<<<Back to Table of Contents
On to Part I: A view of the conventional geoscience information system>>>

Geoscience after IT: Part H

Familiarization with managing the information base

T. V. Loudon
British Geological Survey, West Mains Road, Edinburgh EH9 3LA, U.K.
e-mail: v.loudon@bgs.ac.uk

Postprint of article in Computers & Geosciences, 26 (3A) April 2000, pp. A63-A73

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.

Key Words - 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 (part I, section 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. Geographic 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 overlap).

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. Archived 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 contents, 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 location 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 (**OPAC**s) 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, p6). 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 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.

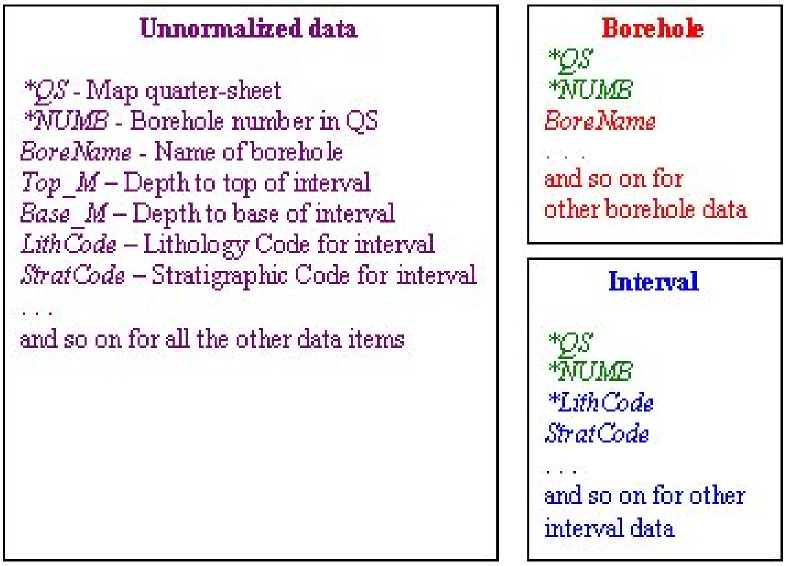
Successful sharing of detailed information depends of a common purpose. Data are more consistent when they are collected in a similar way by similar instruments. The extent of integration may depend on the supporting information technology. Long ago, paper records about the geology of an oil field, for 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 (1999), 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.



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.

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.

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.

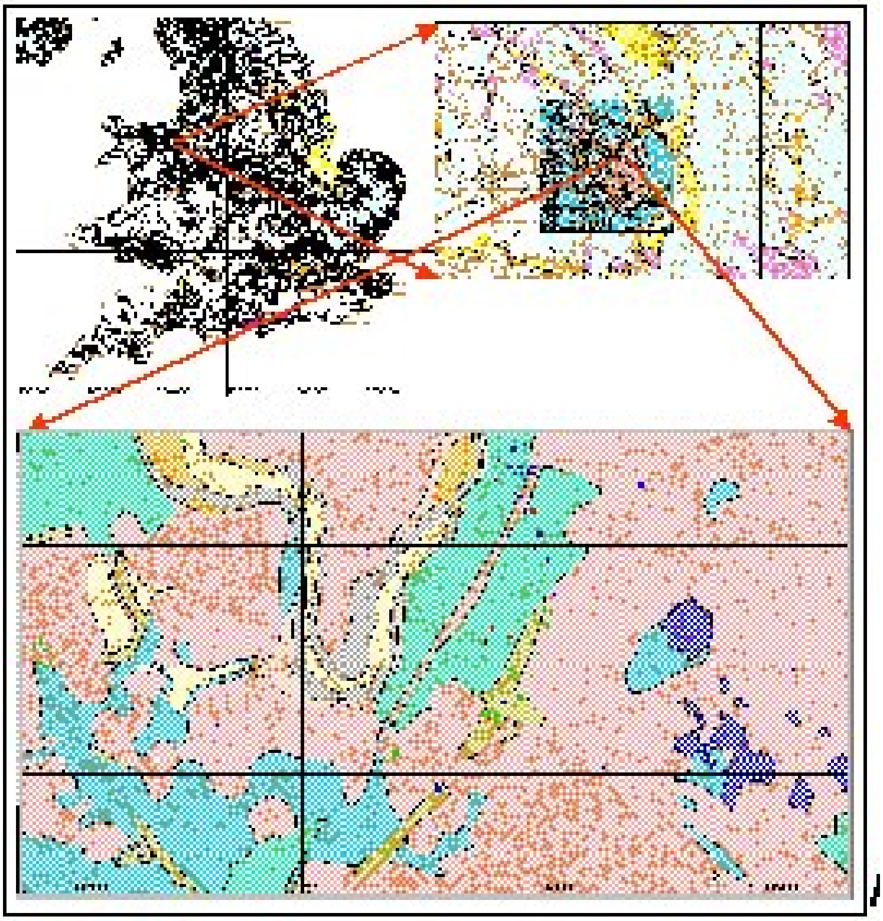


Fig. 2. Spatial search. Within a single topic, such as Drift geology, GIS enables the user to zoom in to an area of interest, seeing how it relates to the broader picture. These extracts are from the BGS Geoscience Data Index. British Geological Survey ©NERC. All rights reserved. Base maps reproduced 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 3d 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.

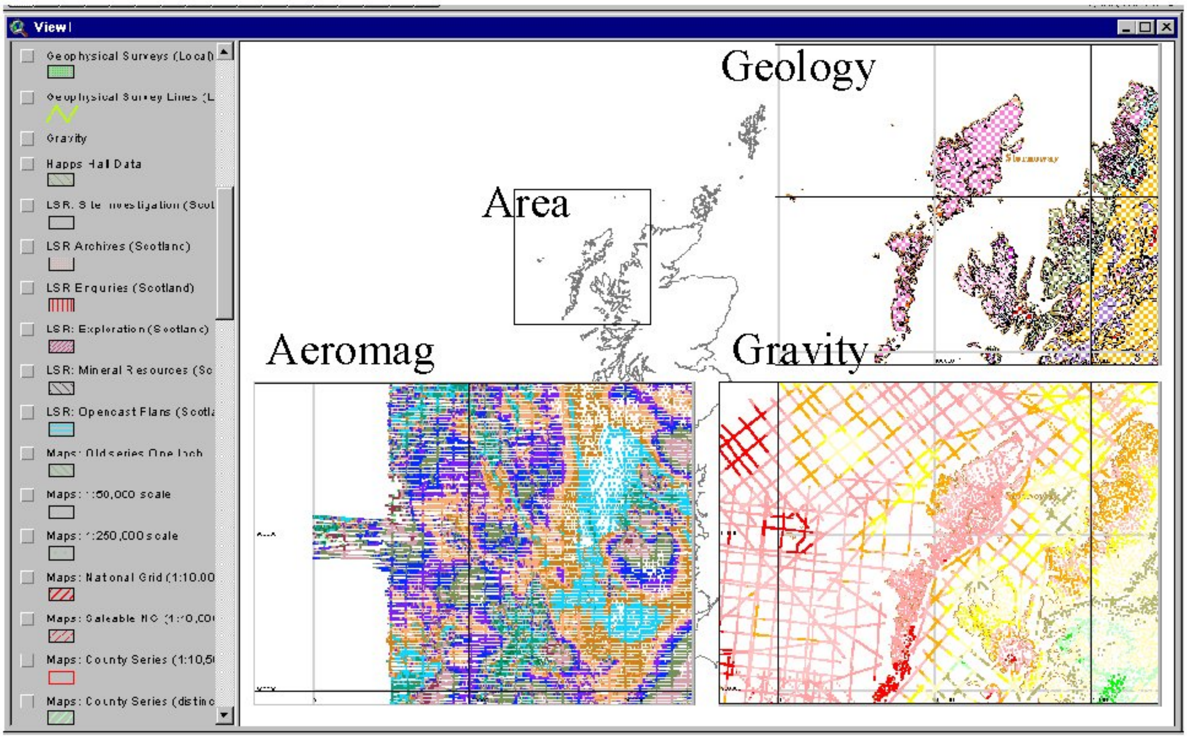

View1
□ Geophysical Surveys (Local)
□ Geophysical Survey Lines (L
□ Gravity
□ Happs Hal Data
□ LSR: Site investigation (Scot
□ LSR Archives (Scotlanc)
□ LSR Erquries (Scotland)
□ LSR: Exploration (Scotlanc)
□ LSR: Mineral Resources (Sc
□ LSR: Opencast Flans (Scotl
□ Maps: Old series One Inch
□ Maps: 1:50,000 scale
□ Maps: 1:250,000 scale
□ Maps: Natonal Grid (1:10.00
□ Maps: Saleable NO (1:10,00
□ Maps: County Series (1:10,5
□ Maps: County Series (distinc
Geology
Area
Aeromag
Gravity
Stornoway

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.

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 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 or Blaha, 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 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 L - M.

6. References

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

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

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

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

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

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

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, 320pp.

Shimomura, Ruth H. (Ed), 1989. GeoRef Thesaurus and Guide to Indexing, 6th edn. 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.

6.1 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/

Disclaimer: 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.

<<<Back to Table of Contents
On to Part I: A view of the conventional geoscience information system>>>