Aaron Averett
Bureau of Economic Geology
The University of Texas at Austin

National Geological and Geophysical Data Preservation Program 2008-2009
Final Technical Report

During the 2008-2009 fiscal year, funded by USGS as part of the National Geological and Geophysical Data Preservation Program, work was performed at the Bureau of Economic Geology (BEG) for the purpose of developing and deploying a computerized system for storing and organizing metadata that describe materials stored in BEG's geologic sample and geophysical log repositories. The generation of metadata for insertion into a national catalog of geologic materials was described as the primary goal, with the creation of a tool for this purpose being necessary. Listed as deliverables in BEG's contract with USGS are a copy of the metadata stored in BEG's Core and Log Database in one of several formats, a list of all of BEG's relevant collections provided as input to a web-based database, and this final technical report describing what was and was not accomplished.

File Metadata

The most important deliverable is the set of metadata representing the contents of BEG's geologic material collections. In this context, those collections are the collection of oil and gas well logs stored in paper digital format in the Geophysical Log Facility, and the collection of cores, cuttings and other physical samples stored in the three Core Research Centers, located in Austin, Houston and Midland, Texas.

Internally, BEG uses a system composed of a Microsoft SQL 2005 Server and ESRI ArcGIS Server to store metadata for these collections, and to make it available to researchers and the public through a web-based user interface, which connects to the ArcGIS server. The award states that acceptable formats for submitting the metadata to USGS include either a comma separated flat-file format, a unique XML format, or Open Geospatial Consortium Web Feature Service (WFS). Of these three techniques, the XML format was chosen because it is the simplest to generate. Code was written in C# as an add-on for BEG's core and log database to generate this XML format. That add-on was subsequently used to create six data files, containing a total of 574,946 records (see table), which were uploaded to USGS's MyUSGS web application for inclusion in the National Catalog. Each file represents the contents of one of BEG's collections as reported in the 2007 fiscal year survey. Many more records exist in BEG's database, but were not included in the metadata delivered to USGS because they are missing information stated as required in the published specifications for the CSV and XML file formats.

In the vast majority of cases, the missing information is a latitude and longitude pair. This often happens with wells drilled long ago. These records usually have location information that is expressed differently, in a manner that is very difficult to geocode. In other cases, the well itself is fairly recent, but the information provided to the curators of the collection about it is sparse, and did not include coordinates, or required them to

somehow be extrapolated from some other sources. Efforts to acquire or extract this information are ongoing.

Filename	Collection ID	Record Count
crc2_export_20090914074342_P496.xml	P496	114,222
crc2_export_20090914074342_P497.xml	P497	9,562
crc2_export_20090914074342_P498.xml	P498	318,565
crc2_export_20090914074342_P500.xml	P500	22,353
crc2_export_20090914074342_P501.xml	P501	111
crc2_export_20090914074342_P626.xml	P626	110,133
	Total Records	574,946

File breakdown- Each file represents one "collection" as reported to USGS in 2007.

Web Feature Service (WFS)

The creation of a "federated system integrated with a shared WFS for querying and transferring metadata records" is stated as a program objective, with the implication that the creation of a web feature service is preferable providing metadata in the CSV or XML file formats. In practice, this was not well supported, however, as unlike the CSV and XML file approaches, no specification for what information the service should include was published. BEG has therefore created a prototype feature service that approximates the XML or CSV file formats, under the assumption that the same information would be required.

The WFS is still in development, but initial tests show considerable promise. Improvements in the ESRI ArcGIS software packages that drive BEG's metadata catalog have made the deployment of such a service possible, and will improve quality of service as they mature. It should be noted that since no other surveys have apparently attempted to set up a WFS, no standard or specification exists. In addition to specififying what data are required, an official specification would need to Establishing specifications and best practices at the federal level for what information WFS services intended for participation in the distributed National Catalog need to provide would go a long way toward encouraging participation in that manner, as there is presently almost no guidance at all.

The prototype feature service can be accessed at http://igor.beg.utexas.edu/ArcGIS1/Services/crc2_wfs/GeoDataServer/WFSServer. It can be accessed with a web browser for testing purposes, but in practice, a proper client such as ArcGIS or Gaia is needed for actual use.

Conclusion

This round of the National Geological and Geophysical Data Preservation Program accomplished several important things. Most importantly, BEG has been able to contribute almost 575,000 samples and well logs to the National Catalog, which was seen as a primary goal. In addition, BEG has made significant progress toward establishing a prototype web feature service for participation in the presently hypothetical federated system that was proposed as a longer-term goal of the program. Finally, procedures, infrastructure and organizational momentum have been established within BEG for supporting ongoing data preservation efforts.