The importance of the 'backend' to online delivery of geoscience information

RJ Ryburn

The love affair geoscientists have had with their PCs leads many to think that a do-it-yourself approach can carry us into the 'dotcom' era. However, the secret to the success of major online businesses is their mastery of the 'backend'—the logical, physical and human infrastructure that is the foundation of web sites. These businesses know that in the long run their customers are best served by building a solid backend. Attractive web pages get customers in, but what keeps them returning is the quality, quantity and timeliness of the content behind the web site. Most successful dotcom companies have restructured, or built from the ground up, to provide the best possible backends. Geological surveys must do likewise to survive in the information age.

'Backend' is a term commonly used for the combination of people, hardware, software and data that lies behind corporate information systems. In this article I apply the term particularly to the corporate database component of the backend conglomerate. The 'front end' refers to the part the user sees—the forms, client software or browser pages that are used to access the backend. The front end is like a butterfly: eye catching but with a short life span, soon to be replaced by better, brighter interfaces. The backend, particularly the data and their logical structure, is made of sterner stuff. Good backends take time and effort to construct, but when done properly should survive for decades, outlasting any number of front ends. The hardware and system software aspects of the backend evolve over time, but data and their logical structure should be made to last.

Geological surveys such as AGSO are very much a part of the dotcom scene as we strive to deliver more and more geoscience information via the internet and web.[1] We are presently constrained by narrow communication channels, but these limitations will be largely overcome in just a few years. Although part of AGSO's output will always be in the form of electronic documents, images, GIS datasets and other project-related datasets, enhanced band-widths will increase demand for integrated, seamless national datasets that can be automatically subsetted and delivered online for specified areas of the Australian lithosphere (figure 1). These standardised national datasets replace the map series that were traditionally one of the main outputs of most geological surveys. They will come with added dimensions, greatly enhanced usefulness, and methods of presentation that are limited largely by the imagination.

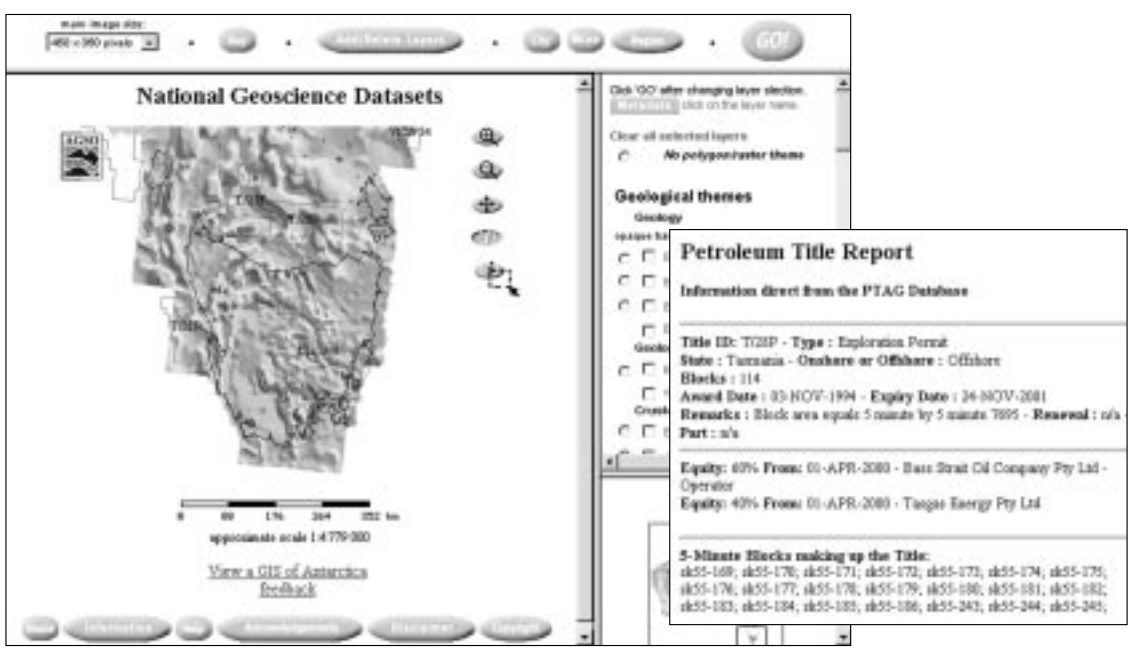

Figure 1. A recent example of integrated online delivery of geoscience information via the web—in this case petroleum exploration titles. The titles layer has been zoomed to the Tasmanian region, and the T/28P title polygon clicked to obtain a report on that title from three tables in the PTAG petroleum accumulations database.
See http://www.agso.gov.au/map/national

Limitations of hierarchical directories

Every PC user is familiar with hierarchical, or tree-structured, computer disk directories. They have been around as long as random-access storage devices—that is, almost as long as the computer itself. Every PC and workstation, and most servers, are organised around tree-structured directory systems. For personal- and project-scale management of information, hierarchical directory systems are still very useful and entirely appropriate.

For large-scale data management, however, tree-structured filing systems are cumbersome, as they permit just one way of classifying and locating files. In corporate computing networks, files can easily be lost in a maze of directories, and file searches are slow and inefficient. Duplication of data is rife, and there is often no way of knowing which version of a file is the most up to date. Other problems concern frequent changes to directory structures, leading to isolation of files relying on pathnames for access. These limitations may not seem serious on a PC, but on the computer networks of large organisations they are major headaches. Similarly, web sites organised around tree-structured filing systems can become difficult to manage once they grow beyond a certain size. Applications become difficult to write, requiring many changes to keep up with fast-changing directory structures.

The evolving backend Relational database management systems (and their forerunners) arose partly to overcome the problems encountered with traditional filing systems. With a relational database the user has many ways of finding data that are efficiently indexed, obviating system-wide file-by-file searches. Similarly, the problem of duplicated data can be largely overcome in a properly designed corporate database system by rejection of inadvertent attempts to re-enter the same information, and clear version control. In reality there are many reasons for using relational database management systems—such as transaction management, business rules, concurrency, security and automatic backup.

Corporate databases have existed almost as long as the computer itself, and as relational databases for the past 20 years. Corporate databases, albeit in a simpler form, were certainly to the fore in the days of corporate mainframe computers with dumb terminal access.

With the rise of the minicomputer in the seventies, and the PC in the eighties, databases were able to migrate to the department, the project and the person. At the same time the idea of distributed computing gained currency, and people liked the relative freedom of being able to do their own thing on their own machine (if at some cost).

The advent of the web in the early nineties gave rise to 'thin-client' computing in which the web browser became the graphical equivalent of the old character terminal, and the bulk of the computing power was put back into the server. Suddenly there was the potential for thousands of terminals anywhere in the world to be connected to your server at little cost.

You could upgrade your system software and applications in the one place, without the need to upgrade all those client systems. The efficiencies to be gained were astounding—the backend was back with a vengeance. The net result has been the reunification of the corporate backend into one integrated system, albeit spread over a number of CPUs. Oracle Corporation, fed up with the inefficiencies of dispersed systems, recently consolidated thousands of their business servers spread around the world into what is effectively one corporate backend at their headquarters in Redwood Shores, California.[2] In so doing they claim they will save a billion dollars.

The rise of object-relational databases

In the early nineties there were predictions that object-oriented databases would take over completely from relational databases. This has not happened, and pure object database systems now form only a small, specialised part of the current database scene. What has happened, though, is that the main relational database systems have gradually taken on some of the more useful qualities of the object-oriented world. A variety of different object types may now be conveniently handled in what are known as universal, or object-relational database management systems. For example, Oracle has acquired the technology for handling images in the database, to the extent that it can now deal with commands like 'get me all pictures that look a bit like this one'—and that is without recourse to textual metadata. In AGSO's case, the ability to handle GIS datasets in Oracle 8i (with or without ESRI's Spatial Data Engine) and to store documents in the database will undoubtedly prove to be very valuable. Images, too, will be increasingly placed inside the database.

At the same time the database is expanding to accommodate a greater variety of data types, the application software that allows one to easily handle these new forms of data is maturing. With Oracle's latest application development tools, one can rapidly construct a web site that includes text, images, documents, video, sound and 'XML' extracts—all from within the corporate database. Once the corporate database includes all these new forms of information in a well-organised and accessible manner, constructing an attractive, highly functional web site becomes an easy and relatively quick exercise. Furthermore, multimedia front ends such as these are now easily modified or expanded to meet client expectations. 'Portals' for use on other web sites can also be projected.

Burgeoning database administration

A consequence of the rapidly expanding backend is that database administration (DBA) is a more complex and critical job than it was a few years ago. The software needed to place corporate information on the web has grown enormously in volume and sophistication, and the variety of information types being stored in the backend for presentation on the web is increasing in leaps and bounds. New products, such as Oracle's WebDB 3.0, allow the contents of the entire web site to be brought within the database. Control of content can then be distributed to the appropriate areas of the organisation with security handled by the database management system. GIS datasets are also moving into the database, as in ESRI's Spatial Data Engine now used in AGSO. We will soon be required to use a document management system, in which all documents and communications are stored within the corporate database system and accessed via the intranet and web. Information can now be supplied via the web encapsulated in the XML data definition language.

Efficient management of a corporate web site requires people with at least three types of expertise. First you need people who are competent with hardware, operating systems, communications,

➥ *Continued page 20*

Continued from page 11

networking, computer security and web servers. You then need good webmasters, database administrators, and application developers. Finally you need people to supply the content—the information to go in the databases and on the web. In organisations of any size it is impossible for one person to be across all these skills; a team is generally needed to run web sites of dotcom organisations. Increasingly, though, it is the database administrator who occupies the pivotal position, and has the most demanding job. In organisations that float on seas of information, the database administrator is the pilot. One must look after the DBA role carefully, or risk disorganisation.

Web delivery of geoscience data

In organisations like AGSO there is a clear distinction to be made between standard national databases and the more ad-hoc datasets and documents produced particularly by research projects. However, both types of information have to be adequately managed as we move towards an era in which online delivery of geoscience information is the norm. Both information types demand that attention be given to the backend part of the total information management environment.

In regard to ad-hoc types of information, I include all traditional documents, papers and publications, most of which will soon be routinely available on AGSO's intranet and/or web site. Some may be handled by a document management system, such as 'TRIM', that can use the Oracle 8i database management system as its backend. Alternatively, Oracle's WebDB 3.0 may be used to place many documents on the intranet and web. Either way, vast amounts of unstructured written material will find their way into the backend database management system in the near future. It is also likely that processed images and project GIS datasets will eventually migrate into the corporate database system, rather than, as at present, be kept in computer files on 'corporate' UNIX disk systems. Almost all datasets that projects currently publish on CD-ROM will probably end up in the corporate database system for online delivery.[3] The demands on the backend will be heavy indeed. The infrastructure to handle the metadata for ad-hoc information is already at hand as the AGSO Catalog.[4]

The formally structured national geoscience databases that have been an integral part of AGSO's information store for more than 20 years are being made available on the web in user-specified chunks. In one mode, the user is able to construct a map of the area in which they are interested by selecting required GIS layers. They should then be able to obtain all the spatially related attribute data that AGSO has in the backend databases (as in figure 1). All laboratory data should be made available, and in the case of geological field data there are thoughts of capturing videos of critical outcrops, in which geologists explain the relationships.

Another mode of use is where other web sites transparently access AGSO data via portals, as if it were their own. AGSO is a participant in the Australian Spatial Data Initiative which has already demonstrated distributed web mapping, with data drawn online from dispersed agencies. The web enables many different data types to be integrated from many sources, but the components are best kept in well-controlled backend databases.

Restructuring for online delivery

One of the main lessons to be learned as we enter the dotcom era is that we must organise properly for online delivery of geoscience data. The task should not be left to individuals scattered among disparate projects. Individual projects can, and sometimes do, set up web interfaces for online delivery of certain types of data, and some of these interfaces are effective. However, such practices 'reinvent the wheel', and time and money are wasted on systems that cannot be readily integrated with other online systems. The web site can easily become an expensive mishmash of disparate systems. We have already experienced instances of this in AGSO.

Information management and delivery is too important a core function in today's geological surveys to relegate to dispersed groups within the organisation. In the current climate it is axiomatic for a dotcom organisation that online information delivery be managed by a well-integrated, multidisciplinary team with good communication skills. Furthermore, this team should act as an ongoing service group, rather than a project. The team must include a strong backend group, specifically those with database administration and development skills. In time, when the systems and procedures for online delivery become established, responsibility can be more devolved. As procedures become routine, the team size can be reduced and many members redeployed into other areas.

There is little doubt in my mind that geological surveys should consider setting up formal information divisions (if they have not already done so) to properly handle the difficult new paradigm of online geoscience information delivery.

Conclusions

- Successful online delivery of geoscience information requires good backend databases.
- Web sites that rely exclusively on hierarchical directory systems do not scale well.
- Object-relational databases, such as Oracle 8i, are built to handle all sorts of information.
- Database administration has recently become a much more complex and critical job.
- Online delivery requires a focused approach to data acquisition and database design.
- Both structured and unstructured types of information must be catered for on the web.
- Geological surveys may need to restructure to better support online information delivery.

References

1. AGSO. 2000. Australian Geological Survey Organisation Online Action Plan. http://www.agso.gov.au/information/online_action_plan.html.
2. Karpain G & Myers L. 2000. Keep it simple: How Oracle consolidated its global infrastructure into a centralised e-business architecture. Oracle Corporation white paper, Sept 2000. http://www.oracle.com/applications/story.html.
3. AGSO. 2000. Australian Geological Survey Organisation 2000–2001 work program. Canberra: AGSO; unpublished.
4. Ryburn RJ, Ross SI & Wijatkowska-Asfaw BS. 2000. Guide to the AGSO Catalog. Canberra: AGSO, record 2000/13.

Dr Roderick Ryburn was Leader of AGSO's Corporate Database Group when this article was written. He can be contacted by phoning +61 2 6291 7362 or e-mailing rryburn@webone.com.au