<<<Back to Table of Contents
On to Part M: Business requirements drive the information system, and provide coherent frameworks>>>

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, U.K.
e-mail: v.loudon@bgs.ac.uk

Postprint of article in Computers & Geosciences, 26 (3A) April 2000, pp. A109-A121

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.

Key Words - 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 Indeed, by the time you read this, some may have been superseded.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 (L 3 - L 5).

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 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 and 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).

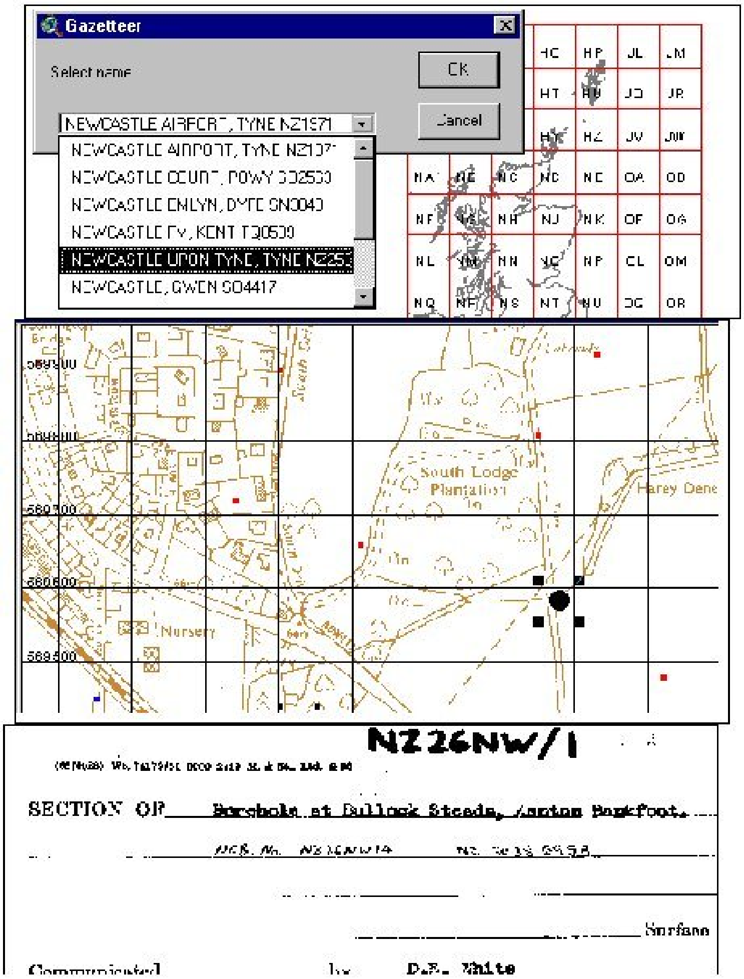
The G7 nations and the European Commission have organized a joint project to provide an information locator service with an emphasis on global environmental information (GILS, 1997). They extended the Government Information Locator Service, which is used in the US Federal Clearinghouses and State agencies, 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 date.

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 overviews 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, 1999) 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.



Gazetteer

Select name

NEWCASTLE AIRPORT, TYNE NZ1971

NEWCASTLE AIRPORT, TYNE NZ1971
NEWCASTLE COURT, POWY SO2530
NEWCASTLE EMLYN, DYFE SN3040
NEWCASTLE FY, KENT TQ0530
NEWCASTLE UPON TYNE, TYNE NZ252
NEWCASTLE, GWEN SO4417

NZ26NW/1

SECTION OF Borehole at Bullock Steads, Kenton Bankfoot.

Surface

Communicated by D.E. 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.

Users, however, may wish to assemble spatial information from many sources, not just from 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 while also being published as an integral part of a larger text-based document.

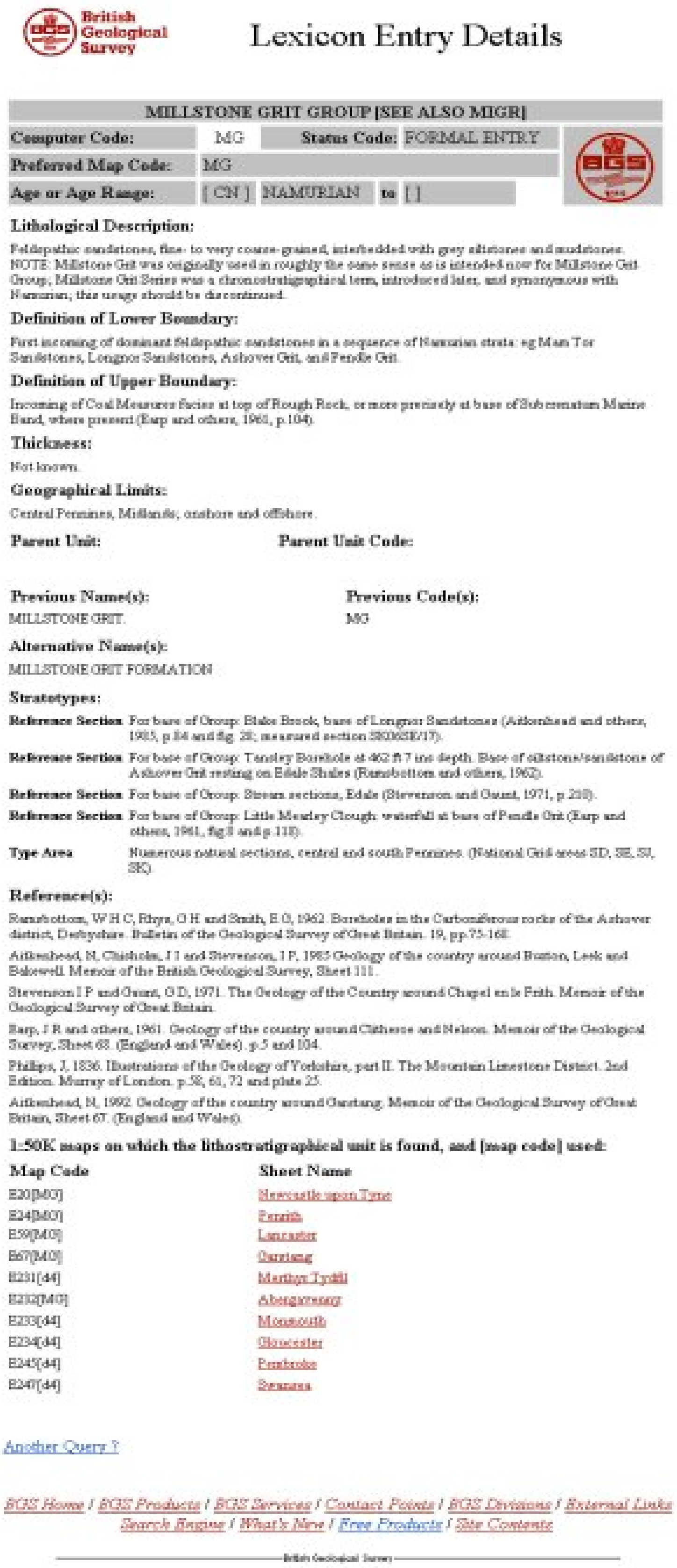
5. Structured data

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, 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.



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.73-168.

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.5 and 104.

Phillips, J, 1836. Illustrations of the Geology of Yorkshire, part II. The Mountain Limestone District. 2nd Edition. Murray of London. p.58, 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

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 - L5) 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”.

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.

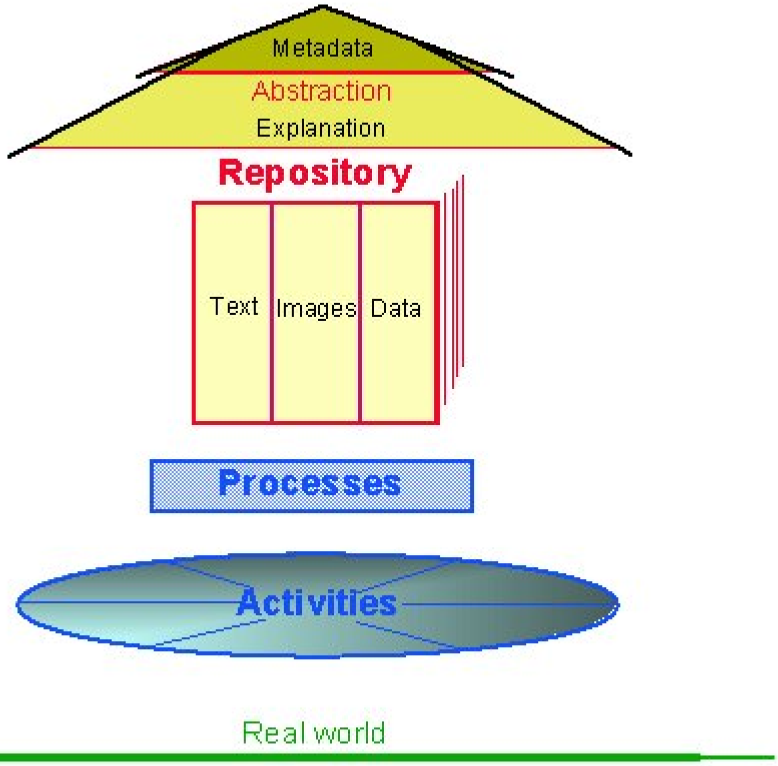
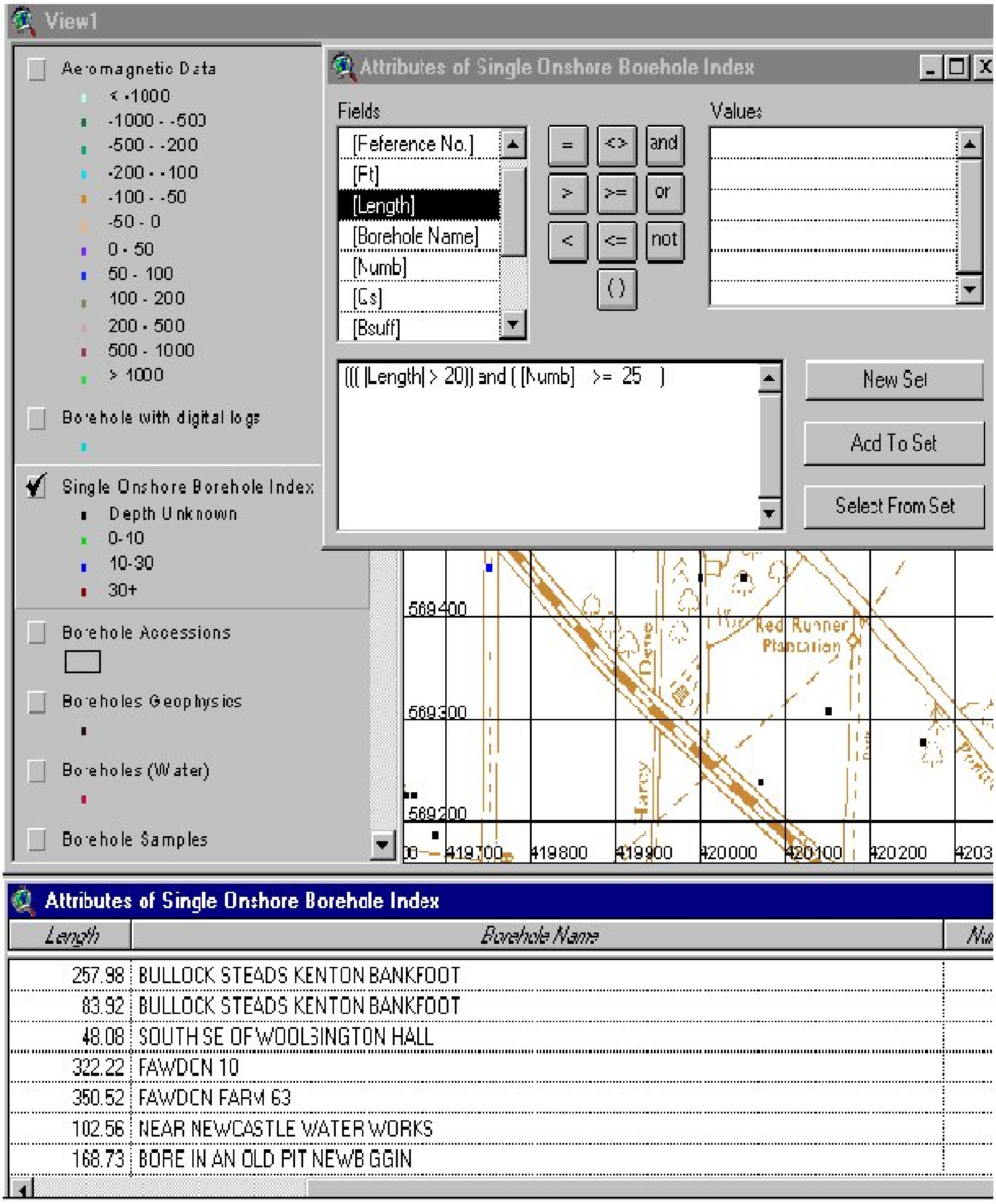


Fig. 3. Some components of the information system. Documents containing various information types are stored in the repository, together with generalized summaries, and metadata which describe the document and define shared vocabulary and standards. Processes to analyze and manipulate the information are shown separately, as are the scientists' activities (see Fig. 1 in part M) which generate and evaluate the documents by investigation of the real world.

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.

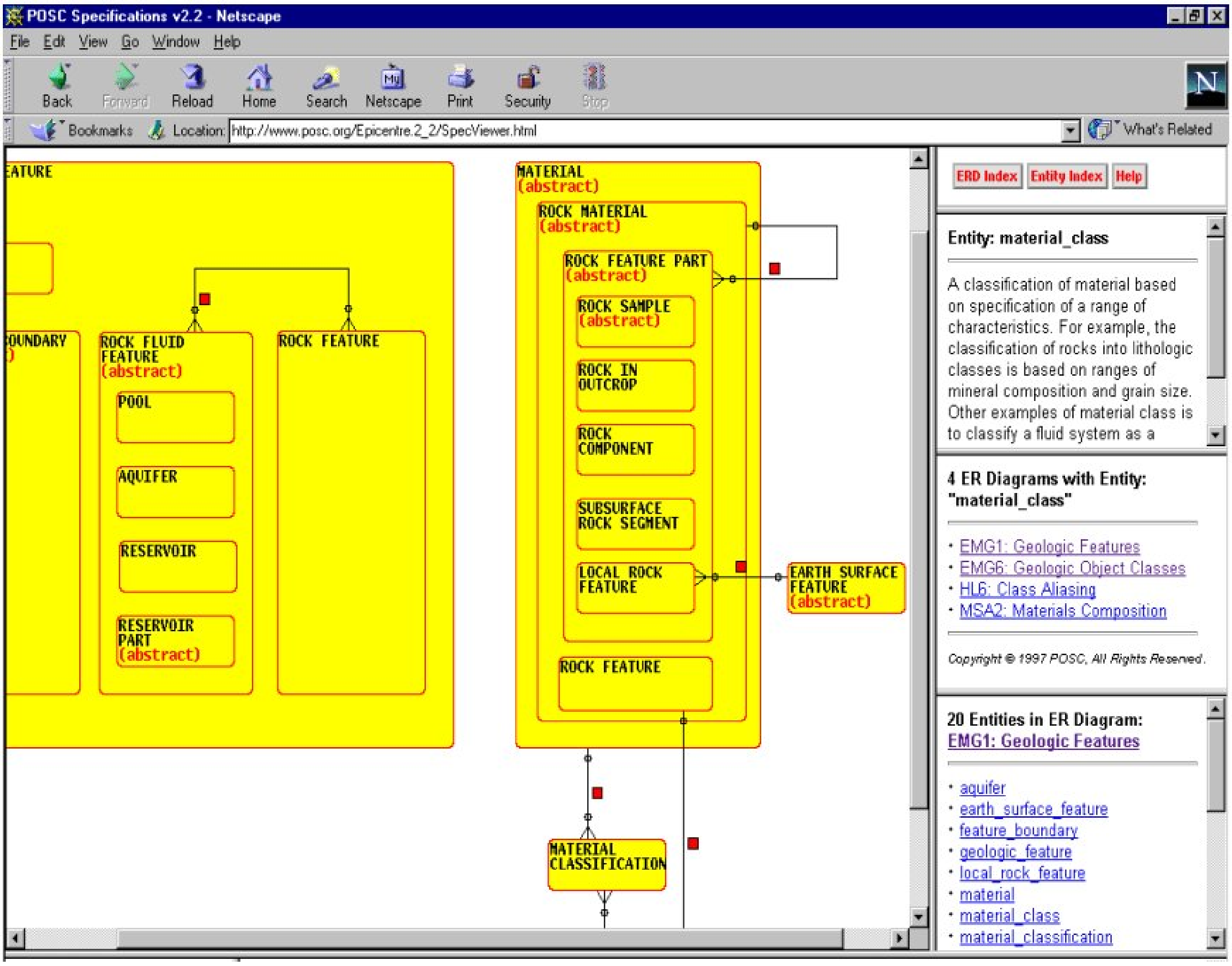


((([Length] > 20)) and ([Numb] >= 25)

Length	Borehole Name	Nu
257.98	BULLOCK STEADS KENTON BANKFOOT	
83.92	BULLOCK STEADS KENTON BANKFOOT	
48.08	SOUTH SE OF WOOL3INGTON HALL	
322.22	FAWDCN 10	
350.52	FAWDCN FARM 63	
102.56	NEAR NEWCASTLE WATER WORKS	
168.73	BORE IN AN OLD PIT NEWB GGIN	

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.

Several software systems may be needed to manage and manipulate the components of a compound document. For example, a document describing a geophysical 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.


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
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
ERD Index
Entity Index
Help

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

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 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 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 retention in archives. Their initial growth may be within a different framework (M 2).

7. References

Albrecht, J., 1999. Geospatial information standards. A comparative study of approaches in the standardisation of geospatial information. Computers & 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. Visualizing Data. Hobart Press, Summit, New Jersey, 360pp.

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.

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 & Geosciences, 25 (10), 1125-1136.

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

7.1 Internet references

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

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. and Guha, R.V., 1998. An MCF tutorial. http://www.textuality.com/mcf/MCF-tutorial.html

Buehler, K., McKee, L. (editors), 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 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

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. 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 1997. http://www.byte.com/art/9707/sec5/art1.htm

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, Netherlands. http://www.ifla.org/II/metadata.htm

International DOI 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 1998. http://www.dlib.org/dlib/july98/07larsen.html

Lynch, C., 1997. Searching the Internet. Scientific American, March 1997. 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 1998, 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 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/

Open GIS, 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, 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, vol 10, no.2, pp 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. and 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

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 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 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/

Web3D Consortium, 1999. The VRML Repository. http://www.web3d.org/vrml/vrml.htm

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 M: Business requirements drive the information system, and provide coherent frameworks>>>