# Data Curation Profile: Agronomy / Land Use

| | |
|---|---|
| **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 12, 2011 |
| **Date of Last Update** | |
| **Version of the Tool** | 1.0 |
| **Version of the Content** | |
| **Discipline / Sub-Discipline** | Agronomy / Land Use |
| **Sources of Information** | • An interview conducted on May 26, 2011 (duration of 1:27). Transcribed into an MS Word document: "[last name]_AgGrad_Interview_JCreview"<br>• A worksheet completed as a part of the interviews.<br>• A sample of the profiled data. |
| **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.<br>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** | None |
| **Licensing** | This work is licensed under a Creative Commons Attribution 3.0 Unported License |

## Section 1 - Brief summary of data curation needs

The largest issue with the data set according to the Graduate Student is the separation of the data from its documentation, which is contained in a physical lab notebook. This separation of data and documentation presents a barrier to sharing the data with others, as the data set cannot be easily understood just from looking at the files themselves. The Graduate Student illustrated these barriers through recounting a time where he shared his data with a graduate student in the Agricultural and Biological Engineering department and the subsequent exchanges that were needed to explain the data sufficiently. The information contained in his physical lab notebook would not be easy to integrate with digital data files.

Issues in managing and organizing the data were discussed as well. The Graduate Student develops a "master file" for his data as prescribed by his faculty advisor, which serves as his official record of the data. His other data files tend to be the raw outputs of data processed in the lab or files that result from statistical analyses. The lab notebook appears to be the primary method through which these files could be related to each other, and by which changes to these

http://dx.doi.org/10.5703/1288284314993

files are recorded. Data are primarily stored on the Purdue network but working copies are frequently transferred to other computers via email and flash drive.

## Section 2 - Overview of the research

### 2.1 - Research area focus

The Graduate Student is working with switch grass that has been planted both on beneficial and marginal agronomic lands to investigate how well it performs in thermo chemical conversion for bio energy as well as bio chemical conversion. According to the Graduate Student, the novel aspect of his research is the performance measurement of switch grass not just on marginal growing lands, but on the same land used to grow annual crops such as a legume, corn or a forage crop (such as alfalfa). The research centers on how the switch grass planted on this land interacts with these annual crops.

### 2.2 - Intended audiences

The audience for this data would likely consist of other researchers interested in bioenergy and land use. These are interdisciplinary subjects and so the data are likely to be of interest beyond the field of Agronomy. The Graduate Student is already seeing interest from researcher and graduate student in Agricultural and Biological Engineering seeking to use his data to test their model.

There may also be interest in the data for use outside of research, through agricultural extension work in particular. One potential use mentioned would be to help guide farmers in growing switch grass for bioenergy purposes .

### 2.3 - Funding sources

Not discussed

## Section 3 - Data kinds and stages

### 3.1 - Data narrative

Data are produced from plant tissue analyses, soil analyses, sugars analyses, an analysis of the bio-mass quality and the yield measurements. The main foci of the plant tissue analysis and soil analysis are phosphorus and potassium content.

The initial data stage is gathering yield data which is derived from harvesting a sample of switch grass (a one meter swath) from particular plots of land at the field station. The samples are weighed individually and in aggregate. They are then dried and weighed again to generate the initial dry matter yield. If needed the process is rerun, sometimes more than once to ensure accuracy. The data that are captured through this process are the yields (in kg/ha and lbs/acre), the sample weights (wet and dry weights of samples and subsamples), and the percent moisture of the sample, according to the plots, the administered treatments of P and K, and by clustering based on particular traits (file: 2010 Yield data.xls). Once the data has been reviewed and the results appear to be accurate the data are added into a “Master Spreadsheet”, which serves as an official record of the data.

Plant fiber, bio-mass and soil samples are then subjected to a variety of processing and analyses in the lab. For example, plant samples are dried once again and ground up to prepare them for further analysis. Information about the lab equipment that is used in processing the data are not kept with the data file, but are documented in the Graduate Student’s lab notebook along with the methodologies he employed. Many of the methods and procedures used are fairly standard to the Agronomy field however deviations do occur in response to issues with the sample or if initial results appear to be outside the scope of what was expected. It is not uncommon for errors to appear in the data and so the data are reviewed by the Graduate Student, as well as the lab tech and his faculty advisor as needed, before they are accepted. These deviations and repeats are recorded in the lab notebook. The amount of detail in the lab notebook is sufficient for the Graduate Student to retrace his steps and repeat the process if necessary. As in the initial stage, after the data has been accepted they are added into the "Master Spreadsheet".

Once numbers have been obtained from the samples collected statistical analyses are performed on the data. Minitab is the primary software used for analyses, and Sigma Plot is used most often to generate "publication-worthy" tables and other graphics, although Excel is sometime used for both of these purposes as well. This is the lengthiest phase of the data lifecycle as the Graduate Student explores the data he has gathered to determine their potential significance, which involve viewing and testing the data in a number of different ways. The outputs generated will include "quick and dirty" analyses, or serve as a means to play with ideas. These types of outputs will likely have little to no direct value for his research, but will still be stored in the Minitab file he has generated.

Finally, the tables that demonstrate the significance of his work are fine-tuned and finalized in SigmaPlot and then added into the publication. The Graduate Student intends to publish primarily in Agronomy journals, but may also publish in bio-energy or bio-mass journals. He noted that some of his data will likely need to be converted into different units of measurement depending on where he submits his article in order to match up with the expectations of the journal.

The Graduate Student inherited two years' worth of data (2008-2009) from a previous graduate student, in addition to generating a year's worth of data on his own. The earlier data was overseen by his current advisor and processed by the same lab tech, using the same processes that were used in processing the data generated directly by the Graduate Student. He reported some minor difficulties in integrating the earlier data with his own, but nothing that could not be resolved by contacting his advisor or the lab tech. Physical plant samples are stored after they are used, and so it is possible to return to the sample and conduct some additional analyses if needed. The Graduate Student did perform a fiber analysis on plant samples taken in 2008 and 2009.

### 3.2 – The data table

| Data Stage | Output | # of Files / Typical Size | Format | Other / Notes |
|---|---|---|---|---|
| | | **Primary Data** | | |
| Harvest | Field data – Plant yields | 1 / 15-20 kb | Excel | In addition to collecting data points from the field, this stage includes harvesting plant tissues, taking soil samples and other physical specimens for processing. |
| Lab Work | Initially, data points are prepared for analysis. Ultimately, a "master file" of data points is generated. | 5 / 125 -175kb | Excel | Data are reviewed and discussed with faculty advisor. Questionable data are re-processed. Accepted data are placed into the "master file" |
| Statistical Analysis | Results of experiments and computations performed | Minitab – 1 / "large" | Minitab, Sigma Plot, Excel, Word | Multiple experiments and computations live within one Minitab file. Some calculations are done in |

| | | | | |
|---|---|---|---|---|
| | | | | Excel. Sigma Plot files are generated to represent the data graphically. |
| Publication | Sigma Plot tables integrated into MS Word | Under development | Sigma Plot, MS Word | The exact methods of representing this data in publications are still a work in progress. |
| **Ancillary Data** | | | | |
| (None) | | | | |

**Note:** The data specifically designated by the Graduate Student for sharing is indicated by the row shaded in gray (the “lab work” row). 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 Master File of his data, an excel spreadsheet, would be the data that are the most appropriate to share with others.

The Graduate Student felt that the Minitab file containing his analyzed data would be of limited use to others as anyone else would likely find the file to be more overwhelming than a useful record of what he did with the data.

### 3.4 - Value of the data

The potential value of the data is two-fold. First, the data could be used by others who are researching bio-energy or the response of switch grass to P and K treatments and other environmental conditions. The Graduate Student himself inherited a collection of data sets from a graduate student who had previously worked with his faculty advisor. It is likely that the data set that he is generating will be used by others in this lab group who will perform different types of analyses on it in the future.

Second, data modelers in Agronomy and related fields would want to be able to test their models against the data that the Graduate Student is generating. He has already shared his data with a graduate student in Agriculture and Biological Engineering (ABE) for use in a model. The Graduate Student and the ABE graduate have had many conversations about the data due to differences in data description, methodologies, interpretation and assumptions in each field. The Graduate Student felt that attaching or associating the methodology used to generate the data to the data file itself in some fashion would clarify the nature of the data and help users better understand and make use of it.

### 3.5 - Contextual narrative

The Graduate Student’s use of earlier data generated by a previous graduate student was facilitated by his advisor requiring all of his graduate students to develop and maintain a “master file” of their data set to serve as its official record. The methodology used was not associated with the data’s “master file”, however when questions on methodology came up the advisor was able to figure them out and answer them fairly readily. If the data were to be used by others outside of the lab group, especially someone outside of the discipline, a record of the methodology used would be needed to understand the data.

The land used to grow the plants has been used by the Agronomy department for quite some time for researcher. However, it was used for other purposes in the past which may account for at least some of the unexpected deviations in numbers from expected results. The anomalies produced in certain plots are accounted for when discovered and tracked informally, through the knowledge and memories of researchers primarily, rather than through formal documentation.

## Section 4 - Intellectual property context and information

### 4.1 - Data owner(s)

This question was not asked directly, but the Graduate Student's responses to other questions stated that he sees his faculty advisors as the ones who will be making the decisions regarding the treatment and disposition of the data. This response strongly implies that he sees his advisors as the owners of the data.

### 4.2 - Stakeholders

This question was not asked directly, but in addition to his faculty advisors, other graduate students in the lab who may make use of the Graduate Student's data might be considered stakeholders.

### 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)

Excel spreadsheets are the primary means of housing and organizing the data. The data are broken up into multiple working files for the purpose of conducting analyses. These working files generally contain summary sets of data, broken up into individual worksheets (tabs) by summary treatment groups or clusters of treatments as determined by the Graduate Student, and the calculations or equations performed on the data.

The Graduate Student also maintains a "Master File" of his data that serves as an official record of the data that he has generated. Once he and his advisor deem the data to be reliable and "official", the Graduate Student may re-organize the data a bit to better integrate it into the existing data in the Master Spreadsheet. The Master Spreadsheet includes separate tabs for the data organization such as: "summary by cluster", "summary by treatment" and "summary by analysis".

### 5.2 - Formal standards used

None. It is unclear from the conversation if the Graduate Student follows a particular protocol for making entries in his lab notebook or not, as this topic was not discussed in the interview.

### 5.3 - Locally developed standards

The technique of generating a "Master File" is a common practice for students working in this lab under the Graduate Student's advisor. Although it was not discussed, it was implied that there may be some common elements in these "Master Files" across graduate students, though this is speculation at this point.

### 5.4 - Crosswalks

This was not discussed directly, though it was implied that crosswalks between data generators in Agronomy and eventual data users in other disciplines (ABE as a specific example) will be needed. This would likely go beyond descriptive metadata to the level of sharing the methodologies behind the production of the data in ways that were easily accessible and understandable by others outside of the lab and/or Agronomy discipline.

### 5.5 - Documentation of data organization/description

All documentation for the Graduate Student's data set is in the physical lab notebook that he generates. According to the Graduate Student, Making the documentation kept in his lab notebook more accessible and associated with the data would be an important element of making his data available to others. Given that the Graduate Student is studying agronomy, the information contained in his lab notebook has been written from the perspective of an Agronomist, and some of the details of the methods used are not listed as most Agronomists would understand these details implicitly. Additional information may be necessary to enable non-Agronomists to understand and make use of the data.

## Section 6 - Ingest / Transfer

Not discussed

## Section 7 – Sharing & Access

The Graduate Student believed that the journals where he would submit his articles do not accept data set as supplementary files.

### 7.1 - Willingness / Motivations to share

The Graduate Student has shared his data with other graduate students in his lab and they in turn have shared their data with him. He reported that he was able to use the data sets that he inherited from a previous graduate student who worked under the same advisor without too much difficulty.

Although his thesis is in Agronomy, the Graduate Student is enrolled in an interdisciplinary research program that includes Agriculture and Biological Engineering (ABE). He has shared data with another graduate student in ABE who is seeking to test a model. Sharing his data with someone from a different discipline was challenging and required a fair number of conversations to work through misunderstandings about the data. The Graduate Student feels that being able to attach or associate his methods with the data in some fashion would have facilitated sharing his data and preempted many of the questions and misunderstandings. Although the Graduate Student documents his methods, they are kept in a lab notebook and are fairly inaccessible.

The Graduate Student is open to sharing this data with others, but stated that this decision was really up to his advisors.

### 7.2 - Embargo

Not discussed

### 7.3 - Access control

Not discussed directly. The Graduate Student did indicate that the decision on what data files should be released, when, to whom and under what conditions were really for his faculty advisors to make.

### 7.4 Secondary (Mirror) site

Not discussed

## Section 8 - Discovery

Not discussed

## Section 9 - Tools

The data that the Graduate Student has identified for sharing are in an Excel Spreadsheet. The Graduate Student has found that he is able to import his Excel data into minitab quite easily and imagines that Excel is compatible with the other major statistical analysis software packages, such as SAS, that are currently used by Agronomists.

## Section 10 – Linking / Interoperability

The Graduate Student feels that anyone seeking to understand his data needs to have an understanding the methodology he used in generating the data. Linking the data set to his (eventual) published articles is one possible approach to providing this information with the 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

The Graduate Student's primary means of storing his data is through his account on the University network.

### 12.1 - Security / Back-ups

The Graduate Student feels that storing his data on his personal account on Purdue's network provides a sufficient level of security for his purposes. He does not use the Agronomy department's space on the network that is open to anyone in the department out of concern that his data could be tampered with or changed accidentally.

He does do some work in labs outside of the Agronomy building. When he needs to work with his data in other labs or from home, he moves his data files through flash drives or by emailing them to himself. Moving data off the network through e-mail or through a flash drive is his primary method of making back-up copies of his data. This is not done on a regular schedule and is directed more by his need to access his data outside of the Agronomy lab than for back-up purposes directly. Presumably the Graduate Student's account on the University's network is back-uped, although this was not brought up by the Graduate Student during the interview.

### 12.2 - Secondary storage sites

The Graduate Student will transfer data files that he has finished working on to his home computer via flash drive so that a backup copy exists outside of the Purdue network.

### 12.3 - Version control

The Graduate Student identified version control as a high priority for him. He is careful to keep earlier versions of data files in case he needs to retrace his steps as he is working on processing the data and conducting his analyses with it. Although, he admits that he could/should be doing a better job in tracking previous versions of files.

Once he and his advisor agree that data are valid and complete they are added into the "Master spreadsheet" and are not revised or adjusted. It is unclear if the provenance of the data points are kept, though it was implied that the Graduate Student's lab notebook might be able to provide some of this information if needed.

## Section 13 - Preservation

### 13.1 - Duration of preservation

Not discussed

### 13.2 - Data provenance

Not discussed directly; see section 12.3

### 13.3 - Data audits

Not discussed

### 13.4 - Format migration

Not discussed

## Section 14 – Personnel

Not used in this profile.