Data Curation Profile – Plant Genomics

Profile Author	J. Carlson
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	October 27, 2009
Date of Last Update	
Version	1.0
Discipline / Sub-Discipline	
Purpose	Data Curation Profiles are designed to capture requirements for specific data generated by a single scientist or scholar as articulated by the scientist him or herself. They are also intended to enable librarians and others to make informed decisions in working with data of this form, from this research area or sub-discipline. Data Curation Profiles employ a standardized set of fields to enable comparison; however, they are designed to be flexible enough for use in any domain or discipline.
Context	A profile is based on the reported needs and preferences for these data. They are derived from several kinds of information, including interview and document data, disciplinary materials, and standards documentation.
Sources of Information	• An initial interview with the scientist conducted in May 2008. • A second interview with the scientist conducted in November 2008. • A questionnaire completed by the scientist as a part of the second interview. • A published article describing the information management system designed and used by the scientist and his lab.
Scope Note	The scope of individual profiles will vary, based on the author's and participating researcher's background, experiences, and knowledge, as well as the materials available for analysis.
Editorial Note	Any modifications of this document will be subject to version control, and annotations require a minimum of creator name, data, and identification of related source documents.
Author's Note	This Plant Genomics data curation profile is based on analysis of interview and document data, collected from a researcher working in this research area or sub-discipline. Some sub-sections of the profile were left blank; this occurs when there was no relevant data in the interview or available documents used to construct this profile.
URL	http://www.datacurationprofiles.org

http://dx.doi.org/10.5703/1288284315011

Brief summary of data curation needs

The scientist and his lab employ an information management system that they designed specifically to manage, organize, describe, and share their data. This system is an integral part of enabling the scientist and his lab to do their work and, once the data has been analyzed by the lab, serves as a publicly accessible repository for the data.

The scientist stated that he has several needs with regards to his data. He needs a means to ensure the persistence of the data for two purposes. One, the scientist wants to be able to link to the specific data set under discussion in a journal article. Two, the scientist wants to provide a formal standardized means for people to acknowledge the source of the data if the data is re-used or published by others.

The scientist is also looking to incorporate services that support collaboration and community development, such as fostering an ongoing dialogue about the data and discoveries stemming from the data. This dialogue could be generated through user annotations, wikis, blogs, etc. that are attached or connected to the data in some way. One specific service mentioned was to generate connections between the data and the relevant literature in the field.

Overview of the research

Research area focus

The research centers on plant genomes; specifically on how a plant regulates and controls certain inorganic materials.

The scientist and his lab group seek to analyze the presence and function of a set of metallic and other inorganic elements that are found within the molecules of plants. Several species of plants are grown by the scientist's lab group under controlled conditions and the presence of these elements is measured to better understand how these elements are regulated and controlled by plants.

Intended audiences

The data is freely accessible and available to anyone to use as they see fit. Generally, the data is designed for use by other researchers working in the area of plant genomics, particularly by plant biologists.

Funding sources

NSF. The scientist is mandated by the NSF to develop a data management plan and share his data with others outside of his lab. He specifically stated that one of the reasons for building his information management system was NSF's data sharing requirements.

Data kinds and stages

Data narrative

The data consists of measurements of the amounts of different inorganic elements from samples of the Arabidopsis plant which are grown under defined conditions (soil, lighting, time, etc. are controlled). Information about the plant variety and the defined conditions are captured as metadata at the onset of planting the seed.

Once the defined amount of time has passed, a sample is harvested from the plant for processing. The processing consists of drying the sample, digesting it in nitric acid, diluting it with distilled water and then running it through an inductively coupled plasma mass spectrometer to measure out the amounts of 18 different elements present in the plant tissue. Metadata about the procedures used and the steps taken are added at each stage of processing. Data from the sample are then generated by the spectrometer as extracted as a .csv file. The .csv file is then manually uploaded into the information management system.

Once uploaded into the information management system the data is processed and reviewed by the analytical chemist who generated the data for quality control purposes. The data are then released into the "private" side of the information management system, where it can be used for analysis by researchers within a lab.

After a period of time (typically six months), with permission of the data owner, the data is released to the "public" side of the database. Once in the "public" side, the data are freely available to anyone.

The information management system is hosted outside of the scientists' lab by the Information Technology unit at the scientist's institution. As a result the scientist did not know the size of the database or of the typical size of the files.

The categories in the "data stage" column listed in the table below were developed by the authors of this data curation profile. The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray.

Data Stage	Output	Typical File Size*	Format	Other / Notes
Raw	Plant Sample	NA	NA	Sample is taken from a plant.
Gathering Descriptive Information	Metadata about the plant sample		postgreSQL	Information about the sample is added into the system through web forms.
Processing	Amount of the 18 ionomic elements present in sample	Unknown – more than 2,000 samples are run daily.	.csv file	Sample run through ICP-MS instrument to generate data.
Ingest	An entry into the information management system (a postgreSQL database)		postgreSQL	Csv file is integrated into the postgres database. Metadata are attached to the data. Data are reviewed before it is fully released into the system.
Internal Release for Analysis	Summaries and graphs of the data		.csv or .pdf	Access to the Data is limited to the researcher or lab group initially
Public Release for Analysis	Summaries and graphs	Database contains data on approximately 125,000 samples. Average file size is unknown	csv or .pdf	Data are typically publicly released after six months of ingest
		Augmentative Data		
Images	Photographs of the plant trays		.jpg	Photographs are used for verification or explanatory purposes. These images are also publicly available, but are not easily accessible

Note: The data specifically designated by the scientist for sharing are in the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

Target data for sharing

In addition to supporting the work of the scientist, the information management system functions as a public repository for the data. Data are typically released for public access 6 months after they were generated as a natural part of the data management lifecycle supported by the information management system. Data are made available as csv or pdf reports (see "contextual narrative" below).

Use/re-use value of the data

The size of the data set (over 1 million samples have been done), its comprehensiveness (all samples are tested for all 18 elements) and the standardized and detailed level of description of the conditions in which the samples were generated (as presented in the metadata), enable scholars to undertake research projects that were previously very difficult, if not impossible, to do.

Contextual narrative

Data collection is highly standardized and adheres to a strict set of procedures. Ingest of the data into the information management system for analysis and eventual dissemination is a critical aspect of the lab's workflow. Data are still being collected.

Although much of the seeds are generated by the scientist and his lab, he accepts seeds and processes samples from other researchers. The data from these 3rd party samples are managed using the lab information management system that he has created, just as the samples from his lab are. Any researcher may send a seed and request that it be grown and processed by the scientist's lab group, as long as they are able to provide the necessary descriptive information about the sample through an "order form" in the information management system.

The information management system provides several types of data outputs as well as options for viewing or downloading the data.

Graphs and charts generated within the information management system:

- Z Score Graph
- % Difference Graph
- Z Score Value
- Weight Normalized Value
    - Note: The scientist indicated that there was no need to save the charts and graphs themselves as they can be generated from the numeric data in .csv format.

Downloadable reports:

- CSV Report – for each line(s) selected this csv file shows the observed values (in ppm), weight normalized values (in ppm), Z score values, percent difference values for each of the 18 elements being analyzed. Data can be manipulated and used to generate charts, graphs, etc.
- PDF Report – for each line(s) selected contains tray information, data analysis settings, tray photos (if any), the observed values, normal values, z score values, a pot summary (in ppm), z score graph and an element percent difference graph (if applicable). Data are static and cannot be manipulated

Other data types:

- Photographs – photos of the plants within the trays are used for verification or explanatory purposes. Photographs may not always be available. Photographs are not attached to the downloadable reports, but can be downloaded manually by right clicking on the image. The scientist expressed some interest in incorporating the data and the image together for a user to download jointly, although this is a low priority.

According to the lab's website, 5 publications have resulted from this data as of this writing.

The information management system and its use have been documented through a peer-reviewed publication and a presentation. Both are publicly available.

Intellectual property context and information

Data owner(s)

The Scientist views himself as the data owner until the data are made publicly available. Once is released into the public domain the scientist feels that it is public information and that it is available for anyone to use as he/she sees fit.

The ownership and rights over the data generated for these 3rd party researchers are not yet formally defined. This system does provide a means for the 3rd party researcher to embargo their data. The standard embargo lasts 6 months, but the Scientist will prolong the embargo period if asked to do so by the 3rd party researcher.

Stakeholders

Stakeholders include the 3rd party researchers who have submitted data to be processed and analyzed by the scientist's laboratory, the IT department at the scientist's institution, and the NSF as the funding agency.

The exact number or identities of the 3rd party researchers are unknown, and their degree of ownership over the data has not yet been formally defined.

The scientist contracts with the IT department at his institution to provide the technical framework for the information management system. This contract includes hosting, maintenance, and running back-ups. The scientist indicated that he migrated the data from a proprietary database format into postgress SQL, an open format, in part because it was a format supported by the IT department.

Terms of use

Currently the data in the lab's information management system that are designated as being publicly accessible are made available without restrictions. The user is asked to give attribution for use of the data in the resulting publication by citing an article authored by the lab group describing the information management system.

No terms of use statement appears on the information management system site and "no terms of use" document accompanies the data when downloaded.

Attribution

The scientist indicated that the ability to cite this dataset in his publications is a high priority for him.

Currently, although the scientist and his lab team request that acknowledgement be given for use of the data, they have no formal procedure or requirements in place to cover attribution over the acknowledgement or attribution of data from information management system in scholarly articles or other outputs. The scientist would like to be able to connect readers of their articles directly to the relevant data sets in information management system. Presumably this would require the assignment of a persistent URL, DOI or other enabling another means of persistence.

Organization and Description of Data for Ingest (incl. metadata)

Overview of data organization and description

The application of standardized metadata to the dataset is a high priority for the scientist. The information management system developed by the scientist is designed to force his lab group and 3rd party researchers to enter the needed metadata as they place an "order" (submit their plant sample for processing and analysis). As a result, much of the metadata needed to understand and use the data has already been obtained and is believed to be well structured and understandable by other researchers.

The forms used to gather the metadata are typically composed of drop-down menus rather than free text boxes. This is done for quality control purposes and to enable a high degree of standardization in the metadata.

Each order is assigned an "order number" which is attached to the seed and used to connect the eventual sample with its metadata. The order number serves as the data's unique identifier.

The scientist did express that he did not yet have all of the metadata that he would like to have about the samples in the information management system, but did not address which specific metadata.

The scientist expressed a strong desire for others looking at or using the data to be able to annotate the data with notes or information as to how the data is/was being used and/or resulting discoveries or questions.

Formal Standards used

The scientist expressed a strong interest in integrating his data sets with others through the use of shared, community-supported ontologies. However, he also stated that the types of ontologies needed for his specific purposes do not exist yet.

A standardized, community-supported controlled vocabulary is used for gene names.

Locally Developed Standards

The information management system supports a locally developed metadata schema that includes information about the plant sample, the conditions under which it was grown, and the resulting data is captured as a part of the workflow within the information management system.

The information management system ensures that a controlled vocabulary is used through employing drop down menus for researchers to describe and categorize their samples.

Crosswalks

The scientist has expressed a desire to use standardized metadata to enable interoperability and sharing of data sets. As he has not yet selected a formal community supported metadata schema for his data, a crosswalk to such a standard is not needed at this time, but may become necessary at a future date.

Documentation of data organization/description

The information management system and its functionality has been described in a published article. This article provides some information as to how the data and metadata are generated, collected and disseminated, however it lacks sufficient detail in and of itself to be used as a basis for curating data.

Ingest

Data is ingested into the information management system manually by a lab technician at the “ingest” stage of the data workflow (see “Data Kinds and Stages” section). The information management system currently serves as the data repository once the data are made publicly available.

Access

Willingness / Motivations to Share

The scientist is motivated to share his data for a couple of reasons. First, the funding from the NSF requires that he share the data. Second, he would like to develop a community of practice for his particular research field. He sees the sharing of data as a means to enable and encourage communications and collaborations between researchers.

In addition to his data, he is also interested in sharing the information management system he has developed as open source software for others to install and use for similar data sets.

Embargo

In the information management system researchers can keep their data private, hidden from public access and accessible only to the researcher (and presumably information management system admin).

Once uploaded into the information management system, data are typically released to the public after six months (though the scientist has stated that this is an arbitrary time frame). Data are not released automatically. The scientist receives an automated notification that data has been in the information management system for six months. The scientist then contacts the data owner and inquires as if it is okay to make the data publicly accessible. Data are released upon receipt of confirmation from the data owner.

Access control

The information management system contains a “private” side and a “public” side. Data that has not yet been moved to the public side is only accessible to system administrators, the scientist’s lab group, and individuals granted access by the system administrators (those who submitted their samples, etc.)

Once the data are “published”, they are in the public domain and available to anyone.

Secondary (Mirror) site

The scientist indicated that access to the data through a secondary site was a high priority for him. Ideally, the information management system would never be offline.

Note on Access

The scientist has stated that the next upgrade to the information management system will enable him to “archive” data. Archiving data will cause the data to be removed from display without removing it from the system itself. The purpose of this feature is to clean up the system by removing experiments that were not completed for one reason or another.

Discovery

The scientist indicated that the ability for people within and outside of his discipline be able to find his datasets easily was a high priority for him. The use of highly structured metadata and controlled vocabulary in the information management system is designed in part to facilitate discovery of the data by others.

The ability for people to discover his datasets using internet search engines such as Google was listed as a high priority by the scientist. The scientist has expressed a strong desire to make the data in information management system available through other access points beyond the information management system itself (the library catalog was mentioned specifically by the scientist as a desirable access point).

The public interface of the information management system allows the browsing of data through tray number, experiment number, and status.

The public interface of the information management system allows for the searching of data through a basic and an advanced search. The “basic” search options of the information management system are available for anyone to use. However, in order to use the “advanced” search options, users of information management system must register by creating a user account.

The basic search consists of the following search boxes: gene type/ATG number, parent line, line name, tray number(s), or order number.

The advanced search adds the ability to search by phenotype. A user can select one of 18 elements and indicate the amount or range of amounts of the selected element the retrieved samples should contain.

The information management system enables the use of Boolean operators to search for data and allows searching for a range of values.

Tools

The ability to connect the data to visualization or analytical tools is a high priority for the scientist. The information management system provides the ability to conduct some calculations (z-scores, % differences) on the data and then to generate charts and graphs based upon these calculations. These are key functions of the information management system.

Data can be exported in a tabular format (.csv) or as a report (.pdf). MS Excel (or a csv reader) would be needed to use this data in its tabular format. Users would need to have Acrobat Reader installed in order to access the data in the report format.

The scientist expressed an interest in adding literature mining functionality to the information management system.

Interoperability

The scientist is generating data for three different types of plants. Although data from each of the three plant types are housed in different databases, the data sets are all connected to the information management system and are all made available through a common interface.

The scientist has expressed a desire to link the data within his information management system to other relevant and publicly accessible data sets. The scientist indicated that support for web services APIs is a high priority for him. Currently, the information management system is enabled to make use of web services to disseminate data and "communicate" between other similar systems to share data. A specific example used was to link his and other data sets by gene names, so that a user could click on the gene name in a database and be presented with any additional data or information related to the gene in the other databases.

The information management system uses a controlled vocabulary to define gene types and has integrated the controlled vocabulary into the information management system. This controlled vocabulary is overseen and curated by a 3rd party.

The information management system lists articles produced by the scientist and his lab group as an RSS feed on the information management system website. The scientist also expressed a desire to be able to link readers of articles that used this data directly to the relevant data sets in information management system.

The scientist expressed a strong desire to integrate literature mining into the information management system to connect the data to the relevant body of literature that exists.

Measuring Impact

Usage statistics

The ability to see usage statistics on the number of people who accessed his data set is a medium priority for the scientist.

The information management system displays statistics on the number of experiments, samples, unique lines, unique genes, orders submitted and order completed. These and other statistics are used in project reports submitted to the NSF and others. The scientist stated that the current statistics do not get him everything he needs, but did not elaborate on needs for specific statistics.

Gathering Information about users

Google analytics is used to gather information about users IP addresses, where users are from and how often the visit information management system. Although the scientist stated that he was largely satisfied with the information he received from using Google Analytics, he did express a desire to learn more information about people who visit information management system and download data. Other than names, the scientist did not mention specifics about the type of information he would like to have.

The scientist expressed a desire to know and keep track of what articles and other outputs have resulted from the data, in addition to the outputs produced by his lab group.

Data management

In addition to serving as a data repository, the information management system is the mechanism through which the data are managed in the lab.

Security / Back-ups

The agreement made by the scientist with his university's IT organization to build and maintain the information management system includes a provision for backing up the data on a regular basis.

Secondary Storage Site

The scientist indicated that a secondary storage site for his data is a high priority, but that a secondary storage site at a different geographic location is not a priority for him.

Preservation

The scientist and the IT unit at his organization have negotiated an agreement in which the scientist rents server space to host the information management system and pays a fee for the maintenance. Maintenance includes migration to the most current version of Postgress. The scientist views this arrangement as enabling the preservation of his data because as long as he pays this fee the data is taken care of by the IT unit.

Duration of Preservation

The scientist does not have a particular time-frame in mind with regards to the preservation of his data. Instead he indicated that “It would be useful (to preserve the data) until all knowledge has been extracted and published.” The scientist mentioned a cost benefit analysis a means to determine the duration of preservation; however he does not currently have a definite means or criteria in place to determine when or under what conditions the data should no longer be preserved.

Data Provenance

Data is time stamped as it is processed and analyzed, though it is not clear exactly when time stamps are issued or which events merit a time stamp.

The scientist indicated that documenting any and all changes that have been made to the dataset over time is a high priority for him.

Data Audits

The scientist indicated that the ability to audit his dataset to ensure its integrity over time is a high priority for him.

Version Control

The scientist indicated that the ability for the repository to provide version control for his dataset is a medium priority for him.

Format Migration

The scientist indicated that the ability to migrate datasets into new formats over time is a medium priority for him as he stated “[I do not] foresee that Prostgress will evaporate in the near future. And, it’ll have to be some really big breakthrough in some other database to make it super necessary to move it.” He does acknowledge, however, that the “database landscape” will change over time and that he will need to be open to the migration of the data in the future.

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.

Primary data contact (data author or designate)

Data steward (ex. library / archive personnel)

Campus IT contact

Other contacts

Notes on personnel