

Data Curation Profile: Agronomy / Biofuels

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	February 27, 2012	
Date of Last Update		
Version of the Tool	1.0	
Version of the Content		
Discipline / Sub-Discipline	Agronomy / Biofuels	
Sources of Information	An interview conducted on May 24, 2011 (duration of 1:04:28).	
Notes	<p>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.</p> <p>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.</p>	
URL	http://www.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 is collecting data that could be of use to others in her lab to augment their own research activities and may have value outside of her lab as well. She believes that data needs to be shared in her research community to better communicate findings and is quite willing to share her own data as an appendix to her eventual publications. However, she is uncertain as to which journals in her field would accept data as a part of the publication. In order to be useful her data would need to be connected to the corresponding article as the article would contain a description of her methodologies in generating and working the data, information that would be needed for others to understand the data. Her data may require additional labeling or descriptions before it could be used by other researchers in her field.

Section 2 - Overview of the research

2.1 - Research area focus

The Graduate Student's research focus is on a particular species of switch grass, *Miscanthus*, and its potential use as a source for biofuel. *Miscanthus* has only recently received much attention and so there is not much known about *Miscanthus* currently. The Graduate Student studies how *Miscanthus* cycles nutrients, nitrogen and carbon particularly, and achieves high yields compared to other perennial crops. Of particular interest is the season cycling dynamics of *Miscanthus* and its effects on the concentrations of nutrients.

2.2 - Intended audiences

There are two primary audiences for her data: one, her fellow graduate students and other researchers directly affiliated with her lab, and two, other researchers who are studying the use of *Miscanthus* or other plants as sources of biofuel.

2.3 - Funding sources

The Department of Energy (DOE) and the United States Department of Agriculture (USDA) both support this research. The Graduate Student was unsure of whether or not the DOE required her to share her data with others (her awareness on the USDA's stance on data sharing was not discussed).

Section 3 - Data kinds and stages

3.1 - Data narrative

The Graduate Student has collected two years' worth of data on nutrients, nitrogen and carbon concentrations, and yield on *Miscanthus* (2009 and 2010). She has also collected this data on two other plants for comparative purposes.

The first stage in her data lifecycle is to gather raw data about these plants from the field. These data are collected by hand and written down onto paper and then transferred into her lab notebook. Plant samples are also collected once a month from each of the 3 types of plants for the 7 month growing season in each year, and are then processed in the lab. Processing includes feeding a sample into a CN analyzer to generate raw data points on carbon and nitrogen concentrations, and drying the plant tissue to get weights for her yield data. The outputs are captured in spreadsheets from the CN analyzer or as a physical "receipt" that lists the data points generated by a spectrophotometer in the wet lab. The Graduate Student prints out these spreadsheets and receipts and pastes them into her lab notebook, alongside of the data she has gathered from the field. This lab notebook serves as the official source of the data, meaning that if discrepancies in the data were discovered later in the lifecycle, the Graduate Student would return to her lab notebook to resolve the discrepancy. The Graduate Student documents the methodologies and references the protocols used in generating the data in her lab notebook alongside of the data. The source of the data (field, CN analyzer, spectrophotometer) are not labeled directly in the lab notebook, but could be determined by the units of measurement and other distinguishing characteristics. Any of the plant samples that remain after processing are dried and stored, with some samples kept in a freezer if needed. These samples are labeled with the plant, date (presumably the harvest date, though this was not made clear), and plot number from where the plant was harvested.

The next stage is comprised of transposing the data that were gathered physically and recorded in the lab notebook and entering them into an Excel spreadsheet. In addition to transferring the data between media, the Graduate Student also synchronizes the data through converting them to a common unit of measurement.

The Graduate Student then performs some additional calculations on the data points to prepare them for analysis. First, she generates an average across the 4 replicate samples that she collects. Once

she has the averages she creates new spreadsheet files, copies the averages into these files, and then begins to make some comparisons between them. If she observes that a particular data point appears to be an outlier she will rerun the sample if possible to increase the reliability of her data. These reruns are labeled through coloring the cell that contains the data point. She will also perform additional calculations on these data points to create a “usable number” for reporting out the results. For example, she generates yield as grams per meter squared but yields are usually reported in tons per hectare. She will calculate out her yields to this unit so that other people can interpret and more easily understand her work in her presentations and papers. Standard curves are also generated at this point and placed with the corresponding data in Excel.

The next stage is analyzing the data. Using a balanced ANOVA analysis, the Graduate Student seeks to identify statistically significant differences in the means between the dates samples were taken or between the 3 types of plants using Minitab software. She could import her spreadsheets into Minitab to conduct this analysis, but tends to copy and paste her data into the software instead as she organizes her data a little differently in Minitab than in her spreadsheets. She will also bring in her raw data on individual reps into the analysis to increase her confidence in the results from analyzing the means. The Graduate Student uses SigmaPlot to develop graphs and charts to include in her presentations or papers, but she also generates visualizations of her data as a part of the process of analysis. Seeing visual representations of her data helps her to determine if anything seems out of place or if she appears to be on the right track.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	Data points in lab notebooks; spreadsheets of raw data points	~200 page lab notebook	Physical; xlsx	Data collection is mostly physical at this point. Even data generated electronically are copied into the lab notebook.
Transposition	Spreadsheets of calculated data	200 / 75 – 100 kb	.xlsx	Data are entered and synchronized with each other for analysis purposes. The # of data files are more or less equally distributed amongst the three plant types.
Calculations & Conversions	Spreadsheets of averages for comparison and reporting purposes	50 / ~100kb	.xlsx	New spreadsheet files are made and some of the data points from the previous stage are copied. The Graduate Student described this stage as both a processing stage and a “cleaning-up the data” stage.
Statistical Analysis & Graphs	Minitab and SigmaPlot files for statistical analysis and visualizations of the data	200 / 25-50 kb	Minitab; SigmaPlot (.jnb)	Data visualizations may serve as a pre-cursor to the analysis in addition to being generated for papers and presentations.

Publications & Presentations	Sigma Plot graphic or table generated in MS Word or MS Excel	5 / (unknown)	Sigma Plot (.jnb); .xlsx; .docx	Visualizations that are developed specifically for inclusion into papers and/or presentations.
Ancillary Data				
(none mentioned)				

3.3. - Target data for sharing

The calculations and conversions stage of her data is the stage that would likely have the most value for others and therefore be the stage that the Graduate Student would be most likely to share. This stage incorporates transformations of the data into units that could be reported out and more easily understood by others.

3.4 - Value of the data

The Graduate Student believes that that her data could be used by others in her lab as a means to augment the research and findings of others and to draw more informed or better supported conclusions. Modelers could use her data to better calibrate their model to better reflect what one would find in real-world settings.

Individuals outside of the Graduate Student's lab may also be able to use her data as a means to better understand her research. They may also be able to run comparative analyses of their own across different climates, different soil types, etc.

3.5 - Contextual narrative

According to the Graduate Student, some of the journals in her field permitted data to be published as an appendix to a paper; however this is not yet a common practice in her field. She also expressed disappointment that many in her field did not provide their data as an appendix in their paper. As a consumer of research papers, the absence of the underlying data leads her to wonder if the research is really representative of what she assumes it is representative of.

In addition to having access to the data generated by others working in her lab, the Graduate Student also expressed an interest in gaining access to the biofuels data being generated at Oak Ridge National Labs.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Not discussed

4.2 - Stakeholders

Not discussed

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)

For the spreadsheets containing her initial data points, the Graduate Student creates a separate tab for each rep that she runs. Within each spreadsheet tab representing a rep she lists the month the sample was taken, the plot number and the type of tissue used for the sample. The averages of the data points are kept in separate tabs in the same spreadsheet.

Within her spreadsheets the Graduate Student uses descriptive column headings to organize her data (worksheet). She also identifies which data points were rerun from her samples by coloring the cell of the data point. The result of the rerun is indicated by a particular color (rerun data points that are kept are one color, rerun data points that are not kept are another color).

The Graduate Student's primary means of organizing her data is to store all her Excel spreadsheets and her Minitab files in one folder. Subfolders are used to organize the data by date and then by analyses. She indicates a relationship between her files through creating file names that evoke an association between one file and another for her. She may include dates in her file titles as well. Ideally, she would like to have her data organized first by year, and then by yield, carbohydrates, proteins and output from the CN analyzer, and finally by her individual analyses. She may eventually reorganize her data around the presentations and papers that she will be generating from her data.

5.2 - Formal standards used

Not discussed

5.3 - Locally developed standards

The Graduate Student uses her own standard for organizing and describing her data, though her practices are informed by her advisor and good lab practice. She describes her data as "polished" to her, but that someone else who wanted to understand and make use of her data would need more thorough labeling than what is currently provided (which is primarily column headings in the spreadsheets and descriptive file names), as well as a copy of the published paper to understand her methodologies.

5.4 - Crosswalks

Not discussed

5.5 - Documentation of data organization/description

The lab notebook serves as the primary means of documentation of the raw data. The Graduate Student does not use the lab notebook beyond the raw data stage. Once the data enter into the transposition stage of the lifecycle documentation is done electronically, though it is not clear if documentation consists of solely of the annotations made in her spreadsheets or if other forms of documentation are used as well.

The Graduate Student feels that her data are documented and described well enough that others in her lab could understand and make use of them. However, she does believe that Agronomy researchers external to her lab would require additional information before being able to understand and make use of her data. At minimum, researchers outside of her lab would need to have access to the methodologies she used and procedures she followed for developing the data set. This information would be captured in the publications that resulted from her work.

Most of her practices in developing and working with her data come from standard operating procedures (S.O.P.s) that have been adopted by her lab. She does not currently link her data to these S.O.P.s or list which ones she is using in her data set. This information may be useful for others seeking to understand and make use of her data in lieu of a published paper.

Individuals without background and training in Agronomy would require additional contextual information beyond what a publication would provide to use this data.

Section 6 - Ingest / Transfer

Not discussed

Section 7 – Sharing and Access

7.1 - Willingness / Motivations to share

The Graduate Student is generally open to sharing her data with others, and of her field making data more openly available. She operates under the assumption that “the more you reveal the better for the scientific community”.

The data generated by the Graduate Student may be useful for others in her lab. Her data on carbon and nitrogen concentrations would be useful for modeling purposes as a specific example. Some of the data generated by other graduate students in her lab would be useful for the Graduate Student to have access to for her own research. Soil and microbial data generated by another graduate student was mentioned during the interview. One of the barriers discussed by the Graduate Student is the lack of a central repository of data being generated by the lab that could be accessed by lab members and collaborators. Currently, data sharing in the lab or amongst collaborators, when it occurs, is done through negotiation on a one to one basis.

The Graduate Student also expressed a desire to share her data as a means to give the readers of her publications a more complete understanding of her research. She envisions that her data could be provided as an appendix to the paper, and that it would probably consist of a more in depth table of the averages across the reps per month, per tissue.

The Graduate Student is not currently planning to share her data outside of including it as appendices to her publications, although she is unsure of whether the funding agency supporting her research requires her to share her data. She has not yet determined which journal she will submit her research to, nor has she investigated which journals permit publication of data as an appendix.

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 several pieces of lab equipment in her research to generate her data, including a Spectrophotometer and a CN Analyzer. One of the steps she takes in working with her data is to convert the results she receives from these machines so that she can make valid comparisons across the data that are produced. She will also perform additional calculations on these data points to create a “usable number” for reporting out the results as a part of her work with the data.

No special tools, other than software that can interpret Excel spreadsheets, are needed to work with the data.

Section 10 – Linking / Interoperability

Linking the data to the published paper would be essential for others seeking to understand or make use of the data. The paper would contain the methodologies she used and describe her data protocols that would be necessary to give the data context and meaning. Ideally, the paper and the data set would cite one another and the citations would serve as active links that could be followed by interested parties.

Most of her practices in developing and working with her data come from standard operating procedures (S.O.P.s) that have been adopted by her lab. She does not currently link her data to these S.O.P.s or list which ones she is using in her data set.

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

12.1 - Security / Back-ups

The Graduate Student makes back-up copies of her data set onto an external hard drive located in her office. Back-ups are performed every two months. Data produced by the CN Analyzer are delivered directly to the lab tech's computer when they are generated. The lab tech keeps a copy of this data.

The Graduate Student keeps her external hard drive in her office which is locked when no one is present. There are no special measures taken for securing her data on her laptop computer other than having to enter a log in ID and password for the laptop.

12.2 - Secondary storage sites

Not discussed

12.3 - Version control

The Graduate Student indicated that version control for her data, specifically tracking any changes made to the data set, identifying why they were made and who made them, would be beneficial to people utilizing the data in the future.

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.