

# Semantic Mediation in the National Geologic Map Database (US)

David Percy<sup>1</sup>, Stephen Richard<sup>2</sup>, David Soller<sup>3</sup>

<sup>1</sup>Portland State University, <sup>2</sup>Arizona Geologic Survey, <sup>3</sup>US Geologic Survey  
<sup>1</sup>Portland State University, Department of Geology, 1721 SW Broadway, Portland, OR 97201

<sup>2</sup>Arizona Geological Survey, 416 W. Congress #100, Tucson, AZ 85701

<sup>3</sup>U.S. Geological Survey, 926-A National Center, Reston, VA 20192

[percycd@pdx.edu](mailto:percycd@pdx.edu)

## Abstract

Controlled language is the primary challenge in merging heterogeneous databases of geologic information. Each agency or organization produces databases with different schema, and different terminology for describing the objects within. In order to make some progress toward merging these databases using current technology, we have developed software and a workflow that allows for the "manual semantic mediation" of these geologic map databases. Enthusiastic support from many state agencies (stakeholders and data stewards) has shown that the community supports this approach. Future implementations will move toward a more Artificial Intelligence-based approach, using expert-systems or knowledge-bases to process data based on the training sets we have developed manually.

## Background

The National Geologic Map Database (NGMDB) "phase three" prototype is a combined effort of the AASG and USGS to harmonize data sets from differing sources and serve them in a format that can be consumed by web browsers or web services. It builds substantially on the project's 10-year effort to develop standards (e.g., contributions to development of the North American Data Model, or NADM) for a common data structure and controlled science terminology for geologic map and science data (SLTT). In recent years, this effort has been incorporated into an international standard mediated by the IUGS Commission for the Management and Application of Geoscience Information (CGI) with has produced the GeoSciML standard.

GeoSciML is a transport mechanism and schema developed under the auspices of the CGI and demonstrated successfully in late 2006 at the International Association for Mathematical Geology conference in Liege, Belgium. In that demonstration, geologic databases from world-

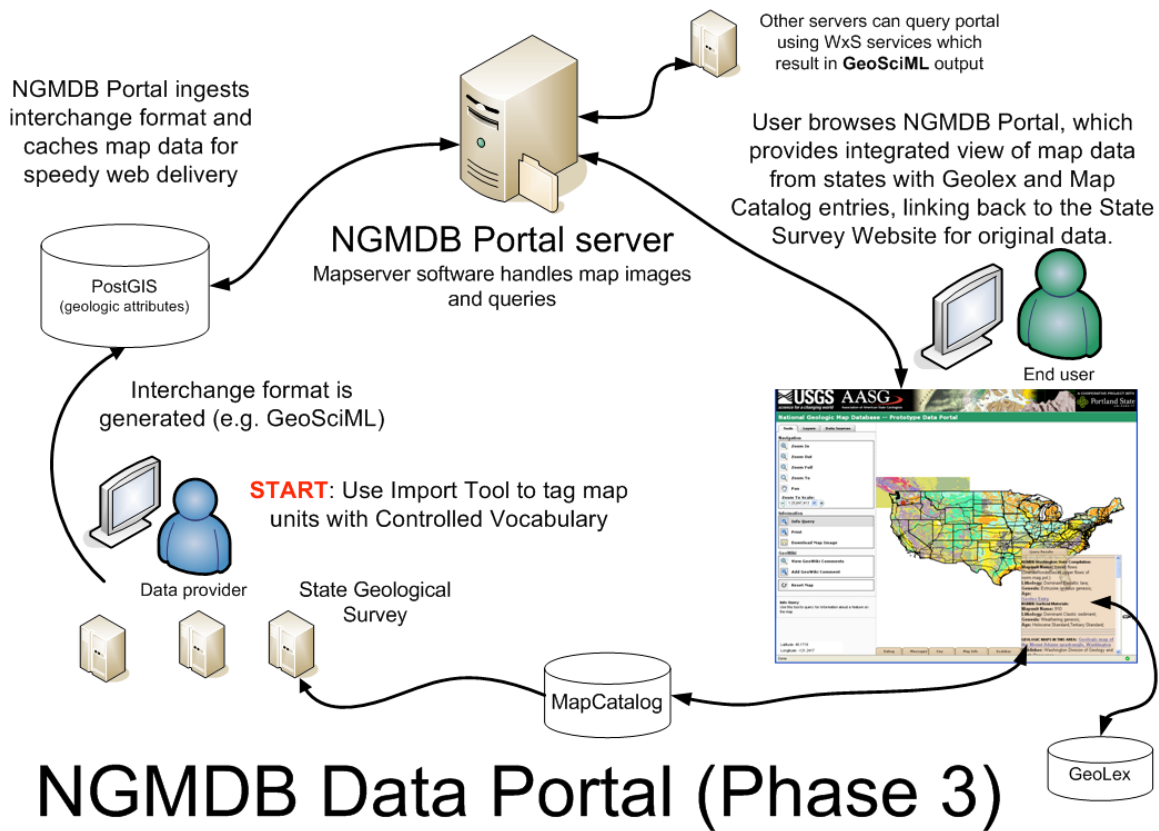
wide participants were queried by a desktop client to show a consistent set of geologic data across disparate data sets from different countries or agencies. It has recently (May, 2007) reached version 2.0 status.

Phases one and two of the project comprised defining the problem-space, building the consistent schemas and terminologies, and compiling databases of stratigraphic lexicon units, and a catalog of scanned maps.

## Phase Three

The phase three prototype of the NGMDB integrates data from Arizona, the Pacific Northwest (Oregon, Washington, and Idaho), as well as several national data sets. A regional meeting was held in which the domain specialists from each state participated in parsing their terminology into our standardized set. Additionally, we demonstrate interoperability with other standards-based services such as the NRCS' soils database and NASA's MODIS satellite data. We demonstrate this all in standard web browsers. For better performance in a web-based client it is usually necessary to cache external data sets locally.

The prototype is enabled by a custom-programmed data import tool that allows the user (domain expert) to match data in fields from an input database source such as Oregon or Washington geologic compilations to the repository database schema used for the web map service. The web repository schema is a subset of content from the NGMDB database design (Richard et al., 2004), including lexicon stratigraphic unit, geologic age, lithology



**Figure 1** – Data processing diagram for National Geologic Map Database. Semantic mediation happens at the step labeled “START”.

composition, genesis, and data source. Fields in the input table are matched to corresponding fields in the repository schema.

This is equivalent to the “registration” activity described in various cyber-infrastructure initiatives in recent years (GEON and BIRN, for example). Subsequently, unique values from each input field are matched to corresponding terms in controlled vocabularies defined by the NGMDB. This is the step that we are calling “manual semantic mediation”. Others refer to this a “light-weight semantic mediation” (Fox, pers comm., 2007).

After matching fields and terms from the input map database to the portal data schema and vocabulary, the import tool outputs the data for input to the database repository for the web server.

Data transport between the import tool and repository uses a GeoSciML document constructed according to our application profile. Figure 1 shows a schematic of the overall process from tagging data with controlled vocabulary to end-users being directed back to the original data, as well as relevant data from other NGMDB data repositories.

As a proof of concept for integrating multiple data sets stored locally or remotely (federated systems), we are developing the capability of reading and writing GeoSciML files as a function of the data import tool. Once the user has created a fully converted data set, it is exported to GeoSciML. This file can then be published on their own site for harvesting by other GeoSciML-aware services. We then import this file into the

full NGMDB data structure and optimize it for speedy web delivery by building generalized (less vertices) versions of shapefiles for use in Mapserver, as well as full resolution versions for display when a user is zoomed in to a sufficient level of detail.

This system is a model for aggregating multiple data sets from many agencies. Some organizations have sufficient resources to set up a WFS server and maintain their own GeoSciML-compliant data which can be integrated into our system via harvesting. Some organizations, however, lacking these resources, could simply provide the data to our project for hosting on the NGMDB site. A third option is to allow organizations access to a virtual server on our system; they will have their own subdomain, for example <http://id.ngmdb.us> (Idaho), and manage their own data as if it existed on their own local server.

Each of these three scenarios is mediated by our custom Data Import Tool, which allows the expert geologist for a region to map their data fields to a common schema and the unique values contained within to controlled science terminology of the NGMDB. A future agent-based crawler could theoretically trawl through these federated data sets and compile derivative data sets based on user criteria. Federated systems, however, are not viable for real-time display, due to the overhead of parsing these large text files.

## Conclusion

We see the current process as a bridging technology to a future implementation in a more AI-like approach using a knowledge-base or expert-system. Each of the term matching exercises that we go through are essentially generating semantic relationships primarily of the form “IsA”. These can then be used to process data sets or documents that the system has not encountered previously. For example, given that we have mapped the terms “basaltic rock” and “basaltic lava” to the higher level concept of “basalt” we could classify documents or map units as being

about basalt, or more generally igneous extrusive rocks, if they contained either of the previous phrases.

We feel that by investing in current solutions, while simultaneously investing in future technological advances, we are able to work the problem-space from both directions.

## References

Richard, S.M., Craigie, J., Soller, D.R., 2004, Implementing NADM C1 for the National Geologic Map Database, in Soller, David R., Editor, Digital Mapping Techniques '04-Workshop Proceedings, U. S. Geological Survey Open-file Report 2004-1451, p. 111-144, accessed at <http://pubs.usgs.gov/of/2004/1451/pdf/richard.pdf>