Presentation of regional geochemical data via Internet Earth browsers

Stephen W. Adcock[1], Wendy A. Spirito[1], & Eric C. Grunsky[1]
[1]*Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario K1A 0E8, Canada (e-mail: adcock@nrcan.gc.ca)*

ABSTRACT: The Geological Survey of Canada (GSC) has large amounts of regional geochemical data which are under-utilised both within and outside the organisation. The situation is being rectified by three overlapping activities: (1) cataloguing of the data holdings, to international metadata standards; (2) downloading of the data in a standardised spreadsheet format; (3) simple visualisation of the analytical data using Internet Earth browsers. Regional geochemical surveys have undergone a continuous evolution in sampling and analytical procedures. The resultant multitude of sample media and analytical methods poses several data management challenges, which are being mitigated by developing XML-based data transformation procedures. The GSC geochemical data management system has provisions for detailed metadata describing every aspect of a sample's history. In order to avoid inappropriate merging of disparate data, the analytical data are separated into many distinct datasets. The recent emergence of free, XML-based Earth browser software makes it much easier to visualise and compare the different datasets, without resorting to sophisticated GIS software.

KEYWORDS: *regional geochemistry, geochemical mapping, web mapping, Canada, New Brunswick*

INTRODUCTION

Many government geoscience organisations have accumulated large quantities of regional geochemical data over the past half-century. Systematic management of the data has posed many problems, due primarily to the continuous evolution in sampling and analytical procedures. Regional geochemical data are typically presented in a spreadsheet format, with each row representing an analysed sample, and each column representing an analytical measurement. This approach works well for individual surveys, but becomes unmanageable for a large collection of disparate surveys. Hence, most organisations have opted to store the data in relational database management systems, and to "normalise" the data to varying degrees. Whilst this makes centralised management of the data much easier, it puts a major obstacle in the way of end-users easily accessing the data.

In the past, accessing these data has usually involved the creation of complex, specialised computer programs, either in-house or via external contracts. Both approaches are expensive and hard to sustain. The recent emergence of XML-based technologies, coupled with the decreasing costs of computers makes it possible to create much simpler data extraction and manipulation procedures.

This paper outlines one approach that is being developed at the Geological Survey of Canada (GSC), where data are extracted from the database and re-cast as KML files that can be displayed in Google Earth or any other KML-aware application. Examples are presented for a recently-completed compilation of till geochemical data from New Brunswick (Adcock *et al.* 2009).

MANAGEMENT AND PUBLICATION OF REGIONAL GEOCHEMICAL DATA AT THE GSC

Geochemical surveys at the GSC are generally conceived and executed as individual entities, optimised for the local geography and for the scientific objectives of the funding project. The only major exception to this approach is the National Geochemical Reconnaissance (NGR) program, which has been collecting and analysing lake and stream sediment samples across Canada since the early 1970s according to a well-defined protocol (Friske & Hornbrook 1991).

Individual surveys are usually published as GSC Open Files, which can be downloaded over the Internet at no cost. Most Open Files include maps of the analytical data, and raw data in the form of a spreadsheet. The maps are typically presented using proportional dot symbology (Bjorklund & Gustavsson, 1987), but other approaches, such as contouring, are sometimes used.

The exact format of each Open File is left to the discretion of the authors. This has resulted in a tremendous variability in content and format over the years, which in turn makes it frustrating and time-consuming for end-users to incorporate the data into their own projects. [For a comprehensive listing of GSC geochemical Open Files, see http://gdr.nrcan.gc.ca/geochem/metadata_pub_gsc_e.php] Even within the GSC, it is hard to get an accurate idea of what data are available. A further complication arises from the close working relationships between GSC scientists and their colleagues in Provincial agencies (projects are co-managed, samples are shared, publications emanate from different organisations).

The unsatisfactory nature of the current situation has been recognised for many years, but efforts to rectify it have encountered many technical challenges which could not be overcome with the available personnel and financial resources. Recent advances in computing hardware and software have led to new approaches and the development of cost-effective solutions.

A long-term project commenced in 2004 to catalogue GSC and Provincial geochemical surveys. The catalogue conforms to 'Federal Geographic Data Committee' metadata standards (FGDC, 1998) and is publicly accessible over the Internet (Spirito & Adcock 2009). Attention is now focussing on managing the analytical data. Several demonstration projects have been completed as a test of the data management system, most notably two collections of till geochemistry (Adcock 2009a; Adcock *et al.* 2009).

DEVELOPMENT OF A STANDARDISED DATA MODEL FOR REGIONAL GEOCHEMICAL SURVEYS

Any generalised approach to managing regional geochemical data must be built on top of a well-designed data model. If the model is well-designed, it should be able to adapt easily to ongoing changes in how geochemical surveys are carried out. Poorly constructed models will be harder to adapt, and will become steadily more difficult to maintain. Prior to the late 1980s, geochemical data modelling at the GSC was not consciously undertaken. The emergence of desktop relational database software in the late 1980s led to a serious effort to construct data models, based on methodologies designed specifically for use with relational database software (Halpin, 1995). The current data model (Adcock, 2009b) has been used successfully in a variety of projects. The most complex of these projects involved the compilation of till geochemical data from across New Brunswick (grouped into 39 surveys, collected and analysed between 1985 and 2006). The final database contains data for 13846 samples, collected from 11841 distinct sites. The database, in MS Access 2003 format, is available as a DVD (or free download) from the GSC or NBDNR (Adcock *et al.* 2009).

The data model is highly normalised, which suits data integrity and manipulation via the SQL language, but creates difficulties for the end-user in terms of accessing the data. Straightforward queries may involve numerous tables which have to be joined together. As a practical work-around, the database includes several very large denormalized tables.

XML-BASED TECHNOLOGIES

Extensible Markup Language (XML), in its various forms, is rapidly emerging as the dominant data format across all computer systems. Coupled with Unicode (UTF-8), it promises to greatly reduce the problems of interoperability between operating systems and software packages. XHTML is being adopted by all large organisations as the preferred “dialect” of HTML, for use on web sites (W3C 2006). Office productivity software developers are switching to XML as the default file format for word processing and spreadsheets (ISO, 2006; ISO, 2008). KML is becoming a very popular XML syntax for geographic data, partly because of the popularity of Google Maps and Google Earth, but also because it is a relatively easy file format for programmers to work with (in contrast to the SHAPE file format, for example (ESRI 1998)). KML was recently endorsed by the Open Geospatial Consortium as a recognised standard for geographic data visualisation (OGC 2008).

The different dialects of XML (XHTML, KML, …) are constrained by XML schemas (W3C, 2004). These schemas are critical to the success of XML. They are used to ensure that an XML file adheres to a well-defined structure. Schemas are themselves XML files, which must conform to the XSD specification. Schema designers are free to develop constraints to varying degrees. Forcing an XML file to be compatible with a tightly-constrained schema frees developers from having to write their own data validation procedures. This leads to a great simplification of data manipulation software.

XML files can be manipulated programmatically by a number of different techniques. One of the most powerful and elegant techniques involves yet another XML-based technology – the XSLT programming language (W3C 2007). An XSLT program is itself an XML file. XSLT was designed for the express purpose of transforming XML files into alternative formats. The language includes many constructs to allow efficient restructuring of the data. This leads to programs which are much smaller and easier to write than equivalent programs in more generic languages.

XSD and XSLT working together provide a very powerful technique to take output from one software package and transform it through intermediate formats, each of which is constrained by its own XSD schema, into a format suitable for end-user display.

XML TRANSFORMATION IN PRACTICE

All modern relational databases include the ability to export tables as XML files. It is usually possible to apply an XSLT transformation to the data as part of the export procedure. In the interest of simplicity and compatibility across different databases, no special transformation was applied to the tables extracted from the New Brunswick till database. Therefore, after exporting data out of MS Access in a generic XML format, the first XSLT transformation involves restructuring the data to conform to a “Geochemical Survey” XML schema, developed at the GSC (Adcock 2009b). The second transformation produces a set of files which conform to the GML schema (OGC, 2007). KML shares many features with GML, and hence the third and final GML-to-KML transformation is very simple.

Separating the data transformation into three distinct steps enforces a completely modular software design. In practice, the data transformation is executed via command shell scripts, using freely available software for both the XSLT transformation and XSD validation. The raw data contained in the 39 surveys in the New Brunswick compilation are exported into 7,000 individual KML files, which can be viewed online at http://gdr.nrcan.gc.ca/geochem.

REFERENCES

ADCOCK, S.W. 2009a. The Canadian Database of Geochemical Surveys: Analytical data for till surveys carried out across Canada by GSC staff. Geological Survey of Canada Open File 5935, 1 DVD.

ADCOCK, S.W. 2009b. The Canadian Database of Geochemical Surveys: Technical documentation for the data model. Geological Survey of Canada Open File 5938, 1 CD-ROM.

ADCOCK, S.W., ALLARD, S., PARKHILL, M.A., PRONK, A.G., & SEAMAN, A.A. 2009. The Canadian Database of Geochemical Surveys: Analytical data for till surveys carried out in New Brunswick by NBDNR and GSC staff, 1985-2006. Geological Survey of Canada Open File 5937, 1 DVD.

BJORKLUND, A. & GUSTAVSSON, N. 1987. Visualization of geochemical data on maps: New options. Journal of Geochemical Exploration, **29**, 89-103.

ESRI 1998. ESRI Shapefile Technical Description. An ESRI White Paper—July 1998. Environmental Systems Research Institute, Redlands, California

FGDC 1998. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.

FRISKE, P.W.B. & HORNBROOK, E.H.W. 1991. Canada's National Geochemical Reconnaissance Programme. Transactions of the Institution of Mining and Metallurgy, Section B, **100**, 47-56.

HALPIN, T. 1995. Conceptual Schema and Relational Database Design. 2nd edition. Prentice Hall, 547pp

ISO 2006. ISO/IEC 26300:2006 Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0

ISO 2008. ISO/IEC 29500-1:2008 Information technology -- Document description and processing languages -- Office Open XML File Formats -- Part 1: Fundamentals and Markup Language Reference

OGC 2007. OGC 07-036 : OpenGIS Geography Markup Language (GML) Encoding Standard. Version: 3.2.1. Open Geospatial Consortium Inc.

OGC 2008 OGC 07-147r2: OGC KML Version: 2.2.0. Open Geospatial Consortium Inc.

SPIRITO, W.A. & ADCOCK, S.W. 2009. The Canadian Database of Geochemical Surveys: Metadata for 600 geochemical surveys across Canada. Geological Survey of Canada Open File 5934, 1 CD-ROM.

W3C 2004. XML Schema Part 0: Primer Second Edition W3C Recommendation 28 October 2004

W3C 2006. XHTML 2.0 W3C Working Draft, 26 July, 2006.

W3C 2007. XSL Transformations (XSLT) Version 2.0 W3C Recommendation 23 January 2007