Award Number 08HQGR0113

National Geological and Geophysical Data Preservation Program

Preservation of Geologic Data and Collections in Illinois: Compilation, Documentation and Planning

Final Technical Report

Fiscal Year 2008

9/1/2008-8/31/2009

Ronald Klass
Information Management Coordinator
Tel. 217-244-5585
Fax 217-333-2830
rklass@illinois.edu

Zakaria Lasemi
Senior Geologist and Head, Industrial Minerals and Resource Economics Section
Interim Director, Earth Resources Center
Tel. 217-244-6944
Fax 217-333-2830
lasemi@isgs.illinois.edu

Illinois State Geological Survey
Institute of Natural Resource Sustainability
University of Illinois
615 East Peabody Drive
Champaign, IL 61820

November 16, 2009

Abstract

This report addresses the following primary NGGDPP Products and Deliverables:

- **Collections Inventory**

  A total of 36 collections were identified and entered into the on-line inventory forms.

- **Metadata for the National Catalog**

  Metadata records were created and uploaded for 5 collections.

- **Long-Range Data Preservation Plan**

  A Long-Range Data Preservation Plan was developed following the NGGDPP guidelines.

Collections Inventory

A total of 36 survey collections were inventoried and entered into the online website survey. Of these, 11 are of type “Physical”, and 25 are of type “Derived and Indirect”.

The survey has an established database of significant collections and data. Collections listed in our database with middle to high significance to the survey were used to facilitate generation of the collections inventory. The established data were augmented by input from other project members all of whom are significantly involved with our collections, as well as input from curators, staff, and management. The inventory is also helping us develop a more comprehensive database of our significant collections.

In a few cases, it was difficult to decide exactly what to define as a “collection” for the purposes of the online survey. This was especially true for our fossils, which are divided and lumped into various similar and dissimilar categories, storage locations and significance.

Geologic Samples Library Inventory

The Natural Resources Studies Annex building contains the Geologic Samples Library, a large climate-controlled warehouse housing most of our physical collections. The legislatively-mandated and well-curated cores and cuttings collection occupies a majority of the floor space. The remaining area contains a substantial amount of cabinets, files, and shelving, holding miscellaneous collections of all types, sizes, and origins. The material has accumulated over the 100+ year life of the survey, much of it moved to new locations or new buildings many times over many decades.

Other than certain significant paleontological collections, little was known or documented about these miscellaneous collections, stored in over 5000 drawers, bins, and shelves. To gain a better overall understanding, we created an inventory of container-level (drawer, bin, or shelf) information. Although not item-level metadata, this nonetheless substantial task was a necessary first step in assessing the miscellaneous Geologic Samples Library holdings.

A University of Illinois undergraduate student in Geology was hired to assist in generating the inventory. A total of 5378 drawers, bins, and shelves were examined, their contents summarized (or confirmed from labeling) and entered into a database along with location information such as building section, cabinet number, and drawer number. The result is a mineable database of groups of generally similar items and their locations within the facility.

This data helped to substantiate or confirm the approximate sizes and types of some of the collections entered into the Collections Inventory. In the future this data will be useful to help locate and break out smaller specific collections from large collections such as “hand samples”, which consists of approximately 475 drawers, bins, and shelves of miscellaneous collections and samples.

Given the long history of the survey, prior documentation on some of these miscellaneous collections may have been temporarily or permanently lost in space shuffles or retirements. Some documentation may still be recoverable and could be the subject of a future project under this program.

Collection Metadata

Metadata for a total of 5 collections was uploaded to the NGGDPP website for this project. We developed an XML metadata generation program using Microsoft Access and VBA that obtains data directly from our Oracle Enterprise Database System.

As a test early in the project, we derived metadata from two existing databases representing the cores collection and the cuttings collection, creating 11975 core and 62552 cuttings metadata records. This exercise also brought to light some issues with the existing databases which are being reviewed and corrected as time allows. We also created databases and corresponding metadata for the following 3 collections.

Metadata for: Electrical Earth Resistivity Survey records

We have a collection of Electrical Earth Resistivity Survey data, reports and maps going back to the 1930's. In general these were used to determine sites with better potential for water wells in a survey area.

In consultation with the collection curator, a custom database was developed and a student hired to assemble and enter data. We began by entering data from historical handwritten summary forms, and included a link to a PDF scan of the form containing additional information. These forms cover from the 1930's through approximately 2000.

In order to capture the newer surveys as well as earlier surveys missed by the historical summaries, the student next began systematically examining each of the paper-based files, beginning with counties where complete metadata was judged most beneficial. Since the paper files were being systematically reviewed we decided to also scan them at the same time; the database includes a link to the scanned PDF reports.

The student completed full review and processing of 6 counties of paper files representing 489 studies by the end of summer break, when the student left the project. In total 3663 locations representing 2334 studies were entered into the database and used to create metadata.

A significant number of studies were found in the paper files that were not in the summaries. We estimate approximately 80% of all EER studies in the paper files are now in the database. We plan to complete the remainder as time allows most likely using internal resources.

Issue: A significant number of records had invalid PLSS section data when converted to coordinates, requiring significant time to research and correct the original location data (where practicable). This was caused by a mixture of issues, from typos on the original report summary to missing data to invalid quarter specification in partial sections to incorrect county. About 60 locations were judged not practicable to resolve, but we insured that each survey had at least one converted PLSS location so that it would be represented in the metadata records.

Metadata for: Lead-Zinc mining district borehole records

These paper records and mine maps were donated to us by a closed Lead-Zinc mining company in Illinois. A student was hired to first organize and determine exactly what we had. The data and maps are in different formats from a number of different companies that were bought up by the last surviving company, making the task more difficult. In addition, data from Wisconsin was separated out and transferred to the Wisconsin Survey.

The student entered location and formation data from a total of 452 boreholes into our Enterprise well and borehole database, 372 of which included geochemical analysis data for Lead, Zinc, and Iron at regularly spaced depths in the borehole and which were used to create metadata records for the geochemical data collection (a few do not yet have valid locations and thus are not in the metadata records).

The location data are currently imprecise, based on PLSS sections and not suitable for precise mapping. The student scanned and is currently (for FY2009) georeferencing many of the maps in order to obtain much more precise locations for the boreholes. This requires significant effort and interpretation as non-standard undocumented reference points were used, perhaps as in-house shortcuts or to help protect trade secrets.

Metadata for: Geologic Samples Library Paleontological Samples, Silurian subset

Since the full collection consists of well over 150,000 specimens, as a pilot project we limited metadata creation to the Silurian portion (approximately 10% of the collection). A student numbered samples and entered data such as locality, unit, collector and geologist notes from sample tags as well as storage location into a custom database developed for this project. Even for this limited subset creating a full set of metadata appears to be a daunting task.

The vast majority of fossil samples are not identified (other than by unit and system). To complete the metadata, a paleontologist must identify each sample and enter the common and scientific name into the database; this will be done by staff as time allows. However, even without specific fossil identification, the data are useful to locate fossils from a specific formation.

At the end of the FY2008 project period, 11086 samples from 460 localities have been entered into the database, and corresponding XML metadata uploaded to the catalog. This represents well over half of the Silurian samples, but is not yet complete. We decided to complete this level of metadata for the Silurian portion of the collection in FY2009 using the same student who has become quite proficient at this task. We are also utilizing other available internal resources to assist with data entry.

Issues: Many metadata sample locations do not yet have a precise latitude/longitude point, generally for one of two reasons: first, some do not list a section, but rather name a specific quarry or locality. Staff members expressed a strong preference that they (rather than the student) derive detailed location information because they are familiar with most of the localities. However until then, the point information for most of these samples is approximate (county or state centroid). The derivation and precision of each sample location is noted in the alternate geometry section of the XML data.

Secondly, about 40% of the localities are outside the state of Illinois, mostly in adjacent states. We did not have a method to generate latitude/longitude points from PLSS data outside of our state. We researched methods to accomplish this and wrote software to interrogate a United States Bureau of Land Management Web service to programmatically obtain points for out of state sections. However, many of the out of state localities do not have section information, and until revised by staff are represented by the state centroid.

Long-Range Data Preservation Plan

A Long-Range Data Preservation Plan was developed following the NGGDPP guidelines. Future NGGDPP program proposals will indicate how planned activities address the plan. This is a starting version and we intend to revise the plan as needed in the future. We are also working towards integrating the plan with existing information management documents.

Comparison of Proposal Goals with Project Accomplishments

The FY08 proposal lists the following products and deliverables:

> 1) a relational database application summarizing the collection compilation and inventory efforts; 2) completion of the online survey for the National Catalog including relevant information from the collection compilation and inventory; 3) a relational database application containing metadata for the various collections identified in the inventory effort; 4) uploaded content to the National Catalog, including relevant metadata for the identified collections; 5) a Long-Range Data Preservation Plan that is consistent with the guidelines set forth by the NGGDP Program.

Items 2, 4, and 5 are deliverables under the program and as discussed earlier were completed. This was our first year in the program and since we did not have a complete collection inventory at the time of the proposal we did not specify how many or which collections for which we would generate metadata. However, we believe the collection metadata generated for this project represents an appropriate and successful use of the resources provided under the contract.

Items 1 and 3 are not deliverables under the program; however, these databases were developed to assist in the creation of related deliverables and continue to be used internally.

General Issues and Comments

Issues with metadata coordinate points

All metadata generated for this project must contain a point expressed as latitude and longitude. Most of our location data are based on the Public Land Survey System which uses a section, township and range from one of 3 principal meridians in Illinois. However, when specifying a location, our geologists have traditionally specified the county rather than the meridian. This non-standard specification requires that meridian be inferred from the township, range and county. Although we have an ArcGIS tool developed in-house that we use to convert our PLSS data to coordinates for various purposes, significant unanticipated effort was expended by the PI in order to generate coordinates for much of our metadata.

Use of our in-house tool requires a number of manual steps, including running ESRI ArcMap on a desktop system referencing our PLSS layers, and selecting various options in the tool. Although parts of the process can be automated, it is still tedious to move and convert data back and forth between temporary files and formats and a database. In addition, the tool requires complete PLSS information including a section number to generate a coordinate, otherwise the data are rejected. Finally, it accepts only Illinois data. However, the in-house tool has two advantages: it is quite precise, utilizing quarters and footages if specified, and will accept a county in lieu of a meridian.

The PLSS location conversion system developed for this project utilizing the Bureau of Land Management Web service is fully automated and works for all PLSS states for which we have metadata. Further, we added additional logic so that a coordinate is always generated regardless of the detail level of the location (if necessary simply the state centroid is returned). However, the BLM site we are using does not convert quarters and footages and thus is less precise, and requires a meridian which is generally not available in our data.

As a temporary but inefficient solution, we split the data and used our in-house tool for Illinois locations with sufficient (section) information, and the new system for all other locations. Unlike Illinois, most of the other states for which we had PLSS locations use only a single meridian so it was easy to derive the meridian based on the state. We could add software to derive the proper meridian for Illinois locations as well, allowing us to use a single automated system for all PLSS metadata locations, but this would result in a substantial loss of precision for many Illinois locations. Thus we are still searching for a better solution to PLSS conversion.

Issues with XML uploads

As an early uploader, we encountered and assisted in tracking down bugs on the NGGDPP upload website. However at the end of the project period, we still encountered some upload issues requiring assistance from USGS staff. While this assistance has always been prompt, we encourage continued improvement of the website and will continue to work with USGS staff as needed.

While possible, updating individual XML files to the NGGDPP site on a regular basis may become time-consuming and problematic. We plan to research the viability of using OGC services such as WFS to minimize our effort to both upload and to update metadata information. This may also simplify synchronization, confidentiality, and other issues with our internal databases.

Summary

We believe we made good use of the resources provided to us to assess our collections and to create metadata for a subset of them. We are continuing metadata creation for FY2009.

Acknowledgements

Several staff members at the Illinois State Geological Survey participated in this project:

- **Don Keefer** was a key advisor and assisted with many aspects of the project at all stages from initial proposal to final report.
- **Tim Young** helped coordinate much of the work for the Electrical Earth Resistivity metadata creation and assisted with the EER database design.
- **Don Mikulic** coordinated much of the work for the paleontological sample metadata creation and assisted with the paleo database design.
- **Jared Freiburg** coordinated much of the work for the Lead-Zinc mining district metadata creation.
- **Bob Mumm** assisted with the Geologic Samples Library miscellaneous collections inventory and with the cores and samples metadata.
- **Rod Norby** (retired) assisted with inventory of our paleontological collections and with the paleo database design.