Data Curation Profile – Plant Nutrition and Growth

Profile Author	J. Carlson
Profile Author	N. Brown
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	November 23, 2009
Date of Last Update	
Version	1.0
Discipline / Sub-Discipline	(Withheld)
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 scientist/scholar's reported needs and preferences for these data. They are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information	• An initial interview with the scientist conducted in July 2008. • A second interview with the scientist conducted in December 2008. • 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 Plant Nutrients and Growth 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/10.5703/1288284315012

Brief summary of data curation needs

The data consist of multiple spreadsheets in MS Excel format, some SAS and Minitab files, and a small number of Power Point files containing images and some descriptive text. The data is potentially useful for multiple audiences, however different audiences would likely have different needs and interests in accessing and using the data. Currently, data description is minimal and would require further development before data could be shared with others, especially if the data is to be accessible to and understood by multiple audiences. The data files can stand on their own, although linkages between the spreadsheets and the gels and blots would be desirable. Use of the data would require attribution to the scientist.

Overview of the research

Research area focus

The research consists of two separate but related studies. The initial study focused on investigating the effects of phosphorus and potassium nutrients on the ability of alfalfa to recover from stress brought on by the winter season or the close cutting done in harvesting. This study was carried out through manipulating and managing the plants environment through controlling the levels of nutrients provided. Instead of using estimations of plant persistence through "above ground" estimations as is the common practice in this type of research, data for this study was gathered through an intensive sampling process, in which plants and their roots were dug up and counted directly.

The early results from the initial study raised a number of questions about commonly-held assumptions about alfalfa survival rates during the winter; the outcomes were contrary to what was expected. As a result, a second project was launched at the same location to examine growth behavior and persistence of alfalfa during the summer season. This second study focused on examining changes in taproot, physiology, biochemistry and gene expression during the summer to help understand alfalfa persistence through the winter. The second study ran for five years and was conducted on a smaller scale study than the initial one, though the sampling done was more intensive than the first study.

Although the studies are closely related and may be cross-referenced when published, they are considered to be separate and distinct from each other.

Intended audiences

Other researchers interested in issues of plant nutrition, growth and persistence. The data may have practical value for farmers, government agencies such as the EPA, agricultural companies, and agriculture extension programs.

Funding sources

A variety of funding sources have sponsored this research over its lifecycle, including his home institution, corporate sponsorships and the USDA. The scientist has not been mandated by his sources of funding to generate a data management plan or share his data with others outside of his research team. He is aware that the USDA is considering such mandates, and would support such a mandate should it come to pass.

Data kinds and stages

Data narrative

Data from both studies in this project consist primarily of field data and plant samples. Variables gathered include the yield and overall health of the plot, the physical characteristics of the plant sample, and the amount of selected nutrients present in the sample. Both studies in the project

have been running for several years and have gathered data at multiple times, resulting in hundreds of data files.

The field data are initially gathered in the field by hand and recorded on printed data sheets. This data is later entered into an Excel spreadsheet.

Plant samples are harvested from the plot. Plant tissues and roots are processed and 30-40 characteristics from the samples are measured in the lab. Data are captured in Excel spreadsheets. Samples are also sent out to a 3rd party to determine the amounts of selected minerals in the sample. The data are delivered back to the lab in Excel spreadsheets.

Samples of the soil are taken and analyzed to determine its mineral composition and investigate the linkages and relationships between mineral composition in soil and plant persistence. This information is entered into an Excel spreadsheet.

Each of the two studies has a Master Spreadsheet, which brings together a summary of the data from the multiple data spreadsheets generated from the field observations and collected samples to identify relationships in the data across time and space. These Master spreadsheets serve as the official record of the data.

Additional spreadsheets are derived from the Master Spreadsheet for use as "working files". The Master Spreadsheet, as the official record of the data, is not meant to be altered or manipulated. These "working files" are used as a means of data sharing and temporary storage within the lab group. Once the data are ready for analysis they are imported into a statistical software package (usually Minitab or SAS) for data reduction or statistical analysis.

The scientist also looks at the presence or absence of proteins in the process of re-growth of alfalfa. Gels and blots are generated in the scientist's lab from the roots of the plant sample and are then digitized into images. These gels and blots images are then inserted into Power Point slides for the purposes of annotation and storage.

Weather data, precipitation and temperature, are also gathered from the state climatologist's office to identify and explain effects of environmental conditions on plant growth and health in this project as needed. These data are processed into weekly, or sometimes monthly, averages and correlated with the plant data as appropriate. Raw data are not kept after processing.

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
Raw	Field Notebook Entries; Samples		Paper; MS Excel	Plant and soil samples are gathered. Information about the field and observations of the samples are recorded by hand and then entered into a spreadsheet.
Processed	Multiple spreadsheets of data from the samples	100-200kb	MS Excel	Samples are processed in a lab and data are collected in multiple spreadsheets.
Integrated	Master Spreadsheet	1 MB	MS Excel	Data from field observations and processed samples are integrated into a Master Spreadsheet which serves as the official record of the data.

Extraction	Working copies of the data	variable	MS Excel	Data are extracted as needed from the Master Spreadsheet into working copies.
Analysis	Data summaries; Tables and Figures	~50kb	SAS; Minitab	
Qualitative	Gels and Blots	~10 MB	MS Powerpoint	The images of gels and blots are inserted into Powerpoint slides to enable their annotation.
Augmentative Data				
Weather	Weekly / monthly precipitation and temperature ranges		MS Excel	Linking the spreadsheet data to weather data is desirable.

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 data identified by the researcher as having the most research value for others are the Master Spreadsheet and the processed spreadsheets.

The researcher is open to sharing some of his other data files. He has been asked by publishers to share the analyzed data as supplemental files when publishing the findings of the research and would consider placing these files in a repository as well. Sharing the gels and blots files is possible as well; however, additional description and documentation would be required for other researchers to understand and make use of this data.

Use/Re-use value of the data

According to the scientist, these data sets represent a more comprehensive examination of plant persistence than is typically available from other data sets on the same subject because of the methodology used to generate the data (frequency of harvesting and analyzing individual plants rather than examining the plot).

Although the data are limited in that they were gathered exclusively from a single location, the data could be used as a point of comparison with other data sets. Other researchers have expressed an interest in repeating the experiment in different locations and conditions (harsher growing conditions, more severe winters, etc.) to see if the findings of the original study are replicated or not.

Making the analyzed data available could serve as a means of validating the research findings, which have been somewhat controversial. The gels and blots from this project may be of interest to other researchers engaged in studying the expression of genes in plants.

Different audiences would be interested in different aspects of the data, and some audiences may find it hard to identify traits of the data likely to be of specific interest to them if they were presented with the data in its entirety. In an ideal situation, traits of the data likely to be of interest to a particular group (livestock farmers, the EPA, etc.) or for a particular use (performance of fertilizer) would be captured in their own files and identified and described accordingly.

Contextual narrative

Data collection has ceased in both studies. Data have been analyzed and the scientist is in the midst of writing up and publishing the results.

The data primarily consist of hundreds of Excel spreadsheets. File size varies but are fairly small overall, with the largest files, the Master Spreadsheets, at around 1MB. Excel is used to house the data as it is easy to work with and share amongst members of the research team, however

Excel is not used for analysis of the data. Data in the Master Spreadsheets serve as the official record of the data and are static. Data in the working files are not as well managed or well documented as the Master Spreadsheet.

Data analysis is performed using SAS and Minitab, although Minitab is preferred as it is perceived to be more compatible with Excel. SAS and Minitab are also used to generate tables and figures of the data. The scientist reports that about ten SAS and Minitab files served as the basis of the findings. These ten would be the files he would consider making available.

The scientist estimates having five Power Point files containing approximately 10 gel and blots images each. Original images from the gels and blots are kept but rarely used as they require the annotation written into the Power Point files to understand and use them effectively. The format of these images was not discussed.

Intellectual property context and information

Data owner(s)

The scientist believes that the data are owned by the public, as public funding (grants) helped support the research. However, he noted that this stance is not universally held by others in his department.

Stakeholders

Stakeholders in this project would be the funding agencies who have provided support. According to the scientist, none of his funders have expressed a direct interest in his data from this project, although the USDA is considering implementing requirements for data management and/or sharing for future grant recipients.

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

Once the data has been processed, normalized and corrected for any errors, it could be shared for anyone to access and make use of (provided attribution is given).

Attribution

The scientist is willing to make his data publicly available to others as long as he receives credit when the data is used. The nature of the attribution desired by the scientist might vary depending on the intended use of the data, and could potentially range from an acknowledgement in the paper, a citation to the data, or at a very high level co-authorship of a paper if the data were the core element of the paper.

The scientist indicated that the ability to cite this dataset in his publications is a high priority for him. In addition to enabling attribution, the scientist believes that enabling the citation of the data will help prevent it from being misinterpreted or misused by others, as the original data can be tracked down if any questions arise.

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

Overview of data organization and description (metadata)

The data are primarily described through annotations made by the scientist or by staff/students at his lab. In the spreadsheet files, these annotations center more on identifying and explaining anomalies in the data than on describing the data itself. Spreadsheet files do however have descriptive column headings that include the unit of measurement (kg, etc.).

Annotations for the gels and blots data serve to explain each “lane” (sample) of data in the image so that it can be understood and used effectively. Annotations listed in the Power Point slides

holding the gels and blots data include the harvest date and how the sample was processed in the lab.

The data files are organized by the two different studies, then generally by year, then by harvest, (there are four harvests within each year at minimum). Within each harvest are data files for plant yield, compositions, physiology, biochemistry, and enzymology. The gels and blots slides are stored along side the spreadsheets, but are not directly related to them through naming conventions, annotations, or any other means.

Formal standards used

No formal metadata standards, ontologies, or controlled vocabularies have been applied to this data.

Locally developed standards

None.

Crosswalks

NA.

Documentation of data organization/description

Given that the findings from the first study were contrary to what was expected, the scientist noted the need for solid documentation for his data to back it up should the quality of the data be questioned. Specifics of the scientist's current documentation practices were not discussed.

Sharing this data with others would require additional documentation of the research and description of the data files. For example, the abbreviations used in the spreadsheet would need to be defined in the spreadsheets. The processing methodologies used in measuring the plant samples and developing the data would also need to be made available in order for others to understand and use the data. Other needed documentation mentioned by the scientist included information about when, where and how the fertilizer was applied and the plants were harvested.

Documentation needed to share the gels and blots data include the harvest date of the plant sample, the amount of potassium, phosphorus and other minerals in the sample, and information on how the sample was processed in the lab.

Ingest

The scientist indicated that both the ability to automate the submission process and the ability to submit data manually into a repository are medium priorities for him.

The data sets are composed of tabular data in spreadsheet files, the data summaries generated using statistical software packages, and the gels and blots captured in Power Point slides. The connection points between these data sets have not yet been firmly established, nor are the files described sufficiently for use outside of the scientist's lab. It's likely that documentation about the methods and procedures used to generate the data will have to be codified or created to make the data understandable and usable by others.

Access

Willingness / Motivations to share

The scientist has shared this data with colleagues in his field whom he has deemed to have a legitimate purpose in using the data. The scientist is generally willing to share his data with others before his work has been published; although he would want to know how the data would

be used before access was granted. His colleagues have had experiences in which they felt their data was used inappropriately by others.

Embargo

Before making his data publicly available the scientist would require a 4 to 6 month embargo on the release of his data. During this time, the scientist would review the data one more time to identify and address any lingering errors.

Access control

Once the embargo period has passed, the data would be made publicly accessible. No further limitations on access would be required.

Secondary (Mirror) site

Having a secondary mirror site for the repository to ensure the availability of the data is a medium priority for the scientist.

Discovery

The scientist indicated that he places a high priority on enabling researchers both in his field and outside of his field to easily find his data. Enabling the data to be discovered through internet search engines is also a high priority for the scientist.

The current organizational structure of the spreadsheet data files - study, year, harvest - needs to be maintained as they represent likely points of access to the data. In addition, the spatial attributes of the data and the types of fertilizer treatments given to the plants are also likely access points.

The points of relationship between the spreadsheet data and the gels and blots should be identified and linked between the data files to enhance discovery.

Tools

The ability to connect the data to visualization or analytical tools was given a medium priority by the scientist. SAS, MiniTab and other statistical analysis programs have been used to analyze these data in the past. The data needs to be made available in (a) format(s) where it would continue to be accessible to these and other statistical programs.

Interoperability

Linking the data to the publications that have resulted from the data is a medium priority for the scientist.

One possible use of this data would be as a point of comparison with other plant nutrition and growth data. However, the scientist stated that enabling direct alignment between his data sets and other data may be difficult as different researchers would take different measurements or apply different practices in their own research. The scientist would like to see others replicate elements of his study in different locations and then to add their data to his own to build a richer collection.

Measuring impact

Usage Statistics

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

Gathering information about users

In addition to counting the number of times the data was downloaded, the scientist is interested in learning about who is accessing the data and what institution they are from. This information would serve as another means to make connections between researchers who may have similar research interests.

Another feature for a data repository that would be useful to the scientist would be a notification system, alerting the scientist when ever the data was downloaded.

Data management

Security/Back-ups

The primary means of data backup has been an external hard drive located in the scientist's office. The scientist has not used his department's network server for back up purposes, but may do so in the future.

Paper copies of field notes are kept in the scientist's office as backup, but are not generally used. Field notes are primarily kept as a means of verifying the research findings in case there are questions.

Students who are making use of the project data for their dissertations are required to submit their data files to the scientist once they have completed their dissertation. These files are kept by the scientist on a back-up hard drive in his office and are not accessed unless questions are asked or information is needed.

Secondary storage sites

Both a secondary storage site and a secondary storage site at a different geographic location are high priorities.

Preservation

Strong concern was expressed by the scientist that the documentation and description of how the data were gathered be preserved along side of the data, as this information is key in enabling others to understand and use the data.

Duration of preservation

The scientist indicated that this data set has lasting value and therefore should be preserved indefinitely.

Data provenance

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

Data audits

The ability to audit the data within the repository is a high priority for the scientist.

Version control

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

Format migration

The scientist is agreeable to the migration of the data out of their current proprietary formats and into open source equivalents such as .csv, provided that the integrity of the data and annotations are maintained. Special care would be needed for the gels and blots if they were to be separated from their Power Point slides and deposited in a repository as images, as the slides hold key the information needed to identify, understand and use them.

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