Data Curation Profile – Water Flow and Quality

Profile Author	J. Carlson
Profile Author	N. Brown
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	October 27, 2009
Date of Last Update	
Version	1.0
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 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 conducted on August 2008. • A second interview with the scientist conducted on January 2009. • A questionnaire completed by the scientist as a part of the second interview. • A published paper explaining the research and the methodology used to gather, process, and analyze the data set in question.
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	The Water Flow and Quality Data Curation Profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct the profile.
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284315017

Brief summary of data curation needs

The primary data sets for deposit are a series of spreadsheets of water flow data over set intervals of time in a tile drainage system and spreadsheets summarizing water flow rates and water quality information on an annual basis. This data has been collected over a 25 year period.

The data would be made available to others for re-use once the scientist has published her findings. The data are not well documented currently and it would likely take a considerable investment to prepare the data for use by others. The lack of documentation is a particular concern of the scientist in sharing her data with others. The scientist would need to receive attribution if the data set is used by others.

Overview of the research

Research area focus

The scientist primarily examines water flow and water quality using a tile drainage system. The scientist has used this data as a part of her research into the impact of drain spacing, soil management practices on nitrate leeching and effects on other substances, and impacts of drainage on crop growth and yield.

Intended audiences

Other researchers in the field, particularly those that are engaged in developing predictive models, would be the primary audience. Farmers or other agriculture professionals may also have an interest in some aspects of her data.

Funding sources

Funding sources in the past have included the USDA and the agricultural research programs of the scientist's institution. The scientist does not currently receive much outside funding. She has not been mandated by her sources of funding to generate a data management plan or share her data with others outside of her lab.

Data kinds and stages

Data narrative

The Scientist collects data on drainage, water flow and water quality from a single location. Data are still being generated.

The raw data are collected both from data logger equipment and manually at the site. Manually collected data are primarily used as a back up in case of equipment failure or for verification purposes. A software program processes the data to discern the rate of water flow over certain intervals of time (typically 6 minutes, 1 hour, and 1 day). The size of the data files vary in size based on the frequency of the data collection. This raw data are then processed, cleaned, and analyzed at the scientist's institution. During this process, missing or erroneous variables are identified and accounted for. For example, data are not collected on the weekends, but are generated through estimation based on other information gathered.

In the intermediary phase, the data are organized by day and time. The data are manipulated using Excel and SAS programs in the "analyzed" stage. While the scientist generally performs data calculations in Excel, she has enlisted the help of statisticians and others to run more sophisticated analyses. The finalized data are typically saved in Excel spreadsheets which are used to generate charts/graphs for use in publications or presentations. Data are backed up at all stages of the data cycle in many formats including lab notebooks, CDs, zip drives, an external hard drive, and the scientist's departmental server.

The categories in the "data stage" 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
Water Flow Data				
Raw	Stream of data from the data logger	<1 MB	.dat	Data are also collected manually for back-up / verification purposes.
Processed	Rates of water flow parsed out into set intervals of time.	3-4 MB	Excel spreadsheet / ASCII	Data are run through a software program (proprietary) that splits it. Data are transformed into a useable format, typically an excel spreadsheet, sometimes ASCII.
Interpolation	Water flow data with corrected tabular data	6 minute flow: ~20MB; Other rates: ~1MB	Excel spreadsheet	Data are checked; missing or erroneous values are estimated or otherwise accounted for. Explanatory notes are included in the spreadsheet itself and/or other documents.
Water Quality Data				
Raw	Water Sample	NA	NA	Water samples are gathered for testing at specified intervals.
Processed	Amounts of nutrients / substances present in water	Unknown	Excel spreadsheet	Samples are run through scientific instrumentation to measure concentrations of particular substances.
Water Flow and Quality Composite Data				
Joined	Water flow rates and amounts of tested substances		Excel spreadsheet	Water flow data and water quality data are joined to varying extended based upon the research question and the type of analysis being conducted.
Analyzed	Analyzed statistics calculated in Excel or SAS	Approximately 21 MB	Excel & SAS (or other statistical program used)	Data are typically analyzed via Excel or SAS. For more sophisticated analysis, the scientist has enlisted the help of statisticians.
Summarized	Summary flow and concentration data for a particular year		Excel spreadsheet	The scientist typically composes a "best of" spreadsheet summarizing her water flow and concentration data for the year.
Published	Charts/Graphs		Presumably .pdf, .doc, or .ppt	Used as tables in publications or slides in presentations.
Augmentative Data				
Back-up data	Notebooks / Print outs			Manually generated data are kept in print in notebooks in the scientist's office for backup purposes.

Weather data	Daily/Monthly precipitation totals			Gathered from a weather station located on-site. Integrated with primary data files when needed for analysis.
Crop yields / Soil permeability / water tables, etc.	Varies			Other data are gathered on an as needed basis, depending on the types of analyses being done. Stored and analyzed separately from primary data sets.

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 scientist would consider sharing the water flow data that has been cleaned and processed, as well as the summarized versions of her water flow and water quality data with others. She would be willing to deposit her data into a data repository and enable public access provided that her concerns were adequately addressed (see “terms of use”, “ingest” and “willingness/motivations to share”).

The scientist indicated she might be willing to share data in other stages (except the raw data), but this was not specifically discussed in enough detail to make a firm conclusion.

Use/re-use value of the data

The data would be very useful for constructing and testing water flow and drainage models. The scientist notes that the data with the most value for others is the concentration and flow data that has already been processed. This data can be used in data modeling research in the agronomy field.

The augmentative data sets may have value for others as well. The scientist specifically mentioned her crop yield data as being of interest to farmers, although this was not discussed in depth.

Contextual narrative

The Scientist has been collecting water flow data for more than 25 years. Water quality data has also been collected over this period. However, with the exception of nitrates, the substances being measured have changed over time. Some of the substances that have been measured include pesticides and major nutrients (such as ammonium, nitrogen, phosphorus, and potassium).

Data are still being generated.

Excel spreadsheets are the typical format employed for storage and use of the data. The scientist also has data in ASCII, SAS (for analysis), printed notebooks (back ups of manually collected data), and Lotus 1-2-3 files (little used legacy data). The scientist has made a conscious effort not to let the size of her files grow to a large and reports that most of her excel spreadsheet files range from 10-20 MB apiece. However, this practice has led to the proliferation of data files. The scientist does not know the exact number of data files she has but estimates that she has “hundreds” of them.

The scientist also collects additional data to augment or enable her analysis of the data. This ancillary data includes weather data (precipitation), crop yields, soil permeability, and the depth of water tables. With the exception of the weather data, these ancillary data sets are not integrated with the water flow or water quality data.

Intellectual property context and information

Data owner(s)

The scientist feels that her institution is the true owner of the data.

In a previous situation in which the scientist has shared data with a colleague for re-use in a model, she indicated that the data belonged to her, while the model used to analyze the data belonged to the other scientist.

Stakeholders

The data stakeholders are the USDA and the agricultural research programs of the scientist's institution, which have funded some of the projects that have enabled the data to be collected.

At this point, it is not clear when the funding was received or what precisely it was used for in terms of generating and using the data. However, according to the scientist, the USDA and the institutions agricultural research programs have no intrinsic interest in the data itself, only that the results are published.

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

If the dataset were to be made available in a repository, the scientist would want to include descriptive information about the data and how it was generated to guard against its potential misuse. The scientist mentioned having a mechanism to indicate that the user had read this information before being allowed to use the data.

Attribution

The scientist would like to be credited in some manner if the data are used by someone else. The scientist indicated that the ability to cite this dataset in her publications is a medium priority for her.

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

Overview of data organization and description

The scientist admits that the data organization and description for the current dataset is insufficient for others to utilize the data. Lack of time and the lack of trained assistance have been the major barriers in her managing and organizing the data.

The scientist has not employed a standardized naming convention for her data and she has mentioned version control as a concern. The lack of such a standard may present a challenge in working with her data, particularly in the selection and appraisal process.

The scientist is interested in developing the metadata necessary to describe her data effectively, and hired person with a PhD in a related field on a part time basis to help make the data more accessible to others.

Formal standards used

No formal metadata standards, ontologies or controlled vocabularies have been employed with this data.

Locally developed standards

The scientist uses annotations within the data file as her primary means of description.

Crosswalks

Not discussed.

Documentation of data organization/description

The primary means of description used by the scientist has been detailed annotations within the spreadsheets themselves. She also has Microsoft Word files containing dataset descriptions that are referred to in some of the spreadsheets.

Ingest

The primary issues surrounding ingest of the data into a repository are tied to when the data would become available. The scientist's conditions for making the data publicly accessible are that the scientist has published all that she has planned to publish using the data and that the data are cleaned-up and described well enough so that others can understand and make use of her data effectively.

The precise timing of when the data should be ingested into the repository was not discussed. If the data were to be ingested before the conditions for public access were met, an embargo would be necessary.

Much of the data are currently structured in MS Excel spreadsheets. The scientist is aware that Excel is a proprietary format and would be amenable to migrating the data into a more open format (.csv or ASCII were discussed) for curation purposes upon ingest into a repository, provided that the explanatory annotations and notes that she has made within the data set are captured, associated with and made available with the data in some fashion.

The scientist indicated that she would prefer to submit her data to a repository herself rather than to have the process be automated.

Access

Willingness / Motivations to share

The scientist would not share raw data outside of her immediate collaborators. The scientist has shared her data before its publication with colleagues and other institutions with whom she has already developed a working relationship. She would be willing to do so again. She feels that this group of colleagues would have enough knowledge and familiarity with her work that they would be able to understand and use her data effectively.

The scientist indicated that she had not completely thought through the questions surrounding who should be allowed to access her data at a particular point in time and that her responses were on the conservative side.

Embargo

The need for an embargo is event-based rather than time-based and rests upon whether the conditions for access have been met (see "willingness to share" above). If these conditions have not been met then an embargo for the data would be required.

Access control

Before the conditions for making her data publicly available are met, access to the data set would need to be strictly controlled. The availability of the data would be limited to those the scientist has identified as trusted colleagues, if the data were to be made available at all.

Once the conditions to make her data publicly accessible have been met the ability to restrict access to authorized individuals would be a low priority.

Secondary (Mirror) site

The ability to access the dataset at a secondary site if the repository is offline is a low priority for the scientist.

Discovery

The scientist indicated that she places a high priority on enabling researchers in her field to find the data, and a medium priority on enabling researchers outside of her field to easily find her data. The scientist places a low priority on enabling the data to be discovered through internet search engines.

The data are primarily organized by date, which presumably would be a key attribute for browsing or searching the data.

Tools

Anticipated use of the data includes statistical analysis and the testing/verification of models. In the past, the data has been analyzed using MS Excel, SAS and other statistical analysis programs. The data needs to be made available in (a) format(s) where it would be accessible to these and other statistical programs. Currently, it is unclear how the data should be formatted or structured for use in modeling software.

The scientist did indicate that the proprietary software from the data logger (from Campbell Scientific) used to generate the data may be required to utilize it. However, in reviewing the other information obtained from the scientist, it is believed that the data logger software in primarily used to generate the data, not to make use of it.

The ability to connect the data to visualization or analytical tools was given a low priority by the scientist.

Interoperability

Developing connections between the data and any publications that have resulted from the data are a high priority for the scientist.

The scientist indicated that support for the use of web services APIs is a low priority for her.

Measuring impact

The scientist did not specifically discuss a need to measure the impact of making her data available to others.

Usage Statistics

The ability to see usage statistics on how many people have accessed the data are not a priority for the scientist.

Gathering information about users

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

Data management

Security/Back-ups

The scientist does not routinely make backup copies of her data. Her primary means of back-up currently is an external hard drive and, less frequently, her department's computer network. In the past she has backed up data on to diskettes, some of which she still has in her possession.

Currently, the manually collected data that serve as a backup to her digital data are filed into a notebook which is kept in the scientist's office. These notebooks are generally not used once the data has been verified, although they may contain notes and annotations that could help to inform or generate the descriptive metadata.

Her primary security concern with placing her data into a repository is that the data not be released before she has completed her work with it and has published the results.

Secondary storage sites

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

Preservation

Duration of preservation

The scientist indicates that the data would be useful for 20 years or more but less than 50 years.

Data provenance

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

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 scientist has migrated data from outdated software (Lotus 1-2-3) to usable formats for her purposes (MS Excel) on an as needed basis. Data that has not been used since it was originally formatted has not been migrated and so some of her data are likely to be in its original, outdated format. Most of the data in outdated formats is likely to be ancillary and of lesser value.

The ability to migrate the dataset into new formats over time is a high priority for the scientist.

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

Although the scientist currently has a part-time employee to assist with data management, that person is not specifically trained in data management or curation.