Workshop on Developing a Database for Metamorphic Petrology (MetPetDB)

20 May 2004
Montreal, Canada

Report of the workshop convener

Frank S. Spear
Rensselaer Polytechnic Institute, Troy, NY 12180

A workshop was held on May 20, 2004 during the annual spring meeting of the American Geophysical Union in Montreal, Canada to discuss the development of a database for metamorphic petrology.

The purpose of the workshop was to bring together a group of active and interested metamorphic petrologists to discuss a number of issues concerning the development of a database for metamorphic petrology. The group was formed shortly following a meeting at the University of Kansas in March, 2003 (Workshop on Cyberinfrastructure (CI) for the Integrated Solid Earth Sciences (ISES), chaired by J. Douglas Walker and Richard Carlson). F. Spear was one of only two metamorphic petrologist at this meeting (the other being Mike Brown, one of the organizing committee members). The recommendations of the ISES-CI Steering Committee included visions and goals for the ISES in the geosciences. With respect to metamorphic petrology, the report stated:

> ***Geochemistry of metamorphic rocks and metamorphic petrology.*** *This effort is starting but has yet to fully organize. This group can establish synergistic relationships with the geochemical efforts mentioned above and structural geology efforts mentioned below. At present, the community must decide upon data standards, reporting information, and just what information is critical. The individual identified for leadership of the working group that will advance this effort is Frank Spear (RPI).*

The primary goal of the workshop, therefore, is to begin development work on a database for metamorphic petrology.

Attendees at the workshop included

Frank Spear (Rensselaer Polytechnic Institute)
Dave Pattison (University of Calgary)
Matt Kohn (University of South Carolina)
Nigel Kelly (University of Edinburgh)
Mike Williams (University of Massachusetts)
Walter Trzcienski (University of Montreal)

John Schumacher (University of Bristol)
Bill Carlson (University of Texas, Austin)
Dave Mogk (University of Montana)
John Brady (Smith College)
Barbara Dutrow (Louisiana State University)
Thomas Foster (University of Iowa)
Darrell Henry (Louisiana State University)
Helen Tomkins (University of Calgary)

In addition to the above members of the working group, invited participants also included:

Tracy Rushmer (University of Vermont: ISES representative)
Kerstin Lehnert (Lamont-Doherty Oceanographic Institute: PETDB representative)
Maria Luisa Crawford (Bryn Mawr: GEON representative)

The workshop began with presentations from Tracy Rushmer, Kerstin Lehnert, and Maria Luisa Crawford. Each invited speaker provided background information on geoscience database and cyberinfrastructure efforts that were going on within each of their respective groups.

Tracy Rushmer (ISES)

ISES (Integrated Solid Earth Sciences) is a grass roots effort formed in order to to facilitate an interdisciplinary, systems-based approach to research and education in solid Earth sciences. The ISES coordinating and steering group is comprised of:

Mike Brown
Art Goldstein
David Mogk
Tracy Rushmer
Basil Tikoff
Ben van der Pluijm

The formation of ISES was triggered by 'Earthscope', which is perceived as incorporating primarily geophysics with little or no geology component. Earthscope is comprised of three components (SAFOD: San Andreas Fault Observatory at Depth; PBO: Plate Boundary Observatory; and USARRAY: a continental scale seismic observatory) the objective of which is to "apply modern observational, analytical and telecommunications technologies to investigate the structure and evolution of the North American continent and the physical processes controlling earthquakes and volcanic eruptions."

The next NSF Earthscope proposal deadline is August 24, 2004 and ISES is encouraging geologists and geochemists to submit proposals. Proposals need to tie in

with the Earthscope mission and the proposed science should be directed towards broadening the impact of the Earthscope mission.

The ISES steering committee and members are working on several fronts:

- Organizing workshops and forums aimed at developing mechanisms to synthesize and integrate across disciplines in the geosciences. Several forums have been held, and more are planned. Forums are timed to precede major meetings (e.g. GSA or fall AGU).
  - Past forums include
    - Geoinformatics and cyberinfrastructure (ISES-CI)
      - database building and tools
    - High precision geochronology (ISES-CHRON)
      - techniques and infrastructure
    - Numerical modelling of lithospheric processes (ISES-CALC)
  - And the next ISES forum will be
    - Rheology (ISES-CE continental evolution)
      - Fall AGU meeting in San Francisco
      - Emphasis will be on rock materials, rheologic behavior and the derivation of constitutive relations applicable to earth materials.
- Organizing workshops aimed at developing a cyberinfrastructure (e.g. databases) in geoscience disciplines.
  - This effort is underway on many fronts (e.g. the MetPetDB workshop; CHRONOS; Earthchem) and is going well.
- Supporting initiatives to integrate research equipment facilities and to provide access to these facilities. Examples include the ISES webpage, a cross country survey of facilities, and the status and availability of geochronology laboratories.
  - ISES Website:
    - facilitate communication and build community
    - currently in the works (depending on funding)
    - housed at Carlton College; Basil Tikoff
- Educating and mentoring the next generation of solid earth science researchers (e.g. graduate students) and organizing summer schools
  - Annual Summer Schools:
    - aimed at junior researchers
    - one week duration
    - anticipated summer schools:
      - rheology
      - geochronology
      - geoinformatics
    - This effort is not going so well because of difficulties in getting funding
- A major problem faced by ISES is that the amount of funding for specific endeavors uncertain

Kerstin Lehnert (PetDB)

- Important URLs:
  - PetDB: (http://beta.petdb.ciesin.columbia.edu/index.jsp)
  - Earthchem: (http://www.earthchem.org/)
- Goals of building a geochemical database:
  - Collecting and federating heterogeneous data
  - Data-sharing and use
  - Data mining and knowledge discovery
  - Visualisation
  - Data preservation
- Steps to creating database
  - Build an archive
  - Serve data – make them accessible
  - Allow people to retrieve only the data they need
  - How to define, catalogue data, etc
  - Develop tools to facilitate data analysis
  - Integrate with other data systems
- Problems in creating database:
  - Clashing of opinions, goals, and perspectives
  - Need to actually start getting the data together
  - Data is dispersed in literature, often not in electronic form
  - Compilations by investigators are time consuming, redundant, often incomplete
  - Missing links among related data
  - Data is lost due to incomplete publication
    - Getting better due to electronic supplements
  - Data is not properly documented

- Earthchem
  - Earthchem mission: Advancing data management in solid earth geochemistry
  - Three components to EarthChem
    - PetDB:
      - Oceanic crust generated at MORs (includes BAB, seamounts, old oceanic crust)
      - Only 'raw' chemical data for rocks, minerals and melt inclusions
    - Georoc:
      - Oceanic islands
      - Convergent margins
      - Large igneous provinces
      - PreCambrian rocks
    - Navdat:
      - Cenozoic igneous rocks from western North Merica
      - Expanding to Canada and Mexico
      - Lots of geochronology data

- All three databases work together.
- There is a wide range of metadata associated with and essential for the analysis included in each database, necessary for user to evaluate data and select the data they want
- Biggest challenge is getting data into database – not only data but also all the extraneous metadata
    - PetDB working with publishers to have a format for data to improve ease of submission of data to database – including enforcing submission of all the relevant metadata (e.g. values plus precision at the least)
    - Funding is a problem – very effort intensive, but need to have people compiling the data who can read papers, understand data, etc.
- For the three databases: chemical analyses for 200,000 samples
- 3 systems have interactive, web-based interfaces
    - allow users to select, filter, view etc data
    - still no way for users to input their data – will be a proposal from Earthchem
    - For example, in PetDB a user may start by selecting a parameter such as
        - Geography
        - Ship/cruise
        - Analytical type
        - Rock type
        - Mineral analyses
- Need to address interoperability of databases
    - agree on common vocabularies
- Sample ID Issues:
    - same names used for different samples
    - one sample has lots of different names in different publications
    - create an international Geo sample number 'ISGN' (like ISBN for books)
    - Need unique ID for samples
        - Solid Earth Sample Registry (SESAR)
- In the future for PetDB:
    - Hold workshops on use and development
    - Improve interoperability among databases
    - Exhibits at conferences (e.g. fall AGU)
    - Add information to databases
        - Pictures of hand specimens
        - Thin section images

Maria Luisa Crawford (Geon)

What is GEON? www.geogrid.org

- GEON is 8 years old
  - US institutions and GSC
  - IT people in Geon – San Diego supercomputer centre
- Mission is to develop technology to advance science
- Geology has 'things' + space + time
  - Interesting IT problem
- Geon is not a database – rather it setup to manage, analyze and mind information from complex databases
- Integrating all the other databases
  - Geochron
  - Geochem
  - Geophysics etc.
- IT Issues:
  - To make data available:
  - Understand and align database formats
  - Capture concepts
  - Semantic mediating – linking diverse data
  - Relies on concept spaces that describe relevant concepts and their relationships
- Ontologies – formal machine processable counterparts of concept spaces
  - need geoscientists to tell the IT people what the concepts are
  - Need to develop equivalent terms and relevant terms (thesauri)
  - Will permit the computer to recognize, for example, that enstatite and opx are the same thing
- Geon will permit questions, search for patterns, identify anomalies, integrate data across disciplines
  - e.g. integrate maps across the US – ask questions related to maps
- Geon is independent of databases – above all these – will access all of them

• Metamorphic Rock Ontologies:

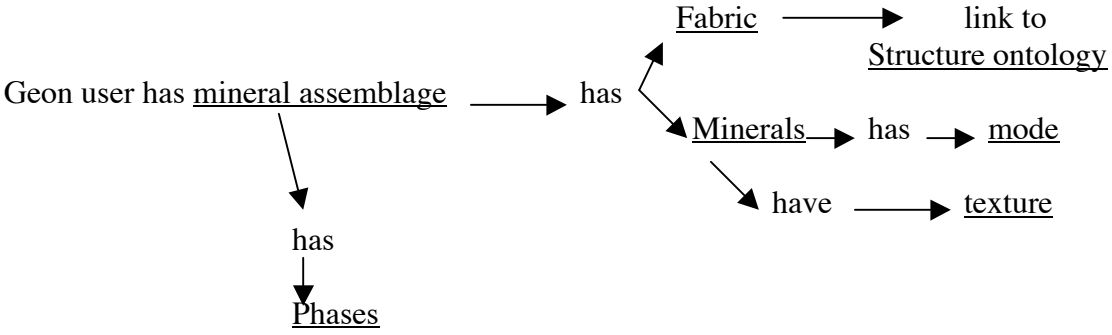
Metamorphic Rocks	*Metamorphic Processes*	*Metamorphic Grade*
Composition	P and T	
Structure	fluids and gases	
Classification	time	
Mineral assemblages	chemical reactions	
Mineral paragenesis		

Data sources

- Minerals
- Texture
- Fabric
- Geometry

- Geochemistry
- Experiments
- Isotopes
- Etc

The ontology establishes not only the vocabulary, but the relationships among the vocabulary. For example:



- Idea is to use Geon experience to design what we need to include in the MetPetDB
  - Geon is not making databases, they are deciding what everyone needs to know to answer a particular question about something specific by accessing all the separate databases that are linked through Geon.
  - Sample labeling again important – combining different databases will mean that Geon may have to have their own sample identifier.
  - When updating database, have to check newly submitted data against pre-existing data – e.g. new analyses on a sample already described in the database.
  - Therefore needs to be a level of minimum information on each sample or analysis
    - Evaluation of terms and meanings:
    - Agree on terminology and boundaries
    - Knowledge changes, therefore ontologies have to keep up with this
  - Computer programs and other tools
    - Needed to link to databases
    - e.g. GIBBS, Thermocalc, Perplex, etc. to evaluate data from databases
    - Geon can house programs, or can provide links to individual webpages
    - Eventually would like Geon to be able to use the programs to answer questions

Afternoon Session

The afternoon session focused on issues specific to the development of a metamorphic petrology database. Some local efforts are already underway, and the group was brought up to date through informal presentations by Dave Mogk, Tom Foster, and Mike Williams:

David Mogk

- Discussed ongoing efforts to develop Digital libraries
  - Think about end-users
  - Classroom use, undergrads
- What is the data?
  - Broad or narrow view?
  - Analyses?
  - Images?
  - Model output?
- Need to subset data
  - Users need to be able to do as they require
- Data is only half the need
  - Toolkit is also required
- Derivative products – e.g. image galleries

Thomas Foster

- Presentation of a metamorphic textures website
  - X-ray maps
    - By date
    - By area
  - Can integrate maps with analyses (GIS)
  - Integrate hand samples with SEM images, analyses, x-ray maps, etc
    - BSE images and data points
  - Gives textural context to published (and other) data, from thin section, mineral scale to outcrop and regional scale
  - Can classify zoning patterns
    - Subjectivity
    - Establish a protocol around this?
  - Images are large – store on a central server or give web address for images on individual servers?
    - Presumably there'll be the funding and set up to have our own servers
  - Sample archiving?
    - Existing samples from areas that can no longer be accessed, or people retire – preserving sample sets

Mike Williams

- Presentation of a tool for the analysis of age data:
  - Catalogue everything relative to a full thin section map
    - E.g. monazite grains in a thin section
    - Has an x,y grid
  - Grain textures (zcn, mz, etc) relative to dates is essential
    - E.g. are they inclusions? In what?
  - Fundamental petrological data
    - People can re-interpret data if they want, but have to have all info there
    - Ages are meaningless without context

General Discussion:

The remainder of the workshop was devoted to a general discussion in which participants outlined specific requirements of a metamorphic petrology database, and formulated a plan for its implementation.

Questions that were addressed in the afternoon session

- Objectives
    - Who is the audience?
    - What do we want the database to do?
        - What types of queries are expected?
- Scope of the database (what data will be stored?)
- Data types
    - Raw data versus derivative data
- Database toolkit
    - What is needed?
- Database development and management
- Where should we go from here?
    - A plan for the future
    - How do we ensure community buy-in?

Objectives

- What is our audience?
    - Potential audience includes
        - Undergrads
        - Graduate students
        - Researchers
        - General geologists
        - Lay people
    - It was decided that the initial focus should be on the development of a metamorphic database as a research tool
    - What do we want the data base to do?
        - What types of queries are expected?
            - There was considerable discussion about this point during the afternoon, and there was general agreement about desired capabilities.
                - We want a database system to manage our own data because the amounts of data collected by a single researcher is growing rapidly and becoming increasingly difficult to manage.
                - We want to be able to access quickly and thoroughly the available data from a particular area including data on metamorphic rocks/assemblages, geochronology, structural data, geophysical data, and any other useful

information. The metamorphic database will therefore have to link to other databases.
    - We want to be able to query the database with queries about mineral assemblages, metamorphic grade, mineral chemistry, etc.
  - It was decided that this question should be asked of the metamorphic and geologic community. The polling will be one of our near future projects (see below).

Scope of the database

- How large should the database be?
  - Considering the growth of computational and storage capacity, it was decided that we should not worry about the size of the database.
- Sources of data to include
  - Do we only catalogue published data?
    - Lots of unpublished, good, useful data
  - The general consensus was that
    - Published data should be included as it appeared in the publication
    - Unpublished data should be included
  - The strategy for published data is relatively straightforward because the amount/types/scope of published data is limited
  - The strategy for unpublished data is not as clear because the types of data are evolving
    - The database should provide tools for analyzing and manipulating data
    - The database should provide accessibility to data by collaborators
    - It should provide accessibility by other researchers when the data are made "public"
    - Need to flag unpublished/non-peer reviewed data
- It is important to consider what we will need in 5, 20 and 50 years hence.
- MetPetDB can be used in place of a supplemental data table in publications, making all data used in a study available for others to examine and use.

Data types

- Two general types of data that the group considered are raw data and interpretive data.
  - Raw data
    - E.g. textural descriptions, mineral assemblages, microprobe data, sample images, etc.
  - Interpretive (or derivative) data
    - E.g. metamorphic grade (or facies), P-T conditions, P-T paths, fluid-rock ratios, etc.

- Both types of data should be included in a database, but require somewhat different considerations.

Raw data

The group spent considerable time discussing what types of raw data should be included in a database. It was decided that the basic unit should be the sample, and the following is the list of important items the group came up with.

Sample information

- Sample ID
- Outcrop information
    - Location (lat – long)
    - Position within outcrop
    - Images of outcrop
- Orientation of sample within outcrop (for oriented samples)
- Minerals in hand sample
- Whole rock chemistry
    - Data type
        - Technique
            - Provenance
- Texture/fabric
    - Images of texture
- Location (curator)
- Rock type (this might be can of worms)
- Date of collection
- Collector
- Field area (region)
- Comments on the sample
- Thin section (a sample may have more than one thin section)
    - Mineral assemblage
    - Minerals (for each we have)
        - Name
        - Chemical data
            - Location of analysis
                - X-Y-Z coordinates in thin section or image reference frame
                - Context of analysis (e.g. core, rim, etc.)
            - Data type
                - Microprobe, ICP-MS, SIMS, etc.
                    - Provenance
                    - Actual data
        - Texture/fabric/habit of mineral in thin section
        - Image of
            - Scale

            - Location
            - Types
                - BSE etc.
        - Mode
        - Inclusion(s)
    - Images
        - Scale
        - Location within thin section
        - Types
            - BSE etc.

Minimum dataset for a sample

It was decided that the minimum amount of information required for a sample to be entered in the database would be

- Location (at least latitude and longgitude)
- Sample ID
- Mineral assemblage

Outstanding questions regarding raw data

- How "raw" should "raw data" be?
    - Reduced microprobe data or
    - Raw counts (peak, background, standards and unknowns)
    - It would be impossible to include raw counts for published microprobe data
    - For new data, it might be possible to include raw counts, but there was no consensus that this would be either necessary or wise.
    - However, the database should permit incorporation of raw counts data for the future.

Interpretive data (derivative products)

It is recognized that many users of a metamorphic petrology database may not be metamorphic petrologists and will be more interested in the interpretations of the data than in the raw data themselves (e.g. for tectonic syntheses).
Interpretive data include

- P-T points (peak)
- P-T paths
- Crystallization ages
- Metamorphic maps
- Mineral assemblages
- Fluid/rock ratios
- Tectonic setting

- Cooling rates
- Exhumation rate
- Deformation style
- Metamorphic grade/facies

Important considerations with respect to interpretive data

- Caveats associated with derivative or interpretive data
    - E.g. Quality assurance
    - E.g. Activity models used for PT work
    - Thermometers used
- Need to be able to look at derivative data and then, if need be, go back to raw data and reexamine
- Tools and/or tutorials should be available to help people make decisions about good versus bad data
- There should be links to source of data
    - E.g. references, laboratory where data were collected

Toolkit

There are many tools (programs) in use by petrologists for the analysis of data, and many of these can be linked directly to the database. These include

1. Recalculation of raw microprobe data, mineral formula calculation
2. Plotting/graphics routines
    a. Ternary, etc. diagrams
    b. Projections
3. Tabulating routines
4. Image analysis
5. Thermobarometry
6. Thermochronometry
7. Linked to GIS – geologic maps, digital topography, etc.
8. Thermodynamic calculations
    a. Pseudosections
    b. Phase diagram calculation
    c. Etc.
9. Annotation tools

Some of these programs have been written by researchers and some are available commercially or as freeware

Additionally, two tools are considered to be absolutely critical for the successful development of a metamorphic database:

- Import/Export tools so that data can be uploaded from existing formats (e.g. spreadsheets) and downloaded for use in other programs

- A tool for linking chemical data to spatial information (e.g. images) and then provides the means to display the data.
    - Mike Williams demonstrated a version of his program to link, display, and manipulate data. While not currently linked to a database, this program appeared to offer many of the desired capabilities.

Database development and management

These topics were discussed briefly. There was concern that the programming expertise and the resources to manage a large scale geologic database are not available. This is an important issue to resolve before significant development can occur and will surely require an integrated effort with other database working groups.

Tasks for the Workshop Group:

- Develop a format for data submission
    - Everyone to submit one sample in the agreed-upon format
    - Analyze results and see if the model works
- Construct a MetPetDB website
    - Provide details of the mission of the working group.
    - Provide links to other efforts such as Geon, ISES, Earthchem, PetDB etc.
    - Darrell Henry to start on website
- Generate an E-mail list of all interested geoscientists so they can be kept up to date on developments
- Develop a prototype database for testing
- Query users for desires
    - Request from all interested petrologists that they submit an example of a database query that they would like to see answered.
    - Post the request on the Geometamorphism and MSA list servers
    - Require a detailed paragraph response
- Plan the next workshop
    - Fall GSA or AGU
    - Have a start made on all the above for next workshop