Data Curation Profile – Carbonate Sedimentology

Profile Author	M. Cragin
Profile Author	M. Kogan
Institution Name	UIUC
Contact	M. Cragin (cragin@illinois.edu)
Date of Creation	October 29, 2009
Date of Last Update	
Version	
Discipline / Sub-Discipline	
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline. Data Curation Profiles employ a standardized set of fields to enable comparison; however, they are designed to be flexible enough for use in any domain or discipline.
Context	A profile is based on the reported needs and preferences for these data. They may be derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information used for this profile	• Interview with scientist (date) •. Follow-up interview with scientist (data) • Data Needs Checklist
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
Author's Note	
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284315001

Brief summary of data curation needs

The data set for deposit would be a package consisting of two-three Excel spreadsheets, annotated photos and microscopy images, and possibly additional files of contextual information. It is anticipated that the data in the spreadsheet has increased value when the photos are retained and paired with this refined observational data (i.e. numeric values recorded for multiple variables, or “parameters” of rock, water chemistry, and microbial data). As such, the files in this data set package would need to be linked together, and these links would need to be maintained over time. Data would be submitted to a repository upon publication of related paper(s). Use of the data set requires attribution.

Overview of the research

Research area focus

As noted above, this research concerns geobiology, which brings together geological, hydrochemical, biological/genomic (microbial and multi-cellular organism activity), and atmospheric data to analyze the impact of microbial activity on carbonization. Data are collected to represent system interactions at various scales of analysis. Data collection occurs at hot springs and coral reef locations.

Intended audiences

In addition to other geobiologists, scientists who study hot springs and coral reefs will also have an interest in these data. The U.S. Park Service is interested in these data, and has a particular interest in location data, and abstracted observation data for education programs in the Parks. Beyond this, there are many audiences interested in the field photos and microscopy images generated by this research; these groups include microbiologists, geologists, chemists, physicists, medical doctors, and architects and educators.

Funding sources

The funding sources for this geobiologist range from governmental institutions like NSF to the natural resources extraction industry, which include petroleum companies. The funding for the data covered in this profile is funded by the NSF.

Data kinds and stages

Data narrative

Geobiology data include four major aspects: geological, atmospheric, biological (genomic) and water chemistry data, making for an aggregate of data sets generated by sensors and human observations. The focus here is specifically on the geological and water chemistry components a much larger aggregate of data sets, which includes other types of observational data (e.g. atmospheric or hydrology data) and microbial genomic data. It was indicated that the total data set size would increase dramatically if the genomic data was factored in.

For the geological and water chemistry research, much of the data are collected by automated sensors; the sensor data are directly uploaded into a Microsoft Excel spreadsheet.

The initial “raw data”, or Replicate Spreadsheet is very large as it contains multiple (i.e. “replicate”) measurement values for each location and each variable (or “parameter”). For reliability and to facilitate accuracy measures, the spreadsheet includes 3-10 observations for each variable. The replicate data are then reduced by combining (usually with statistical averaging) the multiple observations. The resulting “Reduced spreadsheet contains the reduced data, which, along with related images, has been identified as having the most value for sharing.

Photographs are taken at the field sites for purposes of contextualization. Upwards of 300 photos in a day are shot while out in the field. The photos are linked to the parameterized data by the records in the paper field notebooks, and then by related identifiers in the spreadsheets. The notebooks also contain other contextual information, as well as procedures of data collection. A digital camera is used to make copies of the notebook pages as part of the data management strategy.

The categories in the "data stages" column listed in the table below were developed by the authors of this data curation profile. The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

Data Stage	Output	Typical File Size	Format	Other / Notes
Geologic and Water Chemistry Data				
Raw, hand-collected data	Field Notebook Entries	Kb	MS Word	Pages of these notebooks are also photographed daily with a digital camera, as backup. Spreadsheets are named with identifiers in the notebooks.
"raw" sensor data	Quantitative output uploads directly into the Raw Data Spreadsheet			
Raw, or "Replicate" data	Both sensor and hand measured data (generally first recorded in the field notebook) go into a matrix	10 Mb	MS Excel	Data for each variable are generated or collected (i.e. "replicated") 3-10 times; there may be 5-20 variables measured.
"Reduced"	Matrix of statistically averaged values for the replicate data in an Excel spreadsheet	1 Mb	MS Excel	One each for rock, water chemistry, and microbial data
Field and microscope photos	Digital photos, often with hand-written annotations	10Mb/photo @ 300 photos/day for a 5-10 day fieldtrip	.jpg	contextualizes the matrix data
Digital Back-ups of Data				
MS Word files	Typed up copies of the field notes documenting data collection context	Kb	MS Word	Typed up copies of the field notes documenting data collection context
Digital photos of the Field Notebook Pages	Digital photos of the pages of the field notebooks	1-10 Mb	tiff, jpg	Backup of documentation of data collection processes, metadata, and other contextual info

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Target data for sharing

The target data set for ingest consists of several (minimum of three) “Reduced Data” spreadsheets, accompanied by the field and microscope photos that provide detailed contextualization for the quantitative data. Some of these photos have been annotated with hand-written notes or labels for identifying content.

Value of the data

These data sets are developed very carefully with an objective of supporting scientific replication. It is anticipated that they will have high value for longitudinal studies beyond technological change, and upwards of 50 years. A range of user groups have interest in both the quantitative data and the image data. It is notable that the landscape and environment photo sets generated during the fieldwork, beyond those specifically accompanying the water chemistry data, are in great demand.

Contextual narrative

Data are collected over time in a “time series”, but the data set is static in the sense that there data values are not updated, and he “caps” the data collection period. While the scientist talked about the “Replicate Data” as a “million dollar spreadsheet,” based on both the high cost of collection, and use value. However, the data with the most informational value are a compound set that includes (at least) three “Reduced Data” spreadsheets (rock, water chemistry, and microbial activity) and accompanying, annotated photos. He noted that the spreadsheet data are often presented individually in published papers, but then synthesized for interpretation. Publishers vary in their requirements for the presentation of these spreadsheets.

This scientist has a store of 10 years worth of data, which are maintained and used by the scientist and his students. He has managed migration and backup processes within his lab, and feels that the data will be valuable for a long period of time (10s of years). It is uncertain whether this is typical practice in this (sub)field. While this scientist is willing to share the data described here, it would be embargoed until related articles are published.

Intellectual property context and information

Data owner(s)

This scientist views these data as jointly owned with NSF.

Stakeholders

The Stakeholders are the National Science Foundation (NSF), as the agency that funded the collection and analysis of these data, and the U.S. Park Service, which has an interest in scientific research conducted on public land, and also manages the permitting system for external research. There may also be interest on the part of energy-related industries.

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

Once the scientist has published on this data set, it is to be available for anyone to access and use (with attribution).

Attribution

Use of this data set requires attribution. In addition, scientist is exploring approaches to preserving IP for photos and images, including watermarking.

Organization and description of data for ingest (incl. metadata)

Overview of data organization and description

These data are highly structured: spreadsheets have column headers that specify the parameters (variables) measured, and each set of values for a sample has a location and a number; an additional identifier links the spreadsheet data directly with records in the field notebooks. Field photos are also labeled with the data identifier.

Formal standards used

No formal metadata standards, ontologies or controlled vocabularies have been employed with this data. While there are standards that might be applied to these data (e.g. FGDC or ISO 19915), this scientist seemed to be unaware of these even though he does document location data for observational data.

Locally developed standards

Metadata – None.

Crosswalks

n/a

Documentation of data organization/description

The data set is described and organized on multiple levels:

- Field notebooks contain methods information, and identify and describe each sample and photos taken; there is additional contextual information for the field expedition;
- Digital photos (and previously, Xerox copies) of field notebook pages are taken, and MS Word documents are created to record this content of the Field Notebook;
- The data files are organized by sample location and number;
- As noted above, each row in the spreadsheet represents a sample; each column represents a parameter (variable). The column heading identify the parameters. The sample id number indicates date, place and time of data collection.
- The photos are linked to specific samples by the sample id recorded in the field notes; photos are labeled with hand-written annotations and captioned with the sample id numbers.

Ingest

There are several aspects to the deposit of these data: The main controlling factor is temporal, in that data would not be deposited until the time of. These data sets are composed of multi-file packages that include spreadsheets, photos, and possibly additional files of contextual information. It is likely that additional metadata would be required for the data set package. The scientists was uncertain as to whether, or how, to automate ingest of these data sets.

Access

Willingness / Motivations to share

This scientist is generally unwilling to share his data with anyone outside of his research group until just prior to publication. This is a shift for him to a more conservative stance based on recent experiences. He is very willing to share data that will not be published.

Embargo

n/a

Access control

While he does share data with immediate collaborators, this scientist currently prefers strict control over access until publication, at which point he will make it openly available.

Secondary (Mirror) site

Access via a mirror site is a high priority.

Discovery

The scientist's experience is that most people find his data via his publications. He also deposits data or information about data in public sites, such as DLESE and the NSF Research Coordination Network (RCN).

The scientist placed a high priority on the ease with which researchers from within and outside of his discipline would be able to find these data. Also a high priority was the ability to discover these data using Internet search engines.

Tools

MS Excel (or a csv reader) is needed for using these data, and image viewers that can display JPEG and TIF format images. Also, the ability to connect the data to visualization or analytical tools is a high priority for the scientist.

Interoperability

Interoperability of this data set was not directly discussed. The ability for the repository to support the use of web services APIs is a high priority for the scientist.

Measuring impact

Usage Statistics

The ability to see usage statistics on the number of people who accessed this data was identified as high priority by the scientist.

Gathering information about users

Gathering information about the users of his data was not addressed by the scientist.

Data management

Data management is handled "in-house." This scientist has 10 years of data on hand for use by his research group; it has been migrated to new storage at least once, and is backed-up regularly. Current data is replicated and copies kept off site.

Security/Back-ups

Back-up of these data is of high importance to this scientist, as is the ability to audit the data set over time to ensure integrity.

Secondary storage sites

Secondary storage (backup) and off-site storage are both seen as high priority needs.

Preservation

Duration of preservation

These data would be preserved indefinitely; therefore the ability to migrate these data sets into new formats is high priority.

Data provenance

Documentation of any and all changes made over time to the data or data submission package is a high priority for the scientist.

Data audits

The ability to audit the dataset is a high priority.

Version control

Version control is not an applicable concern for respondent since his data are static.

Format migration

Format migration will be necessary as needed for maintaining accessibility of spreadsheet (or csv), image and photo files.

Personnel - This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.

Primary data contact (data author or designate)

Data Steward (ex. Library / Archive personnel)

Campus IT contact

Other Contacts

Notes on Personnel