

EarthChem Workshop on USGS National Geochemical Database

US Federal Center, Denver, Colorado
April 24, 2007, 8:30 am to 11:30 am

Report compiled by Doug Walker and Steve Smith.

Participants:

Doug Walker, Steve Smith, Steve McDanal, Dorothy Trujillo, Jason Ash

Background

EarthChem is a community driven project to facilitate the compilation and dissemination of geochemical data of all types. The project is active at building a home for future data contributions by working with authors, societies, and publishers as well as government organizations. In addition, the EarthChem project responds to community needs to facilitate compiling and serving data.

One of the aspects of EarthChem is to assist contributions of data to the system from existing databases and to provide expertise to other groups assembling databases. A very large geochemical database not yet in EarthChem is the USGS's National Geochemical Database (NGDB). This database contains most of the geochemical data on a large percentage of rock samples collected by the USGS over the last 40 years. NGDB has information for over 400,000 igneous, metamorphic, and sedimentary rocks. The database is currently served by the USGS, but would form a valuable component of the EarthChem system. In turn, EarthChem would provide an important discovery tool for USGS data that would then be able to link back into the NGDB.

To move this process forward, a meeting was held between the EarthChem and USGS workers to identify the process by which NGDB could be brought into EarthChem. The USGS personnel will provide Access database files to EarthChem. EarthChem programmers will then take these into the database using the XML method already in place for NAVDAT, PetDB, and GeoRoc (as well as MexDB when available).

Goals

The goals of the meeting were to address important issues on the ingestion of data from NGDB into the EarthChem One-Stop-Shop database, and to establish links from EarthChem back to the USGS. In particular, the goals are:

- a. Discuss the amounts and types of data to contribute.
- b. Determine how best to get data into EC.
- c. Establish ways to add additional data as it becomes available.
- d. Identify any additional data sets/databases to contribute.

Workshop Summary

The USGS has over 400,000 samples in the NGDB that could be contributed to EarthChem. These samples contain sufficient metadata about method and time of analysis to be very useful to the EarthChem community. These data are contained in an Oracle database at the USGS and in an Access database that is maintained by Steve Smith. Smith is actively working on quality control issues related to these data, and the Access database is the most up-to-date and cleanest version of these data.

These samples can be contributed to EarthChem. A procedure was identified for getting the data from the Access database into the Oracle-based EarthChem One-Stop-Shop. Some additional work will have to be done on cleaning up the rock names in the USGS database, especially considering that there are metamorphic and sedimentary rocks in the database as well. This will be an important addition to EarthChem. Lastly, the NURE-HSSR database (National Uranium Resource Evaluation and Hydrogeochemical Stream Sediment Reconnaissance) was identified as another data set ready to go into the EarthChem system. This is another ~400,000 samples that could go into EarthChem.

Approaches

The USGS NGDB is implemented in two ways. The main database (including many more samples for extensive types of materials) is contained in an Oracle database served at the USGS in Denver. Records for samples of rocks from the database were exported to an Access database that has undergone comprehensive quality control and quality assurance by Steve Smith. The Access database consists of essentially two tables. One table contains the extensive metadata for sample characteristics and age (the Geo Table); the other table presents all of the measured values (elemental and isotopic data – the Determinations Table). The two tables are related by a unique field termed “Lab-ID” which references values to samples.

It would be a difficult and time consuming operation for USGS workers to repopulate the master Oracle database tables with the corrected and controlled information in the Access tables. Alternatively, the Access tables comprise a complete representation of the best values and metadata available for the NGDB. For this reason, the EarthChem and USGS workers have decided that the best approach is to take the contents of the edited tables into the EarthChem database.

The basic approach will be for the USGS workers to prepare an edited version of the NGDB database in Access that can be easily ingested into the EarthChem system. This will require that the USGS make some additions to the database. This mainly involves, for the Geo Table, making sure that every rock description gives the type (e.g., igneous vs. metamorphic) and, to the extent possible, the specific name (e.g., basalt vs. rhyolite). For the Determinations Table, the method information must be as complete as possible, and methods of similar type must be grouped. The USGS will also make sure that the element or ratio analyzed (called Species) is as general as possible. Once these issues are resolved, the data will be sent to EarthChem for conversion to XML and ingestion into

the database. EarthChem will take the data, join by the Lab-ID, and prepare a sample-by-sample output to populate into the database.

To ensure proper credit for the data, EarthChem will clearly identify the USGS data as being provided by the NGDB (in the same way that GeoRoc, PetDB, and NAVDAT contributions are currently noted). The source will be listed as USGS NGDB, the submitter of the samples will be listed as the “author”, and publication date will correspond to the date of sample submission for analysis. These aspects should allow clear credit to be given to the USGS as well as the scientist studying the rock for which data are provided.

Next Steps

Considering the approaches listed above, the participants recommended the following next steps to move the process forward. We anticipate that these steps will be done over the next 3 to 6 months.

1) The USGS Geo Table needs to be modified so that it contains all necessary metadata. Values for the specific name of a rock (basalt vs. granite) will be populated as well at ensuring that the general class of sample is given (metamorphic vs. igneous). To the extent possible, the age of the sample will be defined. Of these, only the general class must be given a specific value; unknown is an acceptable attribute for the other information. Steve Smith will take on this task in collaboration with Steve McDanal.

2) The USGS Determinations Table will omit values based on only partial dissolution of samples for elemental determinations. It will be impossible to evaluate these determinations quantitatively, so they will be eliminated. In addition, many values are reported relative to a detection limit. Currently, the nomenclature for these values is given using several qualifiers. These will be consolidated to a simple “less than” or “greater than” (L or G) designation with a value. Steve Smith will do this task.

3) The USGS Determinations Table will be modified so that it is as simplified and consistent as possible. This will require some lumping of particular “species” or elements and ratios where they are so similar as to be indistinguishable. Methods of analysis will be aligned with the current EarthChem schema where possible; if needed, the schema will be extended to take USGS methods into account. Steve Smith, Jason Ash, and Doug Walker will complete this work.

4) The USGS will group the Geo and Determinations Tables at the sample level so that data for between 10 and 1000 samples are contained in any set of two tables. This will provide a significant, but manageable, contribution of data. Such groupings will be constructed so that the entire NGDB is represented. This will be the responsibility of Steve Smith.

5) Once the Geo and Determinations tables are finalized and grouped, EarthChem programmers will take this information and create the XML files needed to populate the

EarthChem database. This will be done by importing the Access tables into an Oracle database, joining the information base on the Lab-ID, and exporting to the EarthChem XML schema. Once done, the data can be easily imported into the EarthChem database. Jason Ash will take on this task.

6) Additional data available can be added to EarthChem using the same steps outlined above. The new data will comprise groups to be submitted to the database.

Supporting Materials

Tasks and database specifications identified during workshop, given by responsible party. This was listed on white board and agreed to by all parties.

USGS tasks:

For Geo Table:

- Assign values to the following columns if necessary:
 - Specific Name (unknown acceptable)
 - Age (unknown acceptable)
 - Secondary Class (Specific Name) (unknown unacceptable)
- Provide the following columns from existing data:
 - Lab ID (unknown unacceptable)
 - Submitter (Author field on EarthChem schema) (unknown unacceptable)
 - Date Submitted (unknown acceptable)
 - Latitude/Longitude (unknown unacceptable)
 - Source = National Geochemical Database
 - URL back to USGS based on Lab Number
 - Field ID (unknown unacceptable)

For Determinations Table (Chemistry):

- Consolidate species names if necessary.
- Fix methods using EarthChem names if needed.
- Eliminate records where values are based on partial digestion.
- Reduce Qualifiers for detection limits to “L” or “G” (less than/greater than).
- Consolidate tables for different types (e.g., major and trace elements).
- Add needed items/type to enumerations and return to EC.

EarthChem tasks:

- Send enumerations/controlled vocabulary and schema to USGS.
- Convert USGS Access tables to XML.
- Populate EarthChem One-Stop-Shop with XML files.

Time:

- USGS: 3 Months or less to provide data in Access table form.
- EarthChem: Around 2 weeks to populate once data is received from USGS.