Data Curation Profile – Plant Genetics / Corn Breeding

Profile Author	Katherine Chiang
Author's Institution	Cornell University Library
Contact	ksc3@cornell.edu
Researcher(s) Interviewed	Withheld
Researcher's Institution	Cornell University
Date of Creation	3-31-2012
Date of Last Update	
Version of the Tool	1.0 / modified
Version of the Content	
Discipline / Sub-Discipline	Plant Genetics / Corn breeding
Sources of Information	• Initial interview conducted on March 5, 2012. • Second interview conducted on March 13, 2012. • Worksheet completed by the scientist as a part of the interviews. • A sample of the profiled data. • A published paper explaining the research.
Notes	
URL	http://www.datacurationprofiles.org http://hdl.handle.net/1813/29064
Licensing	This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License.

http://dx.doi.org/10.5703/1288284315010

Section 1 - Brief summary of data curation needs

The data are in a series of spreadsheets that are assembled as the field tests progress. They are intended for internal use until the final versions of the data are combined into the tables that appear in an annual publication. The data would need minimal curation to make them useful for subsequent use, The information needed is retained by the researcher as a matter of course.

Section 2 - Overview of the research

2.1 - Research area focus

The College of Agriculture provides a disinterested third party evaluation of corn hybrids for grain yield performance as a service for the growers. "New York Hybrid Corn Grain Performance Trials" (http://plbrgen.cals.cornell.edu/programs/departmental/corn/). Seed companies provide hybrids for testing, and the University sometimes adds hybrids from their breeding programs to the test cycle. They are planted at several locations and data on their performance are gathered and published annually. The goal of the research is to produce reference type data for the seed being tested.

The program has been in existence since the 1930s.

2.2 - Intended audiences

The audiences for these data are:

- growers making decisions on what to grow
- seed companies deciding what to market in New York, and what to breed in the future
- crop consultants on what seed to recommend
- Cooperative Extension offices and educators so they can help growers and breeders
- other plant breeders and plant breeding researchers (both public and private sector)

Secondary audiences are researchers interested in the data for other uses, such as climate change studies.

2.3 - Funding sources

The research is funded by Cornell University's College of Agriculture and Life Sciences, participating seed companies, and Hatch and Smith Lever Act Formula Grant funds. The sharing of the analyzed data is a part of the program. The Hatch and Smith Lever Act funds have no requirements for data management.

Section 3 - Data kinds and stages

3.1 - Data narrative

In spring the program asks the seed companies which seeds they would like tested, offering them a pre-determined number of test locations and replications for a pre-established fee. The program decides on the locations for the tests for each grouping of hybrids. The seeds are planted and the fields are monitored; at harvest the plant yield

data are collected. Those data are analyzed, and then a multi-year analysis is completed.

3.2 – The data table - The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Primary Data - Data Stage	**Output**	**# of Files / Typical Size**	**Format**	**Other / Notes**
Classification data		3 files 5 columns, 20-40 lines	Excel	Hybrids to be tested, specifics on the plot locations and planting replications and randomizations. One file for each maturity grouping (early, medium, late)
Response variables		12 files 12-20 columns 20-40 lines	Excel	Plant stand count, stalk lodging, root lodging, plot weight, grain moisture. Sometimes additional variables if there are significant disease outbreaks like northern or southern leaf blight, gray leaf spot, etc. Sometimes plant vigor ratings (for silage yield potential)
Analysis				Single year yield, percent stalk and root lodging, percent moisture at harvest, plant count See New York Hybrid Corn Grain Performance Trials (NYHCGPT)
Finalized				See CGIFCM
Ancillary Data				
Ancillary Data #1 Planting and harvest dates				Done by location, there can be 3-5 locations per maturity grouping
Ancillary Data #2 Data sheets				Data Sheets are generated for the data that are hand collected in the field (root lodging, disease prevalence, etc.) Those are entered into the spreadsheets.

3.3. - Target data for sharing

The single and multi-year analyzed data are shared.

3.4 - Value of the data

The data are used for growing decisions, they have immediate economic impacts on corn growers using the results. They are also used by the breeders and have economic implications in how the breeders market their hybrids.

3.5 - Contextual narrative

These sorts of yield trials were more common in previous years. Every agricultural experiment station in corn growing states would do these assessments as a service to farmers. Some states still have fully subsidized testing program, others rely on the seed companies paying for testing. With budgets being tighter the seed companies are less willing to pay for the testing and some institutions cannot afford to run the tests.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Cornell is considered the owner of the data, per the Cornell University Standard Product Testing Agreement (CUSPTA).

4.2 – Stakeholders

The companies providing seeds are considered stakeholders and the CUSPTA describes their rights regarding republication of results, publicity and acknowledgement of their support.

4.3 - Terms of use

The seed companies must acknowledge the source of the data in any public comments based on those data. It is also stipulated that the data tables cannot be changed, or data presented partially (e.g. cherry picking from the results) without prior permission from Cornell.
Other users of the data are also requested to cite the data source.

4.4 – Attribution

See terms of use.

Section 5 - Organization and description of data

5.1 - Overview of data organization and description (metadata)

The data are Excel spreadsheets of 20-40 lines and 12-20 columns. Ultimately there are about 54 data files. Currently there is little or no annotation or explanation of the headers (which are abbreviations) because only project staff see the pre-publication data. The header abbreviations are explained in NYHCGPT.

5.2 - Formal standards used

None used, researcher not aware of any standards.

5.3 - Locally developed standards
None, column headers are semi-standard over the years.

5.4 – Crosswalks
None, none needed.

5.5 - Documentation of data organization/description
Each annual report (e.g. NYHCGPT) includes the methodology.

Section 6 - Ingest / Transfer
The data would need additional information about the spreadsheet (column headers, units of measure.) Specifics on the locations of the trials, such as GPS coordinates, would be useful for some reuse scenarios. As none of the files are large they would not need to be compressed. The ability to batch load the files and ease of transfer are high priority. Self submission was a low priority, and automated submission rated as medium.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

7.2 – Embargo
No embargo, the data are meant to be distributed.
The researcher specifically stated that the seed producers would be able to see the full data on their hybrids at any point in the process.

7.3 - Access control
The researcher made no statements about controlling the subsequent use of the data.

7.4 Secondary (Mirror) site
The researcher felt that the need for the data was not widespread, or time sensitive enough to warrant that type of site.

Section 8 – Discovery

The researcher felt the existing channels of data distribution (both online and in print) were sufficient to ensure the farmers, breeders and researchers in counterpart programs in other agricultural extension services were discovering the data. If the Stage 2 response variable data (or an enhanced version of that which included things like GPS coordinates) were available from a repository the researcher thought other researchers might find more interesting questions they could answer if they could “play with the data.” The researcher occasionally uses a USDA dataset on corn yields in New York State, going back to 1896, when looking at trends, and noted that the Cornell dataset is unique in that it is a measure of yield from experimental plots whereas the USDA data are based on surveys of farmers and estimations of yields.

Section 9 – Tools

The university owns the basic farming equipment (planter, combine, etc.) The corn is harvested by combine, there are automated buckets that weigh the corn and test the moisture, they have antiquated software and a data logger that collects those numbers and they are transferred to an antiquated computer. The equipment is not supported by the manufacturers anymore.

The data are all excel spread sheets, the analyses are done in Excel.

Section 10 – Linking / Interoperability

No interoperability needed. There is a url for the online report, publications point to that.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

The researcher knows how many print copies of the report are distributed. They do not collect data on downloads of the pdf from their site, other groups (such as breeders and Cooperative Extension offices) redistribute the pdf. The researcher assigned a high priority to the ability to track data citations but does not do that currently.

11.2 - Gathering information about users

The researcher suspects there is little use of the data by people outside the core group of farmers and breeders.

Section 12 – Data Management

12.1 - Security / Back-ups

The data are stored on various hard drives and thumb drives for backup. The institutional backup service does one of the researcher's machines (as it is in the College administrative offices.) Otherwise backup is done sporadically. There are no security measures taken. The researcher seemed comfortable with the current backup procedures.
The researcher retains the manuscript data sheets in file cabinets for several years in case there are questions about the analysis.

12.2 - Secondary storage sites

No need for secondary data storage.

12.3 - Version control

The dataset goes through many versions as they work through the analysis. Version control is important, and is handled through file names e.g. year+DR for data review, ADJ for adjusted, or numbers for revisions.

Section 13 – Preservation

13.1 - Duration of preservation

The final reports should be preserved indefinitely, there is value in the time series.

13.2 - Data provenance

The report captures the appropriate data provenance. If the data from the Stage 2 were to be reused the citation requirements on reuse would provide the provenance as the citation is to Cornell's program, rather than to an individual.

13.3 - Data audits

Since the data are preserved in print this priority was assessed as a low priority.

13.4 - Format migration

There has been no data migration beyond the transition from manual processes to Excel analyses. The need to migrate the data in the future was given a medium priority.

Section 14 – Personnel

This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.