Cyberinfrastructure for Solid Earth Geochemistry (CSEG) Workshop
Carnegie Institution of Washington, October 3&4, 2003

Workshop Summary:

The CSEG Workshop was convened to review the current status of data management efforts in solid earth geochemistry and discuss ways in which these activities can grow and collaborate to best participate in, and contribute to, the cyberinfrastructure revolution in the geosciences.

The workshop combined geoscientists currently involved in active and developing geochemistry database efforts that include

- PetDB (Petrological Database of the Ocean Floor)
- GEOROC (GEOchemistry of Rocks of the Oceans and Continents)
- NAVDAT (Western North American Volcanic and Intrusive Rock Database)
- GERM (Geochemical Earth Reference Model)
- RIDGE2K/MARGINS Data Management System
- MetDB (Metamorphic Petrology Database)

with information technology experts from organizations involved in IT development in the geosciences including representatives from

- GEON (Cyberinfrastructure for the Geosciences)
- CIESINS (Center for International Earth Science Information Network)
- KGS (Kansas Geological Survey)
- SDSC (San Diego Supercomputing Center)

The workshop began with presentations from existing database groups outlining current status and near future plans. Following the presentations, open discussion in the workshop centered on the tasks needed to improve geochemistry databases so that they can provide the most useable product for a wide user community. These tasks, and the action items identified at the workshop are listed below:

1) Creation of seamless search capability across databases

Users should be provided an information system for geochemical data that will allow them to search seamlessly across all geochemical databases. At present, the databases exist in individual database efforts whose contents are constrained by personnel, programmatic (e.g. the OCE-EAR boundary at NSF and between US and International activities) and budget restrictions. The easiest way to begin to achieve this goal is for the geochemistry database developers to agree on a compatible metadata schema and a controlled vocabulary with eventual publication in an OAI (Open Archives Initiative). This opens the databases to access by a much broader range of search mechanisms employed by the IT community. The three igneous geochemistry databases (PetDB, GEOROC, NAVDAT) recently formed a consortium (www.earthchem.org) that will

assist in reaching this goal. This systemization of schema and vocabulary/ontology will simplify the writing of cross database query capability. The EarthChem consortium adopted this as a high priority action item. The schemata employed by these three efforts are already similar since they all grew from that originally developed by the PetDB and GEOROC efforts (Lehnert et al., G3, 2000). As the content of the databases expands to include data complementary to rock geochemistry (e.g. geochronological data and geochemical data for a wider range of natural materials including minerals, metamorphic rocks, and sediments), the metadata content of the schema will need to be expanded. The current schema, however, is capable of accommodating many additional types of information, including much of the available geochronological measurements. A possible model for this expanded metadata content is described in Staudigel et al., G3 (2003).

Bob Arko of the RIDGE2K/MARGINS database effort at LDEO offered to help condition the EarthChem schemata for OAI publication. He will work with Kerstin Lehnert of PetDB to provide the model for this transformation using PetDB as the first test case. Chaitan Baru of GEON also asked to be involved in this effort both to assist in production of the more general geochemistry schema and the eventual development of a geochemistry ontology based on the metadata content of the schema.

2) How to involve data contributors in the data entry procedure

Presently, data entry into geochemistry databases mostly involves incorporating published data. Personnel employed by the database efforts do this time consuming step. Data entry personnel with a moderate level of geological/analytical knowledge are needed in order to more reliably spot errors in the data and extract the appropriate metadata from publications. The general procedure is to scan a published data table and then convert it to an Excel sheet by optical character recognition. Additional information (sample descriptions, information on techniques, analytical errors, etc.) is then gleaned from the accompanying paper by data entry personnel and entered as metadata. No clear improvement to this approach was offered at the workshop, though there was discussion of adoption of error-checking "wizards" in the Excel sheets to assist data entry personnel in spotting and correcting errors.

Most of the discussion on this topic centered on eliminating the steps described above by receiving electronic files directly from the data producer. The first step is to appeal to journal editors to establish minimal standards for reporting geochemical data. Chemical Geology editors Roberta Rudnick and Steve Goldstein, who have prepared a guide to authors regarding data publication standards for the journal, have already taken steps in this regard. This action resulted from journal-database dialogs that began at the GERM meeting in May, 2003.

Frank Podosek, editor of Geochimica et Cosmochimica Acta, met with Rick Carlson in October to discuss the development of suitable electronic submission criteria for this journal. To further the journal-database interaction, EarthChem will have an exhibitor's

booth at the Fall AGU meeting and will put on database demonstrations for journal editors and publishers.

Workshop participants identified the following four items that are both essential metadata for geochemical databases and are simple requirements for journal contributors:

- Sample Location including GPS data if available
- At least some information on sample age (stratigraphic, radiometric)
- Definition of units and normalization used in data presentation
- Information on data quality (e.g. precision of analysis and data obtained on recognized international reference materials)

If this simple set of information were included along with the geochemical data, and these data were delivered from data providers to database administrators in electronic data files that could be directly input into the databases (Excel or comma-delimited files, not PDF files), the job of data incorporation into the database would be dramatically simplified and errors in data entry reduced.

Another item of critical importance to databases and the scientific method of geochemical studies is a unique sample identifier. Many examples can be given of a sample collected by investigator A, analyzed for major elements by investigator B, for trace elements by investigator C, etc., with the individual data sets reported in different publications, commonly under slightly or completely different sample names or numbers. Ideally, all chemical information on a single sample should be presented by databases as a single entry. Without a unique sample identifier, the prudent database response is to keep the individual data sets separate to avoid combining data for what might not be the same sample. The exact form of a unique sample identifier was not identified at the workshop, but it was clear that some combination of the collector's name, the date of collection, and the position of collection could eliminate this confusing aspect of data entry.

Data providers/creators must be made aware of the wealth of information that can be contained in, and served by, electronic databases. For example, publication of electron microprobe analyses have become increasingly rare as journal space becomes more restrictive, but such analyses can be incorporated easily into electronic databases. This problem will grow as geochemical instrumentation becomes increasingly facile at high-throughput analyses. Close association of database groups with journals could lessen the burden to the journals to archive massive data tables, while simultaneously easing data entry into databases so that the user community can widely and efficiently access the data.

3) Improvements in data query and manipulation options

Users of geochemical databases want to easily obtain the datasets most pertinent to their interests. The difficult aspect of this service is to accommodate users at all levels, from the expert interested in fine-scale details of the data to the non-geochemist who may just

want to obtain a broad idea of the rock types present in a given area. Discussion at the workshop centered on:

- The need for more visual data query options via maps (see section 4 of this report) or simple plots (section 5) to allow users to select data on the basis of visual examination.

- Distinction between queries that are likely to be applied to all sample holdings versus those pertinent only to a restricted portion of the data.

- Production of "precompiled" data sets representing useful subsets of the data holdings

Doug Walker showed a combined map/plot visual interface for the NAVDAT sample holdings that allows users to query individual data points shown on maps or Harker diagrams. When development of this user interface is complete, exchange visits between NAVDAT and GEOROC personnel will allow its implementation either directly on the GEOROC site or on a cross-database query page developed as a part of EarthChem consortium activities.

Discussion on the production of "precompiled" data sets touched on the issue of the difference between simple products that can be extracted directly by queries and "derivative products" that involve scientific input and decision-making. An example of this distinction is the difference between a compilation of major element data for all volcanic rocks in Nevada and an average "primitive basalt" from Nevada. Sample selection criteria for "primitive" require decisions about what compositional characteristics distinguish fractionated from unfractionated lavas, and the elimination of lavas affected by crystal accumulation and/or weathering. Finally, "basalt" may not be the same to all users of the database.

Experience in PetDB and GEOROC suggests that the majority of users are satisfied by simple precompiled data sets, for example, all basalts from Hawaii or from the East Pacific Rise. Such compilations also are potentially important contributions to efforts such as GERM that report to a broader geoscience community summary compositions for different Earth "reservoirs." At present, these compilations are created by database personnel and are not routinely updated as new data enters the database. The compilations that require considerable scientific insight in their creation are perhaps better viewed as scientific outputs rather than database outputs.

The option was explored of creating stored queries that could generate such simple compilation datasets when requested by the user thus providing the most up-to-date sample of the database. Since at least some fraction of these compilations are likely to see discussion in publications, the databases need to provide documentation on how such compilations were produced and what data were included in the compilation so that the results can be checked and verified. Besides providing with queries the version of database sampled, another method of query documentation could be for the databases to

return all references sampled by the query, preferably in proper format to be included in a paper that reports the results of the query. This step would also address the issue of the proper attribution to data providers and counter the increasing tendency to reference data as, for example, "data from GEOROC". Although testimony to the success of the database efforts, this simplified referencing of data sources does not allow proper credit attribution to the investigator who initially produced the data.

4) Creation of map interfaces and other visualization guides

Given the importance of geospatial referencing of most geological information, the ability to compare geochemical data with other geologic and geophysical information that can be presented on maps could offer distinct improvements both in the ability to query the databases and provide avenues to a vast array of data analyses. The workshop identified a number of existing maps that could assist in various analyses of geochemical data. These include:

Ocean
1) Smith and Sandwell 2' map
2) Ridge/Multibeam synthesis
3) SIO Explorer multibeam collection
4) Ocean floor age – University of Sydney

Land Elevation
1) Gtopo30/SRTM30 – 1km resolution.
2) NASA data/NED

Land Geology.
1) World map.
2) US/North America
3) State maps where possible
4) 2° sheets

The most effective way in which to display data on these maps received considerable discussion. At the lowest level, being able to show where samples were collected in comparison to the information on the various base maps described above would provide a potentially important search criteria. The next step would allow the plotted data points to include a variety of sample information retrievable by clicking on a given point. Doug Walker demonstrated this level of information display for the NAVDAT sample holdings. Beyond this embedded information, another useful visualization technique would allow the sample points themselves to convey sample information, for example, point color, shape, or size coding that would correspond to some desired sample parameter (e.g. age, concentration of a selected element, isotopic composition, etc.). A work plan, described in Appendix II, describes the course of action proposed to begin implementation of these map visualization approaches to the various databases.

5) Analytical Tools

Both to extract scientific content and to better guide querying, the ability to manipulate raw data using a variety of analytical tools is essential. Workshop discussion

distinguished between well-developed tools that would assist in data visualization (e.g. plotting routines) from those that would operate on the data to produce a derived data set. An example of the latter would be the MELTS program that uses a thermodynamics database to calculate parental magma composition and fractionation paths for a given input composition. These more complex tools will continue to develop and should not be fossilized by inclusion in the database. Instead appropriately formatted output files should be delivered by the databases for insertion into these stand-alone programs and clearly defined API's (Application Programming Interface) should be developed to allow tool developers to access needed data directly. Possible programs identified that would fit this definition include:

- MapApp mapping program (discussions between Kerstin Lehnert and Bill Haxby (LDEO) will explore what is needed to coordinate database output for this program)
- MELTS (Plans have already been discussed to work with Mark Ghiorso to develop the necessary database-tool communication to allow MELTS to be used with input from the EarthChem databases)
- EC-RAFC (Energy Conserved - Recharge, Assimilation, Fractional Crystallization: Rick Carlson will contact EC-RAFC creators Frank Spera and Wendy Bohrson to explore their interest in developing links to the databases)
- PetroPlot (A petrographic plotting routine developed for PetDB that could be of use for query visualization in all EarthChem databases)
- MatLab - for general statistical evaluation of data in the databases

The discussion on "tools" distinguished those calculations that are likely to be repeated many times by a user versus those that would only rarely be accessed. Calculations likely to be repeated (for example the creation of normative mineral compositions or the definition of consistent rock names from major element analyses) are better done once and the results stored for easy access. More complicated calculations approach the "derivative products" issues discussed previously.

Appendix I: Participants at Workshop

Name	Affiliation	e-mail
Robert Arko	Lamont-Doherty Earth Observatory	arko@ldeo.columbia.edu
Jeremy Bartley	Kansas Geological Survey	jbartley@kgs.ku.edu
Chaitan Baru	San Diego Supercomputer Center	baru@sdsc.edu
Todd Bowers	University of Kansas	tbowers@ku.edu
Richard Carlson	Carnegie Institution of Washington	carlson@dtm.ciw.edu
David Epp	National Science Foundation	depp@nsf.gov
Sonia Esperanca	National Science Foundation	sesperan@nsf.gov
John Helly	University of California, San Diego	hellyj@ucsd.edu
Albrecht Hofmann	Max Planck Institut fur Chemie	hofmann@mpch-mainz.mpg.de
Anthony Koppers	Scripps Institution of Oceanography	akoppers@ucsd.edu
Charles Langmuir	Harvard University	langmuir@eps.harvard.edu
Kerstin Lehnert	Lamont-Doherty Earth Observatory	lehnert@ldeo.columbia.edu
Chris Lenhardt	CIESIN - Columbia University	clenhardt@ciesin.columbia.edu
Kurt Look	Kansas Geological Survey	klook@kgs.ku.edu
Baerbel Sarbas	Max Planck Institut fuer Chemie	sarbas@mpch-mainz.mpg.de
Frank Spear	Rensselaer Polytechnic Institute	spearf@rpi.edu
Hubert Staudigel	Scripps Institution of Oceanography	hstaudigel@ucsd.edu
Sri Vinayagamoorthy	CIESIN - Columbia University	sri@ciesin.columbia.edu
Thomas Wagner	National Science Foundation	twagner@nsf.gov
Douglas Walker	University of Kansas	jdwalker@ku.edu

Appendix II: Workshop Goals and Action Items Developed

1) **Better communication to ensure future compatibility of products**
    a. This workshop as an example
    b. Formation of consortia for groups with similar objectives
        i. PetDB, GEOROC, NAVDAT have formed the consortia EarthChem (www.earthchem.org) as a first step in developing improved interactivity between these igneous rock databases
    c. Visits between groups to accomplish specific tasks
        i. Within EarthChem to develop a graphical (map and basic plotting) interface for improved visualization during data queries
        ii. NAVDAT – GEON communication during the development of more advanced visualization/mapping efforts in EarthChem
        iii. PetDB will work with MARGINS/RIDGE2K (Bob Arko) on development of OAI (Open Archives Initiative) protocol
        iv. Once OAI is available for PetDB/EarthChem, members of EarthChem will work with GEON on the development of an igneous geochemistry ontology
    d. Annual meeting of database producers held in conjunction with:
        i. Geoinformatics Institute – ESRI meeting August 2004 or alternatively GSA – Denver
        ii. GERM meeting 2005 – LDEO
    e. Link with tool providers to develop best method of tool interaction with databases
        i. MELTS - Mark Ghiorso will work with PetDB
        ii. Energy-constrained, recharge-assimilation-fractional crystallization model (Spera and Bohrson, G3, 2002). Contact to be pursued by Rick Carlson.
        iii. Petrological plotting routines - investigation of generalization of Petroplot routine developed for PetDB

2) **Identification of what maps and tools need to be made available in databases**
    a. Potentially Available Maps
        i. Ocean
            1. Smith and Sandwell 2' map
            2. Ridge/Multibeam synthesis
            3. SIO Explorer multibeam collection
            4. Ocean floor age – University of Sydney - Dietmar Müller
                a. Availability to be explored by Hubert Studigel
        ii. Land Elevation
            1. Gtopo30/SRTM30 – 1km resolution
            2. NASA data/NED
        iii. Land Geology
            1. World map. Source needs to be identified
            2. US/North America
            3. State maps where possible

4. 2° sheets

b. Types of map-based queries
    i. Return Lat, long, precision, age, rock type, sample number, link to sample
    ii. Return some data about sample – major, trace, isotope
    iii. Query by Rock type, age, depth, location, chemistry
        1. What are appropriate values to return for map application?
            a. Shapes of points for database
            b. Color for "value"

c. Implementation
    i. Each database send KL vocabulary used in the databases (rock names, geographic names, tectonic settings, chemical items, etc.) – 10/15/03
    ii. Jeremy Bartley – Maps – 10/25/03 - ISES
    iii. Prototype NAVDAT – 10/26/03 - ISES
    iv. EarthChem Web site to announce implementations
    v. Teleconference on aspects on 11/10/03 to figure out where to go with query portal. Define code/parameters coming to each database. Database then must implement query strategy/code. Each database must provide data back in XML or CSV format on agreed schema.
    vi. PetDB depending on telecon outcome, target implementation of 11/30/03.
    vii. Implement user-interactive symbols/stretch – 1/31/04.
    viii. GEOROC implements when possible.
    ix. Share metadata on coverage availability – Ridge, Margins (3/31/04), SeaMount Catalog.
    x. OAI 5/31/04 – Helly provides code and examples. Databases inform each other on what is going on – which samples, where, what is planned.
    xi. Create metadata set for each layer used 5/31/04.

3) Definition of boundaries to databases to ensure complementarity and minimize overlap

a. Agreement within EarthChem and development of data entry priorities
b. Coordination with EarthRef/GERM (avenues for access to synthesis information: path depends on development in EarthChem. Involvement in definition of data entry metadata scheme. Anthony Koppers will serve as contact and information transfer point. Provide link points from reference information to more in-depth search points in EarthChem. EarthChem can point to more extensive reference information in ERR)
c. Communication with CHRONOS particularly regarding geochronology (Anthony Koppers will serve as contact and information transfer point)
d. PetDB/EarthChem communication with MARGINS/RIDGE efforts (happens all the time between colleagues at Lamont)

e. Development of geochemistry OAI and ontology – Institution ID (Connections with John Helly for OAI implementation. Bob Arko will assist particularly with PetDB OAI development. Chaitan Baru involved in schema development)
f. Coordination with GERM to recover "derivative products" from raw databases
g. Serving geochemistry data to GEON grid
    i. Facilitate interoperability (develop ontology).
    ii. As tools and access members defined, if done as web services (API) it makes it easier for others to enter database for use.
    iii. Standard way of getting into GEON environment is GEON grid (authenticate certificate) – front-ending database in a GEON node.
    iv. Entry at 2-levels (registered/non-registered users)
    v. Watch development of authenticate certificate authority (GEON, NSF)
    vi. GEON assistance in defining common schema for EarthChem (relation to OAI, ontology development, web service on top of PetDB)

4) **Identify course for future data entry protocols**
    a. Define near term data reporting standards for journals.
        i. Started at GERM meeting in 2003, continue at personal meetings of opportunity with editors, put on tutorial for editors present at EarthChem booth at AGU and on EarthChem web site
        ii. Minimal data requirements to be communicated to Editors
            1. Sample location and error (method of locating)
                a. Prepare technical brief about sample identification and Georeferencing technique location criteria including GPS configuration (Doug Walker and John Helley to collaborate)
            2. At least rough age information
            3. Standards for data presentation (units, normalization)
            4. Minimum requirements for data quality documentation
                a. precision, standard analyses
                    i. example as recently defined for Chemical Geology
            5. Delivery of author-produced electronic data files that can be directly input into the databases (no pdf/print conversions)
        iii. Work towards definition of unique sample identifier
            1. Community buy-in will take time
            2. CHRONOS faces similar problem, ask how they are dealing with issue
            3. May be premature item to discuss with editors now – wait for SAMPLES archive to develop
    b. Longer term database - data provider relations

i. Work towards more ideal sample representation files to show the richness of information that could be added to the databases
ii. Establish direct connections with data providers
iii. Work with NSF data reporting requirements to ensure contribution
iv. Deal with publishers to fit criteria for data holder
    1. Documentation of longevity
    2. Establish procedure for how things are submitted to database for incorporation.
    3. Provide URL and data entry requirements for the databases

c. Consideration of data entry forms to be used by data providers
    i. Examination/expansion of Anthony Koppers's data entry form
    ii. How to provide a form to the user that will provide value added for their data management to promote their use of the form and simplify eventual incorporation of data into the databases
d. Devise list of checks to compare with data
    i. Totals for major elements, allowable ranges for a given element, smoothness of normalized trace element plots
e. Work towards standardized units for reporting data by the data provider community, in general the issue of establishing a defined vocabulary
f. Develop tests for data errors at different levels of data inspection
    i. Level 0 data – raw data – simple checks on entry form
    ii. Level 1 data – processed data that can be subjected to statistical tests
    iii. Level 2 data – higher level checking looking of systematic offsets in datasets check for consistency with other datasets
    iv. Use of ontologies as a tool for data checking. Development of, for example, rock name versus composition and comparison with ontologies developed for rock names based on mineralogy and/or texture
g. Make it explicit what is done in error checking and let the criteria evolve
h. Flag data points with quality checkmarks

5) How to pay more attention to Education and Outreach aspect of database use

a. Petrology teaching workshop – use of databases in teaching
b. DLESE case study participant
c. Production of precompiled datasets
    i. Static versus dynamically populated
    ii. Versioning for result checking and verification
    iii. How do you reference it?
        1. Publish query to show date and dataset involved in compilation
        2. Deliver the reference set for the accessed data formatted in the requested journal style
            a. Addresses both documentation of precompiled data set and proper attribution to data providers
    iv. Using precompiled files to serve GERM reservoir purposes

6) **Finding the funding framework to allow database expansion**
    a. Across NSF Division boundaries (OCE, EAR)
        i. Continued interaction of PetDB-NAVDAT-ISES
    b. Across International boundaries
        i. GEOROC to explore German IT development priorities
        ii. GERM to evaluate French interest
    c. Measures of database value to user community
        i. Soliciting commentary from users
        ii. Tracking number of users (needs to move to back-end query tracking once back-end access via OGC or OAI is provided)
        iii. Developing a way for databases to appear as "references" in a paper and hence be tracked by Science Citation Index
        iv. Preparation of broadly used "derivative products"