

The evolution of geoscientific metadata

Roderick J. Ryburn¹

The fruits of geoscientists' labours are consigned increasingly to computer files. Although the capacities of electronic media are expanding rapidly, the means of keeping track of all these files is lagging. Knowledge-based organisations like AGSO need the electronic equivalent of libraries to house this information, the analogue of library catalogues to allow us to find critical bits, and the equivalent of librarians to manage the metadata. Files worth keeping must be kept permanently online, referenced by a metadatabase, accessible from the Web, and compliant with changing hardware, software, and data standards.

The way it was

Back in the 'paper epoch' — just 20 years ago, but extending all the way back to the Great Library at Alexandria — scientists strove to record the fruits of their work in papers, diagrams, maps, and books. Apart from verbal reports (lectures, demonstrations, sound and visual recordings, etc), most scientists chose to record their endeavours as paper-based products. It was truly a case of publish or perish! Geoscientists did not differ in this regard, though they tended to produce more maps than most other scientists.

The final resting place for all these paper products was unquestionably a library. Libraries offered the most enduring facility for preserving the scientists' published works and making them accessible to others for the foreseeable future.

As libraries evolved, so too did the means of locating items in their ever-growing collections. Catalogues appeared as the first metadata system for libraries, and later evolved into the card-index systems that most of us still remember. Library catalogues evolved to such an extent that whole university departments were set up to train librarians in the complex subject of library cataloguing. Their evolution continued through the advent of microfilms and microfiche, but it is the rise of computer communications, and the World Wide Web in particular, that has dramatically changed the way we must now think about library catalogues.

The way it is

For the last 20 years, the electronic medium has accounted for the dissemination of an ever-increasing share of scientists' primary outputs. Geoscientists, perhaps to a degree greater than many other scientists, have to contend with a large variety of electronic outputs — mostly in the form of digital computer files. These now include files from wordprocessors, spreadsheets, databases, email, the Web, image and seismic processing, airborne geophysics, radar, marine surveys,

laboratory instruments, GIS, CAD, and specialised 3D mining and petroleum packages.

Some digital geoscientific information now finds its way into corporate databases, which should come with in-built metadata. However, such highly structured data systems will never completely eliminate the need to store less structured static information in files of various sorts. The traditional scientific paper or report will remain a valid output for the foreseeable future, and there will always be a need to represent information that is too specialised to fit into a standardised database system. Although static files may soon be stored in a database management system, rather than in a traditional computer filing system, this does not negate the need for adequate metadata.

The current tendency is to group computer files produced by a project onto a CD-ROM (soon to be overtaken by the DVD, the digital versatile disc). In this form, the information can be stored and catalogued in libraries, as for paper publications. Unlike paper publications, though, there is no guarantee that such CDs will be easily readable in 10 years time, let alone 100 years on. The problem is not so much the longevity of the medium (although this could also be a problem) as the rapidity with which hardware, software, and data structures are currently evolving. Files written by specialised software, such as GIS systems and 3D mining packages, often change their format with software upgrades. Five years on, finding a previous version of a program may be a nuisance. Ten years on it may be virtually impossible, particularly if the operating system has also changed. So, placing files on a CD, or even a DVD, is not the long-term solution it first appears.

The problems we are currently facing are to do with the transition between the old paper-based systems and the new online methods that have yet to become fully established. Libraries are trying to come to grips with the storage of, and access to, digital outputs, but still have some distance to travel before they can claim success. Conventional corporate network systems and disc directories lack the necessary metadata facilities, and files can all too easily be lost in an ephemeral maze of hierarchical computer directories. Right now, there is a significant danger that we may lose much of our current geoscientific output. We need to upgrade our facilities and techniques for handling metadata.

The way it will be

The ultimate solution to the problems outlined above is to ensure that all computer files worth keeping are stored online in conjunction with a corporate metadata system. The metadata system should not only describe the

format and contents of the files but also include pointers to enable the automatic retrieval of files. In storage systems other than small ones, if the metadata do not exist, the file cannot be found, so it might as well not exist. The supply of adequate metadata thus becomes an absolute necessity. The files need not be instantly accessible, but they can reside in automated data warehouses, such as tape 'silos'. Computer object-reference systems — like Open Systems' CORBA (Common Object Request Broker) and Microsoft's DCOM — are now showing the way that digital information can be accessed over the corporate network without the need to know file locations on network discs or tape silo systems. A spin-off is that all information becomes potentially accessible via the Web and thus much more e-commerce friendly.

Only in this way can we be sure that the problem of evolving file formats is easily tackled, and that all information can be kept up to date with respect to current file types, formats, data standards, software, and operating systems. When all files of a certain type and vintage can be automatically identified and located, then it becomes a relatively trivial exercise to extract these files, translate them into the required new format, and replace them in the 'warehouse'. The process can be largely automated, and all information thus kept safe and in currently readable forms. Right now, digital information can routinely be kept more safely than paper information in libraries. With good metadata facilities, we can ensure that the same information becomes immune to the rapid evolution of software, particularly specialised software. The key is good metadata.

Metadata on the web

Mention metadata in connection with the worldwide web — and most people immediately think of search engines, like Yahoo and AltaVista. Wonderful as they are, these search engines are text-based and somewhat hit-and-miss when compared with a properly structured metadata system, such as a library catalogue. Although web access is now virtually mandatory for any comprehensive metadata system, the metadata should reside in a corporate-strength relational database management system. This can now be done fairly easily by providing an attractive web interface (or interfaces) to the corporate metadatabase. Metadata standards and interchange protocols, such as the 'Dublin core' metadata standard and the 'Z39.50' protocol for the interoperability of library catalogues, will eventually lead to transparent access to metadata, irrespective of location or custodianship. Similarly, the Australian Spatial Data Infrastructure (ASDI) initiative

envisages distributed custodianship of spatial databases, with nodes or 'clearing houses' from which the users can query standard metadata in the distributed databases and obtain just the spatial information they require.

The 'AGSO catalog'

In the absence of an off-the-shelf system spanning the full breadth of digital geoscientific information, AGSO has opted to establish its own minimalist system, called the 'AGSO catalog' (Ryburn 1999: *in* AGSO Record 99/24), which is designed to capture metadata on hardcopy output and computer files alike. In addition to providing the outside world with a window into AGSO's outputs, the 'catalog' satisfies our internal requirements for a comprehensive geoscientific metadata system, while tracking the progress and quality of all outputs. A text-based web interface for the products in the 'catalog' can be seen at <http://www.agso.gov.au/databases/catalog/agsocat.html>, but this will soon be joined by a spatially based web interface. Developed from AGSO's former 'Products database', with information on all products sold by the AGSO Sales Centre, the 'catalog' includes metadata on datasets, 'resources', external articles, and products. It also draws

on a much wider community of contributors for additions to its information base. The 'catalog' is seen as the key to future systems of online access to, and distribution and sales of, AGSO's information.

In addition to the 'catalog', AGSO's spatial outputs — e.g., maps, images, and GIS datasets — are in the 'GEOMET' spatial metadatabase, which is on AGSO's web site at <http://www.agso.gov.au/information/structure/isd/database/metadata.html>. This metadatabase closely follows the ANZLIC (Australia New Zealand Land Information Council) guidelines for the transfer of metadata. However, it is generally unsuited to AGSO's non-spatial outputs, and the entry of metadata is too tedious and specialised for general use. The dilemma we face in AGSO is that many of our products bridge the gap between the text-based library world, represented by the 'Dublin core' metadata standard, and the spatial dataset world represented by the ANZLIC guidelines. The 'AGSO catalog' and 'GEOMET' systems go some way towards satisfying both these requirements. Only by making sure that our metadata remain current and well organised in relational databases can we be confident of keeping up with further convergence and evolution in metadata standards.

Conclusions

- Traditional, paper-based methods of handling information are disappearing.
- Geoscientists now produce an enormous range of data and computer-file types.
- Storing away computer files on CDs or DVDs is not a long-term solution.
- Files worth keeping must be stored online or, at least, in an 'active archive'.
- Metadatabases that refer to online files are fast becoming an absolute necessity.
- Corporate metadatabases must have the means to automatically access online files.
- Files must be kept up to date with respect to evolving data standards and software.
- Large organisations must employ professionally qualified metadata custodians.
- Librarians are the traditional metadata experts, but they should expand their roles.
- Corporate metadata must be stored in a relational database with web interfaces.
- The 'AGSO catalog' and 'GEOMET' spatial metadatabases are a good start for AGSO.

¹ Information Management Branch, Australian Geological Survey Organisation, GPO Box 378, Canberra City ACT 2601; tel. +61 2 6249 9605; fax +61 2 6249 9984; email rod.ryburn@agso.gov.au.