Data Curation Profile: Agronomy / Soil Microbiology

Profile Author	Jake Carlson
Author's Institution	Purdue University
Contact	Jake Carlson <jrcarlso@purdue.edu>
Researcher(s) Interviewed	[name withheld], Graduate Student
Researcher's Institution	Purdue University
Date of Creation	September 29, 2011
Date of Last Update	
Version of the Tool	1.0
Version of the Content	
Discipline / Sub-Discipline	Agronomy / Soil Microbiology
Sources of Information	• An interview conducted on May 17, 2011 (duration of 1:42). • A worksheet completed as a part of the interviews.
Notes	The interview and subsequent Data Curation Profile were modified from the default version. The interview was scaled back to focus on identifying the data set and its lifecycle, sharing the data and managing the data. This data curation profile was developed as a part of an initiative to identify and address the data management and sharing practices of graduate students in an Agronomy lab at Purdue University.
URL	http://datacurationprofiles.org
Licensing	This work is licensed under a Creative Commons Attribution 3.0 Unported License

Section 1 - Brief summary of data curation needs

The Graduate Student has developed her own method of organizing her data that has worked well for her overall. Her lab notebook serves as the primary means of documentation for her data, but as it is in a physical form, the information it contains is not easily associated with the digital files that comprise her data set, nor is the information easily accessible to others. The Graduate Student will print out her data and paste them into her lab notebook, a rather cumbersome process. Although the lab notebooks were not observed by the author, it sounded as though the Graduate Student is quite thorough in documenting her work. The Graduate Student also builds in redundancies into her spreadsheets enabling her to trace back the provenance of her data.

The Graduate Student expressed an interest in making use of some of the data generated by other graduate students in her lab, but stated that she did not always know who to ask or how to obtain this data, and she expressed concern over the amount of time and effort it would take for others to be able to share with her. Other graduate students in her lab may feel the same about the data she is generating, although this was not discussed explicitly.

http://dx.doi.org/10.5703/1288284314994

The software she uses presents some barriers to her making use of the data, either due to outdated software / computing resources or inadequate exporting capabilities. MS Word is used as a work around in addressing the later issue.

If her data were to be made available to others, linking the published data to the articles that resulted from the data would be very important to her.

Section 2 - Overview of the research

2.1 - Research area focus

The Graduate Student studies management strategies for bioenergy crops and their effect on soil structure and the sustainability of soil quality. Information about soil structure and quality is determined through preforming a series of complimentary experiments and measures on soil that has been subjected to particular treatments. The experiments conducted include enzyme assays, measuring CO2 respiration activity, quantifying lipids through Gas Chromatography, and molecular analysis. These experiments are then compiled and analyzed for any identifiable patterns and trends in the results that demonstrate the nature or effect of a particular soil treatment. The end result is like an environmental impact assessment on soil.

The Graduate Student envisions at least three publications coming out of her work with soils. The first paper is a soil quality indicator study from a compilation of fifteen experiments on five candidate bio field crops. The second study focuses on the nitrification potential of the soil and an examination of the functional genes that are responsible for different transformations of nitrogen in the soil.

The third study is on determining the bacterial activity taking place on the surface of sorghum and maize roots, and then relating them to the diversity of the functional genes as well as to the characteristics of the root material that support this activity.

2.2 - Intended audiences

The Graduate Student mentioned that other graduate students working with her advisor now or in the future may want to make use of her data for their own purposes. Other researchers studying soil microbiology may also have an interest in her data.

2.3 - Funding sources

Not discussed

Section 3 - Data kinds and stages

3.1 - Data narrative

The Graduate Student's research takes place at Purdue's Water Quality Field Station, a plot of land managed in part by her advisor. The field station has 4 blocks and in each block, there are a total of 12 different treatments applied to the soil. The Graduate Student has gathered samples from 8 of the 12 types of treatments and takes multiple samples at 5 different times across seasons.

The Graduate Student is primarily interested in two things: the composition of the soil and its biomass. Determining soil composition will include characterizing the total carbon content, the PH of the soil, the presence and location of nitrogen. Determining biomass consists of identifying the presence of multiple items such as bacteria, protozoa, and other small microscopic organisms in the sample.

In her research, The Graduate Student demonstrates the strengths of her hypotheses and findings using multiple types of analyses. These analyses generate two kinds of data: numerical and image.

The lifecycle of the numerical data is as follows. In the raw phase of the data lifecycle, The Graduate Student gathers data in the field and takes soil samples from the treatment plots. The soil samples are homogenized for her research purposes, which includes sifting through the samples to remove plant materials and roots. Several sub-samples of data are extracted from the samples and they are subjected to a variety of processes, such as air-drying or freeze-drying the soil, to generate usable data points. Samples are then placed into bags which are inventoried and kept in storage once sub-samples have been extracted from them.

These initial data points are recorded into an excel spreadsheet, or are generated through the use of a Gas Chromatographer or other lab equipment as a .csv file, and then converted into Excel. The Graduate Student then conducts a series of calculations on the raw data points in order to prepare them for analysis.

The process of generating and verifying the reliability of the initial data generates a great deal of raw values and replicates. These raw values are then corrected through replication or calculation to determine the actual data points that will be used in her analysis. These finalized data points are then taken from the multiple spreadsheets that The Graduate Student has generated and are compiled into a single spreadsheet as preparation for analysis. The Graduate Student may generate additional versions of the compiled data file if her advisor recommends looking at the data from additional perspectives or approaches.

The complied data spreadsheet is then transferred to a computer with the statistical analysis software program SAS installed on it. The spreadsheet is imported into SAS for more complex analysis and Sigma Plot for basic statistical analysis and for generating tables, graphs and other visualizations of the data. The raw output from SAS is copied and pasted into an MS Word file. This is done because SAS lacks an easy to use export interface and the data are not easy to review or share with her advisor as a SAS file. The Graduate Student then combs through the output and determines what from her raw data are significant as well as how to best represent their significance (in a bar chart, a table of values, or a scatter plot as examples). The data elements that are found to be significant are extracted and pasted into additional MS Word files, along with any correlating visualizations of the data generated with Sigma Plot. Visualizations from Sigma Plot can be exported into a variety of common image file formats. The Graduate Student exports her visualizations as jpegs as they are easy to paste into MS Word files. These MS Word files and the information they contain serve as the foundation of her eventual publications.

Another aspect of her data set are the gel images that result from her genomic analyses of soil samples. The gel images show each sample as a “lane” with multiple bands across the lane. The pattern of bands in the gel image acts like a fingerprint for a soil sample through demonstrating a binary presence or absence of genetic materials. The gel images are then fed into another software program to perform a cluster analysis based on the identified patterns to generate distinct groupings of samples. The Graduate Student is quantifying the gels by two types of measurements. First, she determines the intensity of the bands. Second, she uses spectroscopy to quantify the bands at a certain optical density.

The images are produced as .sc files, which is a proprietary format produced by the specialized software (“Quantity One”) that generates the image. A copy of the .sc image is then edited using image capture software to enhance its utility (the original image is retained for authenticity verification purposes). Images are converted into .tiff files for the cluster analysis, and into .jpegs for insertion into MS Power Point presentations or MS Word for publications.

The Graduate Student uses representations of both her numerical data and her image data in her papers and publications.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	Spreadsheets and gels (images) of "raw" data points,	Numerical files: More than 100 / range from 1-400kb, Image files: More than 100 / range from 150 kb to 3MB	Spreadsheets are .xlsx and .csv formats. Images are in .sc, format	This stage includes the processing and calculations that are needed to generate usable data points. .CSV files are generated by lab equipment, which are then converted to .XLSX for usability purposes. .SC is a proprietary format
Preparation / Compilation	Bringing together the needed data points into a single spreadsheet / editing images for usability	Numerical:1 file / 270kb Images: (unknown)	.xlsx / ,sc, .tiff, .jpeg	Additional experiments and calculations are performed. Information about soil quality and soil composition is brought together for analysis purposes. Images are edited and converted to other formats.
Statistical Output	Graphs, Tables and Plots	Numerical: 50 + files / (unknown) Image: (unknown)	SAS, Sigma Plot and .docx files	Output from SAS and Sigma Plot are pasted into an MS Word document for use and preparation for publication.
Publication	Publication ready graphs, tables, plots, and gel images	(unknown)	.docx	Summarizations of the data that will appear in publications.
Ancillary Data				
None				

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). Empty cells represent cases in which information was not collected or the scientist could not provide a response.

3.3. - Target data for sharing

The Graduate Student believes that the spreadsheet generated through the "preparation / compilation" stage of her data's lifecycle would have the most value for other researchers. The usefulness of the image data that she generates outside of her own work would be limited to demonstrating the validity of her work if it were to be questioned.

3.4 - Value of the data

The Graduate Student listed one possible use of her data would be to conduct a comparison of soils at different geographical locations. Another possible value would be for conducting a "meta-analysis" of regional differences in soils.

The Graduate Student stated that, unlike many of the other data sets generated in her lab, the data she generates is less interesting to modelers because modelers don't account for soil biology. Her impression is that soil biology is too complicated to boil down into a metric. Modelers are interested in what goes into the soil and what comes out of it, but not as interested in what happens to drive the changes in between, which is what she is focused on.

3.5 - Contextual narrative

The Graduate Student mentioned that the Gas Chromatographer is run off of a Windows 95 computer which forces her to enter the results of GC runs into her spreadsheets manually.

The Graduate Student typically crops her images to make them easier to understand or highlight the important elements, but she keeps all the originals to demonstrate the validity of her work. It's unclear if these original images would have any value for others or how long they would need to be kept for her own purposes.

The Graduate Student mentioned that she would be interested in obtaining data from some of the other graduate students in her lab. Data pertaining to nitrate levels coming out of ground water and the carbon to nitrogen ratio of the above ground plant biomass were specifically mentioned by the Graduate Student. The barriers to obtaining this data include the Graduate Student's uncertainty on who to ask, lack of knowledge as to how developed these data sets are, and the amount of time and effort it would take another graduate student to find the specific data and prepare it for the Graduate Student's purposes. Her lab group and graduate students engaged in related research have recently begun to hold meetings to share their research work and ideas with each other.

The Graduate Student is unaware of how her fellow graduate students manage their own data, but stated that she would need to know some key pieces of information about their data before being able to make use of it. These include:

- Origin - Which plot the data came from and what treatment was applied to the plot,
- Timing - When were the data collected – the sampling time points in particular
- Replication – Whether the data came from a field replicate or a technical replicate

This information would allow the Graduate Student to be able to link her data to the data that she received from another graduate student (although the graduate student may have to do some additional processing of the data she received first).

The data generated by these graduate students came from the same plots of land and are generated by the same lab equipment as the Graduate Student's and so the Graduate Student already has a base level of familiarity with this data. She stated that if she did not share a lab or equipment with other graduate students who were generating data that she wanted to make use of that she would need to know the name of the equipment that the data were run on and the standard operating procedures / methods that were used in generating or processing the data.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Not discussed.

4.2 - Stakeholders

Not discussed, although certainly her advisor would be considered a stakeholder (if not the owner) of her data set.

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

Not discussed

4.4 - Attribution

Not discussed

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

5.1 - Overview of data organization and description (metadata)

The Graduate Student works with two data types: numerical data and images, and organizes them in different ways.

The numerical data are generally kept in individualized excel spreadsheets and are categorized by the type of experiment and then grouped by the (planned) publication that will result from the data. Raw data are kept in separate columns from data that are being complied, calculated, or otherwise prepared for analysis. If the steps taken to prepare the data for analysis merits an explanation (or a reminder of what was done), the Graduate Student inserts a comment into the column heading or cell as appropriate.

Some of the key data points and the information needed to understand the data points may be repeated across several spreadsheets as needed. An example of repeated information is information about the soil samples that would enable the data points to be tracked back to the sample bags of soil kept in storage. Redundancies are also built into the spreadsheets to enable the Graduate Student to trace the linage of her data.

The image files of the gels are named and organized according to the date they were produced. This enables the Graduate Student to connect the images to information about how they were generated in her lab notebook. The Graduate Student typically crops or edits her images to make them easier to understand or highlight the important elements, but she keeps all the originals to demonstrate the validity of her work. All of the file names of the images contain the date on which they were generated to link them with entries in the Graduate Student's lab notebook. The cropped images contain the words "version 2" in their file name for easy identification.

The Graduate Student indicated that her lab notebook is the key for identifying the connections between her spreadsheet data and her image data. The spreadsheet data files and the corresponding SAS data files can be connected to each other through the plot number and the replication number listed in each of the files.

Another method used by the Graduate Student to organize her data and related files is to only use certain formats for particular types of information. Numerical data is kept in a spreadsheet, presentations are all in Power Point, and written documentation such as methods sections, preliminary exams, or proposals is within a word document. Her lab notebook is not electronic but a physical notebook, so it is unclear as to when she transfers information from the lab notebook into MS Word, or Power Point for documentation, publication, or presentation purposes.

The Graduate Student mentioned that the information contained in her planned publications would be sufficient for others with similar training to make use of her spreadsheet data. To her, the publication is the concise representation of your research and should contain enough information that would allow someone else to reproduce your work. Her publications will include a description of her methodology and/or references to manuals that describe the standardized procedure in her field that she used to generate the data. Prior to publication the information needed for others to understand her spreadsheet data would be contained in her lab notebook.

For her image data, the information contained in her eventual publications would not be sufficient for others to replicate her work. Potential users would need to know more about her methodology including how she extracted the DNA, what the PCR conditions were, how the gels were run to obtain the bands, how the gel images were analyzed within the gel compare analyses, and the optimization and tolerance values that were employed.

5.2 - Formal standards used

None

5.3 - Locally developed standards

The Graduate Student developed her own method and procedures for organizing and describing her data files. This includes making annotations, when needed, to describe the processes taken to prepare her data for analysis. These annotations may appear in the data file itself, in the case of her spreadsheets, or may appear in her lab notebook next to a pasted in print out of the data in the case of the images.

5.4 - Crosswalks

Not discussed

5.5 - Documentation of data organization/description

The Graduate Student's lab notebook is the primary means through which she documents her work with this data set. She described her lab notebook as her "work diary" which allows her to back track and check her work in the event that something goes wrong or questions arise later on. She has incorporated writing detailed notes in her lab notebook, including the "recipes" for her solutions to process her samples, into her daily work routine, and often uses it as a checklist of activities. The Graduate Student also reported printing out some of the images of gels that she has generated and taping them into her lab notebook, and then writing up accompanying labels, notes and other needed documentation. Information from her spreadsheets may also be written into her lab notebook to accompany the images.

As with any lab notebook, her entries are in chronological order. This organizational style means that she has to flip back and forth through the book to get a complete picture of what she did for any one particular aspect of her data set or for a particular analysis.

Access to the information contained in her lab notebook would be essential for others to have if they were to try and understand or make use of the Graduate Student's data.

The Graduate Student mentioned that she will print out a paper copy of her data set at the "Preparation / Compilation" stage and place it with her inventoried soil samples, in case anyone at Purdue wanted to re-analyze her data for their own purposes. She takes this step to provide people with enough information to decide whether or not they should keep the soil sample or not. The default information that is captured and associated with each soil in insufficient to determine its potential value.

Section 6 - Ingest / Transfer

Not discussed

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

The Graduate Student expressed both interest and skepticism about sharing data with others.

At one point in the interview she stated that she did not have a clear idea of who would be using her data for another purpose. However, she also indicated a real interest in obtaining and making use of some of the data being generated by other graduate students associated with her lab to augment her own research though she has never actually asked anyone for their data. She also stated that she includes some of her data about the soil samples with the inventoried samples themselves in case others at Purdue wanted to reanalyze the samples.

Later on in the interview she articulated potential uses for her data in comparing soil composition in different sites or geographic locations. She also stated that someone else may want to reanalyze her data using a different type of analysis or just look at her data in a different way. She stated that once her article and data were published that she would not have a problem with other people taking her data and reanalyzing it.

7.2 - Embargo

Not discussed

7.3 - Access control

Not discussed

7.4 Secondary (Mirror) site

Not discussed

Section 8 - Discovery

Not discussed.

Section 9 - Tools

The Graduate Student uses a variety of lab equipment and software to generate, process and analyze her data. Some of the lab equipment software only runs on operating systems that are outdated, such as the Gas Chromatographer is run off of a Windows 95 computer, which means that she has to transfer some of her data into her spreadsheets manually.

The Graduate Student stated that others would need to know the name of the equipment that the data were run on and the standard operating procedures / methods that were used in generating or processing the data in order to feel confident in making use of the data.

Some of the software that she uses, particularly for the image data, produces data in a proprietary format. The Graduate Student does convert most of her data into more commonly accessible formats; however anyone wishing to make use of her original data would need to have a copy of the software that produced it.

Section 10 – Linking / Interoperability

The information recorded in the Graduate Student's lab notebook is her primary means of connecting her spreadsheet and image data. These linkages include her making print outs of the key data for her electronic files and physically pasting them into the notebook.

The bio numeric software that she uses does enable her to convert her images of banding patterns into binary matrices, and then converted again into similarity matrices that can then be entered into SAS as variables. These binary matrices could also be used to generate “band classes” which could be expressed as band class tables. These tables could be used as variables and incorporated into statistical analyses. The Graduate Student will convert her image data into numerical data if she is seeking to do a big composite analysis, but she implied that simply demonstrating the results from the image data in her publications is usually sufficient for her purposes.

The Graduate Student stated that some of the journals in her field do accept supplementary files of data and associate these files with the articles they publish, while other publishers do not offer this service. If the Graduate Student is required or presented with the option of making her data available to others it is very important to her that the data be linked with the publication that resulted from the data so that users can access the data from the publication and vice-versa.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

Not discussed

11.2 - Gathering information about users

Not discussed

Section 12 – Data Management

The Graduate Student does not have a lab computer available for her use. Instead, she keeps her data on her personal computer. When she needs to use SAS or other software that she does not have installed on her computer she transfers her data to a computer that has the software installed.

12.1 - Security / Back-ups

The Graduate Student uses her personal computer as the primary means of storing and securing her data. She does make back-up copies of her data on a personal external hard drive every 10 days or so. She has set up her backup software to provide her with a reminder after 10 days have elapsed since her last back up. She also backs up her data on to her department’s secured network space once a semester or so, although she indicated that this is more driven by the amount and importance of the work that she has done with her data since her last back up than by particular time interval.

The Graduate Student is not terribly concerned with the security practices taken by a repository or publisher once the data and the article that results from the data are published.

12.2 - Secondary storage sites

In addition to her external hard drive and the department’s network space, the Graduate Student uses “Drop Box” as a means to store copies of her data that she deems critically important.

12.3 - Version control

The Graduate Student’s data set is not meant to be iterative. Once she has the final data points in hand she will not be adding to or building on the data set. Therefore, version control is not a priority for her.

Section 13 - Preservation

13.1 - Duration of preservation

Not discussed

13.2 - Data provenance

Not discussed

13.3 - Data audits

Not discussed

13.4 - Format migration

Not discussed

Section 14 – Personnel

Not used in this profile.