Data Curation Profile – Botany / Plant Taxonomy

Profile Author	Sara Rutter
Author's Institution	University of Hawaii at Manoa
Contact	srutter@hawaii.edu
Researcher(s) Interviewed	[withheld], co-PI [withheld], PI
Researcher's Institution	University of Hawaii at Manoa
Date of Creation	June 22, 2011
Date of Last Update	
Version of the Tool	Data Curation Profile Toolkit is 1.0
Version of the Content	1.0
Discipline / Sub-Discipline	Botany / Plant taxonomy
Sources of Information	An initial interview conducted on June 14, 2011 A worksheet completed by the scientist as a part of the interviews.
Notes	
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 primary data sets for deposit into the University of Hawaii's ScholarSpace repository includes:

- Excel spreadsheets with morphological measurements of specimens of *Astelia* (Asteliaceae) found in Hawaii for the delimitation of taxonomic boundaries and construction of species descriptions.

Other data associated with the primary dataset are:
- Scanning electron micrographs of specimens
- Phylogenetic tree constructed from associated DNA sequence data

The data would be made available to others for re-use once the researcher has published the analyses. Because the researcher believes that interest in the morphological data is limited primarily to plant systematics and conservation biologists, making the data accessible to that smaller population is of more importance to her than providing public access.

The data are well-documented in terms of how data were collected, manipulated, and analyzed. The researcher is interested in sharing data after publication for the purposes of comments and annotations by others and possible re-use for other studies.

Section 2 - Overview of the research

2.1 - Research area focus

The researcher is embarked on a project to taxonomically revise *Astelia* (including 46 taxa), an expansion of her dissertation work that recognized 3 species and four proposed varieties of *Astelia* in Hawaii. The data she has used to develop taxonomic descriptions are from morphological measurements and gene sequencing for alignment analyses. Scanning electron micrographs were produced that document seed coat features. The gene sequencing data will be submitted to GenBank.

2.2 - Intended audiences

The data collected in this research project will interest others engaged in monocot systematics; evolutionary biologists (phylogenetic trees); conservation biologists; bio-geographers; and collators of regional flora.

2.3 - Funding sources

There are approximately 16 funding sources, from local sources (University of Hawaii at Manoa Graduate Student Organization and College of Arts & Sciences awards) to a National Science Foundation award for which the scientists are co-PIs, and the American Society of Plant Taxonomists (ASPT).

Section 3 - Data kinds and stages

3.1 - Data narrative

Morphological measurements using digital calipers are made of up to 10 specimens for each taxon and of approximately 61 characteristics. Specimens were borrowed from herbaria located in many different countries.

Herbarium specimens have data associated with them, e.g. collection data, location data. Some specimens have precise Geographic Positioning System location data (those which scientist #1 collected) and others much less accurate location data noted by the collectors. It is important to note that the herbaria specimens are part of a larger project that will ultimately reclassify the entire genus. For the first curation efforts we will focus on the seven Hawaii species. There are believed to be 41 species in the genus.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data	Morphological Measurements of Specimens	Excel (2008) spreadsheet	.xls	Generating morphological measurements of about 61 characteristics, 10,000 rows; (sometimes a leaf length may take seven cells).
Raw	Measurements of approximately 61 characteristics for up to 10 specimens in each taxon. Another spreadsheet has the list of specimens from herbaria around the world, with location and collection information.	Excel (2008) spreadsheet with approx. 10,000 rows	.xls	The spreadsheet file containing the original measurements is saved and not used for subsequent processing.Morphological measurements are made using digital calipers that input data into Excel
Processed	The mean and standard deviation of the mean are calculated for each characteristic over each taxon.	Excel spreadsheet	.xls	
Analyzed	Analysis using R		.csv, .R, .pdf	R scripts can also be shared to replicate analysis
Finalized	Phylogenetic trees		.pdf, .nwk	MrBayes, Paup, RaxML, J-ModelTest
Finalized	Phylogenetic trees and morphological data (evolutionary history of each morphological character)		.pdf, .nex	Mesquite,
Ancillary Data				
Ancillary Data #1	A Word document is created to describe the specimens examined for each taxon. This usually becomes an appendix in the publication, as supplemental material. The data in the spreadsheet are edited to make corrections and the new spreadsheet is saved as a version.	Word, Excel	.doc, .xls	Collecting information associated with herbaria specimens
Ancillary Data #2	Location data		GPS	Data associated with specimens collected by Dr. Birch
Ancillary Data #3	Scanning micrographs used for morphological measurements		.jpg	

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray (the "processed" row is shaded here as an example). 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 scientist is interested in publically sharing the processed, analyzed, and finalized data products once they have been described in scholarly publications. The scientist would share raw data with immediate collaborators, people in the same laboratory with the understanding that research products generated from the data would not be published before the scientist is able to publish. The scientist would share the analyzed data (R analysis) with others in the field, e.g. herbaria from which samples were borrowed, again with the same conditions previously noted prior to the scientist publishing the results.

3.4 - Value of the data

The scientist views the data as having value for scientists from a broad range of biological disciplines but of interest to a relatively small number of people. In taxonomy, research is valued in years by a few people (rather than in months by a larger population).

3.5 - Contextual narrative

The data collected in this project were in support of a Ph.D. dissertation. The dissertation will be available in ScholarSpace, the UH Manoa institutional repository. Some of the data analyses such as the phylogenetic trees can be submitted to TreeBase once published. The scientist has also analyzed DNA sequences through gene alignment; the sequences will be submitted to GenBank.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

The scientist and the PI of the lab believe that the scientist is the owner of the data. The scientist collected the data and thus owns the data.

4.2 - Stakeholders

Stakeholders include the PI of the lab, the lending herbaria and people who have contributed to the project. Because there were over 16 funders, it is unclear whether they would be stakeholders. At the time of receiving funding none of the funders required sharing of data or data management plans. The data are not associated with privacy or confidentiality concerns. The journals in which the scientist expects to publish include the provision of appendices for the list of examined specimens with their provenance.

4.3 - Terms of use (conditions for access and (re)use)

When the data are submitted to the institutional repository the scientist wants a "how-to-cite" note attached to the record so that users will properly cite the dataset. Citations or attribution for use of the data is a high priority.

The scientist noted that the ability to connect her datasets with others and the ability to link the data with publications and other metadata is a high priority.

Usage statistics and information about people who downloaded the data is of some interest.

4.4 - Attribution

The scientist has not deposited data into a repository before but is willing to submit her data once the analyses are done to ensure that the data are accurate and after she has published research based on the data. The scientist is interested in receiving attribution by users of her data and would like the repository to make it clear to others how they should cite the data.

Section 5 - Organization and description of data (incl. metadata)

5.1 - Overview of data organization and description (metadata)
The morphological measurement data are in an Excel 2008 (Mac) spreadsheet. A Word document defines the column headings and another Word document identifies the specimens examined, their provenance and holding herbaria. Methods are in a Word document that is a part of the final publication.

The scientist generates phylogenetic trees using the software packages noted earlier, that are saved in pdf form. Analyses using R are performed to determine species and the R scripts are saved as .R files.

5.2 - Formal standards used
The scientist is familiar with data in TreeBase, which requires files to be submitted in NEXUS format (a file format in which Mesquite and other phylogeny tree software save files).

We discussed DarwinCore for species identification. Species names are highly controlled vocabularies (Note: uBio is a project to link taxonomic name variations).

5.3 - Locally developed standards
The scientist uses a Word document to define terms used in data files.

5.4 - Crosswalks
Crosswalks between any local standards that may exist and formal standards have not been made, nor were they discussed in this interview.

5.5 - Documentation of data organization/description
The publications will describe the methods used to collect the data and a supplemental Word document will list the sources of herbarium material. The spreadsheet contains the information associated with the collecting of the specimens within each taxon. The second stage of the measurement data shows the means and standard deviations of the mean of the specimens within a taxon. These are the data used to generate the phylogenetic trees.

Section 6 - Ingest / Transfer

The need to restrict access for a period of time is a high priority for the scientist. Once she publishes on the data she would like the restrictions lifted, but use must be with attribution. Prior to ingestion of the data the scientist wants to ensure that the data are "clean" and asked about being able to edit the data once they are submitted. She was told that some editing of the data is workable and will not affect the metadata attached to the files.

The scientist understands the vulnerability of proprietary software (in terms of obsolescence) and can submit .csv files. If R scripts are submitted the scientist wants to make sure that these function correctly.

Section 7 – Sharing & Access

7.1- Willingness / Motivations to share
The scientist would share her raw data with her immediate collaborators but would not want to share it with those outside the project because of concerns about the data being mis-used, the possible inaccuracies in the raw data, and the probability that the data would not be useful or comprehensible to anyone outside of the project in that form.

The processed data for which the statistical means have been determined would be shared with collaborators for their personal use.

The analyzed data would be shared with immediate collaborators and herbaria that lent specimens. (Herbaria may need to report on use of their collections.) Funders may require data submission at this stage to measure progress.

After publication the scientist is willing to share her data. Though she was not sure she wanted to be alerted to every download instance of her data, she voiced some concern about how her data might be used. The scientist noted that citations to her dataset and the requirement to use with attribution were high priorities.

The scientist is especially interested in preserving the data about Hawaii species at UH.

7.2 - Embargo
The scientist requires the ability to restrict access prior to publication. She expects to publish in S*ystematic Botany* or a similar journal. The Hawaii taxonomy will go in *Pacific Science* or a more regional journal, and the genetic work will go in a journal like *American Journal of Botany*. She wants the Hawaii species data to be held in the UH repository. TreeBase (after the publication) is also a repository to which she may submit the data and phylogenetic trees. GenBank will be the repository for the alignment data.

7.3 - Access control
The scientist requires the conditions outlined in 7.1 before data are shared. Once those conditions are met, access can be opened to the public.

7.4 Secondary (Mirror) site
A secondary mirror site for her data was a low priority for the scientist. Information about the back-up of ScholarSpace data was provided to the scientist.

Section 8 - Discovery

The scientist noted that the ability of collaborators, those in her field, and funders to find her data was a high priority. The ability for those outside her discipline to find the data was a medium priority. Being able to search over a general search engine or for the general public to find the data was a low priority. She believes that people within her circle of collaborators and funders will be able to find the data through her publications and other communication routes.

Section 9 - Tools

Tools used to make and record morphological measurements included digital calipers that are connected to a laptop (Mitutoyo), Excel 2008 (Mac), light microscope, and scanning microscope micrographs. Tools to analyze the data included R (open source statistical software) and Excel. To create the phylogenetic trees the scientist used several applications, PAUP, MrBayes, RaxML, Mesquite, jModelTest.

Section 10 – Linking / Interoperability

The ability to link the data with publications and Word documents was a high priority. The scientist also saw the ability to merge her data with other datasets as important. At the time of the interview the scientist did not see a need to support the use of APIs.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics
The scientist would like to know download statistics from the UH repository. Of higher priority is a count of citations to the data in other publications.

11.2 - Gathering information about users
The scientist was somewhat interested in knowing who has accessed her data but did not place this as a high priority.

Section 12 – Data Management

12.1 - Security / Back-ups
The scientist currently backs up her data held on a laptop onto a hard drive of another computer and onto a Mac Time Machine (external drive). The raw data worksheet is always copied before any manipulations are done. The files are backed up twice a day. She does not employ password protection or other security measures.

12.2 - Secondary storage sites
The scientist indicated that the UH repository's current back-up measures were sufficient for her data.

12.3 - Version control
The scientist currently makes all manipulations on versions of the raw data worksheet and of the processed data worksheet. The type of changes made to the dataset are recorded (e.g. null values replaced by mean values) for each version of the file. Changes to individual cells are not recorded. Version control is a high priority for the scientist.

Section 13 - Preservation

13.1 - Duration of preservation
Both the scientist and PI described the useful term of this taxonomic data as long-term with no definite end.

13.2- Data provenance
Information that must be maintained with the data includes the species name, links to the phylogeny files (.nex?) must be stable.

13.3 - Data audits
The ability to audit the data over time to ensure integrity is a high priority. [D-Space ensures "bit preservation", whereby a file will remain exactly the same over time, not a single bit is changed. The repository system also uses a checksum, a tool for verifying the integrity of bitstreams.]

13.4 - Format migration
The ability to migrate the data to new formats when needed was noted as a high priority. Also of high priority was backing up the data onto a secondary storage site (as is done with ScholarSpace), and documentation of changes to the dataset over time.

Section 14 – Personnel

14.1 - Primary data contact (data author or designate)
Identifies the data client and provides contact information for this person.
Scientist and PI

14.2 - Data steward (ex. library / archive personnel)
Scholarly Communication Office, UHM Library

14.3 - Campus IT contact
Scholarly Communication Office, UHM Library