Data Curation Profile: Agricultural and Biological Engineering / Eco-hydrology

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	December 2, 2011
Date of Last Update	
Version of the Tool	1.0
Version of the Content	
Discipline / Sub-Discipline	Agricultural and Biological Engineering / Eco-Hydrology
Sources of Information	An interview conducted on May 25, 2011 (duration of 1:50:44). Transcribed. An email dated December 2, 2011 with the subject line "Re: Data Interview Follow-Up"
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	

Section 1 - Brief summary of data curation needs

The graduate student employs multiple types of data to develop and validate a new application of an existing model, SWAT, traditionally used in assessing environmental impacts. Many of these data types are generated by other students and, as no standardized policies or procedures for documenting, describing or organizing the data are in place, it is sometimes difficult for her to understand and make use of the data. The graduate student is aware of the need for her to consider additional uses of her data by others once she graduates but is not sure how to address this need in developing her own data. As a graduate student, she feels that she lacks the knowledge, experience and position to address the need for common approaches and practices to handling data. Talks on these subjects have begun amongst the graduate students in her lab and things are beginning to improve.

http://dx.doi.org/10.5703/1288284314990

Section 2 - Overview of the research

2.1 - Research area focus

The graduate student is researching the impacts on sub-surface drainage where land use has recently changed from the production of commodity crops, such as corn and soy, to bio-feed stocks which include perennial grasses, such as *Miscanthus* and Switchgrass, and fast growing trees. This analysis of sub-surface drainage includes assessing the amount of water and the nutrient quality of the water in drainage. The goal of this research is to test the theory that producing bio-feed stocks generates environmental benefits for the land.

An important element of The graduate student's research is an effort to test the validity of the Soil and Water Assessment Tool (SWAT) in capturing the hydrologic cycle at the field scale. SWAT is an open source model that has been used extensively as a means of documenting environmental impact at the watershed scale. The purpose of SWAT was to produce a model that would not need to be calibrated with specific field data, but could be used to simulate field conditions to try to estimate basic soil erosion, water quality, and other hydrologic functions. The SWAT model draws on four key areas: climate data (including precipitation and heat unit information), soil classification, slope of the topography of the landscape and land use.

Her research data comes from two research stations local to Purdue, the Water Quality Field Station ("WQFS") and the Throckmorton field station ("Throckmorton"). The WQFS has generated detailed data sets pertaining to water quality from actual observations and measurements for 15 or more years. These data sets enable The graduate student to compare a field scale iteration of the SWAT model against the actual data collected, allowing her to study and account for deviations in the performance of the SWAT model with confidence. The expected outcome will be a SWAT model that has been calibrated and validated to operate at the field scale.

2.2 - Intended audiences

An immediate audience for the data generated by the graduate student would be the other graduate students who are conducting research on land use impacts, the production of crops for bio-energy and other research questions based on data generated from the WQFS and Throckmorton stations.

Other audiences would include researchers in Agronomy, Agriculture and Biological Engineering, or related fields who were conducting similar types of research, and modelers seeking to test and validate their own models or applications of existing models.

2.3 - Funding sources

Not discussed.

Section 3 - Data kinds and stages

3.1 - Data narrative

The data used by the graduate student come primarily from two distinct locations: the WFQS and Throckmorton sites. The WQFS site is composed of prime agricultural land, has growing and collecting data on bioenergy crops for many years, and is primarily a research based facility rather than one meant to represent commercial farming practices. The Throckmorton site is composed of more marginal land, has not been used as a site for collecting data on bio-energy crops for very long, but does include some additional data kinds that are not collected by the WQFS, such as surface runoff measurements. The differences between the two sites do present challenges in being able to generate data set that are comparable with one another.

The graduate student generates her own data, but also makes use of the data gathered from other students and faculty. The majority of the data that she collects herself is the Leaf Area Index (LAI) from Switchgrass and *Miscanthus*. The LAI data are generated through taking solar radiation measurements both above and below the plant canopy. These measurements are gathered biweekly over the summer; this year a total of ten sampling times were scheduled. Seasonal summary files of the data are generated at the end of the season. A series of calculations are then applied using the measurements to estimate the canopy structures of the area of the leaf over a certain square area. The instrument used to gather the LAI data produces an Excel 2003 spreadsheet (.xls), which is later converted to Excel 2007 (.xlsx) formatted spreadsheet by the graduate student. The graduate student will also gather additional data if asked to do so by her advisors. For instance, one of her advisors felt that gathering data about the crops through destructive sampling was needed and so she generated this data through this method.

The graduate student mentioned that she is also collecting LAI data points from prairie lands. Collecting these data points go above and beyond what she is likely to use herself in developing parameters for the model, but having these data points available just in case she decides to use them outweighs the time and effort of collecting them. She recognizes that these data points may be of interest to other researchers working at the sites.

The data collected from others are the nutrient composition in the water runoff from the field, water nutrient composition, crop height, biomass, land use, land cover data from the USGS, and climate data from the National Climate Data Center (NCDC) or the Indiana State Climate Office (iClimate). The external climate data is used to fill in the gaps from the weather stations located at WQFS and Throckmorton. The initial stage for these data is acquiring them from other graduates students or faculty working at the WQFS or Throckmorton. Often times, the graduate student will perform additional calculations or processing to the data before it is assimilated into the spreadsheets that contain the data sets she has developed. Calculations may include aggregating data points, taking the average readings of two data sets on the same subject or using the data from one of these data sets to fill in the gaps in the other data set. A particular instance of processing that was discussed in the interview involved converting spreadsheets containing water flow data into .csv files, cleaning out inconsistencies in the data points, running a script to aggregate and average the data by year and by source. This produced a text file (.txt) containing the data variables that can be used for her purposes.

These observational data sets are used by The graduate student in two ways. First, they are used to set the parameters of the model that she is using. Second, once the model is developed, she will compare the observed data with the output produced by the model to determine its effectiveness. Using the individual data points she has collected, the graduate student will generate a curve. The curve will then serve as the means of generating the parameters of the model. This curve would then be included as a part of the eventual publications that result from her work. The graduate student is currently engaged in plotting these curves and has not yet decided how to generate them or handle them. She may keep them with the data used to generate them and simply create a new tab in the spreadsheets, or, more likely, she may create a separate data file for them. The curves may be generated using MatLab, though SigmaPlot and Excel may be used as well depending on the availability and format of existing scripts and programs to generate these curves and her and her advisor's specific needs and recommendations. As a result, the different data types that she is generating or acquiring for this project may be treated differently.

The graduate student stated that the data that she gathers or collects and stores in Excel initially will eventually be converted into a .csv file for use with the SWAT model. Once the parameter curve has been created and plugged into the SWAT model, the model will generate simulated outputs. Those outputs are then compared to the original data to validate the effectiveness of the model. Representations of the parameter curves will be inserted into publications, and the documentation for the SWAT model describes a procedure for doing so.

The graduate student has stated that she has generated a .pdf file containing the definition of key terms for her work from the existing SWAT operator's manual. However, this documentation file is not directly linked or associated with the data files currently.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Data Collection and Calculations	Spreadsheets of Leaf Area Index (LAI) measurements	Approximately 10 files per year / size is generally quite small	.xls,	Data are collected to generate parameters for the model. Raw data and calculated data are often stored in the same spreadsheet.
Data Collection and Calculations from Others	Spreadsheets of water runoff from the field, crop height, biomass, land use, land cover data, and climate data	Varies	Spreadsheet: MS Excel, .csv, Text: .txt, .dat (varies)	Data from others are frequently subjected to processing to enable their use as parameters for the model.
Synthesis	Data points from The graduate student and others integrated into spreadsheets		.xlsx, .csv	The source of the data is not always clear from looking at the spreadsheet.
Parameterization	A curve generated from the individual data points		Matlab, Sigma Plot, .xlsx, .csv	Matlab is used to generate the curve. There is some question about whether to integrate the curve generated into the Excel / .csv files, or keep it a separate file.
Model Calibration Validation				(This stage was not discussed in much detail in the interview.)
Publication	Parameter curves presented in a publication			The SWAT model documentation describes how to present these curves in a publication.
Ancillary Data				
Processing Scripts	Scripts that were used to process the data to make it usable for The graduate student's purposes.			Scripts include those developed by The graduate student and a faculty collaborator or those that were inherited from other students.

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 question of when the data could be shared was a bit of a challenge to answer in the interview. The "raw" data would potentially have the most value for others to use, but the data would have to undergo some processing first in order to get into a usable state. When they are first generated, these data are not continuous and are not at a stage where they can be used by others. The "Data Collections and Calculations" stages in the table above are meant to encapsulate the process of generating the data as well as processing the data to a point where it can be understood and used.

3.4 - Value of the data

The statements made by The graduate student in the interview on the value of the data refer to the value of the many data streams that she is using in her work and not just on the Leaf Area Index (LAI) data that she generates herself. The data sets generated at the WQFS and Throckmorton could potentially be used in many different ways if it were to be made available to others and if they could be synthesized together effectively. Specific instances mentioned in the interview were examining the water quality data with crop growth data to explore questions on the life cycle of the plants grown at these locations, or using data generated on the impact of water quality, subsurface drainage, carbon generation, and air quality on climate change. In addition other crop growth models could be tested or validated using these data sets as well.

3.5 - Contextual narrative

One challenging aspect to using data generated by others is that the person who generated the data did so with a particular purpose in mind. The graduate student is using the data for a different purpose and therefore has to sort through the acquired data set to better understand it, to weed out the information that she does not need, and to conduct further calculations or other processing to get the data into a state where it can be applied towards her purposes. The graduate student reports that this is a common issue in modeling work. Modeling work requires that any uncertainties have been identified and accounted for sufficiently enough to enable the modeler to be able to make statements with a fair degree of confidence. If the underlying data are not what is expected or varying significantly from one time period to the next it creates uncertainty that has to be explained or accounted for somehow. Sometimes these uncertainties can be explained with additional information on methodologies or access to documentation, other times they may involve a significant investment in time and resources to better understand the data. Although it's often the case that modelers will grab data that's freely available without contacting the data creators, it's often not clear who specifically is responsible for the data or to whom questions could be directed.

According to The graduate student the modeling community has not yet developed standardized documentation practices, although there are groups that are working on this issue. Currently, it would be difficult, if not impossible to reproduce the work described in many of the papers produced by modeler because a sufficient level of documentation is not provided within the paper.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Questions on intellectual property were not asked in the interview however, the graduate student indicated that she does not consider herself to be the owner of most of the data she is using for her work. Instead, she is just one more person on a long chain of people who collected it. It is not clear from the interview what, if any, ownership rights she feels she has over the data that she generates herself.

4.2 - Stakeholders

Questions on intellectual property were not asked in the interview, but the graduate student implied that she would have no input in the decision on whether or not to share the data. She did state that she would hope that her and her collaborators would have the opportunity to make use of the data for their own purposes before the data were made available to others.

4.3 - Terms of use (conditions for access and (re)use)
Not discussed.

4.4 - Attribution
The graduate student indirectly indicated that she would like to be cited if her data were to be made available to others, but the question was not asked to her directly in the interview.

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

5.1 - Overview of data organization and description (metadata)
The graduate student stated that she recognizes the importance of describing her data and documenting the sources of the data that she is using, but that she is unfamiliar with processes and procedures for doing so. She has started to document her own actions, particularly in working with a faculty member in developing the code they are using to work with the data, so that others could follow her decision chain. She has gone as far as trying to develop an S.O.P. for documenting her data but feels she has not been as successful as she would like to be in addressing the problems she has encountered in this area. She stated that the field of modeling as a whole is starting to recognize the need for standardization in documentation and description but that efforts to develop these standards are still in their infancy. Having no guidelines to follow, she typically generates her descriptions in an ad hoc manner based on what information she feels may be useful for her and others when looking at the data file later on.

The graduate student uses data and code that were developed by other graduate students or other external sources fairly extensively in her work. These "external" data are integrated into the data that she herself has generated, generally as columns within spreadsheets. The graduate student will often make an annotation when external data are used, she did state that the source of the data is not formally documented as clearly as it should be in her spreadsheets which can make it difficult to trace back to the data source. The scripts that were developed by other students to work with the data often lacks sufficient annotation for The graduate student and others to easily figure out and make use of them. The graduate student is not an expert in programming and really needs a step by step guide to the code to make efficient use of it.

The graduate student includes the site name and the date in the file name as a means of identifying where and when the data were gathered, and as a basic means of version control.

The graduate student indicated that the meter used to collect the Leaf Area Index (LAI) data provides the ability to annotate the data when it is saved. She includes the date and location (including the site and the specific plot where the measurement was done) of the data.

The graduate student stated that in order for someone with a similar background to understand her data they would need to know what it was she was sampling, where the sampling took place, when the sampling took place (date and time range), the method(s) she used to sample it, and definitions of relevant terms and concepts. Her primary means of communicating this information currently are the descriptive column headings in her spreadsheets, but she has also compiled key information about her methods. For example, she has generated a .pdf file containing the definition of key terms for her work from the existing SWAT operator's manual. However, this documentation file is not directly linked or associated with the data files currently.

She has received feedback from others about the documentation she has generated. She reports that others have told her that the documentation she is providing is not sufficient for them

to make use of the data, but she also reports that they are not able to articulate what is missing from her documentation. The graduate student reports that she and her colleagues have begun to discuss issues in documenting and sharing data amongst themselves and others as a part of the parameterization process for the model she is using. However, a number of issues have prevented them from making much progress on this topic. These issues include the number of people attending these meetings (7-8 people), the depth and complexity of the issues involved (the technical issues in particular), and The graduate student's lack of technical proficiency and uncertainty over how to facilitate these meetings. These issues would make addressing how to generate and apply any kind of standardized approach to documenting data at the lab-scale quite difficult.

5.2 - Formal standards used

No formal standards have been applied to these data sets.

The graduate student recognized the value in the application of a formal metadata standard and in following a consistent process, but indicated that she does not possess the knowledge to be able to select or apply a standard to her data sets. She expressed a reluctance to even attempt to address applying metadata to her data sets, stating that graduate students do not have the experience or broader perspective that faculty do and therefore would not be qualified to make those kinds of decisions. Even a well-organized graduate student who diligently documents their work will do so from their own perspective and for their own localized needs.

5.3 - Locally developed standards

The extent and quality of description and organization of the data files used by the graduate student varies widely depending upon the graduate student or other personnel that generated the data. It was implied in the interview that generally there is minimal standardization across the data files that are used by the graduate student and that documentation is generally lacking. One example that was discussed in the interview was the presence of a "control code" in the water flow spreadsheets obtained from another graduate student, but no definition of what function or information the "control code" represented in most of the files. One file did contain a "kind of description" of the control code. The graduate student is currently working on deciphering and interpreting this definition.

5.4 - Crosswalks

Not discussed.

5.5 - Documentation of data organization/description

Documentation of the data is generally done through annotation of a column or cell within the file itself.

Section 6 - Ingest / Transfer

Not discussed

Section 7 – Sharing & Access

The graduate student has received data from a number of graduate students associated with the Water Quality Field Station and the Throckmorton Morton Station. One particular example was discussed in the interview. The graduate student asked another student in possession of data gathered from the WQFS to share water flow and water nutrient composition data with her. The graduate student described what she was looking for, but the student had inherited the data from previous students and was not confident that she would be able to identify what data the graduate student was looking for would be located in these files. Rather than try to identify the data that

the graduate student wanted, the student gave her all of the data files she had to sort through herself. The graduate student did find the data that she was looking for but it was clear that some of the data had been manipulated by other students and no clear record was provided as to what was done, when and by whom. As a result there are some uncertainties surrounding the data and its suitability for the graduate student's purposes.

The graduate student did state that the more recent files were better documented than earlier files she has received from other students.

7.1 - Willingness / Motivations to share

The graduate student feels quite strongly that data collected using public funds should be made publicly available. She did express some concerns about the timing of sharing data, specifically questions around when data would be released with respect to the content or publication timeline of graduate student and/or faculty work.

The graduate student indicated that, at least with some of the data, she is seeking to produce data sets that anyone else could use. She would want the opportunity to make use of the data that she generated for her own purposes before sharing it with others and the faculty she is working with would likely have the same disposition towards the data. She is reluctant to make the data available beforehand as she would feel pressured to hurry up and finish her own work.

With this particular data set, the graduate student estimates that she will use only 60% or so of the data that she is generating in her own work. However, much of the data that she is using was not generated by her and so she does not feel that she has the authority to make the decision on if and when to share this data.

The graduate student is very supportive of sharing data amongst her fellow graduate students at Purdue. She spoke very highly about the overall level of respect that the students using WQFS and Throckmorton data have for each other and for the data they are using, and how this group culture has facilitated sharing data internally. She expressed some concern that this respect would not carry beyond the boundaries of Purdue and that the data may be misused in some way by others if it were to be made available.

The graduate student sees improvement in communication amongst the faculty involved with gathering and using data from the WQFS and Throckmorton sites, but that more could be done. One issue is that researchers tend to develop data sets based on their own immediate and individual research needs rather than committing to ensuring the longevity of the data, due to a lack of incentives, though identifying and implementing appropriate incentives would likely be a challenge given the nature of academic work.

Other challenges in sharing data is a lack of knowledge on what information would be useful for others to know both currently and in the future, and a lack of knowledge on appropriate data standards that could be applied to these data sets.

The graduate student has direct experience with trying to make use of another's work and how frustrating it can be. She and an Agronomy professor attempted to decipher the code generated by a previous graduate student, and even though this graduate student had documented his work quite well, he had not explained the decisions he made in developing this code sufficiently for another to understand and make use of it.

7.2 - Embargo

The graduate student did indicate a desire not to make her data publicly available until after she had completed her work, but did not specify a particular length of time between the completion of her work and releasing her data.

The graduate student did describe an instance when she was interested in obtaining a student's dissertation done at the University of Illinois, but being unable to do so because of a 2 to 3 year embargo placed on making the dissertation available. The information will not be of use to her in 2 years which has caused her some frustration.

7.3 - Access control
Not discussed

7.4 Secondary (Mirror) site
Not discussed

Section 8 - Discovery

Not discussed.

Section 9 - Tools

The Soil and Water Assessment Tool (SWAT) is the model employed by the graduate student in her research (see 2.1 - Research area focus). As of this writing, documentation about the SWAT model is available from Texas A&M University. SWAT is described on this website as "… a river basin scale model developed to quantify the impact of land management practices in large, complex watersheds. SWAT is a public domain model actively supported by the USDA Agricultural Research Service at the Grassland, Soil and Water Research Laboratory in Temple, Texas, USA" (http://swatmodel.tamu.edu/, Accessed Oct. 28, 2011).

The graduate student is testing the applicability of the SWAT model at the field scale and is particularly concerned with the definitions of the parameters used by the model as expressed in the input and output file documentation: "Soil and Water Assessment Tool: Input / Output File Documentation Version 2009". Variables and definitions used to simulate plant growth in SWAT, was highlighted as an example of the type of information used by The graduate student in her work, though other sections of the documentation are used as well. Plant growth is addressed in Chapter 14 of the "Input / Output File Documentation Version 2009" (http://swatmodel.tamu.edu/media/19754/swat-io-2009.pdf, Accessed Oct. 28, 2011).

The data gathered or collected by the graduate student is stored in a number of different formats including: Excel 2003, Excel 2007, Matlab, text files (.txt) and general data format (.dat). The data identified as the target data for sharing is in Excel formats, and so potential users would need to have access to Excel or programs that could interpret Excel in order to access the data.

Section 10 – Linking / Interoperability

The graduate student uses data points that were generated by others at the WQFS and Throckmorton for her own research and integrates them into her data files. There are several challenges associated with integrating her data with data generated by others. First, different students use different means of organizing and documenting their data. Second, she does not have an effective process for identifying which data points came from which student. The graduate student does indicate if she as modified the data she has received from others by making a notation in the file name ("mod"). She may also develop annotations to explain the changes that she has made.

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 mentioned that she was storing her processed data points in the same file as her raw data points, though they are kept in different tabs in the spreadsheet file. She expressed concern that this was not good data management practice; that raw data and "touched" data should be in different files. However, she was unsure of what proper data management practice would consist of for this type of data and her purposes for it. She stated that she takes a best guess in deciding what course of action to take in managing her data set.

The graduate student stated that even in her professional experience before returning to graduate school, she did not follow many particular protocols. She would welcome a defined and articulated Standard Operating Procedure to follow for managing and handling data. Some elements that might be defined in an S.O.P. would be where to save files, when to create new files, how to developing working files and keeping them separate from the official record of the data, labeling data files, what information a "read me" file should contain, what metadata should be developed and what format it should be in and limitations on file size. She believe that having such an SOP to follow would save her time and reduce frustration from her having to develop her own ad hoc procedures on the fly.

12.1 - Security / Back-ups

The graduate student and her faculty collaborators rely on their department IT personnel to ensure the security of her files, which are stored on the department's network, and to conduct regular back-ups of their data files. Although generally giving IT personnel high marks for their service, she did recount an incident where a faculty member that she was working with made an accidental keystroke and deleted a file. When they attempted to access a back-up copy of the file they learned that regular back-ups had not actually occurred and that the file was lost.

The graduate student does not make back-up copies of her data on her own.

12.2 - Secondary storage sites

Not discussed.

12.3 - Version control

Modelers tend to run multiple iterations of a model and so knowing which data sets are associated with which iterations of the model is an important consideration to the graduate student. Currently, she attempts to keep track of versioning by adding the date to the file name or by indicating the data have been revised by adding "rev" to the file name. A better means of version control for her data is a lower priority for her, but it is something that she would like to have.

Section 13 - Preservation

13.1 - Duration of preservation

Not discussed in this interview.

13.2 - Data provenance
Not discussed in this interview.

13.3 - Data audits
Not discussed in this interview.

13.4 - Format migration
Not discussed in this interview.

Section 14 – Personnel

Not used in this profile.