


Data Curation Profile – Geophysics and Seismology/Structural Geology and Neotectonics

Profile Author	Lori Tschirhart	
Author's Institution	University of Michigan	
Contact	ltz@umich.edu	
Researcher(s) Interviewed	[Name withheld]	
Researcher's Institution	University of Michigan	
Date of Creation	September 20, 2012	
Date of Last Update		
Version of the Tool	1.0	
Version of the Content	1.0	
Discipline / Sub-Discipline	Geophysics and Seismology / Structural Geology and Neotectonics	
Sources of Information	<ul style="list-style-type: none"> • Interview with researcher, March 30, 2012 • Follow-up interview with researcher, May 23, 2012 • Data Curation Profile Interview Worksheet completed by researcher • Researcher's Google Site Data portal: [URL withheld] • Publications describing the research related to this profile [URLs withheld] • Researcher's Google Site Data portal: [URL withheld] • Researcher's departmental profile: [URL withheld] 	
Notes		
URL	http://www.datacurationprofiles.org	
Licensing	This work is licensed under a Creative Commons Attribution 3.0 Unported License . (CC BY 3.0)	

Section 1 - Brief summary of data curation needs

The primary data for deposit include geo-spatial data, geochemical and thermochronologic analyses, and physical specimens for rock samples obtained in the Greater Caucasus mountains. The data were produced and/or obtained for a National Science Foundation-funded project in the Greater Caucasus mountains, where rock samples were collected over four field seasons.

The NSF funded this research before a data-sharing mandate was implemented, but the researcher will need infrastructure to meet data management needs for upcoming projects. Since this work is at a mature stage, the researcher is interested in using this project as a test for how data will be managed for future grants.

The data file includes GIS databases of sample locations and geologic information, digital field photographs, analytical data from samples (thermochronologic, geochemistry), home-grown computer software, model files, sample storage and processing information, purchased map datasets, and a list of physical samples listing details such as sample number, brief description, type of rock, environmental setting, information about current storage location and any analysis performed.

A data file of geo-spatial data and geochemical analyses is maintained locally, but over 600lbs. of rocks must be put somewhere, archived, and kept track of, and their location made public for any other researchers who might want to use them. An existing NSF mandate requires geologists to archive all physical samples in perpetuity.

The researcher wants a supported way to openly share most data (especially physical samples and raw analytical data), connect the data to publications and other scholarly outputs, and make data discoverable by researchers in the field (via search engines and other means).

Occasionally, complex copyright questions inhibit the open sharing of some ancillary data. Where complex legal and ethical issues exist, the researcher needs a flexible infrastructure that allows such materials to be shared selectively (for example, with researchers within his university, as needed and as law allows).

Section 2 - Overview of the research

2.1 - Research area focus

The researcher is interested in plate tectonics, dynamic earth processes that cause continental plate interior deformations, and plate motion distribution. Field observations, geologic mapping, thermochronometry, structural analysis, stratigraphic studies, and collaborations with other earth science specialists inform his work.

For this project, the researcher and his graduate student used low-temperature thermochronometry to investigate the timing, rates and onset of deformation within the Greater Caucasus mountains to better understand their contributions to deformation across the Arabia-Eurasia orogen. Thermochronometric sample analysis and modeling is performed to glean thermal histories that inform this investigation.

2.2 - Intended audiences

Other researchers in the field, practicing professionals in the oil and gas industries, and hazard assessment professionals would be the primary audience for this data.

The data may have value for those interested in understanding the tectonics of mountain belts, those involved in oil and gas exploration and thermal evolution of the Caucasus, and those involved with probabilistic seismic hazard assessment.

2.3 - Funding sources

National Science Foundation (NSF) (small grant)
Civilian Research and Development Fund (provided pilot funding)

Section 3 - Data kinds and stages

3.1 - Data narrative

Data collection for this research project is iterative. Data exists in various stages continually.

Thermochronometric analysis requires samples for analysis, and the initial data stage requires the collection and indexing of raw samples for later analysis. Samples are collected by hand in the field using hammers and chisels. Information such as latitude, longitude, elevation, sample site notes, and digital photos are recorded at this stage. This information is recorded using notebooks, tablets, and/or specialized GPS devices that run GIS and word processing. 100s of data files exist at this stage, with each file averaging between 1-10MB in size. Data are represented in ArcGIS (.shp), .txt, .doc, and .jpg file formats. Geologic maps are sometimes purchased and used in this stage.

The second data stage involves tracking the sample and collecting analytical data. Each sample has its own time-temperature history. Some analysis is done locally and some samples are sent to other labs around the country for analysis. Thermochronology techniques are used at this stage to measure the radioactivity of the samples. Known amounts of parent and daughter elements present in each sample serve as temperature-sensitive geologic clocks that reveal the time a clock was started at a particular temperature. This analysis is used to understand the speed at which the sample rock moved toward the surface of the Earth. The speed allows the researcher to understand the rates of mountain building and mountain growth. Selected rock samples are physically disaggregated, photographed, and analyzed. 100s of new data files are created at this secondary stage, mostly .txt files describing the analysis that has been conducted on each sample. This analysis will reveal mineral composition, information about rings, and information about the types of processing performed on each sample. Each file averages a size less than 1MB. Data in the second stage can exist in previously listed file formats (.shp, .jpg, .txt), but may also exist as .xls, MySQL, and proprietary data file formats.

The third data stage occurs when the data is used for modeling and interpretation. Multiple data points for a mountain range are collected to come up with a single, consistent, geologic model that explains all these data. Tectonic or geologic scenarios are imposed and compared to the data to find the best interpretations. A modeling scheme called “inversion” is used to create a thermokinetic model. The model specifies how Earth will deform. 100s to 1000s of small data files exist at this stage, with each file averaging a size less than 1MB and probably less than 100 KB in size. The file formats may include .txt, .xls, .ps, .pdf, .ai, .Rdata, .py, and Pecube files. Text files represent the model, Excel files show model results, and graphics files display the results visually. Statistical software R creates the Rdata files.

Finally, the fourth data stage involves publication and dissemination. 10s of data files are created at this stage. Each data file has an average size of up to 20MB. These files are represented in .xls, .docx, .pdf, .tex, .ai, .eps, and .jpg file types and include research articles and published data sets. GSA and AGU provide repositories for storing data sets related to their associated research articles.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Sample collection and indexing	Rock samples, Field note entries	100s / 1MB-10MB	ArcGIS (.shp), .txt, .doc, .jpg	Samples are collected by hand in the field using hammers and chisels. Sample site information and digital photos are recorded using notebooks, tablets, and/or specialized GPS devices that run GIS and word processing.
Tracking the sample and collecting analytical data	Time-temperature history of each analyzed sample	100s / <1MB	.shp, .jpg, .txt, .xls, MySQL, proprietary data files	Selected rock samples are physically disaggregated, photographed, and analyzed. Some samples are never analyzed, some are analyzed once, and some are analyzed many times.
Modeling and interpretation	Thermokinetic models and interpretation	100s-1000s / <1MB (and probably less than 100KB)	.txt, .xls, .ps, .pdf, .ai, .Rdata, .py, and Pecube files	Text files represent the model, Excel files show model results, graphics files display the results visually.
Publication and dissemination	Research articles, data sets, and other publications	10s / up to 20MB	.xls, .docx, .pdf, .tex, .ai, .eps, .jpg	
Ancillary Data				
Ancillary Data #1	Geologic maps			

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray (the “processed” row is shaded here as an example).

3.3. - Target data for sharing

The researcher would share most post-publication publicly. On average, a time period of 2-3 years passes between sample collection and publication.

Collected samples (first stage) would be shared initially with fellow researchers prior to publication, but would be shared with anyone post-publication. A NSF mandate requires geologists to archive all physical samples in perpetuity.

Analytical data (second stage) would also be shared initially with fellow researchers prior to publication, but would be shared with anyone post-publication.

Models and interpretation thereof (third stage) would be shared with immediate collaborators pre-publication. Post-publication, the researcher would share with experts in the field, but not globally, because data could be misinterpreted.

Publications (fourth stage) would be shared with anyone post-publication.

3.4 - Value of the data

Data has potential value to professionals in the researcher's field. The data may further understanding of the tectonics of mountain belts and of the thermal evolution of the Greater Caucasus mountains.

The data has potential value for practicing professionals in the oil and gas communities related to the exploration for oil and gas.

Hazard assessment professionals may value the data for probabilistic seismic hazard assessment.

3.5 - Contextual narrative

Where the researcher would share data, he would do so under the condition of acknowledgment.

Disciplinary publishers sometimes provide data repositories. The researcher states that GSA (Geological Society of America) has been providing a system to relate data sets to research articles for a very long time and that they are one of the best and most flexible data repositories. Their repository allows researchers to contribute any file type to their data repository. AGU (American Geophysical Union) began as a repository for machine-readable files related to seismic data, but in the last couple of years they have become more flexible.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

The researcher identifies the PI (principal investigator) and the University of Michigan as owners of the data generated by this research project.

4.2 - Stakeholders

NSF (National Science Foundation)
CRDF (Civilian Research and Development Fund)

4.3 - Terms of use (conditions for access and (re)use)

Requirement details of funding agencies can be nebulous. NSF requires a written data management plan as a condition of funding. These plans should describe the kind of data researchers anticipate collecting, how the data will be curated, and how the data will be distributed for re-use. This applies to everything from digital data to actual, physical samples. In particular, NSF wants the researcher to deposit data into accessible repositories, and they want data to be kept for some length of time.

The data set related to this research is not bound by any privacy or confidentiality concerns.

4.4 – Attribution

A requirement that others cite this data set if they were to use it in their research is a high priority for the researcher. Citing relevant publications and NSF support for the project are conditions for sharing the data.

Section 5 - Organization and description of data (incl. metadata)

5.1 - Overview of data organization and description (metadata)

A file structure is established on a data server.

The researcher considers the ability to make the data accessible in multiple formats a low priority. Users of this data set will likely be using the same formats. In instances where different formats are required, data is presented in a way that allows users to transform it into any necessary format. For example, much of the data is represented in Ascii text files, which data users should be able to manipulate to meet their format needs. In cases where data is presented in proprietary formats, potential data users would be using the same software.

Currently, the researcher places a medium priority on the ability to apply standardized metadata from his field or discipline to the data set. No metadata standards exist yet for analytical data sets. People are working to develop those standards now so his answer may change with the availability of standards. For the time being, the researcher would only want the ability to apply metadata standards to data such as geographic data where samples have been collected and analysis results of those samples.

5.2 - Formal standards used

KML metadata has been applied to geographic data.

5.3 - Locally developed standards

Not discussed by the researcher.

5.4 - Crosswalks

Not discussed by the researcher.

5.5 - Documentation of data organization/description

A file structure is established on a data server. The file structure used to organize and describe data is described in field and lab notebooks.

Section 6 - Ingest / Transfer

Before data can be ingested into a repository or otherwise transferred out of the researcher's direct control for curation purposes, the data files need to be cleaned up, the file structure needs to be organized, and an archival format must be picked.

The researcher places a high priority on the ability to submit data to a repository himself because it allows an opportunity to perform quality control on data before uploading. A medium priority is placed on the ability to batch upload this data into a repository. The researcher would prefer to upload all files at once using a zip file. This ability is a higher priority because it creates a system

that is easy to use and lowers barrier to access. The researcher mentioned that his graduate student produced 153 GB for this research, including huge, high-resolution geologic maps, and that this file volume is not uncommon for this type of research.

The researcher places no priority on a process that would automate submission to repositories because the process removes an opportunity for the researcher to do a quality control check on the data before submission.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

The researcher shares his data widely, and much is available to everyone post-publication. The researcher is motivated to share data because there is an impact on the way science is done when research data gets connected with publications and other outputs.

7.2 - Embargo

The researcher does not require embargos, but makes most data available after publication. He would consider depositing data pre-publication if a sufficient embargo period was provided.

7.3 - Access control

The ability to restrict access to the data set to authorized individuals is a low priority for the researcher. Most data is shared openly. Since third stage data can be misinterpreted, we would only share that post-publication data with experts.

7.4 - Secondary (Mirror) site

The ability to access the data set at a mirror site is a low priority for the researcher.

Section 8 - Discovery

The researcher understands that most data sets are discovered by Internet search or by references in published article that refer to a repository. Therefore, the ability to easily discover this data set using Internet search engines is a high priority.

Likewise, the ability for researchers within the discipline to easily find the data set is a high priority. The ability for researchers from outside the discipline to easily find the dataset is a medium priority, and the ability of the general public to easily find the data set is a low priority for the researcher.

The researcher suggests that beyond repositories, the institutional library home page, the institutional home page, or home pages from departments within the institution may be useful places to highlight institutionally produced data sets and foster data set discovery. If departmental web pages devoted space to a research section or to institutional repository links, so that potential data users could link from the departmental web page to researchers to their data sets, that would foster discovery. If space were devoted to data set discovery on the institution's home page - even in broad categories such as science and medicine – it would allow people to discover data that way.

However, it is likely that for the kind of data we are discussing, the people seeking it already know what they want to find. It's unlikely that people who discover this data while looking for something else will find it to be useful to them.

The researcher also stresses that departments must know how to provide links to repository data from their own departmental web pages.

Finally, the researcher wants to use repositories that are open access, that have robust search capabilities, and that are indexed to enable easy discovery of data. The researcher describes a repository (unnamed) that provides good geospatial indexing. Repository users can draw a polygon on a map and query the database to learn what data is provided in the region represented by the polygon. Users can also search against a name in the United States Geographic Name and Information Service. For example, a researcher could search White Mountains, California and the resource will know the physiographic region and return data available for that region.

The researcher refers to North American Volcanic Database (www.navdat.org) and the National Geologic Map Database (ngmdb.usgs.gov) as two other attractive repository models and lists features that make the repositories impressive.

NavDat users can obtain all kinds of volcanic rock information including analytical data. The database allows search by name, by map, and by age of rock. Search results are presented as a map of the data that the resource provides.

NGMDB indexes every US geologic map ever produced, almost every state geologic map ever produced, and many unpublished theses maps. The database also allows search by names of places, and users can search on a map. Information about scale and type are provided in search results, and a KML overlay that will allow users to search graphically in Google Earth is in development.

Section 9 - Tools

The following tools are used to generate the data:

- Mass spectrometers are used to provide raw measurements of isotopes that researchers turn into age information.
- Cluster computer nodes are used to run models that predict the ages of the surface of the Earth.
- LabView is the code that takes isotopic data from the mass spectrometer and turns it into an age.
- HeFTy and QTQt are two thermal modeling software programs specific to this sub-field of geology. They use different methods to model the time/temperature path of analyzed minerals. Comparisons are made between the models produced with each method.
- ArcGIS is used in sample location plotting and visualization. It allows for spatial and integrational visualization with digital geologic maps and topography.
- Python is computer code. The researcher's graduate student wrote computer code for mountain belt evolution models in Python.
- R is statistical software that analyzes output from Python models.

Potential users of the data will require computers, Pecube (geology modeling software), LabView, HeFTy, QTQt, local software produced by researchers involved in the work, Python, R, and Excel to look at the raw data.

The researcher uses Google Code Repository (code.google.com) to upload code, keep track of versions, and let others access, download, and use the code.

Section 10 – Linking / Interoperability

The ability to connect researcher data with publications or other outputs is a high priority because it has an impact on the way science is done. The researcher wants his data connected to and referenced by the publication in which it is described. In particular, if people reuse this data for their own studies, they must know how to reference, cite, and connect it to the publication in which it was described.

The ability to support the use of web services APIs is a medium priority. Some people would be able to connect geographic data and display it in Google Earth or other visualization software along with other geographic data sets from the same region.

The ability to connect or merge data with other data sets is a medium priority due to a lack of accepted standard metadata that could be used for the distribution and description of this kind of data, but one is likely coming. When it does, this will become a higher priority.

Section 11 - Measuring Impact

11.1 - Usage statistics and other identified metrics

The ability to see usage statistics on how many people have accessed the data is a medium priority. The researcher has a moderate interest in knowing how often the data set is being accessed and used, but compared to many other data sets, he anticipates that usage and access is going to be very low, so bandwidth and server issues are not anticipated.

11.2 - Gathering information about users

The researcher puts a low priority on the ability to gather information about the people who have accessed or made use of this data. The sub-discipline is a small community and the researcher can probably guess who makes use of the data, but there is little need to track which individuals are using the data. The researcher has no concerns about the data being exploited for commercial purposes because it has little use outside of academic research. However, if the data were used for commercial purposes, he would want to know who was using the data.

Section 12 – Data Management

The researcher uses a RAID data server, Vice Versa Pro, and code.google.com to manage the data.

12.1 - Security / Back-ups

The data is stored locally on a secure password-protected data server. The researcher does not want the server to hold both public and restricted materials on the same server for security reasons.

The researcher currently makes back-up copies of his data. Incremental back-ups are performed nightly and full back-ups are performed monthly.

The researcher has shared finalized data with repositories in the past. He would prefer that data stored in repositories be backed up monthly.

12.2 – Secondary storage sites

There is a need to have the data secondarily stored on a public-facing database, and the researcher would like the local institutional repository to meet this need. To become a viable option, the local IR must become more robust to accommodate the needs of final repository to make them accessible to the public, easily found within search engines, and migrated to new data formats where appropriate.

12.3 – Version control

The ability to enable version control is a medium priority. Though it would be helpful, it is also available via Google's code repository. The software written to support the research is evolving and exists in several versions. For much of the raw analytical data, there is no version control.

Section 13 - Preservation

Physical samples and raw analytical data are the most important data to preserve, manage, and maintain over time. All other data associated with this research could be re-generated. However, most data users would want an interpretation of the raw data.

13.1 - Duration of preservation

Were it preserved, the data set would have value for the researcher and others for an indefinite amount of time. Physical samples must be curated in perpetuity to meet NSF funding requirements. They can always be reused.

13.2 - Data provenance

Documentation of any and all changes that were made to the data set over time is a high priority for the researcher. Transparency is important, especially if mistakes have been corrected.

13.3 - Data audits

The ability to audit this data set to ensure its structural integrity over time is a high priority for the researcher so that regular checks could be conducted for corrupted files and corrupted disc segments.

13.4 - Format migration

The ability to migrate data sets into new formats over time is a medium priority because formats do change.

13.5- Secondary storage site in a different geographic location

Secondary storage is a medium priority. Much effort went into collecting the data. Having two storage sites is a redundancy built into departmental practice. One set of hard drives and one set of tapes are stored in the researcher's building. A second set of hard drives and a second set of

tapes are stored at a second campus building. The researcher assumes a secondary site would be in a different locality. It is not a major concern, but could be for different regions.

Section 14 – Personnel

14.1 - Primary data contact (data author or designate)

[Name Withheld]

14.2 - Data steward (ex. library / archive personnel)

[Name Withheld]

14.3 - Campus IT contact

[Name Withheld]

14.4 - Other contacts

[Name Withheld]