Data Curation Profile – Cornell University, Plant Response to Herbivory

Profile Author	Sarah Wright
Author's Institution	Cornell University
Contact	Sarah Wright, sjw256@cornell.edu
Researcher(s) Interviewed	Witheld.
Researcher's Institution	Cornell University
Date of Creation	March 27, 2012
Date of Last Update	
Version of the Tool	1.0 /modified
Version of the Content	
Discipline / Sub-Discipline	Plant Response to Herbivores
Sources of Information	• An initial interview conducted on February 23, 2012. • A second interview conducted on March 1, 2012. • A worksheet completed by the scientist as a part of the interviews. • A published article describing the data collection process and experimental conditions.
Notes	Questions were added to the DCP V.1.0.
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.

Brief summary of data curation needs

The scientist uses a master spreadsheet to organize his data by experiment and by year. His data is primarily tabular, but also includes large chemical analysis files in proprietary formats. The scientist's current system works well for him, but he would also like to deposit his retrospective data associated with publications into a repository and to develop a workflow to continue the practice with future publications. This will require some work, and he is willing to devote some time during his upcoming sabbatical to the organization and preparation of his data sets for deposit.

The scientist is also very excited about the potential of data sharing as a way to give new life to the research by spurring further discussion and interactions over the data. He feels that sharing his data helps him by increasing the interest in his research, and therefore the number of citations his publications receive.

Overview of the research

Research area focus

The scientist studies the mechanisms and ecological consequences of plants' induced responses to herbivore damage.

The scientist and his lab group are trying to better understand the impact of herbivore-induced changes in volatile organic compounds emitted by plants on pollinators. Using a single species of plant growing in the field, they perform correlative surveys of fruit set and pollination with herbivory. They also perform bioassays and field choice experiments with multiple treatments.

Intended audiences
For the most part, the scientist thinks that the data would be of interest to people both inside and outside the field, but mostly to those studying applied plant-insect interactions.

For a subset of the data, the scientist can see potential for wider interest, perhaps in the pharmaceutical realm. However, he feels that as long as pharmaceutical companies would rather invest in making new versions of old compounds than in identifying new compounds, this subset has very little use outside of his own studies. Therefore, the scientist feels that the value of this subset of his data to others is in the future, not in the present.

Funding sources
The scientist receives funding from the NSF and Cornell University. The NSF now requires a data management plan as well as encourages data sharing and preservation beyond the life of the grant, but the researcher hasn't applied for funding since these requirements were put in place.

Data kinds and stages

Data narrative
The data consists of field survey data and bioassays measuring herbivore damage, pollination, fruit set under different experimental treatments (addition of herbivores, treatment with MeJA (plant hormone), removal of visual cues, etc.) Information about plant location and habitat description are collected as ancillary data at the outset of the experiment.

In the lab, data is also generated by subjecting plant samples to analysis using coupled Gas Chromatography – Mass Spectrophotometry (GC-MS). The raw GC-MS data is analyzed using the instrument specific proprietary software to measure the area underneath the peaks for specific known Volatile Organic Compounds (VOCs). The peak area data is then entered into an Excel spreadsheet along with the field survey data, ready for statistical analysis.

Statistical analysis of the data is performed using StatView, and tables and graphs are prepared for publication using Origin data analysis and graphing software.

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
		Primary Data		
Raw - Field	Field data from survey & bioassays	5 / 50-100 KB	MS Excel 2007	Field data is collected for each plant
Raw - Lab	Samples are subjected to analysis using Gas Chromatography – Mass Spectromete	16 / 2 MB	Proprietary instrument-specific Saturn GC-MS files	Samples are subjected to GC-MS analysis

	r (GC-MS)			
Analyzed - Lab	VOC amounts entered into Excel spreadsheet	1 / 50 KB	MS Excel 2007	Specific volatile organic compounds (VOC) are measured in the samples by measuring peak sizes in GC-MS data
Analyzed – Field & Lab	Statistical analysis	9 / 20-50 KB	SVD (Statview)	
Finalized		8 / 30-50 KB	Origin graph	Figures are prepared for publication
Ancillary Data				
Ancillary Data #1	GIS data			Plant location is recorded using geographic coordinates
Ancillary Data #2				

Note: 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.

Target data for sharing

With the exception of chemical analysis data, the researcher's data is available to anyone immediately upon publication. The scientist is willing to share the chemical analysis data with others in his field. This restriction is based on the scientist's perception of its limited usefulness to those outside of his field.

For all of the data, the scientist places no conditions on sharing his data, although he would like to be cited in the papers.

Value of the data

The scientist thinks that the field data would be of interest to people both inside and outside the field, but mostly to those studying applied plant-insect interactions.

In the case of the chemical analysis data, the scientist feels that the raw data is the only valuable component because the full spectrum collected in the GC-MS measures much more than just the VOCs that the scientist is looking at. The scientist would share the raw data with others in his field, but feels that the impact might be limited by several circumstances. First, the raw data is available only in very proprietary format, and not many people possess the software necessary to access and analyze the chemical analysis data. Second, pharmaceutical companies would rather invest in making new versions of old compounds than in identifying new compounds. If this changes in the future, then the scientist can see the potential usefulness of his chemical analyses to those outside of his field.

Contextual narrative

The data collected by the scientist is very tightly tied to the scientist's publications. Experimental context is complex, and may not be easily captured other than by linking publications to the data.

The scientist also feels that data sharing is a good thing to do after publication, and results in higher citation counts for his papers. The scientist stated that the plant biology community is very pro-data sharing because of this effect on citations. Data isn't shared pre-publication because of the fear of being scooped.

Intellectual property context and information

Data owner(s)
The scientist considers himself the owner of the data.

Stakeholders
The scientist receives funding from the NSF and Cornell University, but doesn't specifically identify funding agencies as stakeholders.

Terms of use (conditions for access and (re)use)
The scientist defines no conditions for the use of the data, however he would like to be cited.

Attribution
The scientist considers the ability to cite the dataset in his own publications as well as requiring others to cite the dataset a high priority. Data citation is important to him, and he would like that to be a requirement of using his data, although he is unsure how one would go about ensuring that that occurs. The ability to create a basic, public description of (and a link to) the data was also deemed a high priority because it would make the data more discoverable, and more easily cited.

The scientist saw very little use (rated it not a priority) in the ability to restrict access to a data set. He would rather make it widely available.

Organization and description of data (incl. metadata)

Overview of data organization and description (metadata)
The scientist's data is organized in Excel tables linked by a meta-table. In the meta-table, every experiment is described and notes have been added to column headings describing specific treatments in detail.

The scientist is happy with the current organization, and feels that it is sufficient for another person with similar expertise to understand and properly use the data.

Although he is happy with the organization and description, he would like to develop a standard procedure for depositing his data associated with publications into a repository. He plans on going on sabbatical in the fall and would like to work on depositing all of his data retrospectively, with the goal of making this a standard operation in his lab going forward.

Formal standards used
The scientist is not aware of any metadata standards that would apply to his data. He rated the ability to apply standardized metadata from his field or discipline to his datasets a low priority.

Locally developed standards
The scientist's meta-table includes rich description of the each experiment according to a set of defined inputs, but it is not really a standard.

Crosswalks
Since the scientist doesn't use metadata standards, he expressed no need for crosswalks, however he did place high priority on the ability to make his data accessible in multiple formats. This applies especially to the chemical analysis GC-MS data. He is especially concerned with proprietary formats becoming inaccessible when instruments and their accompanying software are no longer supported by the manufacturer. If there is an alternative way to preserve this data, he is unaware of it.

Documentation of data organization/description

Treatments and novel experimental designs have been described thoroughly in peer-reviewed publications.

Ingest / Transfer

The scientist doesn't currently ingest his data into a repository, but he believes that the data would require very little preparation in order to be ready for ingest. He might have to adjust some table and spreadsheet titles to make them user-friendly. He does this when he prepares tables and graphs for publication, to make experimental treatments more easily understood, so he feels that some renaming of the original data tables to match the publication tables and graphs is required. The chemical analytical data would need much more preparation – it would need to be transferred into a format that is compatible with common software, since it is currently stored in instrument-specific proprietary formats.

He considers it a high priority to submit the data to a repository by himself, and would also like to have the ability to batch upload and to transfer the data to a permanent data archive (although he wondered why the original repository wasn't permanent). Of medium priority was the ability to automate the process – it would be nice, but not necessary.

Sharing & Access

Willingness / Motivations to share

The scientist is very willing to share the data once he has published the paper, the sooner the better, because he believes that it increases visibility of his research and encourages future citations.

Embargo

The scientist does not require an embargo. He thinks that science should work via "the fast and frequent exchange of knowledge."

Access control

The scientist sees no reason to restrict access to his data. In fact, the scientist is excited about the potential for social interaction around his data sets. He sees discussions spurring new ideas including new uses of his data as well as new research directions, and would like to have as many opportunities to show and share his data as possible. He sees this as the biggest benefit of depositing his data in a repository.

Secondary (Mirror) site

The scientist rated the provision of a mirror site a medium priority. He can wait if the repository is offline.

Discovery

He imagines that people would find his data set via an internet-based search engine like Web of Science or Google Scholar, linked from the citation in the original publication. He feels strongly that the link between the data and the related publication(s) is very important, and would like any repository to provide the link and to make that linking easy.

Although he would like his data to be available to anyone without restrictions, he sees very little need for the general public to be able to easily find the data set. Highest priority for him is the ability of researchers within his discipline to find the dataset, and for the dataset to be discoverable using internet search engines. The scientist considers the ability of researchers outside his discipline to find the data set to be of only medium priority.

The scientist feels strongly that the ability to create a basic, public description of (and provide a link to) his data is an important way to increase discovery. He sees a parallel between this and the literature databases he currently searches when he's performing a meta-analysis, and thinks it would be easier to pull together the data for a meta-analysis if he could search a repository or registry of data sets. He feels that basic metadata should include when and where of data collection in addition to basic description and the basics necessary for citing the data (author, title, publication date – like a literature citation). The scientist thinks that it would also be nice to have the ability to "find more data like this," and would like to be able to search using descriptive metadata like keywords, species, and geography.

The scientist would also like to be able to collect his data sets according to project (like an NSF project, addressing a research question) – which may continue across years, with data sets added as they are collected.

Tools

Most of the scientist's field data is organized in MS Excel, and can be viewed using MS Excel or other spreadsheet software. The laboratory chemical analysis data is generated using several different instruments including Gas Chromatography – Mass Spectrometry (GC-MS), High Pressure Liquid Chromatography (HPLC), Liquid Chromatography – Mass Spectrometry (LC-MS), Spectrophotometer. All of these instruments export the raw data in proprietary formats, but Chemstation software can be used to view both the HPLC and LC-MS data. The software necessary to use the chemical analysis data is expensive, but is standard among researchers in the scientist's field.

The scientist is also very concerned about the continued accessibility of his chemical analysis data. Backwards compatibility is not a priority of the companies that produce the instruments and accompanying software, and he is already forced to keep an old computer running an obsolete Operating System in order to access some of his older data. The GC-MS he is currently using is no longer supported, and he knows that there will be no future updates to the software he uses for analysis. He is not aware of any other way of saving the raw data files, and worries about losing the valuable data that is only accessible in this data-rich format.

Linking / Interoperability

The scientist publishes frequently in Ecology and other journals that accept supplemental information, although the method may vary, with data included either in the form of SI or as Appendices. The scientist feels strongly that linking between the data and the related publication(s) is very important, and would like any repository to provide the link and to make that linking easy.

The scientist considers the ability to cite the dataset in his own publications as well as requiring others to cite the dataset a high priority.The ability to create a basic, public description of (and a link to) the data was also deemed a high priority because it would make the data more discoverable, and more easily cited. He feels that basic metadata should provide the detail necessary for citing the data (author, title, publication date – like a literature citation), and it would be nice if the repository could generate the citation to be used. The scientist thinks that it would also be nice to have the ability to "find more data like this," and would like to be able to search using descriptive metadata like keywords, species, and geography.

The scientist considers the ability to support web services or APIs a high priority, and would like to be able to use such services to display data sets on his own laboratory web site in a dynamic and linked manner.

The scientist also considers the ability to connect or merge his data with other data sets a high priority, especially for creating meta-analyses. This is also related to his desire to have a function to "find more data like this." He thinks that this would be an important method for compiling long-term data that might otherwise set in different silos. "No one gets a 16-year grant from the NSF anymore," so it's difficult to find these data across years.

Measuring Impact

Usage statistics & other identified metrics
The scientist considers the ability to see usage statistics and to gather information about the people accessing and using the data to be of only medium priority, although he can see the use for administrative reporting. He would like to be able to track data citations, as data citation is important to him, and he considers this the real measure of the value of his data. The potential service he is most excited about is the ability to track and show user comments on his data. He felt that this would potentially provide new life to the data, and could serve as a global virtual lab meeting. After some thought, he decided that he would like the ability to turn comments on and off, in case they got out of control.

Gathering information about users
The scientist was not very interested in gathering information about users, assigning it only medium priority, but could see some use in collecting the usual IP-address based location information.

Data Management

The scientist uses a master spreadsheet to organize his data by experiment. This system works well for him, but he would also like to deposit his retrospective data associated with publications into a repository and to develop a workflow to continue the practice with future publications. This will require some work, and he is willing to devote some time during his upcoming sabbatical to the organization and preparation of his data sets for deposit.

Security / Back-ups
The scientist currently makes back-up copies of his data every 3 months.

Secondary storage sites
The scientist currently keeps a copy on a hard drive stored in a different geographical location in case of fire or other local disaster.

Version control
The scientist considers the ability to enable version control a high priority and currently keeps versions by appending the date of the update as part of the master spreadsheet file name.

Preservation

The scientist believes that the raw data, including tables of field experiment data and the raw files from chemical analyses, are the most important parts of the data to preserve.

Duration of preservation
The scientist believes that his data sets should be preserved indefinitely.

Data provenance

The scientist felt that documentation of changes made to the data set over time was only of medium priority, primarily because he didn't feel like any changes should be made.

Data audits

The scientist feels that data audits to ensure structural integrity over time are high priority in order to guarantee continued accessibility.

Format migration

Format migration is a high priority for the scientist, especially in the case of the chemical analyses, which are especially problematic since the most data-rich format, the raw files, are also in proprietary formats which rapidly become extinct when the instrument companies cease to support the software and/or instruments.

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.