A Universal Geochemical Survey Data Model

Adcock, S.W., Grunsky, E., Laframboise, R. and Spirito, W.A.

Geological Survey of Canada, Ottawa, ON, CANADA

Abstract

The Geological Survey of Canada has amassed a large quantity of geochemical data over the past 40 years. These data have been collected from lakes, streams, soils, glacial sediments, plants, bedrock, etc. The field samples were analysed by a wide range of dissolution, and total recovery analytical methods following a variety of sample preparation procedures. The resulting data were managed and catalogued in an ad hoc manner; often without the metadata that provide the necessary context for interpretation and data integration requirements. Over time, this resulted in confusion and the loss of essential information for effective geochemical data interpretation and archival purposes.

A strategy was adopted that focused on the development of a single corporate data structure. This model is based on a conventional relational database model system (RDBMS). The core of the model consists of 11 tables which contain entities that are clearly recognizable by a geochemist (surveys, sites, samples, etc.). The core tables are supported by about 60 peripheral tables. These tables, labeled as "SHARED" and "DERIVED" form the practical basis of the model. Shared tables contain a unique list of identifiers for sample type, method of preparation, size fraction, and other characteristic properties for any given survey. Derived tables contain "flat file" views of the data that are used by other processes, such as internet map servers, which require information for viewing geochemical data in a geographic context as well as simple views of the analytical results for visual browsing by geochemists.

The model is being adopted by several other agencies across Canada, as part of the Canadian Geoscience Knowledge Network (CGKN). By coordinating activities across these agencies, to ensure that information in the SHARED tables is kept internally consistent, we are able to allow end users to simultaneously query the data holdings of the different agencies. Software is being developed that will allow easy input and output, using XML and GML.

Introduction

Beginning with the pioneering work of R.W. Boyle in the 1950s (Boyle, 1967; Boyle and Garrett, 1970; Brummer et al. 1987), the Geological Survey of Canada (GSC) has amassed a huge quantity of geochemical survey data. For the purpose of this discussion, a geochemical survey can be defined as an activity carried out over a particular geographic area, over a particular time period, collecting a particular set of sample media.

Geochemical surveys are variable in their spatial extent and sampling strategies. Survey extents can be local, limited to sampling an area of less than 1 km^2 or they can be very large, covering several thousand to several hundred thousand km^2. Sampling strategies can also vary between fixed grid to random, based on the sample design strategy and the availability of sample material. Typically, a survey covers a few hundred square kilometres, takes about 2-4 weeks to complete, and collects between one and five different sample types. The total number of samples collected is usually a few hundred.

Chemical analyses may be carried out in the field, but usually the samples are subjected to some laboratory-based physical sample preparation process, before they are chemically analysed. Surplus sample material may be archived and re-analysed at later dates. Digital data management is undertaken on a survey-by-survey basis, and is the responsibility of the project leaders. The data are most commonly managed using spreadsheet software and more recently using relational databases. However, there are no corporate "best practice" guidelines for how to manage the data.

The GSC's single most important geochemical dataset is the "National Geochemical Reconnaissance" (NGR), which is an ongoing activity that began in the mid 1970s (Friske and Hornbrook, 1991). The dataset is comprised of about 200 distinct surveys. To date, approximately 200,000 stream and lake sediment, and water samples have been collected from across the country. Sampling and analytical procedures within this dataset have stayed remarkably consistent over the years. Yet, in addition to this large and very homogeneous dataset, there is a tremendous diversity of smaller datasets, spanning the whole range of sample media.

Beginning in the 1960s, the GSC has made several attempts to systematically archive the digital data derived from these surveys. With the notable exception of the NGR, none of these initiatives attained any longevity, and there was no corporate process to ensure that the information was properly catalogued and archived. Whilst the NGR data are relatively secure, a significant amount of the GSC's legacy geochemical data are at considerable risk of being lost.

Efforts to remedy this situation began in the late 1990s through:

(a) ensuring that new data are properly managed as they are acquired;

(b) ensuring that as much legacy data as possible are rescued; and

(c) enabling access to the data over the Internet, focussing on geospatial queries.

A significant amount of geochemical data has been stored in spreadsheet form. This practice has the following problems –

(a) There is no consistency in the spreadsheet format. It requires human intelligence to interpret the contents of the spreadsheet. Automated processes to manipulate and analyse the contents of the spreadsheets would be very difficult to create and maintain;

(b) The metadata associated with the survey is not usually stored in the spreadsheet. The metadata may be scattered through several paper reports. Some metadata may never be recorded; and

(c) There is a risk that inadvertent manipulation of columns or rows can destroy the integrity of the data.

Earlier work by Adcock and Laframboise (Friske et al., 1991; Dunn et al., 1992) had convinced us of the feasibility of creating a generic data model that would be capable of holding the vast majority of the GSC's geochemical data. A working group was set up within the GSC, with participation from Provincial Geological Surveys via the "Canadian Geoscience Knowledge Network" (CGKN) initiative, to define the data model. The model, based on a relational database structure, has gone through several iterations, and continues to evolve steadily. However, the core features of the model have remained stable.

The advantages of a single data model compared to the spreadsheet alternative are obvious from a corporate perspective, managing many datasets from many scientists. They are less obvious from the perspective of an individual scientist, working on a single dataset. Part of our challenge is to "sell" the data model to GSC scientists.

Data Model Framework

The data model was designed using object-role modelling (ORM) techniques (Halpin, 2001). Additionally, the business process of conducting a geochemical survey was defined. The two approaches complement and reinforce each other. The business process can be defined (with minor simplifications for clarity) as follows, with ORM objects identified by ***bold italics***:

1. A ***principal investigator*** initiates a ***project***;
2. As part of the project, a geochemical ***survey*** is carried out;
3. In the course of the survey, a number of ***sites*** are visited;

4. At each site, ***samples*** are collected;

5. ***Laboratory samples*** are derived by physical processing of the material collected in the field;

6. Portions of the prepared material are sent to one or more analytical laboratories (***analytical laboratory samples***). These samples are bundled together, to form ***analytical sample bundles***. Control Reference and blind (analytical) duplicate samples (for quality assurance) may be added to the sample bundles;

7. The analytical laboratory performs the requested ***analyses***; and

8. The results of the project appear in various ***publications***.

These eight steps capture the essential relationships between the core objects in the data model. The full physical implementation of the data model encompasses over 60 tables, but the core of the model is very simple.

Multi-agency interoperability

Geochemical data are not managed by a single group within the GSC. The data are highly decentralised. Any universal data model must take into account the diverse needs and capabilities of these groups. Additionally, we wish to encourage other agencies to adopt the model. The generic data model briefly outlined above addresses the fundamental conceptual issues in creating a universal data model, but there are also several pragmatic issues that need to be addressed.

The physical implementation of the data model aims to be independent of any particular hardware or software platform. It has been tested against multiple versions of Oracle™, SQL Server™ and Access™. It achieves this independence primarily by relying on a reduced set of fundamental data types, which are derived directly from the data types supported by the XML Schema Definition (XSD) Language. A side-effect of this RDBMS independence is that the physical implementation cannot rely on stored procedures within the database.

As a pan-Canadian initiative, it is essential that the database be accessible in both English and French. The database does this by avoiding any unnecessary reliance on language-specific fields. Wherever textual information is recorded in the database, there are two fields defined – one for each language. This approach is inherently multilingual. It would be straightforward to add support for any language which can be represented by the UNICODE character set.

In itself, adoption of a common data model should avoid a lot of resources being dedicated to the same task by different agencies, resulting in several very similar, but not quite identical, data models. But we can extend the power of a universal data model by going an extra step, and ensuring that everybody who implements the model also uses the same science language.

Science language issues

If the science language is coordinated across multiple databases, which are all using the same data model, then it is possible to issue the same query against different databases, and expect consistent results. For example, the databases could be simultaneously queried for all Au-INAA data obtained from stream sediments. This implies that all the databases must be consistent in their use of the terms “Au”, “INAA” and “stream sediment”.

The data model enforces this consistency by defining a set of “SHARED” tables, whose contents are identical on every node in the “distributed database”. These SHARED tables make up over half of the tables in the model. Most of the SHARED tables are dedicated to identifying the characteristics of the analytical methodologies (concentration units, instrumentation, etc.). Populating these SHARED tables is not always easy. The model requires classification systems for sample media and analytical

methodologies. In both cases, we have not found any existing classification schemes that we could build on. Instead, we are building schemes from scratch, capitalising on scientific expertise within the GSC.

A single portal for Canadian geochemical data

If the model is generally adopted by geoscience agencies across Canada, then there will be major benefits for all participants.

Software developed by one agency will be usable without modification by any other agencies. Work is currently underway at the GSC to develop generic data loading software to facilitate the loading of legacy data into the model. This software, which is being built as a web service, should be useful to all participants.

The data model is being integrated into "Laboratory Information Management Systems" at the GSC, using business process software to keep track of individual samples, as they are received from the field, and sent out to commercial laboratories for chemical analysis. Again, we are developing software at the GSC to ensure that data are immediately stored in the database as they are acquired, with QA/QC checks wherever appropriate.

A WWW query tool could be built that will simultaneously query all databases conforming to the model. Adding additional databases to the query interface would require no additional software to be written. A prototype of such a tool has already been created, and work is in progress to update it, for compatibility with Geography Markup Language (GML) and other Open GIS Consortium (OGC) initiatives.

Security and performance issues

Exposing databases to the Internet raises many security issues. Increasingly, geoscience agencies are hiding their primary databases behind corporate firewalls, and exposing only "DERIVED" subsets of the databases outside the firewall. The data model defines a set of denormalised tables which present a simplified view of the data, suitable for access by WWW-based query tools. An additional benefit of creating these denormalised tables is that individual queries can target just one table, rather than multiple normalised tables. This results in much faster query times. The derived tables can be updated on either a fixed schedule (perhaps daily), or whenever the original tables are updated.

Support for metadata standards

One of the derived tables has been designed to be directly mappable to the FGDC GEO profile for metadata. Each survey that is stored within the database is represented by a single record in this derived metadata table. It is straightforward to expose this table to a Z39.50 server, thus making the database's contents easily accessible to metadata search engines. Unfortunately, there is no metadata profile that is well-suited to geochemical data, which limits the usefulness of the service.

Extensibility

Although the core of the model handles the large majority of the data to be stored in the database, it does not address certain needs. Scientists carrying out individual surveys often record observations that are highly specific to that survey. The observations may apply to the survey as a whole, to individual sites, or to individual samples. Therefore, the model explicitly allows for additional tables to be created which are specific to a particular survey, or group of surveys. For example, a survey that collects lake sediments may record the depth of the lake. Clearly, this observation would be meaningless for a survey that collected balsam fir twigs. The number of survey-, site-, or sample-specific observations that have been made in the past is huge. Attempting to account for all such observations in the model would be hopeless.

In database parlance, the model allows a developer to implement local subtyping for the survey, site and sample tables, by “separation” (Halpin, 2001, p.426).

Additional information

A web site is being maintained for project participants at http://geochem.cgkn.net. The site is password-protected, but guest logins are permitted. Amongst the resources available at the site are a schema diagram for the entire model, and an Access database populated with a test dataset of NGR lake sediment and water samples. The site contains links to the myriad computer technologies mentioned briefly in this paper.

References

Boyle, R.W. (1967): Geochemical prospecting - retrospect and prospect. Transactions of the Canadian Institute of Mining and Metallurgy, vol. 70, pp. 1-6.

Boyle, R.W. and Garrett, R.G. (1970): Geochemical prospecting, its present status and future. Earth Science Reviews, vol. 6, no. 1, pp. 51-75.

Brummer, J.J., Gleeson, C.F. and Hansuld, J.A. (1987): A Historical Perspective of Exploration Geochemistry in Canada - The First 30Years. Journal of Geochemical Exploration, vol. 28, pp. 1-39.

Dunn, C.E., Adcock, S.W. and Spirito, W.A. (1992). Geochemical Surveys in Nova Scotia Using Spruce Bark. *In* Program and Summaries, Sixteenth Annual Open House and Review of Activities; D.R. MacDonald and K.A. Mills, Editors; Nova Scotia Department of Natural Resources, Mines and Energy Branches Report 92-4, p.59.

Friske, P.W.B., Adcock, S.W. and McCurdy, M.W. (1991). Development of a Digital Data Base for Canada’s National Geochemical Reconnaissance Geochemical Data: Progress, Applications and Outlook. Geological Survey of Canada Current Activities Forum 22-23 January, 1991, Program with Abstracts, p.8.

Friske P.W.B. and Hornbrook E.H.W. (1991). Canada's National Geochemical Reconnaissance Programme; Transactions of the Institution of Mining and Metallurgy, Section B, vol 100, pp .47-56.

Halpin, T.E. (2001) Information modelling and relational databases. Morgan Kaufman 763 p.