Data Curation Profile – Soil Ecology

Profile Author	M. Cragin
Profile Author	M. Kogan
Profile Author	W. A. Collie
Institution Name	Illinois
Contact	M. Cragin (Cragin@illinois.edu)
Date of Creation	September 2, 2010
Date of Last Update	
Version	
Discipline / Sub-Discipline	Soil Organics
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	• An initial interview with the scientist, (April 2008) •. A second interview with the scientist, (January 2009) • A questionnaire completed by the scientist as part of the second interview (January 2009)
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	This soil ecology data curation profile is based on analysis of interview and document data, collected from one researcher working in this research area. Some sub-sections of the profile may be left blank, and this occurs when there are no relevant data in the interview or available documents used, and this will be noted in that section.
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284315014

Brief summary of data curation needs

This high-value data set identified for deposit combines observational parameters and calculated data (including means and variance of), and has been error-checked and cleaned. This tabular form data is held in a Microsoft Excel spreadsheet. Since this data could be represented both in the spreadsheet format and in the more generic comma separated value (csv) format, the scientist believes that the data should be made available in multiple formats to support re-use. The deposited data set is seen to have significant re-use value, and should be preserved indefinitely.

An embargo period of two (2) years is required. The scientist would like attribution when these data are reused by others, requiring a readily available citation for the data set as part of any related metadata or repository record. Access to analytical and visualization tools, as well as web APIs, would be useful for this kind of data. The scientist stated that this type of data carries privacy and confidentiality concerns, as it can include content (such as GIS data) that would identify land owners or other individuals who have responsibility for the soil in specific land areas.

The scientist also noted that there is a general uncertainty in her field about which data to keep, and to prepare and submit for public access.

Overview of the research

Research area focus

This scientist works in the field of soil science, specifically focusing on the soil ecology and organic matter. One area of her research concerns sustainable agriculture which seeks to understand the impact of various treatments on soil fertility. She approaches her research from a systems perspective, analyzing various ecosystem elements that affect soil characteristics. This scientist often work in close collaboration with a statistician or biometrician, during processing and analyses of the data.

Intended audiences

Various environmental scientists (soil scientists, agronomists, others interested in ecosystem dynamics), as well as modelers. People overseeing environmental policies were also identified as having an interest in these data, and particularly those interested in carbon management.

Funding sources

State of Illinois and the US Department of Agriculture (USDA) are the primary sponsors of this research.

Data kinds and stages

Data narrative

Field data are collected by hand (physical samples from the field; qualitative data in field notebooks) and with digital instruments. Lab work generates additional data, when soil samples are analyzed further; processes include “baking” the soil, and spectral analysis. The “raw” data are entered into a table (most often MS Excel; sometimes directly into SAS statistical software); much of the data undergo some type of processing to provide “factored” values for the continuous variables under study, including pH, phosphorus and potassium, bulk density, soil moisture, soil respiration, plant available nitrogen, and organic matter. Other variables are based on qualitative measurements, such as for soil texture, and these are variables that support classification of the samples.

Instruments used in the field collect data that are stored on-board (the instrument), and then transferred or uploaded to a computer. Some data are recorded by hand and added to the spreadsheet by hand. The initial data from these instruments generally exists in an instrument-

specific proprietary format. The proprietary files are rather small (e.g. in kilobytes), and are kept by the scientist as one type of “raw” field data. Once uploaded, the instrument data is transferred into an Excel spreadsheet, where it gets combined with manual measurements taken in the field (and recorded on paper), lab measurements, and experimental conditions.

These measurements are combined into either or a single spreadsheet file with multiple worksheets, or a series of spreadsheets (generally 4 to 5 about 1 Mb each),. It is common for the data to get combined into a single spreadsheet farther along in the research process. Since each workbook (spreadsheet initially) is about 1 Mb, the combined file adds up to about 5 Mb’s, although this varies. The data in the combined spreadsheet are checked for errors and verified using statistical techniques. Data points identified as outliers during the statistical analysis are excluded from the data set. The corrected spreadsheet is the data form that is most likely to be shared because the scientist beliefs it to be most useful for other users. The scientist noted that it is possible to tell the experimental design of the study in the way that the data are set up in the spreadsheet.

The corrected data is imported into SAS statistical analysis software, where additional transformations take place. SAS statistical analysis produced a model, which is statistically verified against the data. Both the model and SAS output are important outcomes of the analysis. Since the size of SAS output is generally extensive, the scientist often keeps most important excerpts of the output in an MS Word file.

Sometimes scientist also performs also performs spectroscopic methods for analysis of the soil organic matter. The results of this analysis are kept in tabular form, though the scientist did not specify which application is used to manage these data.

Data Stage table on following page

Data Stage	Output	Typical File Size	Format	Other / Notes
Raw	Non-instrumented filed measurements	(unspecified)	Paper	Hand-written in the field or lab notebooks; some entered into the spreadsheet, other is contextual for analysis
Raw	Instrument, sensor measurements	Small, kBs	Instrument-specific proprietary format (not specified)	Some instrument data Including sensors) are produced in a proprietary format. The proprietary software often performs some basic transformations on the data before they are transferred into the spreadsheet.
Initial digital matrix	spreadsheet	Small, kBs	MS Excel	Data from notebooks and sensors are entered into a spreadsheet
Combined	All the field and lab measurements and experimental conditions	<1Mb each, up to 4-5 files, merged about 5 Mb	MS Excel	The experimental conditions other non-instrumented data from the field notes, and the lab measurements are added to the sensor / instrumented data (likely from multiple workbooks)
Corrected	Data checked for human error and verified by statistical techniques		MS Excel	Data gets checked against the various sources for human error. It will also include some statistics, e.g. mean, range, and variance. The data points that are statistically determined to be outliers are excluded from the data.
Output from SAS statistical software	Selected material from SAS runs and "model" formula	20 pages	Word file	Important pieces of the SAS output (which are usually very large) get stored in a MS Word document as a distilled analysis output; the model is the statistical formula applied to the data that generated the reported results.
Augmentative Data				
Infrared spectroscopy (IR) or Nuclear Magnetic Resonance (NMR) Spectral data	tabular file	Small, kBs	Excel or native tabular file	Tabular (matrix) data representing the spectrometric analysis of soil organic matter Data may be reduced with values exported to the combined data file

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

Prior to publication, the scientist is willing to share early the "combined" spreadsheet, which contains the values for commonly used variables or, "universal measures" of basic fertility

characteristics of the soil, such as pH, PK, texture, and organic matter - primary nutrients and textures. These data have been error-checked and cleaned, and are in an accessible spreadsheet form; the data are considered valuable because it is ready for use by multiple user groups. Other, more specialized data are not shared until after publication.

Use/re-use value of the data

The scientist believes the data in general is valuable for reuse and should be kept indefinitely. This is particularly relevant for data from long-term or collaborative projects. These more valuable data sets are kept more carefully and organized more diligently by the scientist.

Contextual narrative

These data sets are static in the sense that data collection is a finite process for each experiment or project that is represented in a spreadsheet. However, much of the data could be combined with other data (e.g.: data taken at same site at a later date) to form a larger, integrated data set that would support new analysis.

The importance of the data depends on the perceived re-use value placed by the scientist on each particular data set. Data derived from a more longitudinal or large-scale collaborative project are considered to be more important by this scientist, and thus is requires more careful organization and preservation.

Intellectual property context and information

Data owner(s)

The scientist believes she owns these data, in the sense that she has possession and control over it. However, she is aware that data produced by publically funded research has implications for broader access, such that she is legally obligated to share the data when asked. The scientist acknowledges that some public access to the data is required at some point in time, and noted that Intellectual Property (IP) is a concern for collaborative research in this discipline. Ownership, holding, access and publishing issues are rarely (if ever) negotiated prior to the start of a project, or even during the course of the collaboration. In practice, it seems, again, that the physical possession of the data is a determining factor in ownership and outcomes related to access, sharing, and publishing.

Stakeholders

- Policy makers working on environmental regulation
- Scientists working on the cap-and-trade system, carbon offset techniques
- Immediate collaborators
- General public

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

The scientist is concerned with data misuse or misappropriation, since she had been taken advantage of in a recent collaboration. However, she is willing to share data at two points in the research process, with the common soil variables spreadsheet available once these data are cleaned, and data reported in publications that would be available after an embargo period. The data in the former instance is generally shared upon request.

Attribution

The scientist would like to be credited in some manner if the data is used by someone else, but did not give specifics.

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

Overview of data organization and description

The scientist states that the data organization and description for the current data set is sufficient for others to utilize the data. The variable names are used to label columns in the spreadsheet containing the dataset; other relevant metadata are included in any publication. However, the full accounting of the experiment (or observation), field site, methods, data collection and processing is distributed throughout several field and lab notebooks, and often also in computer files. The scientist reports that data are maintained locally for future access, but that management of data files and related materials is not always handled in the same manner across data collection periods.

Formal standards used

While Ecological Markup Language (EML) is emerging in the field, this is not used at this time. There are no ontologies or controlled vocabularies employed with this data set.

Locally developed standards

None

Crosswalks

None

Documentation of data organization/description

As with any tabular (or matrix) data, this data set is organized in row and columns. Columns commonly contain classification and continuous variables, while rows contain repeated measures. The column headers contain description of the variables below. The left-most columns will have the classification variables and parameters, such as treatments, dates, replications, “rings,” “blocks,” and depth of sampling. To the right of these columns, typically there will be the measurement variables (such as pH, temperature, moisture).

Such organization is a standard in this field. It is so common that the scientist suggests that people sometimes forego including any variable description since the organization itself is seen to be sufficient for the practitioners in the field.

Ingest

The scientist indicated the ability to submit the data to a repository herself was a low priority, as was having the submission process be automated.

Access

Willingness / Motivations to share

The scientist is willing to share the data with the trusted colleagues within or outside her research institution. This idea of the “trusted relationship” was emphasized strongly by the scientist, who said that this is even part of any consideration for sharing the data with members of her own department. The researcher stated that trusted colleagues are distributed, some within her department, at the institution, and then at other institutions. The scientist emphasized that personal relationship with people requesting the data matters a great deal to her, as issues of trust factor in to her decisions. The researcher makes these decisions on the person-to-person basis.

This scientist’s willingness to share is also dependent on the research interests of those requesting the data. One consideration is whether or not they are interested in doing something

significantly different from the work she is undertaking, so there would be no competition for a particular use of the data; in this case the scientist would be open to sharing.

As noted above, the scientist believes negotiating of terms in collaborative research is extremely important. She suggests that negotiation often does not happen before or even during the collaborative research process. She states that most researchers in her field are uncomfortable with even discussing such topics, which leads to ambiguity and sometimes misuse. As the scientist has been taken advantage of in one of her recent collaborative projects, she is now very concerned with negotiating ownership and terms of use in such cases.

Embargo

The scientist needs data embargoed after publication in order for her to explore it all. She would ideally like about 5 years of embargo, but realizes that people would like the data much earlier, especially in the case of more valuable and long-term studies. Thus, the scientist would request 2 year embargo as a compromise between her scientific and needs and those of research community.

Access control

If data are deposited prior to publication, the scientist would like access to it to be strictly controlled, limited to the immediate collaborators and those scientists she has identified as trusted colleagues. This indicates a need for the scientist to have active control over the access mechanisms.

Once the data becomes public (following embargo), she believes that the data should be available to everyone; at that point in time, the ability to restrict access to authorized individuals would be low priority.

Secondary (Mirror) site

The scientist indicated that the ability to access the data at a mirror site if the main site is offline is a low priority.

Discovery

The ability for others to discover this data through the Internet search engines is a high priority, particularly for researchers within this field; this has a lower (medium) priority for researchers outside of her field.

Tools

The need to connect the data to visualization or analytical tools was identified as a high priority by the scientist. The data are stored in a spreadsheet or a comma separated file, so MS Excel or csv reader would be sufficient to open the files.

Web service APIs were also indicated by the scientist to be a high priority for this kind of data.

Interoperability

While the scientist did not talk about the need for these data to be interoperable, it is clear that measures for many variables would be useful for longitudinal studies or aggregating data sets.

Measuring impact

Usage Statistics

The ability to see usage statistics on how many people have accessed the data is a low priority for the scientist.

Gathering information about users

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

Data management

Security/Back-ups

Data are stored on several local machines in the scientist's lab and office. The data are backed up locally by people working in the lab. The scientist indicated that everyone in the lab also keeps their most important data on their personal flash drives. This creates some back-up redundancy between various lab machines and researchers' flash drives. Sometimes the departmental server is used for the back-up purposes. When lab computers are replaced, they are kept as archival storage for the data.

The scientist indicated that at the moment there is no security implemented for this data, except for the physical access to the lab. The researcher stated that no encryption or other advanced security forms are necessary for this kind of data.

Secondary storage sites

A secondary storage site is a high priority for the scientist; however a secondary storage site at a different geographic location is a low priority.

Preservation

Duration of preservation

The scientist indicated that the data would be valuable indefinitely. On the other hand, the scientist suggested that longevity of the data depends on the perceived importance of the particular data set. The data derived from long-term studies is generally kept longer.

Data provenance

Documentation of any and all changes made to her data over time is a high priority for the scientist. However, the scientist would prefer that editing the data set was not allowed in the context of repository, thus alleviating the need for tracking the changes. If editing is allowed, the scientist believes they should be kept in different versions of the data set, so the original data is never lost.

Data audits

The ability to audit the dataset within the repository is a medium priority for the scientist.

Version control

Version control of data within the repository is a high priority for the scientist.

Format migration

The ability to migrate the dataset into new formats over time is a high priority for the scientist. The tools needed to access the most informationally valuable form of data are widely available (MS Office) and even open source alternatives exist (i.e. Open Office).

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