Data Curation Profile: Agronomy / Grain Yield

Profile Author	Marianne Stowell Bracke
Author's Institution	Purdue University
Contact	Marianne Stowell Bracke <mbracke@purdue.edu>
Researcher(s) Interviewed	[name withheld], Graduate Student
Researcher's Institution	Purdue University
Date of Creation	November 9, 2011
Date of Last Update	
Version of the Tool	1.0
Version of the Content	
Discipline / Sub-Discipline	Agronomy / Grain Yield (Sorghum)
Sources of Information	• An interview conducted on May 27, 2011 (duration of 0:59:50). • A worksheet completed as a part of the interviews. • A sample of the profiled data (ACRE repeats; grain yield calculations) • 2 Data Dictionaries created by The Graduate Student
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 initial issue of data curation was "inheriting" data and methodology from a previous graduate student. This data required The Graduate Student to create a data dictionary to define what each column in the Excel files meant and to write out the steps of the methodology used in planting, harvesting, and analyzing. Data gathered is basically the same year to year. After figuring out what all the information she inherited meant, she had to go through and look for anomalies that had to be changed or accounted for in some way.

Section 2 - Overview of the research

2.1 - Research area focus
The Graduate Student looks at three efficiencies in growing sorghum: water use, nitrogen, and radiation use. For instance, she compares how much nitrogen was used in relation to biomass yield, sugar content yield, and fiber yield.

2.2 - Intended audiences
Others in her lab, agronomists, modelers, sorghum breeders and others studying biofuels and land use

2.3 - Funding sources
The Graduate Student mentioned that the Department of Energy (DOE) supplied the funding for this research.

Section 3 - Data kinds and stages

3.1 - Data narrative
This data is all gathered from plots at the ACRE farm. Each plot contains 2 treatment rows and 2 border rows, which serve as the replicate for the treatment row it is alongside. There are 10 plant lines, 9 sorghum and 1 corn. There are 3 applications and 1 control. Yield is measured for wet and dry weight, and total yield is calculated using number of plants in a row, row area and weight. After grain yields were calculated, analysis of fiber, sugar and starch follows. The yield and harvest data comes from someone else, but the fiber and sugar and starch composition is her contribution to the overall research. Additionally, she will look at the radiation but she had not begun that at the time of the interview.

The sugar and starch data goes through a spectrophotometer and is compared against standards the lab uses. That data is entered into Excel. Then it is plotted against the calibration curve, giving the total sugar content from the plant. (This gives you 3 columns: the initial readout, after you calibrate, and multiplying it by the biomass = total sugar content).

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Inheriting existing data set	Field data – Plant yields	6 data sheets	Excel	This was data from several years prior, by year. Includes weather, starch, sugar, biomass, and grain yields for sorghum and corn.
Review & Repeat			Excel	Data are reviewed and Questionable data are re-processed.
Harvesting & Processing		6-7 sheets per year/small	Excel	Harvest also includes subsample (which isn't defined). The steps for harvest and processing are detailed in the Graduate Student's Data Dictionary.

Statistical Analysis		6-7 sheets per year/small	R, Excel	
Ancillary Data				
Weather				Not reviewed at time of interview, but anticipated to be needed

3.3. - Target data for sharing
The Graduate Student believes that the aggregate data (by year or by trend) and results of analysis would be worth sharing. She does not believe that anyone would be interested in the various pieces of raw data that go into the aggregate or results.

3.4 - Value of the data
The data could be valuable to others in her lab (agronomists), sorghum breeders, and modelers. Also DOE or others doing biofuels and land use research. The value is in the fact that similar data has been collected over many years, and will continue to be collected.

3.5 - Contextual narrative
This data had been gathered over several years and appears that it will be continued to be collected into the future.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)
This question was not asked directly, but the Graduate Student sees continuity in the data from grad students before and will be passed along to grad students after her. It can be assumed that the data belongs to the lab.

4.2 - Stakeholders
Stakeholders include other researchers in her lab; the breeding group that developed the sorghum she is using; and potentially the DOE as funders of the research.

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. The file names include year + date + researcher name + identifying modification. These are kept in different folders for different analyses.

5.2 - Formal standards used
Since she inherited the data gathering from others she is continuing to use their methodology.

5.3 - Locally developed standards
The Graduate Student developed a data dictionary.

5.4 - Crosswalks
This was not discussed.

5.5 - Documentation of data organization/description
The Graduate Student created 2 data dictionaries that are saved on the lab computer (and shared with us). The first one, the Graduate Student Data Dictionary, defines each column title in the Excel file. For each column title, there is a definition, a calculation, old units, and new units. The other data dictionary briefly describes the methodology in three parts: planting, harvesting, and analysis for yield. This also includes a chart that shows how graphically how the plots are laid out, including replicate and border lines and the parts that are harvested for testing. She created these after inheriting data for previous years. She interviewed the appropriate people to capture this information. The data collected and methods used are the same from year to year.

Section 6 - Ingest / Transfer

Not discussed

Section 7 – Sharing & Access

The Graduate Student could not foresee any the journals she would publish her articles in would want to publish any of the data in its raw forms – only analyzed and displayed in graphs, chart, or tables.

7.1 - Willingness / Motivations to share
The Graduate Student is part of a chain of graduate students that have worked on this data, and anticipates that she will not be the last. She discussed using data and methodology shared from others in her lab, and presumably she would be willing to reciprocate. From the interview it can be assumed that she would be generally open to sharing, but she also does not see the raw pieces of her data as having much value to a wider audience.

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 data that the Graduate Student has identified for sharing are in an Excel Spreadsheet (with compatibility up to 2010). She uses R to run the statistical analysis, then copies and pastes those numbers back into Excel. She selected R because she had used it previously while getting her Masters.

Additional tools identified in gathering the data include a spectrophometer, weather/precipitation gathering, fiber analysis instrument, tool that measures leaf area, and a balance/scale.

Section 10 – Linking / Interoperability

The Graduate Student believes that anyone could understand her data or methodology now by reading the data dictionary she created.

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 her data is on the lab computer on campus, which is administered by the Department of Agronomy. She also has an external hard drive that she uses occasionally to provide additional back-up.

12.1 - Security / Back-ups
The Graduate Student feels that the storage on the lab computer in the Agronomy Department is adequate. Thus she relies on whatever procedures are already in place to ensure back-up and security of data. She does occasionally take pieces of the data on her laptop to work on at home. She then emails the updated data files back to herself or occasionally transfers her it a jump drive.

12.2 - Secondary storage sites
She does have an external hard drive that she uses for back-up. This is done intermittently, not on a set schedule.

12.3 - Version control
Version control is somewhere between a medium and a high priority. When she takes pieces of the data home to work on, she is concerned about reconciling the old and the updated files. When she makes significant modifications to a file, she creates a newly named file, but also saves older versions so that she can jump back if she needs to.

Section 13 - Preservation

Not discussed

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.