations are set within a framework of current ideas. Thus the neptunists, believing all rocks to have been precipitated from a primitive ocean, could not have been expected to interpret correctly, or even to observe, the features which identify Salisbury Crags (I 3) as an igneous sill. During systematic examination of outcrops, however, unexpected features may be spotted which throw additional light on the nature of the rocks. Their significance may derive from analogies with observations elsewhere, or like Hutton's unconformity, with present-day processes. In many cases they would not be noticed except by a trained geologist aware of their possible significance, just as graded bedding, sedimentary structures or trace fossils must frequently have been visible to, but overlooked by, earlier generations of geologists.

Kuhn (1962), in his work on *The Structure of Scientific Revolutions*, distinguishes between "normal" science and revolutions in science. Normal science is based on a well-established view of a science in which the practitioners share the same exemplars or paradigms. Results are addressed only to professional colleagues, whose knowledge of a shared paradigm can be taken for granted, and who prove to be the only ones able to read the papers addressed to them. The **paradigm** comprises universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners. When the individual scientist can take a paradigm for granted, he need no longer in his major works attempt to build the field anew, starting from first principles and justifying the use of each concept introduced. That, as Kuhn dismissively remarks in his textbook, can be left to the writer of textbooks.

The need for experimental work, according to Kuhn, arises from the immense difficulties often encountered in developing points of contact between a theory and nature. Observation and experience can and must drastically restrict the range of admissible scientific belief, else there would be no science. But they cannot alone determine a body of such belief. "The paradigm provides a map whose details are elucidated by mature scientific research, and since nature is too complex and varied to be explored at random, that map is as essential as observations and experiment to science's continuing development" (Kuhn, 1962, p. 108). Three classes of problem — determination of significant fact, matching of facts with theory, and articulation of theory — constitute the literature of normal science. Research can be seen as a strenuous and devoted attempt to force nature into the conceptual boxes supplied by professional education. Once the reception of a common paradigm has freed the scientific community from the need constantly to re-examine its first principles, the members of that community can concentrate exclusively upon the subtlest and most esoteric of the phenomena that concern it.

One strong, but false, impression is likely to follow: that science has reached its present state by a series of individual discoveries and inventions that, when gathered together, constitute the modern body of technical knowledge, in a process often compared to the addition of bricks to a building. That, Kuhn claims, is *not* the way that science develops. Discovery commences with the awareness of anomaly — nature has somehow violated the paradigm-induced expectations that govern normal science.

Kuhn quoted an experiment by two psychologists, Bruner and Postina, who asked subjects to identify playing cards on the basis of a very brief glimpse. The experimenters introduced occasional cards of anomalous color, such as a black four of hearts. This was identified as the four of hearts or sometimes as the four of spades. Without awareness of trouble, it was immediately fitted to one of the conceptual categories prepared by prior experience. When the brief glimpses were extended to a somewhat longer exposure, however, the subjects suffered acute distress and some broke down in confusion. Similarly, the emergence of new theories is generally preceded by a period of pronounced professional insecurity, generated by persistent failure of the puzzles of normal science to come out as they should.

When the profession can no longer evade the anomalies that subvert the existing tradition of scientific practice — then begin the extraordinary investigations that lead the profession at last to a new set of commitments, a new basis for the practice of science. A scientific theory is declared invalid only if an alternative candidate is available to take its place. A new theory is always announced together with applications to some concrete range of natural phenomena; without them it would not be seen as a candidate for acceptance. It is seldom or never just an increment to what is already known. "... schools guided by different paradigms are always slightly at cross-purposes. At times of revolution, the scientist's perception of his environment must be re-educated — in some familiar situations he must learn to see a new gestalt. Thereafter, the world of his research will seem, here and there, incommensurable with the one he had inhabited before" (Kuhn, 1962, p. 111). This intrinsically revolutionary process is seldom completed by a single worker and never overnight. Kuhn quotes Max Planck: "a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it".

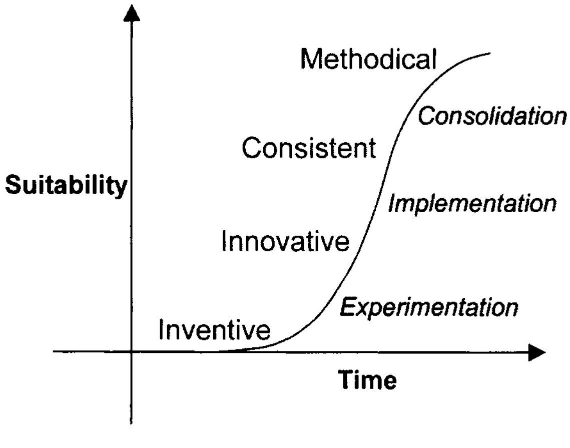


Fig. 1. Learning curve — the development of new technology. Techniques improve slowly at first, but initial success leads to investment and rapid growth, curbed eventually as innovation to a mature system shows diminishing returns.

Kuhn concerns himself largely with development of theory, but claims that the distinction between novelties of fact (discoveries) or novelties of theory (inventions) is artificial. He also points out that it may be the endurance of instrumental commitments that, as much as laws and theory, provide scientists with the rules of the game. Computer support, for example, makes possible a systems view of geoscience and some reformulation of traditional geoscientific reasoning on a more rigorous mathematical basis. Fundamental change to the information system can have far-reaching effects on the way science is conducted.

1.3. Dynamics of change

A plausible picture of the development of new technology is the so-called learning curve or S-curve like that of Fig. 1 (compare Coad and Yourdon, 1991). The vertical axis represents some measure of the appropriateness, success or value of a new development, say, the automobile, the telephone, or the computer. The first stage of invention and experimentation proceeds slowly until initial successes create interest, investment and consequent rapid development. Subsequently, as the technology matures, progress is slowed again by the law of diminishing returns. By then, the main framework of the system is fixed, and change is limited to slow, marginal improvement.

Consider, for example, the procedures of geological mapping which took shape in the early nineteenth century. An initial phase of discovery and invention was followed by experimental diversity and rapid progress as new methods of representing knowledge of spatial characteristics were developed (Rudwick, 1976). This in turn gave way to slow, systematic consolidation, building resistance to further change. Mapping methods were refined to the point where enhancements

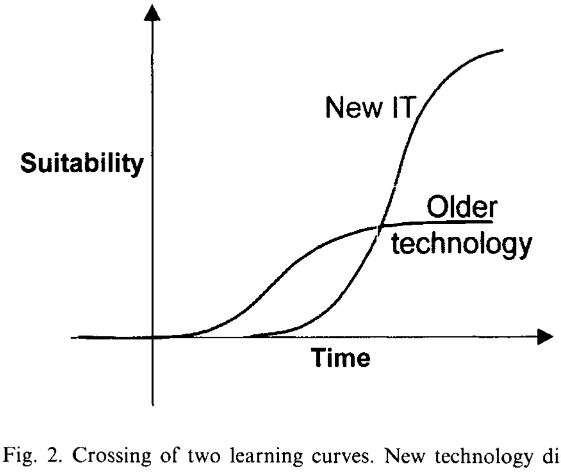


Fig. 2. Crossing of two learning curves. New technology displaces the old when clear benefits appear.

became marginal, and consistency was valued above innovation. Thereafter, newcomers to the craft were trained to follow conscientiously a set of well-established procedures. A blinkered view is a positive asset when knowledge-based work gives way to a rules-based approach. Fig. 1 suggests desirable qualities in the practitioners at different stages of a maturing technology.

Largely unheeded by the traditionalists, new supporting technology is set to sweep aside assumptions on which their hard-won skills were based. Technology tends to displace existing procedures rather than support an entirely new departure. There are at least two distinct S-curves: one showing the development of the older technology, the other the new (see Fig. 2). During its initial development, the newer technology is unlikely to be competitive with the old. Not until the new technology has reached the stage of rapid growth do many workers in the field see benefit in adopting a new approach. Those able to accept new ideas may move ahead, supported by the new technology. Earlier information may need to be reworked or lost. But skilled workers who were selected for their conscientious dedication to repetitive routine may be psychologically unwilling to adjust. There is discomfort and risk in moving from a mature technology to a fast-developing one.

The information system as a whole is based on technologies that are being systematically superseded by computer-based techniques. Measured by cost-effectiveness, the S-curves may already have crossed. But the curves are a gross over-simplification. There is no single paradigm shift. Change occurs at all levels of detail, and may have knock-on effects and implications for other levels. There are many strands of information technology applicable to various branches of geoscience. No single, smooth curve can be examined to see where we stand.

A better analogy might be a mountain which is surrounded by many foothills and shrouded in impenetrable fog. The obvious strategy in aiming for the summit is to go upwards. But when you apparently reach the top (because every direction leads downhill) you may merely be on one of the lower foothills. A research environment can be much like this, but with many workers throughout the world starting from different points and following different routes up the mountain. Despite the fog, research workers can get an idea of their relative success by shouting to one another, or more conventionally, communicating through conferences and the scientific literature. The individual who has reached a sub-optimal peak and hears shouts from above can take a bearing and proceed in the direction of the sound. Pushing the analogy a little further, the researcher who is comfortably atop a low foothill might consider carefully before discarding cherished ideas to make a long and dangerous traverse, aiming for a higher foothill which, on arrival, might turn out to have been abandoned in its turn. This imposes a degree of stability in the system. For most research workers, only major benefits justify a change of direction. In these circumstances it is invaluable to gain some idea of the lie of the land.

Inventions and new techniques, as well as new discoveries, can displace earlier commitments despite inevitable resistance to change. Blackmore (1999) coined the term **memes** to describe ideas, skills, habits, stories, songs or inventions that are passed from person to person by imitation. Taking a meme's eye view, she suggested how meme's evolved, meeting the prerequisites of evolution — variation, selection and heredity. If the process of science is seen as developing and testing models, then it is to be expected that, although the main body of geoscience knowledge is unlikely to be overturned in the foreseeable future, it will undergo continual amendment. Hypotheses will be disproved and new ideas emerge. The fittest, as selected by the scientific community, will survive.

Within their own areas of specialization, scientists may actively seek inconsistencies in the paradigm, attempt to disprove proposed explanations and align the model to their own concepts, thus encouraging diversity and evolution of ideas. Outside their specialism, they are more likely to accept a consensus view. The impact of replacing a model and the knock-on effects on the remainder of the knowledge base are determined by the model's scope and relationships. In geoscience, ideas and methods change at all levels of detail, sometimes with knock-on effects creating minor incremental shifts and partial inconsistencies that ripple gradually through the information system — paradigm drift rather than paradigm shift.

1.4. Reconciling ideas

All individuals presumably have their own unique view of reality, based on their personality, training and experience. By arranging and classifying the stream of sensory experience that impacts on short-term memory, they develop their episodic memory of events and their relationships, and the semantic memory defining their current world view (I 4). The ambiguity and inconsistency of human thought make it possible for one individual to hold incompletely defined opinions and a selection of alternative, possibly incompatible views. This has the advantage that a group of individuals can align their ideas with each other and reconcile their views when required for the purpose in hand. Their reconciliation probably does not extend far beyond the requirements of that purpose, and may not conform in every details. But it enables you to read and understand this without necessarily believing a word of it.

Communication requires a common **frame of reference**, that is a shared viewpoint or set of presuppositions, which can only be developed through education and training. The diversity of ideas and the difficulty of understanding subjects outside one's own chosen field suggest that alignment of ideas is incomplete, partial and specific. Examples can readily be imagined among the group of students examining the outcrop (I 3). **Knowledge** (justified true belief) can be seen as the product of a **social system**, that is, how people perceive each other and their shared activities.

"No two people have a perception of reality which is identical in every detail. In fact, a given person has different views at different times. ... But there is considerable overlap in all of these views. Views can be reconciled with different degrees of success to serve different purposes. By **reconciliation**, I mean a state in which the parties involved have negligible differences in that portion of their world views which is relevant to the purpose at hand. ... For the purposes of survival and the conduct of our daily lives (relatively narrow purposes), chances of reconciliation are necessarily high. ... But the changes of achieving such a shared view become poorer when we try to encompass broader purposes, and to involve more people. This is precisely why the question is becoming more relevant today: the thrust of technology is to foster interaction among greater numbers of people, and to integrate processes into monoliths serving wider and wider purposes. It is in this environment that discrepancies in fundamental assumptions will become increasingly evident" (Kent, 1978, pp. 202–203).

It seems that the information system must cope with overlapping information from different sources, and with many, possibly contradictory, versions of the same ideas. It must provide mechanisms for retaining ideas in their historical context, and for individual users to sift out their own reasonably consistent working views, without losing sight of the alternatives. It must allow the geoscience community to evaluate, select appropriate models, and build evolving views into their current paradigm.

2. Themes and problems

The potential benefits of IT determine future directions, and a number of themes emerged from earlier chapters. One theme was integration. In conventional systems, the scientist must cope with the separate interface and different content of a book, a map, a discussion group, a seminar, or a field study. Windows on a computer screen can offer a coherent view of the various information types, without undue delays and without librarians or booksellers. Material can be filtered for relevance to the specific user and displayed appropriately. With matching procedures, the user can edit the result, and add new information as text, images or data, again with few delays in making the information available and with a reduced need for human intermediaries. The screen on the desktop has many advantages. The map user can select the topics for display, pan to the areas of interest and zoom in or out to the level of detail required. The reader of text can follow references on the spot, search for keywords, and highlight passages for future reference.

If the benefits are clear, however, it is also clear why they are for the most part potential rather than immediate. The coherence of a well-planned document can be lost in a maze of hyperlinks. The accuracy of short-term memory cannot be brought into play when there are long delays in access over the Internet. Information well-printed on paper is more attractive, more convenient to handle, has sharper resolution and is easier to read than anything on a screen. For detailed study, therefore, a printed version is desirable. It can of course be prepared on a desktop printer, even copying the original page layout if desired. Full-size maps are more difficult, as specialized printers of large size and high resolution are required to produce a good copy. Having selected the appropriate area, scale and topics on the screen, either small extracts can be prepared on a standard printer, or a full-size copy can be prepared by an in-house print shop or by a cartographic bureau. However, if conventional products are to hand, they are likely to be more convenient and of better quality than their new-fangled equivalents.

There are also more crucial problems. In geoscience, digital information for remote access scarcely exists outside the petroleum industry and some large organizations. Much information on the World Wide Web is too ephemeral for bibliographical reference, and, for commercial reasons, digital maps are of limited availability. Globalization, in the sense of worldwide exchange of information regardless of discipline boundaries, promises efficiency gains by reducing redundancy in information holdings and offering rapid access to comprehensive information resources. It depends on widely accepted global standards, but comprehensive standards for geoscience are not in place. There is considerable inertia in the system.

Another theme was that of finding more flexible and rigorous expressions of the scientists' conceptual models. Quantitative, statistical and three-dimensional spatial models were mentioned as more complete and precise representations of scientists' ideas. Annotated photographs and video clips, keyed to the model, can help to connect the interpretation with observations. As methods of communication, they all fail if users lack the equipment or skills to receive the message.

The full benefits depend on the system as a whole being centered on IT, and therefore are also affected by the inertia just mentioned.

The theme of metadata was seen as important for a number of reasons. To understand information, you must know how it was collected, and users therefore need access to project metadata. There is also a tendency for different authors to attach slightly different meanings to terms used in papers and maps. It could be helpful to users to have metadata with precise definitions of the objects, and indications of where authors deviate from the standard definition. This can be achieved by hypertext links from the information *to* the metadata. Hypertext links *from* well-structured metadata can also help in searching for relevant material. Starting from the list of topics and relationships in the metadata, it should be possible to trace paths to treatments of these ideas in the literature. Furthermore, if documents indicate their dependence on earlier ideas and background theory through hypertext links, then they are positioned on a map of concepts, which could guide readers to relevant papers within their current understanding. As new ideas are introduced, or old ideas questioned, the links could show the knock-on effects and the ripples of change. Again, however, there is a snag. Links with HTML on the World Wide Web are one-way links to locations, not the necessary two-way and multiple links joining persistent objects.

Yet another theme was the evaluation of contributions to the knowledge base. In some areas, such as the oil industry or in some geological surveys, the quality of data is assessed through rigorous procedures of documentation, checking and evaluation. The user may have more confidence in information coming from a 'brand name' of this kind. The editing and refereeing procedures of scientific journals should serve the same function in a broader setting. The relationship of 'quality' to the intellectual foundations of the science is obscure, however. If there really is a set of ideas and studies generally seen as shared exemplars of how things are done in geoscience, there is no obvious mechanism for identifying that paradigm. Presumably individuals develop their own unique world views from rather fuzzy, overlapping and contradictory ideas. The paradigm appears to be a rather subtle concept, where the practitioners feel they know what is what and pass on the knowledge by nods, winks and tone of voice. A more explicit means of evaluating ideas would help them to evolve efficiently. It should involve the users as well as the providers of information. Again, it can only be part of the wider paradigm shift.

A final theme is the business context, the issue of why geoscientists behave as they do, and what forces drive them. In that sense, business decisions will drive change. Inertia comes from the huge investment in earlier systems, and the commitment of scientists and organizations to existing methods. Many have much to lose and little to gain from change. But on the high slopes, vast commercial enterprises are shifting their ground. Already, IT has brought about local changes throughout geoscience. Technical problems are being overcome. The potential to make money by reducing costs and improving efficiency may have the effect of gravity on a snowfield. When the avalanche finally starts to accelerate, it is only geoscientists who can determine whether or not the fallout benefits themselves and their science. That seems a good reason to list the potential benefits, and consider where we want to go before we arrive.

3. User requirements

A reference list of desirable IT features in the overall geoscience information system can focus ideas and even serve as an idealized check-list for new systems. It must, however, be used with caution. Some features are not economic at current prices and some may not be available at all. Features that require long-term availability may conflict with rapidly evolving technology (L 6, L 6.3). Other imply a change of attitudes which may take many years or may never happen. Most can be provided by other means. For example, user training or specialist support can reduce the need for user-friendly systems. Such a decision can have knock-on effects on other aims, however. For example, if you decide that computer specialists should run the programs, this may rule out interaction between the user and the program. It follows that a broad appreciation of overall developments and their interdependence is needed to make good decisions about even a small subsystem.

If you draw up a **wish list**, that is, a list of features you would like to see in your own system, ask yourself what is practicable and how it can be achieved. If you draw up a **user requirement** — the basis for a contract with IT specialists to supply specific facilities or services — compare the estimated costs of the system over its lifetime with those of alternative solutions. It is unwise to lead the field where no-one will follow. It is unwise to follow others into a dead end. With these provisos, here, for future reference, is an annotated summary of some desirable features in an IT-based information system.

Aide-memoire for a user requirement

The information system includes recorded information and the processes that assemble information and build knowledge. Most geoscience knowledge is held in the minds of scientists, who communicate through the ***user interface*** *with the rest of the system.* ***Repositories*** *store and manage recorded information for access by the orig-*

inators and others. ***Processes*** *manage, manipulate and present the information, helping the user to understand its significance and make decisions. The* ***business*** *context determines the objectives of geoscience investigations, and the deployment and management of resources to achieve them.*

3.1. User interface

Methods of accessing and supplying information should suit the users' ways of working and be easy to use, accepting and delivering appropriate information as, when and where required.

1. *User-friendliness.* A consistent, simple user interface that matches the scientist's way of thinking, including memory levels and modes of thought, should be used throughout. The system should support the specific needs of individual scientists; their joint needs within a workgroup; and communication with the world at large. The information system should be structured to reflect the processes of building knowledge from information.
2. *Coherence.* The interface should be compatible with: other systems in geoscience and other disciplines; the business context; any geoscientific instrumentation that passes data to the system. It should switch readily between browsing existing information; adding new information; and editing, analysis, manipulation and presentation.
3. *Control.* Originators of information, who best understand its significance and their procedures, should be able to determine the form and context of their contributions, and make them available without delay. Users, who best understand their own requirements, should be able to customize the interface to select information to meet their own specific needs and determine the form of presentation.
4. *Middleware.* To maintain a straightforward and familiar user interface, middleware (L 2) should hide the complexities of distributed systems and software such as GIS and DBMS. It should assist access to powerful tools such as SQL or graphical selection.
5. *Hypermedia.* Hypermedia links should allow different types of information (text, spatial, tacit, structured) to be closely associated at all levels of detail, both in collecting the information and presenting it to the user. Within a narrative account it should be possible to embed spatial information and quantitative evidence and reasoning, supported by visualization. It should also provide links to experts for advice, and pointers to archived cores, samples and specimens. An interwoven fabric of ideas should be supported through linkages between objects, processes and metadata.

3.2. Repository

The repository should provide safe, long-term custody of information with ready access to comprehensive, appropriate, current, coherent and testable records.

1. *Integration.* The repository may be partitioned according to information types, separating, say, documents, GIS and database for efficient management. But recognizing that only a shared framework can make communication possible, the partitions should share well-structured metadata, with models which link objects regardless of where they are stored or how they are represented.
2. *Connectivity.* The structure should support complex reasoning, including abstraction and generalization, through a network of links and cross-references among information in all its forms, including pointers to that held in the users' minds. From the palimpsest of overwritten and updated stories, it should be possible to extract material filtered by source and topic, and to drill down as required to detail, to supporting data, or to less popular, conflicting or older views. Dependencies between ideas should be recorded to ensure that change at any point can trigger knock-on effects.
3. *Redundancy and reusability.* The system should rely on connectivity rather than replication, for efficiency and to minimize confusion when changes are made. Later versions should be able to incorporate parts of the earlier by reference rather than repetition. Separation of metadata, objects and processes, should reduce redundancy and increase reusability.
4. *Granularity.* Microdocuments and markup languages, such as XML (L 6.2), should make it possible to handle narrative information in smaller discrete portions (finer granularity). Existing systems for handling spatial and structured data (GIS and DBMS) lend themselves to fine granularity, and can thus complement the detail of subdivided text.
5. *Flexibility.* It should be possible to identify rival paradigms, versions and views and to discover their different implications. From the same knowledge base, a range of software systems (interpreters) should support such activities as training; retrieving observations, interpretations or processes; developing ideas; and exploring analogies.
6. *Integrity.* The evolving structure must cope with past, present and future knowledge. Versions should be frozen on acceptance and retained as necessary for historical reasons, preferably with linkages to show their relationships with the metadata of the time. The system should be able to maintain valid current and historical references while coping with changing ideas, alternative versions and new information (including knock-on effects). Links

should not be left dangling when objects are deleted.

7. *Legacy information.* Legacy information should be accommodated in its original form, together with any updated version where value has been added, for example by digitization and markup. The user should be able to inspect current views or views at some previous time and explore the development of ideas.
8. *Disposal.* A clear disposal policy, which does not compromise the integrity of the system, should be defined and followed for ephemeral and obsolete material. Access should not be compromised by changing technology, nor safe custody by business priorities.
9. *Context and framework.* Project objectives and design features that assist interpretation and evaluation should be recorded. A document should be put in context by indicating the standards followed, together with a note of any divergence, or else carry a full data description. It should be possible to identify the background theory and previous work on which a document depends. These indicate the knowledge required to understand it, and thus its comprehensibility for a particular user. Information assembled by information communities and editorial boards (M 2) should provide coherent frameworks that strengthen the structure of the knowledge base as a whole.
10. *Metadata.* Standards, and the metadata in the computer repository, are analogous to human semantic memory. They create guidelines for organizing the information, and are essential for retrieval and coordination. Widely accepted standards should be followed, so that information can be more readily and more widely shared to greater effect. Metadata should reflect the ideas of the geoscience community as a whole, appropriately controlled through committees that consult widely (L 5, L 6.1).
11. *Evaluation.* The results of evaluation should be generally available, supported by techniques such as quality assessment and branding. It should be possible to record different evaluations, and thus reflect changing opinions. Although older ideas should be retained, it is the fittest ideas that should survive and be the most obvious and accessible to the user.

3.3. Processes

Comprehensive procedures should be available for acquisition, storage, processing, delivery and presentation of information.

1. *Search techniques.* The system should simplify the process of identifying and reaching all recorded information relevant to the individual's needs, by organizing material within a clear browsable structure, and by offering comprehensive search procedures with indexes, summaries, keywords, spatial search, structured query language, and hyperlinks.
2. *Interaction.* The advantages of the computer's precise rules-based activities and the user's fuzzy but extensive background knowledge should be combined in interactive processing.
3. *Analogies.* Explanation by analogy relies on the human mind, with its background knowledge and capacity for inference and intuition. A lengthy learning process is involved, imprecision and ambiguity being the price of flexibility of thought. The computer should help the scientist to detect and explore a wide range of analogous situations under interactive control.
4. *Reconciliation.* The system should respond to the opinions and views of individual users, recognizing that these overlap extensively but seldom coincide. It should support negotiation to align and reconcile (but not obliterate) alternative versions.
5. *Representation.* Computer systems should make it possible to express conceptual models more fully, tied more clearly to the evidence on which they are based, in a form shared by the geoscience community as a whole. They should provide effective representation of geoscience knowledge through computer-based models and processes, such as visualization and statistical and spatial models.
6. *Tacit knowledge.* Processes should be available to communicate tacit knowledge by showing the learner how to do things by example, demonstration and practice. For example, annotated photographs and video clips can help to show the procedures and locations of observation at an outcrop, allowing the reader to repeat the original procedures, and confirm the results or otherwise.
7. *Abstraction.* Information should be available at different levels of detail or abstraction. Where possible, the process of abstraction should be automated, but in many cases will require human judgment and intervention. Standard levels of detail should be available to simplify comparison and integration with other datasets, as is current practice with maps at standard scales.

3.4. Business aspects

The geoscience information system should be relevant, profitable, and efficient in meeting the business needs. The business context of a project is relevant to the scientific interpretation and should be recorded.

1. *Reduced costs.* The need for paper publications and

their management in numerous libraries should be reduced through client/server communication.

2. *Disintermediation.* Dependence on intermediaries and consequent delays should be reduced by computer support for word, data and image processing, and for search and retrieval. The complex tasks of managing a store of scientific information should be eased by computer support and indexing.
3. *Delayering.* A directed flow of business and scientific information should support better decisions, simpler management structures and more efficient working. By making relevant information available to all concerned, conclusions should be reached more rapidly, layers of management can be eliminated, and groups and individuals empowered to make decisions within the constraints of the system.
4. *Project management.* The information system should be brought closer to business requirements by linking it to project management and control. Project management and business aspects should be closely linked to the scientific system, as they bear on the planning and execution of each investigation. Descriptions of past, present and proposed projects should be available to help users to interpret the results and be aware of current developments. Within an organization, the information system strategy should be incorporated into the broader business plan.
5. *Standards.* The system should discourage unnecessary barriers to communication by recognizing the value added through compatibility and adherence to standards.
6. *Outsourcing.* It should be possible to delegate some activities, such as information management, to a specialist organization. Assessment for quality, including adherence to standards, should ensure an efficient service to many users.
7. *Intellectual property rights.* The reward system relies on intellectual property rights, which should be protected. Access should be controlled if necessary by entitlement indexes and encryption.
8. *Incentives.* Participants should be motivated to drive forward all aspects of the system in a coordinated manner, by appropriate incentives and giving credit where it is due. Charging systems should be implemented where appropriate.

References

Blackmore, S., 1999. In: The Meme Machine. Oxford University Press, New York 264 pp.

Chamberlin, T.C., 1897. The method of multiple working hypotheses. Journal of Geology 103, 349–354 Reprinted in 1995.

Coad, P., Yourdon, E., 1991. Object-oriented Design. Yourdon Press, Englewood Cliffs, NJ 197 pp.

Kent, W., 1978. Data and Reality. North-Holland, Amsterdam 211 pp.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago 172 pp.

Laszlo, E., 1972. The Systems View of the World. Braziller, New York 131 pp.

Minsky, M., 1981. A framework for representing knowledge. In: Haugeland, J. (Ed.), Mind Design. MIT Press, Cambridge 368 pp.

Rudwick, M.J.S., 1976. The emergence of a visual language for geological science 1760–1840. History of Science 14, 149–195.

Internet references

POSC, 1997. POSC Specifications — Epicentre 2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/Epicentre.2_2/SpecViewer.html.

PERGAMON

Computers & Geosciences 26 (2000) A109–A121

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part L. Adjusting the emerging information system to new technology

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

Coherent development depends on following widely used standards that respect our vast legacy of existing entries in the geoscience record. Middleware ensures that we see a coherent view from our desktops of diverse sources of information. Developments specific to managing the written word, map content, and structured data come together in shared metadata linking topics and information types. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Middleware; Digital object identifier; Interoperability; Ontology; Metadata

1. Staying in the mainstream

Having suggested some long-term user requirements (part K, section 3), we need to find a way forward which does not put earlier work at risk and leaves room to change course as future trends emerge, securing each step before taking the next. We look at some work in progress that takes a long-term view, although rapid development means that it is too early to predict which ideas will eventually prevail. Indeed, by the time you read this, some may have been superseded. Nevertheless, we can learn from them, and with citation indexes or other tools to trace forward references, they can still be a useful point to start looking for the best current solution.

To be cost-effective, the systems must follow widely used standards. The casual user simply cannot afford to learn techniques which are not of general application. An information system must be updated periodically, migrating along paths supported only by established IT suppliers. For both reasons, it is better not to stray from the mainstream of information technology development.

In the mainstream, we can detect the influence of three major tributaries, each from a separate source. They spring from the text-based information of publishers and librarians; the images and spatial models of geographers and cartographers; and the structured data of knowledge and databases. Different technical approaches characterize each tributary (L3–L5).

2. User interface and middleware

As chronicled in *Byte* (see, for instance, Orfali et al., 1995), it seems to be widely accepted that communication will continue developing within a client/server framework, as this makes it possible for each user to access a wide range of information sources maintained

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00047-9

by many providers. This applies within an organization where information is shared by cooperating groups through an intranet, as well as between organizations.

The graphical user interface is evolving into a network user interface (Halfhill, 1997). This has the potential to mediate among diverse repositories, access distributed objects and assemble information from many sources. It can incorporate earlier developments, such as SQL databases and groupware, as well as document management and geographic information systems. A layer of software, sometimes referred to as **middleware**, can be introduced to shield the user from the complexities of the underlying software. It enables a consistent user interface to control a range of diverse systems. Where a complex interface is needed because of the complexity of the operations, the middleware may be bypassed to tackle the problem on its own terms.

The widely adopted point-and-click user interface to the network seems appropriate for access to much geoscience information. A browser can link to narrative text, spatial data and interpretations, structured databases, computer models, references to material and links to experts. However, browser software based on HTML is inadequate for many purposes. For example, in order to integrate narrative, spatial and structured data, we might make use of separate interworking windows for the different information types. In this way, the user could view, say, a report, map and database side by side, or iconize a window when it is not required. The information in the different windows should share definitions of objects, so that when, say, an outcrop is described in the text, its location can be highlighted on the map. Descriptions of fossils found there could be illustrated by annotated photographs. The windows' contents should be synchronized, perhaps through a joint table of contents, so that when a new topic is introduced in the text, the map changes to match, and vice versa. This opens the prospect of handling compound documents with fully integrated information types (J 1.8, L 6). Web pages currently rely on HTML for most linkages. Because of the need to integrate information types and maintain two-way links, it is too limited for a full geoscience network. The more versatile XML — like HTML, a subset of SGML (E 6) — is an obvious future candidate for Web publication. It can provide a consistent user interface, mediating among the various retrieval systems.

3. Text-based information

Computer-mediated communication can cost much less than conventional publication (B 1). The calculations take no account of the costs of computer networks, application systems, and training, any more than teaching users to read is included in publication costs. Potentially, however, there are also important scientific advantages. We saw earlier (B 1) how publishers were attempting to extend the idea of a scientific journal, by providing hypermedia features. Other electronic journals such as *D-Lib* (D-Lib, 1995) offer more or less conventional content, but are published on the World Wide Web. Some, such as *Byte.com* (1994), provide extracts from printed journals, and most major publishers offer at least tables of contents on the Web (H 2). Some, such as PROLA (described next), attempt to provide a preprint, library and archive service. For obvious reasons, IT journals are in the forefront, but all scientific literature is in the line of IT fire (Butler, 1999).

Parts of the physics community, notably in high-energy physics, have made rapid progress in moving to electronic publication. Thomas (1998a,b) reviews the progress of the Physical Review On-line Archives Project (PROLA), and similar activities can be monitored at various Web sites.

There are three elements to the PROLA vision. The first is the preprint server, which provides rapid publication of results with open access and the opportunity for readers to record comments. This has now been in successful operation for some years. The second element is the peer-reviewed, edited journal. This is seen as essential for offering validated, certified statements of accepted progress. The authors need this as a measure of the value of their contributions, which may determine their career prospects. Readers need it to reassure them that the material is of value and widely accepted. The edited journal can be published electronically, probably with a companion paper copy for continuity and to meet the needs of libraries.

The third element is the electronic archive of past published papers, with facilities for browsing, searching and database retrieval. The electronic archive requires constant support ånd updating, partly to maintain links and references to and from older articles, but mostly to keep up with technical advance. Frequency of access to each document can be recorded as a useful guide to readers, and could be extended to take their evaluations into account. Logically, publication would consist of adding each new article to the archive, rather than placing it in a separate electronic journal. But back in 1999 that stage had not been reached.

So-called **legacy** information, collected in the past according to earlier standards, can be converted to an electronic form. Conventional printed publications can be scanned page by page, and stored, transmitted and displayed or printed as an image of the original. For many purposes, this will be adequate. Full text can be searched, edited and formatted, if need be, by optical character recognition (OCR) from the image, keyboarding from the original, or reusing the initial word processing if it is available (C 5). If required, the original layout can, at a cost, be preserved. Also at the cost of additional human effort, the original text can be marked up (D 6) for more detailed reference. Well-known projects include Project_Gutenberg (1999), which stores digital text of old documents and JSTOR (1995), which digitizes journals from the humanities. Their methods, contents and costs are described on the Web. Copyright is a significant constraint on these developments.

Existing publications must be preserved in their existing form, but in many cases could also be reworked and included in a more comprehensive information system. For example, by archiving current reports in SGML, it becomes easier to categorize small parts of a report separately, and thus to link them precisely to related documents and metadata. Present-day definitions and models for geoscience can only be created by specialists, and are likely to remain distinct from those of other disciplines. However, specialists from other subjects must be able to access and understand geoscience metadata and vice versa. Procedures for recording definitions and models should therefore conform to global standards. We noted however (K 1.1) that, for good reasons, meaning depends on context. The full subtleties of meaning of old records may never be translatable into modern usage, but must continue to rely on human interpretation.

Having obtained electronic documents, the next step is to consider how they can be organized within a repository. The technical design of a digital library is reviewed by Arms (1995), and set out in more detail by Kahn and Wilensky (1995) and Arms et al. (1997). Just as a conventional research library stores more than just books, so the digital library will store many types of digital material, including text, pictures, musical works, computer programs, databases, models and designs, video programs and compound works containing many types of information. Unlike a conventional library, the digital library can supply information which is not identical to that held in store. For example, a subset of data may be retrieved from a database, or a stored figure field may be supplied as a contour map or a perspective view. Because the library functions differently, some new terms are needed.

In the Kahn–Wilensky architecture, items in the digital library are called **digital objects**. They are stored in one or more **repositories** and identified by **handles**. Information stored in a digital object is called **content**, which is divided into **data** and information about the data, known as **properties** or **metadata**. The repositories must have unique names, and the digital object handles must also be unique. Their names must therefore be authorized by designated **naming authorities**. Depositing and accessing objects is accomplished using a defined **repository access protocol**. A **transaction record**, associated with the digital object, can record transactions, such as the time and date of deposit and of each request for retrieval, the identity of the requesting party, and any applicable terms and conditions, including amount and method of payment. A **mutable** digital object, unlike an **immutable** one, may be changed in certain ways after deposition, and may be designed to change with time.

The unique identifier or handle is itself a complex topic because, unlike the Uniform Resource Locator (URL) for accessing Web documents (E 4), it must persist for a very long period, probably much longer than the computer system or the organization that created it. It must be independent of the location at which the information is stored, compatible with earlier identification systems such as ISBN (H 2), and capable of evolving to meet long-term future needs. It should be able to identify fragments, composites, copies and versions of the information. These issues are discussed by Paskin (1997) and Green and Bide (1998). The Association of American Publishers has collaborated with the work described earlier to specify a **Digital Object Identifier** (International DOI Foundation, 1999) in an important initiative to track copyright ownership of electronic publications.

Web search engines help the user to locate relevant documents (Lynch, 1997), but tend to reflect words rather than their significance. The sad tale is told of a search for a project leader named Dr Cook (SHOE, 1999). A search for a combination of "Cook" and the project name yielded nothing. Searching for "Cook" alone provided over 200,000 documents covering everything from haute cuisine to a New Zealand Strait. Unlike libraries, the Web was not designed to support the organized publication and retrieval of information. A more structured search is possible using metadata to help users to locate relevant information, and to assess its reliability and suitability for their purposes. An annotated list of current Web documents on metadata is available (IFLA, 1995).

The Dublin Core (DCMI, 1998) is a leading candidate for recording metadata that helps users to find items on the Internet — the equivalent of the rules for a library's card index catalog. It is a cut-down equivalent of cataloging schemes currently used by librarians (Miller, 1996). It includes such information as subject, title, author, publisher, date, spatial and temporal coverage, and is intended to be simple enough for the author to supply the required metadata. Links can be included to documents which define the terms used. Rust (1998) mentions some limitations. It is one of several metadata packages, for example, for terms and conditions, archival management, administrative metadata, which will evolve to support the digital library as modules within the Resource Description Framework (Miller, 1998).

The G7 nations and the European Commission have organized a joint project to provide an information locator service with an emphasis on global environmental information (GILS, 1997). They extended the Government Information Locator Service, which is used in the US Federal Clearinghouses and State agencies, and renamed it the Global ILS (Christian, 1996). GILS, which is built on the Z39.50 standards mentioned in H 2, is designed to make it easier to find objects, in electronic or any other form, including documents, people and specimens.

The examples in this section suggest how geoscience can follow mainstream developments that stem from conventional document handling. Publishers and librarians are extending the concept of a document to include electronic content, thus altering ideas about what constitutes publication. During the transitional period, geoscientists may have to learn again how to find information and present their results, not once but many times.

4. Spatial information

Geoscience information is generally linked to geographic location, and catalogers regard this as an important aspect of the metadata and an aid to retrieval. The librarians' approach has been to catalog geographical areas by name or by enclosing rectangles specified by maximum and minimum coordinates. Some services, such as the Spatial Information Enquiry Service (SINES) run by the British Ordnance Survey, followed the same route. Although it adds value by bringing together many sources, the copies of metadata supplied by the information holders soon get out of data.

Geographic Information Systems (GIS) can handle the precise boundaries of spatial objects. Their three-dimensional form can be interpolated and stored (Gocad, 2000) and made available through standard interfaces such as VRML (Moore et al., 1999; Web3D Consortium, 1999; E 6). The main GIS vendors offer products that make it possible to visualize these objects as maps available to a Web browser. Information is available on their Web sites (Culpepper, 1998). It is therefore possible to give general overview of the geographical distributions of datasets on the World Wide Web, and for the user to select points or objects for retrieval of additional information. It can also be possible to provide more detailed information from a local GIS using the same user interface. Given adequate bandwidth and an appropriate system design that ensures that the user is not overwhelmed with needless detail, electronic delivery of maps (EDINA, 1999) and satellite imagery (Microsoft, 1998) is set to proliferate. The Web sites of the geography departments of well-known universities give references to other examples. The illustrations (Fig. 1) from the British Geological Survey geoscience index show how the user can zoom in on an area of interest, select an item and obtain additional data about it.

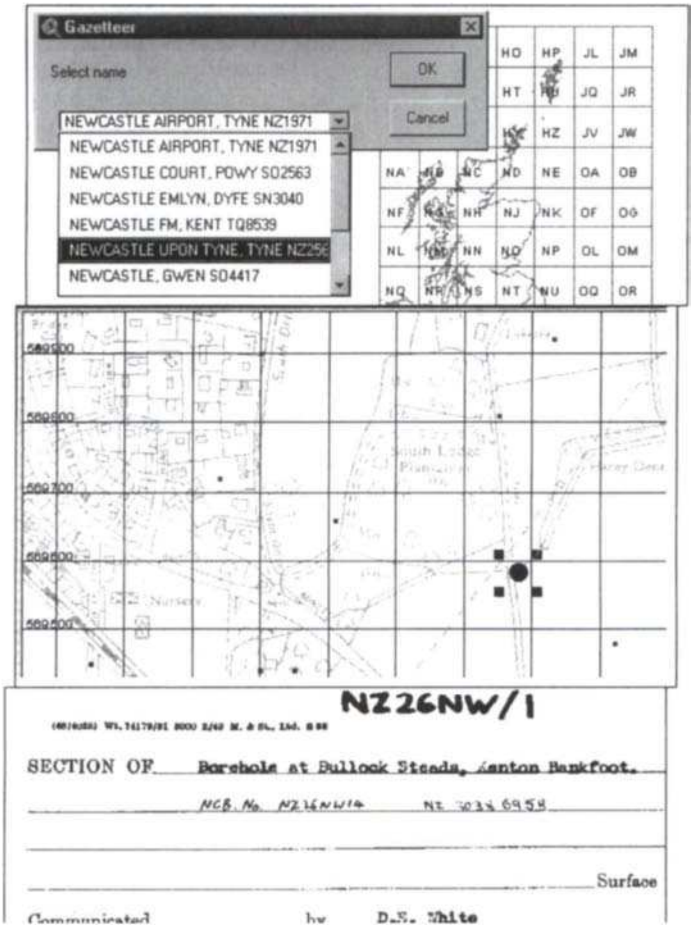
Users, however, may wish to assemble spatial information from many sources, not just for one proprietary system, and to manipulate that information with GIS facilities on their own client computers. As with library documents, problems arise in finding and assessing data because of inadequate metadata, and problems of obtaining and integrating datasets because of inadequate middleware and failure to conform to standards. Current standards are reviewed by Albrecht (1999) and Huber and Schneider (1999). Standards for representing geologic map information are being extended through a collaborative effort led and documented by the United States Geological Survey (1998).

The United States government is funding a National Spatial Data Infrastructure (Federal Geographic Data Committee, 1998) as part of their National Information Infrastructure. The creation of the National Geospatial Data Clearinghouse (1999) is part of this activity. Its aim is "to make data easier to find by supporting the evolution of common means to describe and share geospatial data sets". The data sets and metadata are held and maintained by those responsible for them, but accessible through the common standards. Other national counterparts, such as the UK National Geospatial Data Framework, propose a similar approach (NGDF, 1999).

The Open GIS Consortium (1996) is a consortium of the major GIS vendors and users which is working on the development of **middleware** (L 2), to isolate users from the details of lower layers of software. They aim to provide an Internet interface which is "not limited to the hyperlink and scrolling page mode of operation typical of Netscape, but supports the rich windowing graphics familiar to GIS users". They have prepared a detailed guide which includes a full account of the underlying concepts (Buehler and McKee, 1998). It sets out a framework for **interoperability**, defined as "a user's or a device's ability to access a variety of heterogeneous resources [data and programs] by means of a single, unchanging operational interface". The aim is that geospatial objects and the computer processes to manipulate them, obtained from many sources, should all work together, supplying results to any of a wide range of desktop clients. They have developed the Open Geodata Interoperability Specification (OGIS) — "a specification for object-oriented definitions of geodata that will enable development of true distributed geoprocessing across large networks as well as development of geodata interoperability solutions" (Schell et al., 1995).

US Military proposals point to a significant divergence from the librarians' approach (Larsen, 1998; GeoWorlds, 1998). One proposal is, for reasons of cost and efficiency, to replace their current huge volume of documents (maps, images and terrain models) with a "framework" spatial database with global coverage including the ocean floor. They intend that users should express their requirements in terms of area and topic, rather than named publications and other products. The response will provide data for the required area at the resolution and for the topics required. These could include imagery, terrain models and "features" traced from the original imagery, such as roads, rivers, and population centers. Although where possible the basic data is highly detailed (up to one-meter resolution from orthorectified photography), it would usually be supplied in a compressed form of appropriate resolution, generated from the scale-free basic data. The database could thus no longer be regarded as a library of discrete documents. An example of such an approach, coping with heavy usage of a large database, can be seen in TerraServer (Microsoft, 1998).

The flexibility of handling spatial data within a GIS means that it must bulk large in the future of geoscience. Internet links to Web browsers already provide worldwide access to GIS systems, which are becoming more robust and easier to use. There is some conflict between the discrete documents described in Section 3 and the potential to explore spatial data across project boundaries. There are corresponding problems in regarding a contribution to a GIS as a publication. In principle, however, a segment of a GIS could remain in that environment



NEWCASTLE AIRPORT, TYNE NZ1971

NEWCASTLE AIRPORT, TYNE NZ1971
NEWCASTLE COURT, POWY SO2563
NEWCASTLE EMLYN, DYFE SN3040
NEWCASTLE FM, KENT TQ8539
NEWCASTLE UPON TYNE, TYNE NZ25€
NEWCASTLE, GWEN SO4417

NZ26NW/1

(40/4058) Wt.74179/91 8000 2/62 M. & St. Ltd. 8 88

SECTION OF Borehole at Bullock Steads, Kenton Bankfoot.

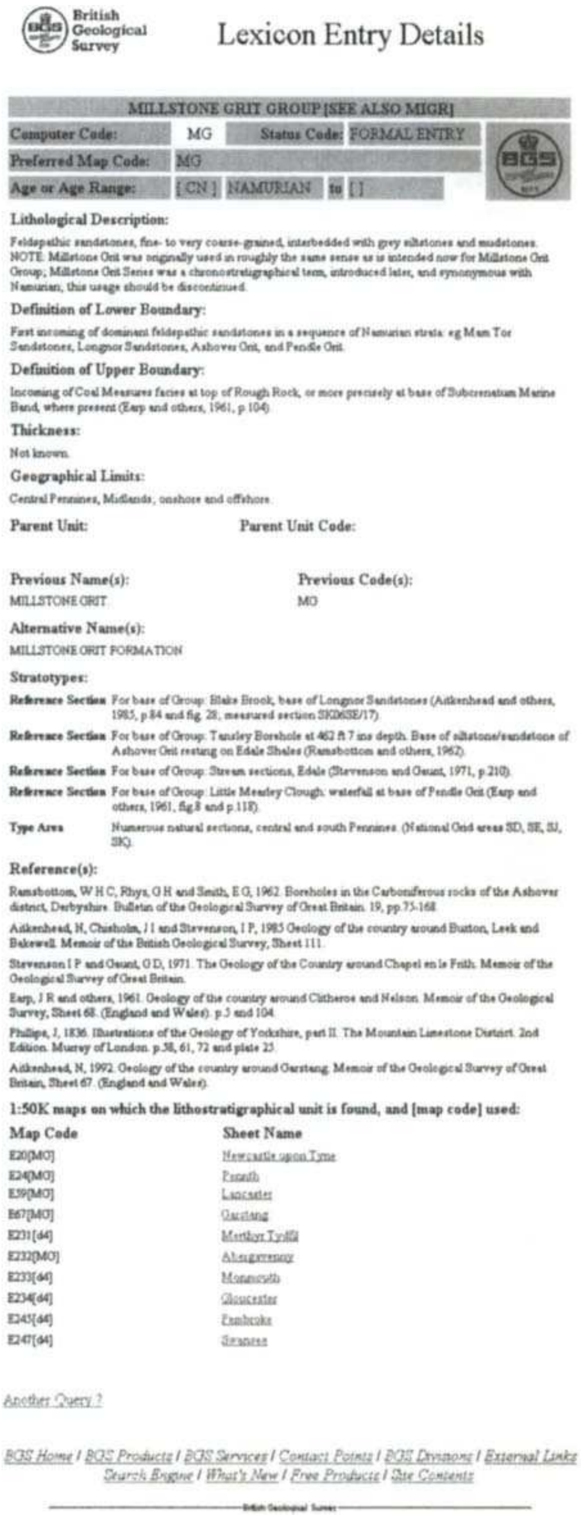
NCB. No. NZ26NW14 NZ 2034 6958

Surface

Communicated by D.S. White

Fig. 1. Finding data with a spatial geoscience index. The area of interest is selected from an index map or a gazetteer. Specific topics, here borehole locations, are selected for display on the detailed map. Information referring to an individual item, such as scanned images of a borehole log, can then be displayed in their spatial context. Extracts from the BGS Geoscience Data Index. British Geological Survey ©NERC. All rights reserved. Base maps reproduced by kind permission of the Ordnance Survey ©Crown Copyright NC/99/225.

while also being published as an integral part of a larger text-based document.



British Geological Survey

Lexicon Entry Details

MILLSTONE GRIT GROUP [SEE ALSO MIGR]

Computer Code:	MG	Status Code:	FORMAL ENTRY
Preferred Map Code:	MG		
Age or Age Range:	[CN] NAMURIAN	to	[ ]

Lithological Description:

Feldspathic sandstones, fine- to very coarse-grained, interbedded with grey siltstones and mudstones. NOTE: Millstone Grit was originally used in roughly the same sense as is intended now for Millstone Grit Group; Millstone Grit Series was a chronostratigraphical term, introduced later, and synonymous with Namurian; this usage should be discontinued.

Definition of Lower Boundary:

First incoming of dominant feldspathic sandstones in a sequence of Namurian strata: eg Mam Tor Sandstones, Longnor Sandstones, Ashover Grit, and Pendle Grit.

Definition of Upper Boundary:

Incoming of Coal Measures facies at top of Rough Rock, or more precisely at base of Subcrenatum Marine Band, where present (Earp and others, 1961, p 104).

Thickness:

Not known.

Geographical Limits:

Central Pennines, Midlands; onshore and offshore.

Parent Unit: **Parent Unit Code:**

Previous Name(s): MILLSTONE GRIT

Previous Code(s): MG

Alternative Name(s):

MILLSTONE GRIT FORMATION

Stratotypes:

Reference Section	For base of Group: Blake Brook, base of Longnor Sandstones (Aitkenhead and others, 1985, p 84 and fig. 28, measured section SK06SE/17).
Reference Section	For base of Group: Tansley Borehole at 462 ft 7 ins depth. Base of siltstone/sandstone of Ashover Grit resting on Edale Shales (Ramsbottom and others, 1962).
Reference Section	For base of Group: Stream sections, Edale (Stevenson and Gaunt, 1971, p 210).
Reference Section	For base of Group: Little Mearley Clough: waterfall at base of Pendle Grit (Earp and others, 1961, fig 8 and p 118).
Type Area	Numerous natural sections, central and south Pennines. (National Grid areas SD, SE, SJ, SK).

Reference(s):

Ramsbottom, W H C, Rhys, G H and Smith, E G, 1962. Boreholes in the Carboniferous rocks of the Ashover district, Derbyshire. Bulletin of the Geological Survey of Great Britain. 19, pp 75-168.

Aitkenhead, N, Chisholm, J I and Stevenson, I P, 1985 Geology of the country around Buxton, Leek and Bakewell. Memoir of the British Geological Survey, Sheet 111.

Stevenson I P and Gaunt, G D, 1971. The Geology of the Country around Chapel en le Frith. Memoir of the Geological Survey of Great Britain.

Earp, J R and others, 1961. Geology of the country around Clitheroe and Nelson. Memoir of the Geological Survey, Sheet 68. (England and Wales). p.3 and 104.

Phillips, J, 1836. Illustrations of the Geology of Yorkshire, part II. The Mountain Limestone District. 2nd Edition. Murray of London. p.38, 61, 72 and plate 25.

Aitkenhead, N, 1992. Geology of the country around Garstang. Memoir of the Geological Survey of Great Britain, Sheet 67. (England and Wales).

1:50K maps on which the lithostratigraphical unit is found, and [map code] used:

Map Code	Sheet Name
E20[MG]	Newcastle upon Tyne
E24[MG]	Penrith
E59[MG]	Lancaster
E67[MG]	Garstang
E231[d4]	Merthyr Tydfil
E232[MG]	Abergavenny
E233[d4]	Monmouth
E234[d4]	Gloucester
E245[d4]	Pembroke
E247[d4]	Swansea

Another Query ?

BGS Home / BGS Products / BGS Services / Contact Points / BGS Divisions / External Links
Search Engine / What's New / Free Products / Site Contents

British Geological Survey

Fig. 2. Metadata for a stratigraphic name. British Geological Survey ©NERC. All rights reserved. More on the BGS Stratigraphic Lexicon at http://www.bgs.ac.uk/scripts/lexicon

5. Structured data

Within a project, data (including quantitative and indexing information) are often collected as tables. This encourages consistency, with the same variables being measured or recorded in the same way at many points. Detailed metadata, with definitions and operational procedures, can help to ensure that the data are collected consistently (H 3). Each project, however, has its own business setting and background. Therefore, there may be subtle as well as major differences between projects, which make it difficult to compare their results. The metadata can help to translate between alternative terms and thus aid integration of data sets, although they do not provide the deeper understanding that can be gleaned from written accounts. Global projects, for instance, in seismology, geomagnetism and oceanography, rely on detailed standards so that many investigators worldwide can contribute to a shared database.

Workers in machine intelligence have carried this process further, with the aim of creating large knowledge bases, which not only contain information, but also the means of making logical deductions from it. As part of this an "**ontology**" is prepared, defined as "a specification of a conceptualization" (Gruber, 1997). A **conceptualization** is "an abstract, simplified view of the world that we wish to represent for some purpose". The ontology defines the objects, concepts and other entities, and the relationships between them. It is analogous to the data dictionaries and data models (H 3) that define the terms in a database and their relationships. In geology, for example, one might expect to find a definition of, say, Millstone Grit, in terms that the Stratigraphic Lexicon might use, some means of defining its hierarchical and positional relationships within the stratigraphic column, and an indication of the scope, validity and provenance of the term (Fig. 2).

Ontologies are an experimental means of labeling Web documents, using Simple HTML Ontology Extensions (SHOE) (SHOE, 1999), in order to make searches by web robots and intelligent agents more effective. Ontologies also appear in ambitious schemes, such as Ontolingua, for knowledge sharing and reuse (Stanford KSL Network Services, 1996). A large working implementation of such an approach, involving a metathesaurus giving information about specific concepts and a semantic network defining relationships, is described at the US National Library of Medicine web site (National Library of Medicine, 1998).

A less rigorous scheme for assembling definitions of concepts is the virtual hyperglossary advocated by Murray-Rust and West (1998). Glossaries can be submitted and revised on any subject from any source, subject to editorial scrutiny. It is accepted that vocabularies overlap, and words do not necessarily carry the same meaning, in different subjects. The words are arranged in alphabetical lists: click on the word for its definition, relationships and other relevant information and references. Its bias is towards organic chemistry, and there are many molecular diagrams of nodes and links: point to the node to see the name of the component, click to see its definition. There is clearly an analogy with entity-relationship diagrams.

The most coherent and extensive data model to include aspects of geology and geophysics is the Epicentre Model (see Fig. 5) of the Petrotechnical Open Software Corporation (POSC, 1993), much of which is now available on the Web (POSC, 1999). POSC is a consortium where major oil companies are represented, together with some IT companies, surveys and other organizations. An objective is to save many tens of millions of dollars every year by sharing information repositories, and accessing data more efficiently. This requires standards for interoperability in oil exploration and production data. The Epicentre Model has a number of sub-models for such topics as: spatial models, geographical referencing, cartography; stratigraphy (litho-, chrono-, bio- and seismo-); materials and substances, rocks, minerals and fluids; stratigraphical and seismic interpretations; geophysics (seismic, gravity, magnetic, electrical); wells, downhole logs, samples and cores; remote sensing; organizations, documents, personnel and activities; equipment, procedures and inventories; reservoir characteristics; computer facilities, software, users and data administration. Data dictionaries and entity-relationship diagrams are used in all of them to provide a definition of the common currency in which geologists express their ideas. The information is also supplied on CD-ROM for those with uncomfortably slow Internet links. The model is compatible with more general international standards, and can thus support searching and integration of data within and beyond geoscience.

As with data in a GIS, quantitative measurements may be held within a rigorously structured database. The database may contain contributions from many sources that meet the standards defined in the metadata. They may be referenced from a text document, thus being fully reviewed and seen as part of a publication. Computer programs can follow similar procedures. For example, the International Association for Mathematical Geology makes the programs and data described in their publications freely available for downloading to the user's computer (IAMG, 1995). We catch a first glimpse here of geoscience documents, published complete with links to their electronic appendages, placed in their business, spatial, and quantitative context through shared standards described in metadata.

6. Integration

Future information technology should have no boundaries, and therefore few features specific to geoscience, whose needs should be identified and met within the mainstream. Levels of human memory, such as semantic, episodic and short-term, have their counterparts in the information system.

6.1. Sharing metadata

At a semantic level, we have seen (L 3–L 5) how metadata developed. From the library background came the concepts of the digital library architecture and of a classification of knowledge, for cataloging documents and searching by concept or keyword. From geographic information systems came the spatial model for describing the location of objects in space, their spatial pattern and relationships, and the active map for spatial search. From database management came data dictionaries, data models, structures to reduce redundancy, and query languages for retrieval by categories and quantitative values. From knowledge base work came the ontology to "specify a conceptualization". As each group generalizes their work into a wider IT context, the cataloging systems, data models, spatial models and ontologies begin to overlap and amalgamate. Examples, notably from POSC (1999), show how a shared framework can operate and how users can benefit from large-scale, collaborative projects.

Metadata are concerned with standards; classification and nomenclature; patterns of investigation; and data models and definitions of object classes. Object classes (H 5) form a hierarchy, classes at lower levels inheriting properties from those at a higher level. A Millstone Grit object, for instance, would inherit appropriate properties that applied to the Carboniferous as a whole (see Fig. 2). Hierarchies of terms are familiar in geological classifications, for example, in paleontology, petrography, lithostratigraphy, chronostratigraphy, and in spatial subdivisions. Each of these can be regarded as a topic, and a data model (H 3) can depict the relationships of classes within that topic (see Fig. 5). At a higher hierarchical level, another data model might show relationships between topics. Internationally accepted definitions of objects and processes, their relationships, and the hierarchy of object classes, are all vital to a widely shared understanding of the geoscience record.

The definitions and characteristics of geoscience object classes are (or should be) the same regardless of information type or mode of representation. A formation, a fossil, or a logging tool, should be the same whether it is illustrated in a diagram, drawn on a map, listed in a register or described in a report. Metadata should be kept distinct from documents recording scientific findings. This allows more appropriate management and more flexible communication and reuse.

6.2. Linking topics

A striking feature of the POSC Epicentre Data Model is its separation into self-contained topics. Each data model represents one topic within the information base, and should therefore provide users with access routes to information which reflects their specific interests. For example, a spatial model might be appropriate where information was required about a particular point or area. A data model for paleontology would be appropriate where a particular species is of interest. The two models should be usable together where fossils of that species in a particular area are required. The business model (where business is used in the broad sense to identify the objectives and procedures for a study) might also narrow the search by guiding users to studies with similar objectives to their own.

Within a project, links between topics tend to involve interpretation, often by comparing visualizations of spatial models, each arising from a different topic, and relying on human perception, intuition and background knowledge. For example, data from a seismic survey might be assembled and processed to provide a contour map of a seismic horizon. Downhole logs might provide a similar map of a nearby formation top, and the two maps might be compared by eye. Individual seismic values, however, are not compared with individual well picks (G 2).

The spatial patterns and relationships of the two topics are of interest, although deciphering each pattern is a task performed largely within the topic area. Nevertheless, the life of the geoscientist is made much easier by an interface which is similar in all topic areas and enables results from different topics to be assembled and compared as compatible spatial models (G 2). Spatial models which describe geometric forms in terms of points, lines, areas and volumes can be positioned relative to the Earth. The geometric objects can then be linked to geological or other features, so that, for example, a line represents a borehole, and surfaces represent the formation tops that it intersects.

An object describing a formation could be linked (with reference to a stratigraphic model) to formations above and below, and to broader, narrower and related stratigraphic units. It could be linked (with reference to a spatial model) to adjacent, smaller and larger areas. This would make it possible to move from summary to detail or vice versa, on the basis of level of spatial resolution, stratigraphic discrimination or both. At the cost of a more structured and therefore less flexible framework, repetition, redundancy and conflict within the information can be reduced.

The tools for doing this are preliminary analysis to match the information to a coherent structure, and markup languages to implement that structure. The Extensible Markup Language (**XML**) makes it possible to categorize information, such as sections of a report, by tying them to metadata, thus superimposing ontological classifications on the sections of text (Bosak, 1997). XML also provides a means of building objects into more than one hierarchy, thus making the traditional concept of a self-contained document unnecessary. Instead, reports explaining maps, for example, could avoid internal boundaries, like the seamless map (L 4), with documents created as required for specific areas, topics and resolutions. The Meta Content Framework (**MCF**), which uses XML, explores such a framework, aiming to structure Web hypermedia to make it "more like a library and less like a messy heap of books on the floor" (Bray and Guha, 1998).

6.3. Linking information types

Obvious in the user interface, but extending to processes and repositories, is another distinction — by **information types**. Text documents dominate the literature. Maps and stratigraphic tables in large format are published separately and independently. Data that support the written or mapped interpretation may be archived, frequently as a computer file, and made available on request, rather than appearing in full in the scientific literature.

Fig. 1 of part I is redrawn as Fig. 3 to show these components of the information system. The user interface is divided by information type into three windows. It represents one of a large number of documents collected for different purposes, each held separately in the repository. We can visualize them lying behind the representative. In the higher levels of the repository area in the diagram are the metadata and the more generalized information arising from abstraction and explanation of the datasets. Beneath the repository are shown the tools for processing the information, possibly learned techniques or computer programs.

The components of the system are seldom totally distinct. Data cannot be entirely separated from explanation, and abstraction is an essential part of observation (B 4.2). Overlap is even more obvious in other cases, such as between information types. Maps may be published separately, but are likely to include text comments and possibly tables of data. Conversely, maps are included as diagrams in books and reports. Processes and data are frequently inextricably joined. The picture of the information system is therefore misleading if taken too literally. It is an idealization that has significant features in common with reality. It is a metaphor or model (J 2.2) which may yield useful insights. The diagram is obviously not part of a rigorous analysis, but can be regarded simply as an aid to remembering the chosen components and their relationships.

It should be possible to search across information types. For example, it should be feasible to: define an area on an electronic map; find the formations within it; retrieve text descriptions of the formations; locate boreholes intersecting them; retrieve their logs from an image repository, and formation thicknesses and contouring software from a database (Fig. 4).

At the semantic level, metadata can define object classes and describe their relationships. At the episodic level (I 4), occurrences (instances) of objects are linked together, along with processes, for a different purpose — to tell a story (J 1.2). They are linked within a document, where '**document**' is defined broadly to include any combination of multimedia in which a collection of objects and processes are tied together for some purpose, probably referring to a single project (D 6). A sequence of events linking the objects may be recorded in narrative text. The quantitative values of their properties or composition may be tabulated as datasets, analyzed statistically (F), visualized graphically (Cleveland, 1993) and thus made available to accurate short-term memory. Their location, form and spatial relationships in geological space-time may be shown as three-dimensional images and maps, regarded as just another form of visualization (MacEachren, 1998; Kraak, 1999; Sheppard, 1999). Other forms of multimedia, such as video, may identify and illustrate other characteristics. The compound document may include any or all of these, possibly following different modes of thought (J 1.7), in synchronized windows that can be viewed side by side.

Several software systems may be needed to manage and manipulate the components of a compound document. For example, a document describing a geophysi-

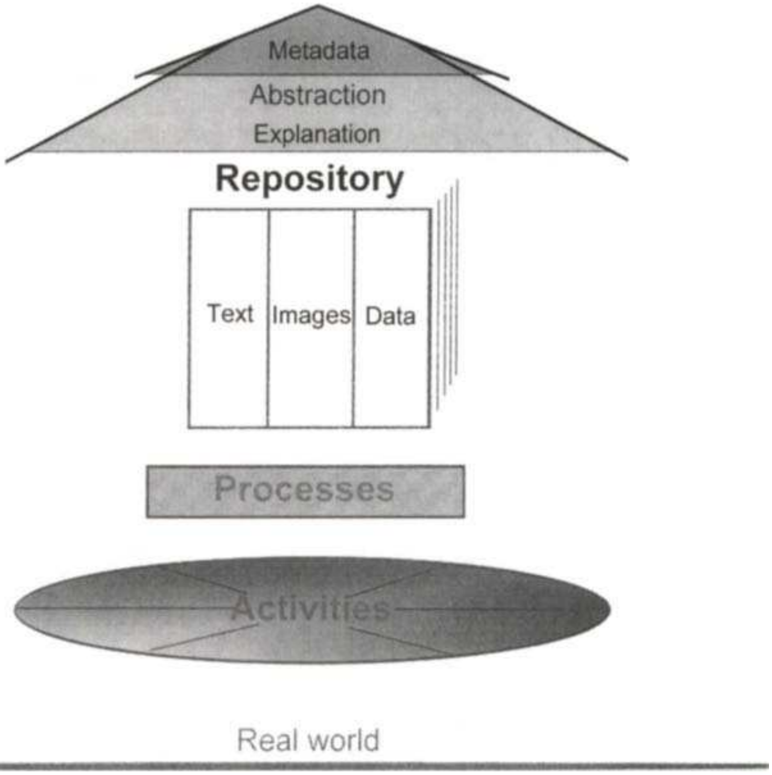
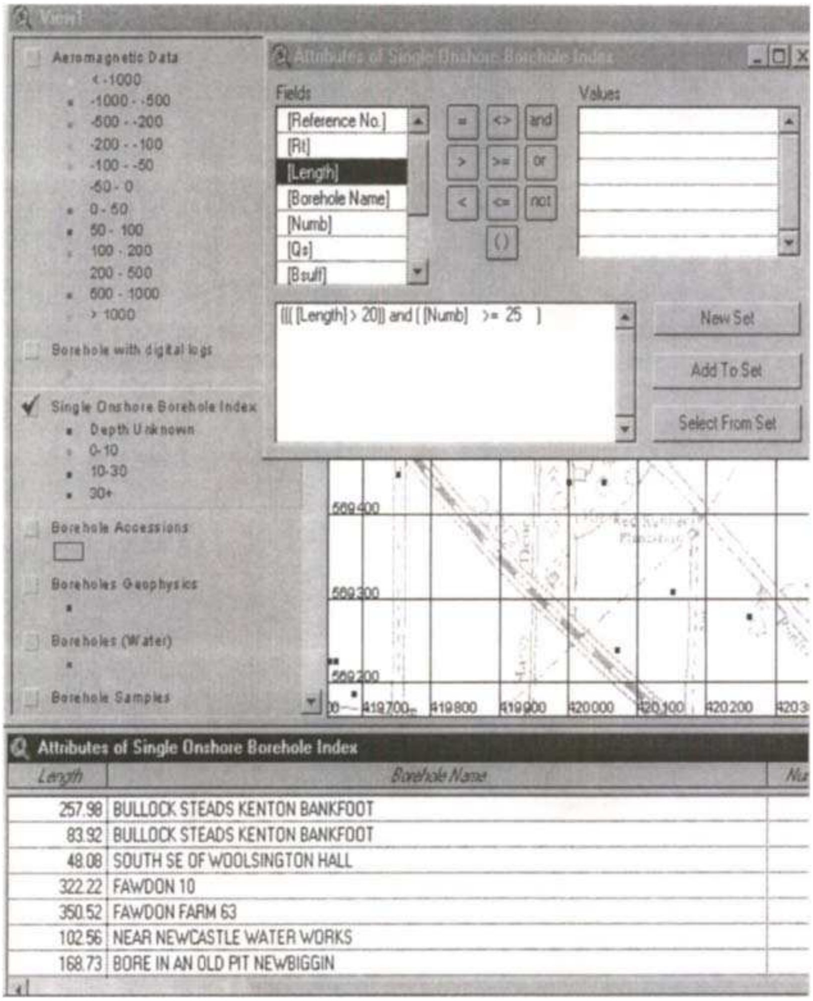


Fig. 3. Some components of the information system. Documents containing various information types are stored in the repository, together with generalized summaries, and metadata which describe the document and define shared vocabulary and standards. Processes to analyze and manipulate the information are shown separately, as are the scientists' activities (see Fig. 1 in M) which generate and evaluate the documents by investigation of the real world.

cal survey might include text held in a document management system, spatial models held in a GIS, and data held in a relational database. Examples of software tools that might be required include: project management software, entitlements register software, a document management system, RDBMS and ODBMS, GIS, application programs (maybe Java-mediated), hypermedia systems. The information types could be managed separately but linked as a single, higher-level object. This could be seen as a tradable object, available to others as a self-contained item, containing appropriate application programs and information about charges and availability.

The future scenario that emerges is of the geoscientist working within a well-defined standardized framework of concepts, terms and definitions. Documents, perhaps written in a specialized dialect of a markup language (J 1.8), weave together records of observations and interpretations in the context of one or more data models. Narrative text, spatial data and interpretations, structured data, computer models, references to material and links to experts are handled together and the results communicated to any desktop. Hypermedia provide the flexibility for integrating different information types and different modes of thought. The ability to follow threads of reasoning through all information types in the document should be matched by the ability to clarify their significance by instant access to appropriate metadata. Citations from the metadata should provide the opportunity to follow up other references to similar objects, or to explore relationships within



(([Length] > 20)) and ([Numb] >= 25)

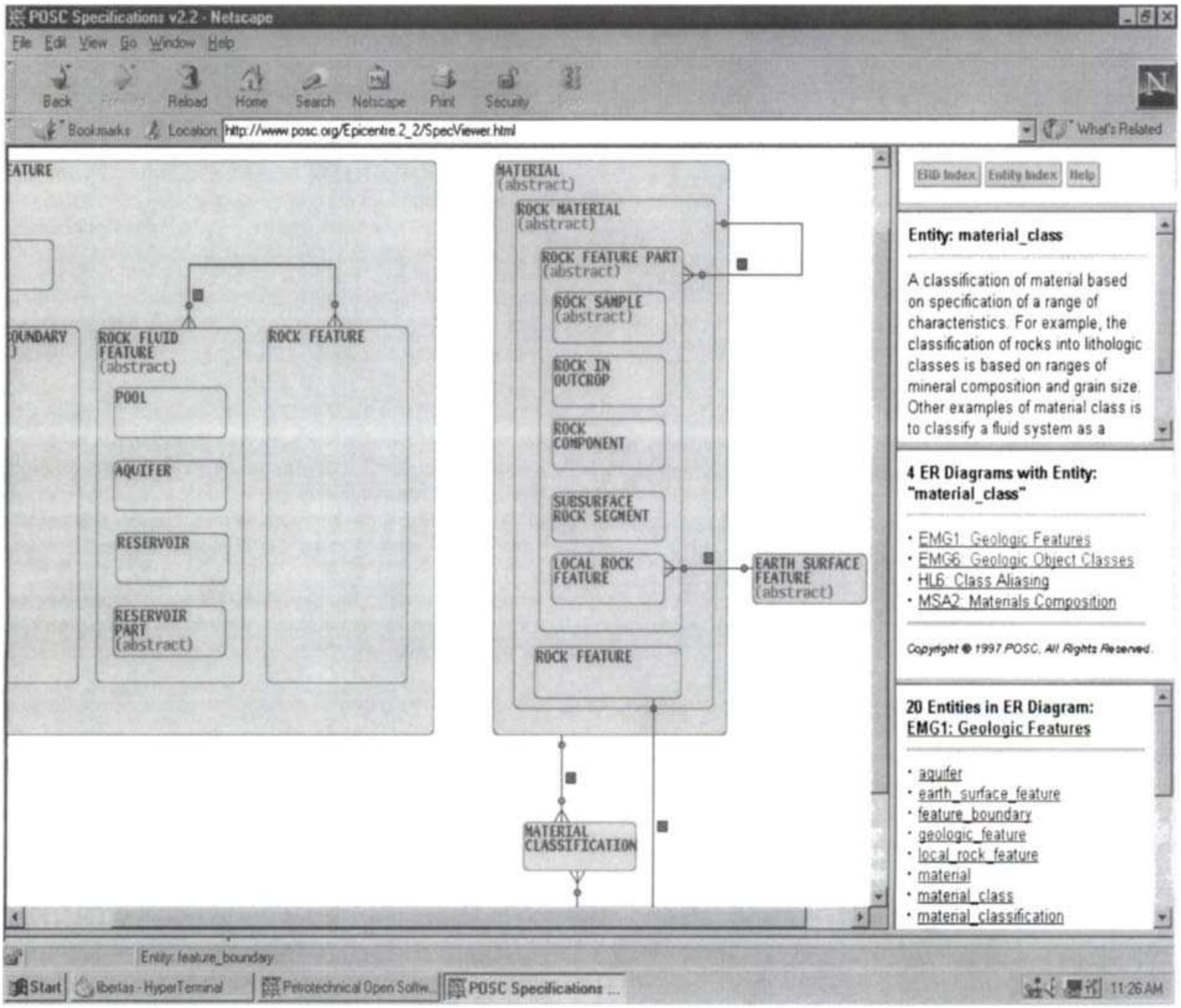
Attributes of Single Onshore Borehole Index

Length	Borehole Name	Nu
257.98	BULLOCK STEADS KENTON BANKFOOT	
83.92	BULLOCK STEADS KENTON BANKFOOT	
48.08	SOUTH SE OF WOOLSINGTON HALL	
322.22	FAWDON 10	
350.52	FAWDON FARM 63	
102.56	NEAR NEWCASTLE WATER WORKS	
168.73	BORE IN AN OLD PIT NEWBIGGIN	

Fig. 4. Retrieving data with GIS and DBMS. Some GIS systems, such as ArcView used here with the BGS Geoscience Data Index, make it possible to combine topic selection, spatial selection and SQL queries, displaying the results on the map. British Geological Survey ©NERC. All rights reserved. Base maps reproduced by kind permission of the Ordnance Survey ©Crown Copyright NC/99/225.

the metadata to identify related object classes (see Fig. 5). The rapid delivery of information through IT allows the use of accurate short-term human memory to control computer procedures by interaction, based on the user's fuzzy but extensive background knowledge. Use by non-specialists could be aided by access to metadata and software agents, possibly reducing the need to rewrite the same material for different audiences.

Unfortunately, maintenance costs for compound documents are high, because technology is on the upward leg of an S-curve (see Fig. 1 of part K). The rapid evolution of technology means that records must be continually modified to match new standards and software. Librarians are accustomed to books and journals, printed with stable technology, which retain their original, usable form for many decades with negligible maintenance costs. Techniques for handling electronic text are well established, but few publishers or librarians have experience of managing documents which also require support from GIS, DBMS and other software systems. Until IT reaches a more stable state, this must slow the acceptance of compound documents and make it inadvisable to rely on their


POSC Specifications v2.2 - Netscape
File Edit View Go Window Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.posc.org/Epicentre.2_2/SpecViewer.html
What's Related
ERD Index Entity Index Help
EATURE
OUNDARY
ROCK FLUID FEATURE (abstract)
POOL
AQUIFER
RESERVOIR
RESERVOIR PART (abstract)
ROCK FEATURE
MATERIAL (abstract)
ROCK MATERIAL (abstract)
ROCK FEATURE PART (abstract)
ROCK SAMPLE (abstract)
ROCK IN OUTCROP
ROCK COMPONENT
SUBSURFACE ROCK SEGMENT
LOCAL ROCK FEATURE
EARTH SURFACE FEATURE (abstract)
ROCK FEATURE
MATERIAL CLASSIFICATION
Entity: feature_boundary
Start
libertas - HyperTerminal
Petrotechnical Open Softw...
POSC Specifications ...
11:26 AM

Entity: material_class

A classification of material based on specification of a range of characteristics. For example, the classification of rocks into lithologic classes is based on ranges of mineral composition and grain size. Other examples of material class is to classify a fluid system as a

4 ER Diagrams with Entity: "material_class"

- EMG1: Geologic Features
- EMG6: Geologic Object Classes
- HL6: Class Aliasing
- MSA2: Materials Composition

Copyright © 1997 POSC. All Rights Reserved.

20 Entities in ER Diagram: EMG1: Geologic Features

- aquifer
- earth_surface_feature
- feature_boundary
- geologic_feature
- local_rock_feature
- material
- material_class
- material_classification

Fig. 5. Diagram from the POSC Epicentre model. Various entities, or object classes, are grouped into topic diagrams. This is part of one diagram (EMG1: Geologic Features) illustrating the Epicentre 2.2 Data Model. When you move the mouse over entity boxes or relationships, adjacent frames offer definitions, examples, and cross-references to other occurrences in the overall model and to other entities within the topic. You can move freely between the diagram and text accounts of the entities and their components, or to more general or more detailed documentation. Reproduced by permission of the Petrotechnical Open Software Corporation. More at http://www.posc.org

retention in archives. Their initial growth may be within a different framework (M 2).

References

Albrecht, J., 1999. Geospatial information standards. A comparative study of approaches in the standardisation of geospatial information. Computers and Geosciences 25, 9–24.

Butler, D., 1999. The writing is on the web for science journals in print. Nature 397 (6716), 195–200.

Cleveland, W.S., 1993. In: Visualizing Data. Hobart Press, Summit, New Jersey 360 pp.

Huber, M., Schneider, D., 1999. Spatial data standards in view of models of space and the functions operating on them. Computers and Geosciences 25, 25–38.

Kraak, M.-J., 1999. Visualization for exploration of spatial data. International Journal of Geographical Information Science 13 (4), 285–288.

Moore, K., Dykes, J., Wood, J., 1999. Using Java to interact with geo-referenced VRML within a virtual field course. Computers and Geosciences 25 (10), 1125–1136.

POSC, 1993. Petrotechnical Open Software Corporation, Software Integration Platform Specification. Epicentre Data Model, version 1, vol. 1: Tutorial. Prentice-Hall, Englewood Cliffs, New Jersey.

Internet references

Arms, W.Y., 1995. Key concepts in the architecture of the digital library. D-Lib Magazine, July. http://www.dlib.org/dlib/July95/07arms.html.

Arms, W.Y., Blanchi, C., Overly, E.A., 1997. An architecture for information in digital libraries. D-Lib Magazine, February. http://www.dlib.org/dlib/february97/cnri/02arms1.html.

Bosak, J., 1997. XML, Java, and the future of the Web. http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm.

Bray, T., Guha, R.V., 1998. An MCF tutorial. http://www.textuality.com/mcf/MCF-tutorial.html.

Buehler, K., McKee, L., 1998. The OpenGIS guide: Introduction to Interoperable Geoprocessing. http://www.opengis.org/techno/guide.htm.

Byte.com, 1994. Byte.com. http://www.byte.com.

Christian, E.J., 1996. GILS: What is it? Where's it going? D-Lib Magazine, December. http://www.dlib.org/dlib/december96/12christian.html.

Clearinghouse, 1999. Information resource page (Federal Geographic Data Committee). http://www.fgdc.gov/clearinghouse/index.html.

Culpepper, R.B., 1998. Weave maps across the Web 1998 edition. http://www.geoplace.com/gw/1998/1198/1198map.asp.

DCMI, 1998. Dublin Core metadata initiative, home page. http://purl.oclc.org/dc/.

D-Lib, 1995. D-Lib Magazine. The magazine of digital library research. Corporation for National Research Initiatives, Reston, Virginia. http://www.dlib.org.

EDINA, 1999. EDINA Digimap: Online Mapping Service. http://edina.ed.ac.uk/digimap/.

Federal Geographic Data Committee, 1998. NSDI (National Spatial Data Infrastructure). http://fgdc.er.usgs.gov/nsdi/nsdi.html.

GILS, 1997. Global information locator service. http://www.g7.fed.us/gils/index.html.

GeoWorlds, 1998. GeoWorlds home page. http://lobster.isi.edu/geoworldspubli/.

The Gocad Consortium, 2000. http://pangea.stanford.edu/gocad/gocad.html.

Green, B., Bide, M., 1998. Unique identifiers: a brief introduction. http://www.bic.org.uk/uniquid.

Gruber, T., 1997. What is an ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Halfhill, T.R., 1997. Network-centric user interfaces are coming to PCs as well as to network computers. Byte, July. http://www.byte.com/art/9707/sec5/art1.htm.

IAMG, 1995. Computers and Geosciences Editor's Home Page. http://www.iamg.org/CGEditor/index.htm.

IFLA, 1995. Digital libraries: metadata resources. International Federation of Library Associations and Institutions, The Hague, Netherlands. http://www.ifla.org/II/metadata.htm.

International, D.O.I. Foundation, 1999. The Digital Object Identifier System. http://www.doi.org/articles.html.

JSTOR, 1995. Journal storage: redefining access to scholarly literature. http://www.jstor.org/.

Kahn, R., Wilensky, R., 1995. A framework for distributed digital object services. Document cnri.dlib/tn95-01, Corporation for National Research Initiatives. http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html.

Larsen, R.L., 1998. Directions for Defense Digital Libraries. D-Lib Magazine, July/August. http://www.dlib.org/dlib/july98/07larsen.html.

Lynch, C., 1997. Searching the Internet. Scientific American, March. http://www.sciam.com/0397issue/0397lynch.html.

MacEachren, A.M., 1998. Visualization — cartography for the 21st century. International Cartographic Association Commission on Visualization conference, May, Warsaw, Poland. http://www.geog.psu.edu/ica/icavis/poland1.html.

Microsoft, 1998. Microsoft TerraServer. http://terraserver.microsoft.com/default.asp.

Miller, E., 1998. An introduction to the Resource Description Framework. D-Lib Magazine, May. http://www.dlib.org/dlib/may98/miller/05miller.html.

Miller, P., 1996. Metadata for the masses — describes Dublin Core and means by which it can be implemented. Ariadne (the Web Version) Issue 5 (ISSN: 1361-3200), September. http://www.ariadne.ac.uk/issue5/metadata-masses/.

Murray-Rust, P., West, L., 1998. Virtual hyperglossary (VHG). http://www.vhg.org.uk/.

National Library of Medicine, 1998. Fact Sheet: UMLS (Unified Medical Language System) semantic network. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html.

NGDF, 1999. National Geospatial Data Framework. http://www.ngdf.org.uk/.

Open G.I.S., 1996. Intergalactic geoprocessing middleware. GIS World, March. http://www.opengis.org/techno/articles/mdleware.htm.

Orfali, R., Harskey, D., Edwards, J., 1995. Intergalactic Client/Server Computing. Byte, April. http://www.byte.com/art/9504/sec11/art1.htm.

Paskin, N., 1997. Information identifiers. Learned Publishing, vol 10, no. 2, pp. 135–156 (April). http://www.elsevier.com:80/inca/homepage/about/infoident/Menu.shtml.

POSC, 1999. POSC Specifications — Epicentre 2.2, upgrade to version 2.2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/.

Project_Gutenberg, 1999. Sailor's Project Gutenberg Server, home page. http://www.gutenberg.org/.

Rust, G., 1998. Metadata. The right approach. An integrated model for descriptive and rights metadata in e-commerce. D-Lib Magazine, July/August. http://www.dlib.org/dlib/july98/rust/07rust.html.

Schell, D., McKee, L., Buehler, K., 1995. Geodata interoperability — a key NII requirement. White paper submitted to NII 2000 Steering Committee, May. http://www.opengis.org/techno/articles/nii2000.htm.

Sheppard, S.R.J., 1999. Visualization software brings GIS applications to life. GeoWorld, March. http://www.geoplace.com/gw/1999/0399/399life.asp.

SHOE, 1999. Simple HTML ontology extensions. http://www.cs.umd.edu/projects/plus/SHOE/index.html.

Stanford KSL Network Services, 1996. Sites relevant to ontologies and knowledge sharing. http://ksl-web.stanford.edu/kst/ontology-sources.html.

Thomas, T., 1998a. Physical Review Online Archives (PROLA). D-Lib Magazine, June. http://www.dlib.org/dlib/june98/06thomas.html.

Thomas, T., 1998b. Archives in a new paradigm of scientific publishing: Physical Review Online Archives (PROLA). D-Lib Magazine, May. http://www.dlib.org/dlib/may98/05thomas.html.

United States Geological Survey, 1998. Digital geologic map data model. http://geology.usgs.gov/dm/.

Web3D Consortium, 1999. The VRML Repository. http://www.web3d.org/vrml/vrml.htm.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A123–A132

COMPUTERS & GEOSCIENCES

Geoscience after IT Part M. Business requirements drive the information system, and provide coherent frameworks. Epilog

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

Abstract

The roles of participants in the information system are changing, and this is reflected in their business groupings and motivation. IT brings greater flexibility to the record, but without a coherent framework, cyberspace becomes a chaotic sludge of trivial ephemera. Cataloging and indexing, peer review by editorial boards and the disciplined approach of information communities can impose the necessary order and standards. Metadata and data models can help to maintain a clear structure for geoscience. Business aspects link the objectives of the investigator to the framework of the science, defining the logic of reorganization and providing incentives to drive the system. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Catalogs; Editorial boards; Information communities; Information system strategy; Business aspects

1. Activities, participants, roles and driving forces

Subsystems are selected to minimize their interactions (part I, section 2.2). Nevertheless, much of the interest lies at the interfaces. Scientific investigations are conducted at the interface between the real world and the information system, drawing information from the repositories, testing or extending it by observations and measurements in the real world, and returning with conclusions that may be added to the knowledge base. The scientists' activities (D 7) are usually described by verbs, such as investigate, integrate, explain, curate, communicate. During a project, there is at least one cycle of **activities** (applying processes to objects), such as plan, undertake desk studies and field work, analyze, report, review, possibly return to additional study of the literature, more field work, and so on. Fig. 1 shows them within a circle, to avoid an arbitrary beginning or end.

The investigators and the users of the information interact with many other **participants** (those playing a part in the operation of the information system). Managers may define the business setting, possibly standing in as proxies for other stakeholders such as customers or stockholders. In an educational environment, supervisors may interpret the views of professors and other academics. Contact with the recorded knowledge base is likely to be through **intermediaries** (who assist the user or contributor to interact with some aspect of the system), such as librarians, information scientists, booksellers and curators. Collection of data may be assisted by laboratory staff, instrumentation experts, field and laboratory assistants. Recording the results

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0098-3004(00)00048-0

may involve typists, data-entry specialists, reviewers, editors, referees, curators, catalogers, database administrators, printers and publishers. The managers or supervisors are likely at all stages to advise, monitor progress, and ensure that the results conform to the intended objectives. The recommendations of standard organizations have a bearing at all stages. In addition, scientists interact informally with others working in a similar area or topic, with customers, and with those who can supply more detailed information or who are involved in broader studies.

Technology reduces the dependence of authors and users on intermediaries, such as those just mentioned (see also B 1, M 2.1). As the information industry changes in response to new technology, the new participants tend to be described in terms of their roles, and some of the old categories can be merged with the new. The roles include: clients; users; information owners, keepers and suppliers; database and repository managers; Internet service providers; network (communications and delivery) operators; webmasters; application developers; standards setters and quality assessors. The client/server mode of operation places responsibility for storing and serving the information with the originator, probably delegated to a proxy, and the form of presentation with the client who reads the information. The considerable support from the IT industry and the service suppliers, the / between client and server, is a vital, often neglected, component, but is not our subject here.

The participants work together in **business groupings**, such as oil companies or academic faculties, sharing broad objectives and ways of working. The organization is likely to be staffed with a range of experts to meet the demands of the projects it undertakes. It probably provides financial support and shared facilities, such as access to records, libraries, laboratories and computing facilities. The participants are themselves likely to be represented as objects in other databases, such as staff lists and personnel files. These may well contain information useful to the scientist, such as an e-mail address or the name of a student's supervisor. The need to communicate crosses all boundaries and cannot be limited by system definitions.

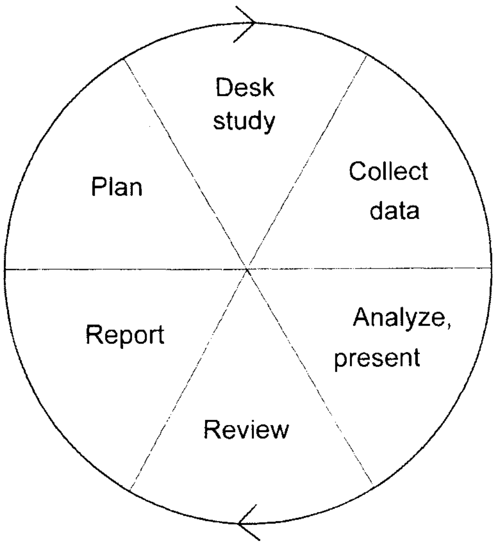


Fig. 1. Some activities in a project. A cycle of activities that may be followed more than once during an investigation.

Information is driven through the system by powerful forces. Curiosity or commercial gain may be the initial incentive for investigation. A desire to share the results, perhaps to gain kudos, promotion or scientific standing, can carry the process forward. Without such forces, it is unlikely that the system would operate at all, and their identification (preferably without undue cynicism) is an important part of analyzing the system. The **motivating factors** that drive the participants (Herzberg et al., 1993) include opportunities for achievement, recognition and advancement, and the chance to exercise creativity and take on responsibility. In any change to the system, motivation must be kept in mind to ensure cooperation in the new development. Its form may determine, for example, whether information sharing or information hoarding seems more attractive (I 8.2, M 3.2).The motives of, say, the scientist and the publisher may conflict, creating tension within the information system that has to be managed by negotiation. The information system is a social creation involving many disparate contributors in a shared activity. It will succeed or fail (Peuquet and Bacastow, 1991) depending on the motivation of all concerned.

2. Frameworks for models

Recorded information has in the past been formalized and fragmented into maps, reports, databases, archives and collections. We are now at any early stage in the global development of the hypermedia knowledge repository also known as cyberspace. Contributors of information are not constrained by the form of final presentation. This can be decided by users to meet their specific requirements. Users can bring together information from many sources, align differing ideas, employ visualization techniques, and present the results as they see fit. However, although the system can accommodate individual contributions on their own terms, that merely transfers to users the task of finding and integrating information from different sources. Flexibility is obtained at the cost of a clear structure.

From the desktop we reach a vast pool of diverse knowledge. Unfortunately, it seems at times like a chaotic swirling sludge of trivial ephemera. Metadata and the associated standards are essential to sharing knowledge, but do not in themselves provide the framework for organized thought. We can understand words without recognizing a coherent story. To bring order to this chaos we need a framework that reflects the structures of our conceptual models.

Projects generally aim to gather new information or develop new ideas. Again, a clear structure might make the process more efficient. In reporting the results, there should be no need to repeat large amounts of what has already been recorded. Instead, linkages should as far as possible give an indication of the dependency on earlier work and earlier ideas. This can be done by the author in the course of preparing the new document. Given a well-structured context, it should be possible to indicate precisely what has been assumed and where the ideas depart from those generally accepted. This would help to assess knock-on effects when ideas or definitions change. It could also indicate to the reader, in a detailed and objective way, the level of knowledge needed to assimilate the new ideas contained in the document, and offer precise guidance on where additional background can be found. It should not be necessary to complete a project before making results available to others, as fragments can be linked in to the existing context, and new versions replace old ones as the work proceeds. Nevertheless, to some extent projects supersede previous studies and therefore cannot be rigidly constrained within a preexisting framework.

Three parallel approaches to a more coherent framework come to mind.

1. One is cataloging. For example, bibliographic information for project documents could be held in a library catalog; or new hydrocarbons data could be recorded, placed in a repository and cataloged following POSC standards. The catalog provides structure and a means of access. The reputation of the data supplier or archive manager gives some assurance of quality.
2. A second approach, well suited to exploratory projects, is extension of the current scientific literature (I 6), to embrace new mechanisms of delivery while retaining the structure and evaluation imposed by the editorial board, as described in L 3. This implies that each document is largely self-contained and self-explanatory. As previously mentioned, archiving problems may arise with multimedia documents (L 6.3).
3. A third approach is for an information community to define a general model (M 2.3) for structuring relevant knowledge within the scope of their activities, treating individual investigations as subprojects within a unified framework that evolves as ideas develop.

The three approaches, considered in M 2.1 and 2.2, are not mutually exclusive. The same object could be relevant in more than one framework, and can readily be shared by hypertext reference. For example, the description of a fossil might be archived only once, but referenced from a geological survey model, an oil exploration repository and a paleontological journal. In this fast evolving field, it is likely that these and many other frameworks for shared knowledge will be explored. As members of the geoscience community, it is our task to drive this evolution — to understand and appraise, encourage or condemn.

2.1. Realigning responsibilities

Archiving has in the past been the responsibility of libraries. The copyright in the documents, however, is generally retained by their publishers, who could create an electronic archive of them as a source of significant future income. In due course, such archives would almost certainly improve the service and reduce the cost. Much of that saving would come from the reduced archiving role of the libraries. Libraries, however, are at present the main channel of government funds in purchasing conventional publications. The publishers may be reluctant, for the time being, to upset their main customers. In a travesty of the market place, publishers may even offer electronic copies of science journals at a higher price than the paper version on the grounds that even if they cost less to produce, they offer more to the purchaser.

If some of the functions of libraries are at risk from electronic archives, the same could be said for publishers (see Butler, 1999). Authors may be tempted to publish their own papers. In the past, this would have condemned the paper to obscurity. In the future, cataloging information is likely to be part of the document or digital object, and thus the responsibility of the object's owner. It is retrievable by search engines, some of which, for efficiency, would copy the cataloging information into an index and possibly extend it. Readers could thus find and retrieve the paper independently. But there are problems. The quality of an unrefereed document has not been assessed, and there is no indication, other than the author's reputation and affiliation, of its scientific standing. There is, therefore, little kudos for the author, and no guidance for the reader about whether its findings are valid or significant. Furthermore, without some assurance that the paper will be widely available within a long-term archive, it cannot be seen as part of the permanent scientific record.

Solutions to these problems are in the hands of editors rather than publishers. Editors and referees are more likely to be concerned, as at present, with whether contributions deal with an appropriate topic at a suitable level, meet the house style and agreed standards, evaluating their relevance and quality, and ensuring that they are original, inoffensive, and give credit where it is due. Readers then have the task of assessing the "brand names" of the editorial boards, just as they would expect today to have a view on the quality of a particular journal.

The role of the publisher, on the other hand, may be subsumed into repository or archive management, concerned principally with organizing and maintaining an information system and making its contents available. Some scientific societies have taken on the role of publisher, for example, the Institute of Physics (1999). In some fields, large commercial publishers may dominate, because of their financial resources, global reach, wide coverage of many disciplines, marketing skills, and above all their copyright of existing content (ScienceDirect, 1999). In the long run, costs must be recovered for access to an electronic repository, and the profit potential, particularly if charges are visible to the reader, may prove controversial.

Features which users might look for in such an information system (K 3) include:

- *stability* — there should be a clear and credible commitment to long-term preservation, access and maintenance of all information;
- *usability* — provenance, ownership of intellectual property rights, and terms and conditions of use should be clear;
- *fairness* — charges and conditions of use should be seen as fair, reasonable, competitive and consistent;
- *reliability* — the user should be able to assess the accuracy and quality of all information;
- *comprehensiveness* — within the demarcated scope of the service, the user should be confident that all likely sources are included or referenced, including the most recent work;
- *convenience* — the system should be easy to use through a consistent, familiar interface, and provide a rapid and efficient response;
- *clear structure* — available and relevant material should be easy to find and have pointers to related information.

The Digital Object Identifier (L 3) is a basis for such a system, building on the existing scientific literature. The digital object is the scientific paper, supported by entries in indexes and catalogs. A consortium of publishers has taken an initiative (ScienceDirect, 1999) to supplement, and potentially replace, paper copies with items archived electronically, for example in SGML. Versions for browsing and printing, for example in HTML, XML and PDF, are generated from the archive and accessible through the publisher's gateway, which controls access and imposes charges if required. A wide range of refereed literature, with indexes, abstracts and catalogs, can thus be made available from the desktop. The flexibility and ease of use of the Web browser complement the authority, structure and permanence of the scientific literature. The digital objects can of course be hyperlinked, and references reached by a click. As they share the same desktop interface, the formal literature can link to ephemera, detail, and work in progress recorded on the World Wide Web. Equally, the literature can have links to and from the spatial models described in the next section.

The scientific paper of around 5000 words is a convenient length for downloading to the desktop, and appropriate for marshaling and presenting a coherent view of a specific argument. More extended accounts, such as topic reviews and books, are normally arranged in chapters dealing with separate aspects of the subject. The chapter, rather than the book, might be seen as suitable for cataloging as a retrievable unit, analogous to the scientific paper. Electronic archives should focus on the content not the container. Older distinctions based on the format of presentation, such as book, serial or reprint, are likely to blur. The general scientific literature, however, is likely to be archived as sets of discrete objects or documents, and this may be inappropriate for some of the tightly structured work of information communities.

2.2. The information communities

The OGIS Guide (L 4) points out that each scientist has a unique view of the world, and that this makes communication more difficult (Buehler and McKee, 1998). They identify **information communities** — collections of people who, at least part of the time, share a common world view, language, definitions and representations — and explore possible means of communicating between such groups. An example might be NOAA, the Department of Mines and the USGS, each with their own objectives, methods, terminology and standards. Understanding the concepts of an information community can be helped by a strong framework of data models with clearly defined terms and relationships.

Valid interpretation across community boundaries is likely to depend largely on the background knowledge of the human interpreter. This is not available, nor likely to become available, to the computer. As Kent (1978) pointed out, language is a powerful tool to reconcile different viewpoints, and a basis for communicating background knowledge both of large concepts and of the details of a single object. Written explanations are therefore needed in close association with spatial and data objects at every level of detail.

We can already see in the World Wide Web the emergence of a global knowledge network, using hypermedia to express and relate ideas. There is a clear distinction between cross-references among objects, which call attention to some relationship or analogy, and the tightly linked conclusions that emerge from a project based on a single coherent set of background assumptions, objectives and working methods. Within the loose global linkages of the hypermedia knowledge repository there are more tightly organized structures of geoscience information managed by defined information communities.

Large information communities, such as geological surveys, already publish many of their own findings, and should find this easier in an IT environment. Their internal refereeing and quality assessment procedures, their copyright ownership, their brand name and reputation among their customers are already established. They should therefore be able to meet most of the users' criteria in 2.1.

As well as their own findings, a survey may hold data originally collected by external organizations for various purposes, such as site investigation or mineral assessment. The accuracy of the data is variable and cannot appropriately be judged by the survey. Provided this is made clear, however, and the source and ownership of data sets are clearly identified in the metadata, the user can evaluate them against the quality-assessed survey view of the same area, and vice versa. The survey is adding value by making the information available in context. There are benefits to the contributor in placing records within the structure of a repository where the costs of initial design, installation, marketing and maintenance are spread across many users. There are benefits to the repository in achieving more comprehensive coverage by accepting external contributions.

The requirement for up-to-date information seems to conflict with the need for a permanent record of earlier views. This can be overcome by archiving date-stamped previous versions, or by retaining the ability to reconstruct them from journalized changes, as generally only a small part of a document or data set is superseded. The task of maintaining versions should not be underestimated, for while it can be readily handled in a prototype, it could be the dominant issue in a production system (Newell et al., 1992).

In fast developing technology, the lead organization tends to keep moving ever further ahead of the pack, because users prefer a single mainstream solution. The leader can set standards while others inevitably fall behind. A **winner-take-all** situation develops, to be broken only by user dissatisfaction, by competitors using technology more effectively or catching a new wave of technological advance, by financial muscle, by political interference or a combination of these. Even within a small niche, such as a country's geology, users may prefer a single source of survey information. All users can then work on the same basis, and different areas can readily be compared.

On those grounds, a survey or similar organization (indeed any group dominating its niche and working to shared, comprehensive standards) can be well placed to maintain its market position. It just needs to stay in the forefront of technology, keep in line with changing standards, and satisfy customers and politicians. Because information technology bridges national and disciplinary boundaries, standards must be international and standards within geoscience must be consistent with those in related fields. Close collaboration with a range of other organizations is therefore essential. Some organizations can share information system resources through "extranet crossware" (Netscape, 1999). Both sides gain from the links (**win–win**), as well as customers benefitting from good service at reasonable cost.

An information community exists because its members share objectives and are organized to find an integrated solution. A geological survey (I 8.1) is an example of an information community, one of its roles being to assemble basic geological information about an area in a form which can be used in many other applications (rather than being collected separately for each project). The conventional means of achieving a coherent overview is to publish a standard series of maps with accompanying memoirs and explanatory reports, all to consistent standards. As this is firmly embedded in old technology, surveys have had to review their work from first principles. The British Geological Survey, an example of a medium-sized survey, considered the issue of their basic geoscientific model (Ovadia and Loudon, 1993) as described in the next section.

2.3. A basic regional geoscience framework

The earlier description of the geoscience information system (I 2.3) gave some impression of its scope and form, but said little about its scientific content. The triangular image of increasing abstraction in M, Fig. 3 hints that there is some shared, general model — the paradigm that geoscientists have at the back of their minds. If so, there should be a route from a single set of observations at the bottom of the triangle, such as a soil profile, linking upwards at higher levels of generality through the entire body of existing knowledge. Indeed, the claim to be a science suggests that the body of knowledge should be coherent and internally consistent. A greatly simplified overview is required to provide an overall structure into which observations and ideas can be fitted, and from which relevant information can be retrieved.

The framework of a general geoscience model can help to bring order to a multitude of investigations whose varied business aims lead to a diversity of approaches. A single, coherent, general model can specifically address the area of overlap and thus help to avoid unnecessary duplication. The task of developing that model and sharing the results can appropriately be assigned to an identified information community, such as a geological survey. The need for cooperation with related information communities is illustrated by, for example, the links between topographic and geological mapping.

Geological, topographic and related surveys worldwide have developed such models of national aspects of geoscience. Examples can be found in Australia (Australian Geodynamics Cooperative Research Centre, 2000), France (BRGM, 2000), Canada (Lithoprobe, 2000) and the United Kingdom (Adlam et al., 1988). Their concern is to convey knowledge of the consequence of geological and related processes, states and events in geological time and space. Their findings, which have a strong spatial element, have generally been expressed as maps and reports on specific areas. Geological maps may list the various rock units present in the area (**classification** and **nomenclature**), and by relating their location to a topographic base map, show their spatial distribution (**disposition**) at or near the earth's surface. Drift and soil maps may show the disposition of sequences of units. Orientation measurements, intersections with the topography, and cross-sections give an impression of the three-dimensional form, sequence, shape and structure (**configuration**) of the units. Generalized sections and text comments give an indication of the lithological and petrographical **composition** of the material. Specialist maps might give information about the geochemical composition of the material, the geophysical **properties** of the rock mass or the geotechnical properties of individual units. Paleogeographical maps and palinspastic reconstructions can be used to express a view on their formation and **historical development**. Symbols on the map may show wells, traverses, measurement stations, outcrops, and collection localities as points where **evidence** was gathered to support the conclusions.

Many aspects, such as detailed descriptions and accounts of processes, can be addressed more satisfactorily in text than on a map. The paleontologist studying a single fossil, or the seismologist studying an earthquake, may indeed consider a general geoscience model to be irrelevant. But their findings are ultimately related to some framework of space and time, viewing the fossil as a component of the material of a rock unit, throwing light on its history; and the earthquake as an event resulting from the reaction of the material to its properties and stress history.

Reports and maps are often closely associated, but perhaps maps give clearer pointers to a general model, because their graphical symbolism, uniformity, and the need for worldwide coverage require a formalized and consistent approach. However, the conventional map is a product of a particular technology. We are looking for an underlying model which refers to the scientific concepts, not the technical solutions, for our interest is in how technology can evolve to fit scientific needs (see Laxton and Becken, 1995). The concepts must be as free as possible from their form of presentation.

In a **general geoscience spatial model**, the objects of interest are the earth and parts of the earth, such as rock units or their bounding surfaces. The aspects of interest just mentioned are their disposition, configuration, composition, properties, history and evidence.

The underlying concepts are familiar. They address issues analogous to those that might worry a three-year-old child on looking into a dark room.

- What is in there and what is it called? (Object classification and nomenclature.)
- Where is it? (Disposition.)
- What does it look like? (Configuration.)
- What is it made of? (Composition.)
- What does it do? (Properties.)
- How did it get there? (History and geological processes.)
- How do I know? (Evidence and business aspects.)

We try to develop and convey the knowledge (held in our brains) of states, processes and events, sequenced in time and patterned in space, which we believe may account for our observations within our accepted world view. The types of model with which geoscientists are concerned largely determine the unique characteristics of their information system. In particular, the spatial model (G 2) is the key, not only to the disposition and configuration of objects, but also to understanding many of the relationships of their composition, properties and processes. IT may offer radical improvements in implementing the framework.

Where a strong framework and good retrieval techniques are in place, the survey model can tie into contributions from external projects, like a commentator adding footnotes to an existing story. Ideally, it should support interwoven stories dealing with any relevant topic, tied to geological space and time and the object-class hierarchies defined in the metadata. Data models define the scope and relationships of the topics considered, and provide a structure for storage and retrieval of information. The content may be complete for all the subject areas and topics, although the level of detail and date of the last revision will inevitably vary. Here is a context where contributions can be evaluated, stored and found when required, and a means of reconciling information obtained for differing business purposes. Conflicting and changing views can be held side by side, for evaluation by users.

Models such as this can provide firmly structured areas embedded in the more flexible hypermedia knowledge repository. Some information communities, such as oil companies, have more clearly defined business requirements, and precise ideas about the geoscience information required to support them. Academic studies, in contrast, may have fewer preconditions, and a need to follow ideas wherever they may lead. They must choose different points on a trade-off. On the one hand, well-defined structures and consistent standards bring reduced redundancy, increased relevance and efficient access. On the other hand, the scientific literature offers greater flexibility and ability to cope with change. The cost is greater repetition and greater effort to comprehend the diversity of ideas and modes of expression. The scientific literature is already evolving to offer hypermedia documents within distinct topic areas overseen by editorial boards.

We thus see the development of a flexible hypermedia knowledge repository. Within it, structured areas are provided by information communities of all sizes and forms, from individual businesses to consortiums of business partners, geological surveys, editorial boards for geoscience literature, and the organizations that help to establish, formalize and encourage the use of standards.

3. Business aspects

Any geoscience project is embedded in some kind of business — mineral development, civil engineering, survey, education, research, or whatever — which sets its objectives, resources and time scale. All the information systems that deliver information for the business, including the geoscience information system, are changing because of IT. Most businesses follow a yearly cycle of reviewing progress, deciding priorities, allocating funds and so on, according to a business plan. Feeding information into this, and therefore synchronized with it, there may be an **information system strategy** which supports the business objectives (CCTA, 1989). It sets out a plan (for the various parts of the business) for development of the information systems, policies, programs of work and IT infrastructure. While the strategy may be the responsibility of an IT department, geoscience needs must be taken into account and fed through to the business plan at the appropriate time. The geoscience manager who wishes to take full advantage of IT must therefore keep the business aspects in mind.

The unpredictable course of technology will itself respond to business needs. Views on mainstream developments in IT can be culled from the Web pages of the major software suppliers, such as Microsoft, Oracle and Sun (their Web addresses can be found by inserting their names between www. and .com). Those with a more academic approach may be more interested in open source software, such as Linux or GNU (place between www. and .org for their Web addresses). Such codes can be amended for specific applications and much of the software is free. Today's standard procedures may be by-passed by tomorrow's technology, and planning should therefore be flexible and kept under continual review.

3.1. Organizational consequences

In areas such as word processing, computer-aided design and Web searching, computers can assist users to carry out tasks which otherwise would require, say, a typing pool, publisher, drawing office, and library. Some changes to roles in the organization were considered in M 1. In general, intermediaries between the originator and the user of information can either be eliminated (**disintermediation**), or given a changed role, for example in providing advice on design and layout or development of standards, or in providing information systems maintenance and training.

Computer support can assist project planning and monitoring. Because computer-mediated information can be made available rapidly and widely within an organization, employees can respond to plans and requirements within a less complex management hierarchy (**delayering**) and with greater independence of individuals and groups.

Rather than regarding collection and management of information as closely linked activities, with data collectors responsible for looking after their own results, **standards** provide flexibility to combine or separate the responsibilities as appropriate. Information can be maintained by the originator during the course of a project and still be available to others over network links. Without necessarily altering the standard format of the information, it can in due course be passed to the control of a repository for long-term security.

Large amounts of data can be analyzed by computer provided they meet uniform standards. Where detailed standards are in place, many groups from many organizations can contribute shareable information. Data can be stored and managed in a shared repository. For example, POSC (L 5) has assembled standards that enable data from many sources to be shared through local and international repositories, where the task of data management is handed over (**outsourced**) to specialists. The result is huge savings to individual companies, and generally more reliable access to information.

3.2. Cost recovery

With most scholarly publication and government-funded survey, the main costs are incurred in prepublication research. Even the costs of publication relate mostly to preparatory work before the first copy is printed (B 1). The effect of electronic delivery is to reduce the initial publication cost and almost eliminate the costs of supplying subsequent copies, as printing costs fall on the user. The costs that might eventually be recovered include digitizing and storing the information, a contribution to the cost of its acquisition, and the overhead cost of maintaining the system standards and metadata. As mentioned earlier (M 2.2), success is helped by dominating the chosen market. It is therefore important for charges to be kept as low as practicable with the aim of attracting the largest possible number of customers. Customers require comprehensive information, and a viable system will need rapid growth both in terms of number of customers and amount of information. A prolonged period of free access while the service is being established, followed by gradual introduction of charges, is the pattern followed, for example, by most electronic journals and newspapers. Their large capital investment has no short-term return.

Registration of users can enable a repository to identify customers, find out which areas are of most interest and keep customers informed of relevant developments. It also ensures that the user is aware of the terms and conditions of supplying the information. The casual or one-off user can be allowed to bypass much of the registration procedure. For organizations that are heavy users of the information, a monthly or annual invoice could be convenient, covering all staff from that organization. This could either be a flat rate at levels related to usage, or based on the total of list prices for all objects accessed. For the large user, fixed amounts are simpler, but the occasional user may prefer to pay per object, and ways of transferring small amounts of cash for such transactions are being developed. For sums of more than a few dollars, charge cards are a possible alternative. The latest news of **charging procedures** can be found on the Web (see, for example, Schutzer, 1996; Herzberg, 1998).

Incentives are the driving force of an information system. An obvious incentive is money — the metric of utility space. As a creature of market forces, money can help to balance supply and demand. As an appendage to tradable objects, it can encourage sharing, not hoarding, of information. For example, a repository might charge a fee to depositors of information in order to recover the cost of managing and storing the information. The user of the information might also have to pay, to recover costs of dissemination and to pass on a royalty to the depositor. Academic susceptibilities might prefer a subsidized repository with payment in kudos not cash. Either way, there are incentives for all concerned to behave in a socially desirable manner.

To ensure that authors can benefit from their creativity, the law recognizes **intellectual property rights (IPR)**, such as **copyright**. This covers the author's rights to acknowledgement (paternity), to avoid alteration by others (integrity) and to royalties from sale of the work. The ease of copying electronic documents puts IPR at risk; see Lejeune (1999) or section 5.1: legal issues in Bailey (1996). This is one impediment to electronic communication. So-called trusted systems have been developed, but not yet widely adopted, which enable the information supplier to control information distribution as never before (The Economist, 1999). Another problem is the difficulty of calculating value. Devising a simple but effective pricing mechanism involves compromise. For example, consider what some economists call **network externalities**, that is, activities that support, benefit and extend the system as a whole, rather than individual users.

You may recall Mr Bell's problem. He invested and built a telephone, but had no-one to call. There is a snowball effect. The more people own phones, the more useful each one is, provided of course that they all follow suitable standards to make communication possible, and their phone numbers are widely known. There might be profit for Mr Bell in setting up a telephone company and selling services; but it is then to his advantage to encourage and subsidize the network of lines and exchanges, the availability of directories, and to reward the initial subscribers until the snowball effect takes over. These network externalities are a necessary development cost to him, not a profit. Above all, he must remain locked into the dominant standards. Someday his telephone might be linked to others throughout the world. I suppose he could have made an alternative decision to give away telephones and profit from the sale of directories, but the customer's perception of value and the difficulties of the protecting market share must be taken into account.

Standards and metainformation, which describe what information is available, what it is useful for, how to get it and what it means, can be regarded as network externalities in the information system. They enhance the value of the main body of information. The more widely known and accepted they are, the more the overall value is enhanced. There is, therefore, a case for making metainformation readily and freely available to all, or even paying for its dissemination (advertising). It follows that standard setting bodies need external funding from members or governments.

Another quirk of the system reflects the difficulty of the first purchaser of a telephone. The high cost and unreliability of the untried device are matched by the tedium of being able to chat only to Mr Bell. Initial involvement with a radically new information system, as contributor or user, has similar drawbacks. Being a pioneer is a mug's game — much better to wait until the systems are grooved in and most information transformed. For the rational individual, the clever strategy appears to be to wait until the last minute before leaping onto a new development curve, and so, for a while, governments, not wishing to be bypassed by history, offer subsidies for new developments. Rational organization grab them, for an organization changes more slowly than an individual, with more to gain by being ahead of the field. They invent ways to motivate staff and customers — and the attractions of the rational employee fade in comparison with the one with knowledge of IT.

4. Epilog

Like a canal navigator watching an iron horse steam by, like a railroad engineer sighting a horseless carriage, the geologist viewing images on a computer screen is witness to a paradigm shift. Unrecognized assumptions lose their validity, and things will never be the same again.

The electronic tools now fashioning the geoscience knowledge base open a door to the unknown. We can explore it only by trial and error, past experience our only guide, and so we rely on metaphor, just as Hamlet, faced with a binary decision on whether to be (or not), took refuge in a stream of metaphors — slings, arrows, sea, sleep, dreams, rub, coil — desperate to throw light from familiar concepts on a scenario he could not fully grasp.

Scientists, like poets and prophets, are accustomed to metaphor, and for the same reason. As individuals, microcosms of the universe, they seek to explore the larger whole. If a fragment of a hologram can reproduce a full image, albeit with loss of resolution, perhaps the lesser can know (imperfectly) the greater, through insights stemming from considered experience. The information technologist uses explicit metaphors — think of the screen as a desktop, the rectangular area as a window. The scientist more often formalizes the metaphor as a model, and harnesses the power of mathematics.

Working geologists give little thought to their global metaphor or world view. They might accept that on the one hand is the real world, existing quite independently of their science. On the other hand is a geoscience knowledge base, the shareable record of observations and ideas from unnumbered contributors, representing selected aspects of that real world. Moving between them are scientists, now observing or testing hypotheses in the real world, now studying or adding to the knowledge base, carrying in their minds additional ideas, too tentative, ephemeral or complex to add to the formal record, but maybe shared, in part, among their workgroup.

Information technology is shaping and transforming not the real world, but the Shadowlands of the knowledge base: no longer remote, but all pervasive. The landscape, the rocky outcrop, the hammer blow, the shattered specimen, can be shared as a visual record, available, on the instant, in Patagonia, in Perth, in Pocatello, Idaho. The network of scientific reasoning, the web of discourse painstakingly assembled over so many decades, is electrified. It is traversed by electronic agents, unhampered by boundaries of discipline or place, retrieving and delivering data to the desktop — the fruits of a multitude of endeavors, filtered for relevance, displayed for easy visualization, formatted for local manipulation and integration.

Unfamiliar metaphors and models thrive in a rebuilt knowledge base, forcing change to investigational design. Broad views across global information are supported by powerful analytical software. The players of the information industry — the students, professors, surveyors, consultants, authors, editors, referees, cartographers, publishers, librarians, booksellers, archivists, curators, customers and readers — assume new roles within changed business groupings. Earth scientists, like all makers of maps and suppliers of information, must review their methods and rationale.

But look away from the screen, step outside the door, and little is altered. The geoscience community is split. For most, there has been no revolution, the case for change has still to be made. Those captivated by new technology have set their own agenda, and largely been ignored by the traditional practitioners who take for granted their pens, paper and printing press.

Increasingly, these tools are being displaced by their electronic successors as new technologies intertwine with the knowledge base. The vision developed here has a myriad of structures, the object-integration platforms of individual geoscientists, metaphorically floating above the real world through the objects and models defining cyberspace. Controlled from the desktop, the platforms change content as they roam, and level of detail as they rise and fall. Structured by metadata and based on a shared paradigm, each supports its user's view. Objects are assembled to interact with the user's knowledge and with skills honed by evolution in the human brain. It is a place for scientists to explore ideas and embody their findings in new objects

— discussed, evaluated, and launched in cyberspace — where through variation, selection and heredity, ideas evolve and the favored survive.

References

Adlam, K.A.McL, Clayton, A.R., Kelk, B., 1988. A 'demonstrator' for the National Geosciences Data Index. International Journal of Geographical Information Systems 2 (2), 161–170.

Butler, D., 1999. The writing is on the web for science journals in print. Nature 397 (6716), 195–200.

CCTA, 1989. The Information Systems Guides. Wiley, Chichester.

Herzberg, F., Mausner, B., Snyderman, B.B., 1993. The Motivation to Work. Wiley, New York 157 pp.

Kent, W., 1978. Data and Reality. North-Holland, Amsterdam 211 pp.

Laxton, J.L., Becken, K., 1995. The design and implementation of a spatial database for the production of geological maps. Computers & Geosciences 22 (7), 723–733.

Newell, R.G., Theriault, D., Easterfield, M., 1992. Temporal GIS — modeling the evolution of spatial data in time. Computers & Geosciences 18 (4), 427–434.

Ovadia, D.C., Loudon, T.V., 1993. GIS in a geological survey's migration strategy. In: Proceedings of the 5th National AGI Conference, Birmingham, UK, pp. 3.12.1–3.12.4.

Peuquet, D.J., Bacastow, T., 1991. Organizational issues in the development of Geographical Information Systems: a case study of US Army topographic information automation. International Journal of Geographical Information Systems 5 (3), 303–319.

The Economist, 1999. Digital rights and wrongs. The Economist 353 (8128), 99–100 17 July 1999.

Internet references

Australian Geodynamics Cooperative Research Centre, 2000. 4D geodynamic model of Australia. http://www.agcrc.csiro.au/4dgm/.

Bailey, C.W. Jr, 1996. Scholarly electronic publishing bibliography. University of Houston Libraries, Houston, 1996–99. http://info.lib.uh.edu/sepb/sepb.html.

BRGM, 2000. Le programme national de recherche scientifique pour l'imagerie géologique et géophysique de la France en 3D. http://www.brgm.fr/geofrance3d/geofrance3d.html.

Buehler, K., McKee, L., 1998, The OpenGIS guide: Introduction to Interoperable Geoprocessing, http://www.opengis.org/techno/guide.htm.

Herzberg, A., 1998. Safeguarding digital library contents: charging for online content. D-Lib Magazine, January 1998. http://www.dlib.org/dlib/january98/ibm/01herzberg.html.

Institute of Physics, 1999. Sources, Journals. http://www.iop.org/jo.html.

Lejeune, L., 1999. Who owns what? The Journal of Electronic Processing 4 (3) http://www.press.umich.edu/jep/04-03/glos0403.html.

Lithoprobe, 2000. Lithoprobe: Canada's National Geoscience Project. http://www.geop.ubc.ca/Lithoprobe/public/aboutlp.html.

Netscape, 1999. Building applications in the Net economy. http://developer.netscape.com/docs/wpapers/index.html.

Schutzer, D., 1996. A need for a common infrastructure: digital libraries and electronic commerce. D-Lib Magazine, April 1966. http://www.dlib.org/dlib/april96/04schutzer.html.

ScienceDirect, 1999. ScienceDirect: providing desktop access to the full text of more than 1000 scientific, medical and technical journals published by the world's leading scientific publishers. http://www.sciencedirect.com/.

PERGAMON

Computers & Geosciences 26 (2000) A133–A137

COMPUTERS & GEOSCIENCES

Geoscience after IT
Part N. Cumulated references

T.V. Loudon

British Geological Survey, West Mains Road, Edinburgh EH9 3LA, UK

1. Bibliographical references

*References to the **World Wide Web** are in Section 2.*

Addis, T.R., 1985. Designing Knowledge-Based Systems. Kogan Page Ltd., London, 322pp.

Adlam, K.A.McL., Clayton, A.R., Kelk, B., 1988. A 'demonstrator' for the National Geosciences Data Index. International Journal of Geographical Information Systems 2 (2), 161–170.

Agterberg, F.P., Cheng, Q. (Eds.), 1999. Fractals and multifractals (special issue). Computers & Geosciences 25 (9), 947–1099.

Albrecht, J., 1999. Geospatial information standards. A comparative study of approaches in the standardisation of geospatial information. Computers & Geosciences 25, 9–24.

Audi, R., 1998. Epistemology: A Contemporary Introduction to the Theory of Knowledge. Routledge, London, 340pp.

Baker, G.L., Gollub, J.P., 1996. Chaotic Dynamics: An Introduction. Cambridge University Press, Cambridge, 256pp.

Barton, C.C., La Pointe, P.R. (Eds.), 1995. Fractals in the Earth Sciences. Plenum Press, New York, 265pp.

Beer, S., 1967. Cybernetics and Management, 2nd ed. English Universities Press, London, 240pp.

Blackmore, S., 1999. The Meme Machine. Oxford University Press, New York, 264pp.

Blaha, M., Premerlani, W., 1998. Object-Oriented Modeling and Design for Database Applications. Prentice-Hall, Upper Saddle River, New Jersey, 484pp.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modelling with GIS. Pergamon, Oxford, 398pp.

Brachman, R.J., Levesque, H.J. (Eds.), 1985. Readings in Knowledge Representation. Kaufmann, Los Altos, 571pp.

Briner, A.P., Kronenberg, H., Mazurek, M., Horn, H., Engi, M., Peters, T., 1999. FieldBook and GeoDatabase: tools for field data acquisition and analysis. Computers & Geosciences 25 (10), 1101–1111.

British Standards Institution, 1963. Guide to the Universal Decimal Classification (UDC). British Standards Institution, London, 128pp.

Buchanan, G.R., 1995. Schaum's Outline of Theory and Problems of Finite Element Analysis (Schaum's Outline Series). McGraw-Hill, New York, 264pp.

Butler, D., 1999. The writing is on the web for science journals in print. Nature 397 (6716), 195–200.

Buttenfield, B.B., McMaster, R.B. (Eds.), 1991. Map Generalization: Making Rules for Knowledge Representation (Symposium Papers). Wiley, New York, 245pp.

CCTA, 1989. The Information Systems Guides. John Wiley & Sons, Chichester.

Cattell, R.G.G., 1991. Object Data Management: Object-Oriented and Extended Relational Database Systems. Addison-Wesley, Reading, Mass. 318pp.

Chamberlin, T.C., 1897. The method of multiple working hypotheses. Journal of Geology. Reprinted in 1995, Journal of Geology 103, 349–354.

Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, New Jersey, 360pp.

Coad, P., Yourdon, E., 1991. Object-Oriented Design. Yourdon Press, Englewood Cliffs, NJ, 197pp.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York, 349pp.

Davis, J.C., 1973. Statistics and Data Analysis in Geology: with Fortran Programs. Wiley, New York, 550pp.

Encyclopedia Britannica, 1973 (Ed.) William Benton, Chicago.

E-mail address: v.loudon@bgs.ac.uk (T.V. Loudon).

0098-3004/00/$ - see front matter © 2000 Published by Elsevier Science Ltd. All rights reserved.

Foley, J.D., 1994. Introduction to Computer Graphics. Addison-Wesley, Reading, Mass., 559pp.

Förster, A., Merriam, D.F. (Eds.), 1996. Geologic Modeling and Mapping. Plenum, New York, 334pp.

Gallagher, R.S. (Ed.), 1995. Computer Visualization, Techniques for Scientific and Engineering Analysis. CRC Press, Boca Raton, 312pp.

Garfield, E., 1983. Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. Wiley, New York, 274pp.

Gilbert, G.K., 1896. The origin of hypotheses, illustrated by the discussion of a topographic problem. Science, N.S., 3, 1–13.

Goodchild, M.F., 1992. Geographical Data Modeling. Computers & Geosciences 18 (4), 401–408.

Graham, I., 1994. Object Oriented Methods, 2nd ed. Addison-Wesley, Wokingham, 473pp.

Griffiths, J.C., 1967. Scientific Methods in Analysis of Sediments. McGraw-Hill, New York, 508pp.

Grunsky, E.C., Cheng, Q., Agterberg, F.P., 1996. Applications of spatial factor analysis to multivariate geochemical data. In: Förster, A., Merriam, D.F. (Eds.), Geologic Modeling and Mapping. Plenum, New York, 334pp.

Henderson, P., 1993. Object-Oriented Specification and Design with C++. McGraw-Hill, Maidenhead, Berks., 263pp.

Herzberg, F., Mausner, B., Snyderman, B.B., 1993. The Motivation to Work. Wiley, New York, 157pp.

Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J., 1997. Global Positioning Systems: Theory and Practice. Springer-Verlag, New York, 389pp.

Houlding, S.W., 1994. 3D Geoscience Modeling: Computer Techniques for Geological Characterization. Springer-Verlag, New York, 309pp.

Huber, M., Schneider, D., 1999. Spatial data standards in view of models of space and the functions operating on them. Computers & Geosciences 25, 25–38.

Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, Oxford, 561pp.

Jones, C.B., 1989. Data structures for three-dimensional spatial information systems in geology. International Journal of Geographical Information Systems 3 (1), 15–31.

Kent, W., 1978. Data and Reality. North-Holland Publishing Company, Amsterdam, 211pp.

Kraak, M.-J., 1999. Visualization for exploration of spatial data. International Journal of Geographical Information Science 13 (4), 285–288.

Kreyszig, E., 1991. Differential Geometry. Dover, New York, 352pp.

Krumbein, W.C., Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill Inc., New York, 475pp.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. The University of Chicago Press, Chicago, 172pp.

Lancaster, P., Salkauskas, K., 1986. Curve and Surface Fitting. Academic Press, London, 280pp.

Laszlo, E., 1972. The Systems View of the World. Braziller, New York, 131pp.

Laxton, J.L., Becken, K., 1995. The design and implementation of a spatial database for the production of geological maps. Computers & Geosciences 22 (7), 723–733.

Leatherdale, W.H., 1974. The Role of Analogy, Model and Metaphor in Science. North-Holland, Elsevier, Amsterdam, 276pp.

Loudon, T.V., 2000. Geoscience after IT. Elsevier, Oxford, 140pp.

MacEachern, A.M., Kraak, M.-J., 1997. Exploratory cartographic visualization: advancing the agenda. Computers & Geosciences 23, 335–343.

Mandelbrot, B.B., 1982. The Fractal Geometry of Nature. Freeman, San Francisco, 460pp.

Mark, D.M., Lauzon, J.P., Cebrian, J.A., 1989. A review of quadtree-based strategies for interfacing coverage data with Digital Elevation Models in grid form. International Journal of Geographical Information Systems 3 (1), 3–14.

McCrone, J., 1997. Wild minds. New Scientist, 156 (2112), 26–30.

Minsky, M., 1981. A Framework for Representing Knowledge. Reprinted, pp. 95–128. In: Haugeland, J. (Ed.), Mind Design. MIT Press, Cambridge, 368pp.

Moore, K., Dykes, J., Wood, J., 1999. Using Java to interact with geo-referenced VRML within a virtual field course. Computers & Geosciences 25 (10), 1125–1136.

Mulvany, N.C., 1994. Indexing Books. University of Chicago Press, Chicago, 320pp.

Newell, R.G., Theriault, D., Easterfield, M., 1992. Temporal GIS — modeling the evolution of spatial data in time. Computers & Geosciences 18 (4), 427–434.

Ovadia, D.C., Loudon, T.V., 1993. GIS in a geological survey's migration strategy. Proceedings of the 5th National AGI Conference, Birmingham, UK. pp. 3.12.1–3.12.4.

POSC, 1993. Petrotechnical Open Software Corporation, Software Integration Platform Specification. Epicentre Data Model, version 1. Volume 1: Tutorial. Prentice-Hall, Englewood Cliffs, New Jersey.

Peuquet, D.J., Bacastow, T., 1991. Organizational issues in the development of Geographical Information Systems: a case study of US Army topographic information automation. International Journal of Geographical Information Systems 5 (3), 303–319.

Pinker, S., 1997. How the Mind Works, Norton, New York, 660pp.

Playfair, J., 1805. Biographical account of the late Dr James Hutton, F.R.S.Edin. Transactions of the Royal Society of Edinburgh, Vol. V.-P.III. Reprinted 1997, in James Hutton & Joseph Black. RSE Scotland Foundation, Edinburgh, Scotland.

Popper, K.R., 1996. Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, London, 431pp.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in Fortran — The Art of Scientific Computing, 2nd ed. Cambridge University Press, Cambridge, 963pp.

Reyment, R.S., Jöreskog, K.G. (Eds.), 1993. Applied Factor Analysis in the Natural Sciences. Cambridge University Press, New York, 371pp.

Rogers, D.F., Adams, J.A., 1976. Mathematical Elements for Computer Graphics. McGraw-Hill, New York, 239pp.

Rudwick, M.J.S., 1976. The emergence of a visual language for geological science 1760–1840. History of Science 14, 149–195.

Shimomura, R.H. (Ed.), 1989. GeoRef Thesaurus and Guide to Indexing, 6th ed. American Geological Institute, Falls Church, Va.

Snyder, J.P., 1987. Map Projection — A Working Manual. United States Geological Survey Professional Paper 1395. Government Printing Office, Washington.

Strang, G., 1994. Wavelets. American Scientist 82, 250–255.

Swan, A.R.H., Sandilands, M., 1995. Introduction to Geological Data Analysis. Blackwell Science, Oxford, 446pp.

The Economist, 1999. Digital rights and wrongs. The Economist 353 (8128) (17 July 1999), 99–100.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Mass., 499pp.

Turcotte, D.L., 1992. Fractals and Chaos in Geology and Geophysics. Cambridge University Press, Cambridge, 221pp.

Van Lehn, K. (Ed.), 1991. Architecture for Intelligence — The 22nd Carnegie Mellon Symposium on Cognition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Watson, D.F., 1992. Contouring: a guide to the analysis and display of spatial data. Computer Methods in the Geosciences, 10. Pergamon, Oxford, 321pp.

Wendebourg, J., Harbaugh, J.W., 1997. Simulating oil entrapment in clastic sequences. Computer Methods in the Geosciences, 16. Pergamon, Oxford, 199pp.

Whitehead, A.N., 1911. An Introduction to Mathematics. Thornton Butterworth, London, 256pp.

Whitehead, A.N., 1929. Process and Reality. Cambridge University Press, 429pp. Reprinted 1969, Free Press, New York.

Worboys, M.F., Hearnshaw, H.M., Maguire, D.J., 1990. Object-oriented data modelling for spatial databases. International Journal of Geographical Information Systems 4 (4), 369–383.

Wyllie, P.J., 1999. Hot little crucibles are pressured to reveal and calibrate igneous processes. In: Craig, G.Y., Hull, J.H. (Eds.), James Hutton — Present and Future. Geological Society, London, Special Publications, 150, pp. 37–57.

2. Internet references

The references to the ***World Wide Web*** *have a citation date of January 2000. If they can no longer be found under the http reference, it may be possible to locate them (or alternatives) by looking up keywords in a Web search engine. Some may have paper versions included in the main reference section.*

Amazon.com, 1996. Welcome to Amazon.com. http://www.amazon.com/.

Arms, W.Y., Blanchi, C., Overly, E.A., 1997. An architecture for information in digital libraries. D-Lib Magazine, February 1997. http://www.dlib.org/dlib/february97/cnri/02arms1.html.

Arms, W.Y., 1995. Key concepts in the architecture of the digital library. D-Lib Magazine, July 1995. http://www.dlib.org/dlib/July95/07arms.html.

Australian Geodynamics Cooperative Research Centre, 2000. 4D geodynamic model of Australia. http://www.agcrc.csiro.au/4dgm/.

Bailey, C.W. Jr, 1996. Scholarly electronic publishing bibliography. Houston: University of Houston Libraries, 1996–99. http://info.lib.uh.edu/sepb/sepb.html.

BGS, 1998. British Geological Survey home page. http://www.bgs.ac.uk/.

BRGM. Le programme national de recherche scientifique pour l'imagerie géologique et géophysique de la France en 3D. http://www.brgm.fr/geofrance3 d/geofrance3 d.html.

Biblio Tech Review, 1999. Information technology for libraries. Z39.50 — Part 1 — an overview. http://www.gadgetserver.com/bibliotech/html/z39_50.html.

Bosak, J., 1997. XML, Java, and the future of the Web. http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm.

Bray, T., Guha, R.V., 1998. An MCF tutorial. http://www.textuality.com/mcf/MCF-tutorial.html.

Buehler, K., McKee, L. (Eds.), 1998. The OpenGIS guide: Introduction to Interoperable Geoprocessing. http://www.opengis.org/techno/guide.htm.

Butler, J.C., 1996. Another node on the Internet for those with interests in geosciences, mathematics and computing. http://www.uh.edu/ ~ jbutler/anon/anon.html.

Byte.com, 1994. Byte.com. http://www.byte.com.

Christian, E.J., 1996. GILS: What is it? Where's it going? D-Lib Magazine, December 1996. http://www.dlib.org/dlib/december96/12christian.html.

Clearinghouse, 1999. Information resource page (Federal Geographic Data Committee). http://www.fgdc.gov/clearinghouse/index.html.

Computers and Geosciences, 1997. Computers & Geosciences Online. http://www.elsevier.com/locate/compgeosci.

Culpepper, R.B., 1998. Weave maps across the Web 1998 edition. http://www.geoplace.com/gw/1998/1198/1198map.asp.

DCMI, 1998. Dublin Core metadata initiative, home page. http://purl.oclc.org/dc/.

D-Lib, 1995. D-Lib Magazine. The magazine of digital library research. Corporation for National Research Initiatives, Reston, Virginia. http://www.dlib.org.

EDINA, 1999. EDINA Digimap: Online Mapping Service. http://edina.ed.ac.uk/digimap/.

Federal Geographic Data Committee, 1998. NSDI (National Spatial Data Infrastructure). http://fgdc.er.usgs.gov/nsdi/nsdi.html.

GILS, 1997. Global information locator service. http://www.usgs.gov/public/gils/gils1p.html.

GeoWorlds, 1998. GeoWorlds home page. http://lobster.isi.edu/geoworldspubli/.

Ginsparg, P., 1996. Winners and losers in the global research village. Invited contribution for conference on electronic publishing in science held at UNESCO HQ, Paris, 12–13 February 1996. http://xxx.lanl.gov/blurb/pg96unesco.html.

The Gocad Consortium, 2000. http://pangea.stanford.edu/gocad/gocad.html.

Goldfinger, C., 1996. Electronic money in the United States: current status, prospects and major issues. http://www.ispo.cec.be/infosoc/eleccom/elecmoney.html.

Graham, L.A., 1997. Land, sea, air: GPS/GIS field mapping solutions for terrestrial, aquatic and aerial settings. GIS World, January 1997. http://www.geoplace.com/gw/1997/0197/0197feat.asp.

Graps, A., 1995. Amara's wavelet page. http://www.amara.com/current/wavelet.html.

Green, B., Bide, M., 1998. Unique identifiers: a brief introduction. http://www.bic.org.uk/uniquid.

Gruber, T., 1997. What is an ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Halfhill, T.R., 1997. Network-centric user interfaces are coming to PCs as well as to network computers. Byte, July 1997. http://www.byte.com/art/9707/sec5/art1.htm.

Hepner, G.F., Sandwell, D.T., Manton, M. (Eds.), 1998. Earth Interactions Journal. http://earthinteractions.org/.

Herzberg, A., 1998. Safeguarding digital library contents: charging for online content. D-Lib Magazine, January 1998. http://www.dlib.org/dlib/january98/ibm/01herzberg.html.

IAMG, 1995. Computers & Geosciences Editor's Home Page. http://www.iamg.org/CGEditor/index.htm.

IFLA, 1995. Digital libraries: metadata resources. International Federation of Library Associations and Institutions, The Hague, The Netherlands. http://www.ifla.org/II/metadata.htm.

IFLA, 1998. Citation guides for electronic documents (Style guides and resources on the Internet). International Federation of Library Associations and Institutions, The Hague, Netherlands. http://www.ifla.org/I/training/citation/citing.htm.

ISO, 1999. ISO 690-2, Bibliographic references to electronic documents. Excerpts from International Standard ISO 690-2. http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm.

Ingram, P., 1997. The Virtual Earth: a tour of the World Wide Web for earth scientists. http://atlas.es.mq.edu.au/users/pingram/v_earth.htm or http://teachserv.earth.ox.ac.uk/resources/v_earth.html.

Institute of Physics, 1999. Sources, Journals. http://www.iop.org/jo.html.

Institute for Scientific Information, 1999. Home page with information on ISI citation databases. http://www.isinet.com/.

International D.O.I. Foundation, 1999. The Digital Object Identifier System. http://www.doi.org/articles.html.

International Telecommunications Union, 1999. IMT 2000: A vision of global access in the 21st century. http://www.itu.int/imt/.

JSTOR, 1995. Journal storage: redefining access to scholarly literature. http://www.jstor.org/.

Kahn, R., Wilensky, R., 1995. A framework for distributed digital object services. Document cnri.dlib/tn95-01, Corporation for National Research Initiatives. http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html.

Kasdorf, B., 1998. SGML and PDF — why we need both. The Journal of Electronic Publishing 3 (4). http://www.press.umich.edu/jep/03-04/kasdorf.html.

Larsen, R.L., 1998. Directions for Defense Digital Libraries. D-Lib Magazine, July/August 1998. http://www.dlib.org/dlib/july98/07larsen.html.

Lejeune, L., 1999. Who owns what? The Journal of Electronic Processing 4 (3). http://www.press.umich.edu/jep/04-03/glos0403.html.

Library, of Congress, 1999. The Library of Congress standards. http://lcweb.loc.gov/loc/standards/.

Lithoprobe, 2000. Lithoprobe: Canada's National Geoscience Project. http://www.geop.ubc.ca/Lithoprobe/public/aboutlp.html.

Lynch, C., 1997. Searching the Internet. Scientific American, March 1997. http://www.sciam.com/0397issue/0397lynch.html.

McCrone, J., 1999. Going inside — the neuronaut's guide to the science of consciousness. http://www.btinternet.com/~neuronaut/index.html.

MacEachren, A.M., 1998. Visualization — cartography for the 21st century. International Cartographic Association Commission on Visualization conference, May 1998, Warsaw, Poland. http://www.geog.psu.edu/ica/icavis/poland1.html.

Microsoft, 1998. Microsoft TerraServer. http://www.terraserver.microsoft.com/default.asp.

Miller, E., 1998. An introduction to the Resource Description Framework. D-Lib Magazine, May 1998. http://www.dlib.org/dlib/may98/miller/05miller.html.

Miller, P., 1996. Metadata for the masses — describes Dublin Core and means by which it can be implemented. Ariadne (the Web Version) Issue 5 (ISSN: 1361-3200), September 1996. http://www.ariadne.ac.uk/issue5/metadata-masses/.

National Library of Medicine, 1998. Fact Sheet: UMLS (Unified Medical Language System) semantic network. http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html.

Murray-Rust, P., West, L., 1998. Virtual hyperglossary (VHG). http://www.vhg.org.uk/.

NGDF, 1999. National Geospatial Data Framework. http://www.ngdf.org.uk/.

NISS, 1999. Library OPACs in HE [Higher Education in UK]. http://www.niss.ac.uk/lis/opacs.html.

NLfB, 1999. Die Bohrdatenbank von Niedersachsen (in German). http://www.bgr.de/z6/index.html.

Netscape Communications Corporation, 1997. White paper — CORBA: catching the next wave. http://developer.netscape.com/docs/wpapers/corba/index.html.

Netscape, 1999. Building applications in the Net economy. http://developer.netscape.com/docs/wpapers/platform/index.html.

OMG, 1997. The OMG (Object Management Group, Inc.) home page. http://www.omg.org/.

Odlyzko, A.M., 1994. Tragic loss or good riddance? The impending demise of traditional scholarly journals. http://www.iicm.edu/jucs_0_0/tragic_loss_or_good/html/paper.html.

Odlyzko, A.M., 1996. On the road to electronic publishing. Euromath Bulletin 2 (1) 49–60. http://www.research.att.com/~amo/doc/tragic.loss.update.

Open G.I.S., 1996. Intergalactic geoprocessing middleware. GIS World, March 1996. http://www.opengis.org/techno/articles/mdleware.htm.

Orfali, R., Harskey, D., Edwards, J., 1995. Intergalactic

Client/Server Computing. Byte, April 1995. http://www.byte.com/art/9504/sec11/art1.htm.

POSC, 1997. POSC Specifications — Epicentre 2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/Epicentre.2_2/SpecViewer.html.

POSC, 1999. POSC Specifications — Epicentre 2.2, upgrade to version 2.2.2. Petrotechnical Open Software Corporation, Houston, Texas. http://www.posc.org/.

Paskin, N., 1997. Information identifiers. Learned Publishing 10 (2) 135–156 (April 1997). http://www.elsevier.com:80/inca/homepage/about/infoident/Menu.shtml.

Project_Gutenberg, 1999. Sailor's Project Gutenberg Server, home page. http://www.gutenberg.org/.

Rust, G., 1998. Metadata. The right approach. An integrated model for descriptive and rights metadata in e-commerce.· D-Lib Magazine, July/August 1998. http://www.dlib.org/dlib/july98/rust/07rust.html.

SHOE, 1999. Simple HTML ontology extensions. http://www.cs.umd.edu/projects/plus/SHOE/index.html.

Schell, D., McKee, L., Buehler, K., 1995. Geodata interoperability — a key NII requirement. White paper submitted to NII 2000 Steering Committee, May 1995. http://www.opengis.org/techno/articles/nii2000.htm.

Schutzer, D., 1996. A need for a common infrastructure: digital libraries and electronic commerce. D-Lib Magazine, April 1996. http://www.dlib.org/dlib/april96/04schutzer.html.

ScienceDirect, 1999. ScienceDirect: providing desktop access to the full text of more than 1000 scientific, medical and technical journals published by the world's leading scientific publishers. http://www.sciencedirect.com/.

Seaman, D., 1999. About Standard Generalized Markup Language (SGML). http://etext.lib.virginia.edu/sgml.html.

Sheppard, S.R.J., 1999. Visualization software brings GIS applications to life. GeoWorld, March 1999. http://www.geoplace.com/gw/1999/0399/399life.asp.

Stanford K.S.L. Network Services, 1996. Sites relevant to ontologies and knowledge sharing. http://ksl-web.stanford.edu/kst/ontology-sources.html.

Thomas, T., 1998a. Physical Review Online Archives (PROLA). D-Lib Magazine, June 1998. http://www.dlib.org/dlib/june98/06thomas.html.

Thomas, T., 1998b. Archives in a new paradigm of scientific publishing: Physical Review Online Archives (PROLA). D-Lib Magazine, May 1998. http://www.dlib.org/dlib/may98/05thomas.html.

United States Geological Survey, 1998. Digital geologic map data model. http://geology.usgs.gov/dm/.

Universal Library, 1999. Numerical recipes on-line. Hosted by Carnegie Mellon University. http://www.ulib.org/webRoot/Books/Numerical_Recipes/.

Varian, H.R., 1994. Recent research papers of Hal R. Varian. http://www.sims.berkeley.edu/~hal/people/hal/papers.html.

Web3D Consortium, 1999. The VRML Repository. http://www.web3d.org/vrml/vrml.htm.

This Page Intentionally Left Blank

PERGAMON

Computers & Geosciences 26 (2000) A139–A142

COMPUTERS & GEOSCIENCES

Index

PII: S0098-3004(00)00070-4