Data Curation Profile – Biophysics

Profile Author	Dianne Dietrich
Author's Institution	Cornell University
Contact	dd388@cornell.edu
Researcher(s) Interviewed	Withheld
Researcher's Institution	Cornell University
Date of Creation	2012-03-30
Date of Last Update	n/a
Version of the Tool	1.0 / modified
Version of the Content	1.0
Discipline/Sub-Discipline	Biophysics
Sources of Information	Initial interview conducted on March 9, 2012. Second interview conducted on March 13, 2012. A worksheet completed by the researcher as part of the interviews.
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.

Brief summary of data curation needs

The researcher stressed the importance of having refined data available and linked to publications. One component of the target data from this dataset would be the protein crystal structures, which are currently in a disciplinary repository and are connected to the published literature. This data can be effectively used by researchers in other disciplines to build off the researcher's work. For this highly specialized example, the value for general public is seen to be very low however, due to the highly specialized nature of the data.

Data generated throughout the project that support the published findings represent a small fraction of the total data generated. While the data that support the published findings have potential for sharing, the researcher stressed the amount of effort needed to make that available and the high potential for misinterpretation by those without the group's specialized knowledge.

The total sum of data generated by the lab does not exceed available resources (for storage and management). The researcher noted ongoing needs to secure raw data (so that it is only accessible to the lab group) and keep it preserved indefinitely for the group's internal use.

http://dx.doi.org/10.5703/1288284314999

Overview of the research

Research area focus

The researcher is in the area of Biological Physics and the lab group studies physical interactions of proteins and develops tools to observe these phenomena. The data that are the focus of this profile were acquired using small angle x-ray scattering in order to examine how proteins respond to a light stimulus. Additionally, for this project, the team collaborated with researchers in the Chemical Biology department.

Intended audiences

Researchers within and outside the field are potential audiences for the data. For researchers working with light-activated proteins, the new model proposed by this research could provide a new insight for their work. Researchers outside the field may be interested in applying the technology the researcher's lab group developed (instrumentation) to their own research.

Funding sources

The NIH and NSF are the primary funders for the researcher. The NIH requires deposit of manuscripts into PubMed Central as a condition of funding and the NSF requires a Data Management Plan to be submitted with grant proposals.

Data kinds and stages

Data narrative

To start, the protein is prepared and raw data are acquired at the beamline. (The researcher also mentioned research pertaining to building the apparatus to collect raw data, but those specifics were not discussed in the interviews.) Typically, two graduate students are in charge of collecting data during the initial experiment. Approximately 100-1000 image files are produced in the initial data stage with a detector that uses a CCD-like camera. Each detector produces slightly different output, but generally, the data produced are close to TIF, and each file is a 1K x 1K pixel image. Immediately after acquisition, there is a quality control check, assessing the data on objective (not scientific) criteria. At this point, raw images or Matlab matrices – ASCII files that contain the reading from each pixel of the original image file – of the data may be used to make this determination. Before concluding the experiment, it is assured that enough good-quality data have been collected to proceed with the more detailed analysis. Once the experiment is over, there may be one more quality control check to pull out any files that can't be used for analysis (because something went wrong with the apparatus during the experiment). At this point, the files may still be images, or they may be in Matlab matrices.

The next processing stage focuses on building models from the experimental data. The data files are clustered together and signal-averaged. At this point, the researcher produces curves that represent the values in the reduced data. The number of files reduces significantly, due to the clustering. From the data, the researcher proposes models to describe what has been observed in the experiment. Next, specialized software takes these curves and models to generate three-dimensional molecular structures. The number of files at this stage of processing increases, because the model structures have now been included. Further processing (using software) allows the researcher to interpret the model in terms of the underlying biological process to validate the consistency of the model.

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

Data Stage	**Output**	**Typical File Size**	**Format**	**Other/Notes**
Raw data acquisition	Approximately 100-1000 image files.	10^6 pixels per image file.	Format dependent on detector used, best approximation is TIF format.	Instrument is built by researcher and team; instrumentation details are separate publications.
First pass analysis	Approximately 100-1000 files, either images or Matlab matrices.	10^6 pixels per image file, or the same amount of data in a Matlab matrix.	TIF or Matlab matrix file.	The purpose of this stage is to perform objective quality control, to eliminate unusable data. (For instance, the protein aggregated or the laser wasn't turned on.) This step can be done by either assessing the original image files or examining processed Matlab matrices.
Detailed processing	Approximately 100-700 files.	10^6 pixels per image file, or same amount of data in a Matlab matrix.	TIF or Matlab matrix file.	Quality control is also performed at this stage.
Building models	Approximately 20 files	Data from 10^6 pixels of an image file in a Matlab matrix.	Matlab matrix file.	The data size is the same as the image/Matlab files above because data points from multiple images/matrices are averaged together.
Fitting the data to models	Approximately 60 files	Uncertain, but each file is in the megabyte range.	Output from ATSAS (software for small-angle scattering data analysis produced by European Molecular Biology Lab)	This stage generates a three-dimensional model structure.

Validation of model* / Publication	Validated model, other representation of data, such as the protein crystal structure from the protein used in the experiment	Uncertain, but each file is in the megabyte range	Same as above for model, PDB for protein crystal structure	

***** Validation of model from a distinct experiment, not a direct result of one measurement

Note: The data specifically designated by the scientist to make publicly available are indicated 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

The data that have been refined and translated into a “common language” understood by others are likely the target data for sharing: for example, the protein crystal structures that are linked to the publications are already made public on the web in a repository that is well-known to researchers in this field. This data are valuable and can be used by others working in related fields. The researcher noted that a small percentage of the raw data directly supporting published results have some sharing potential, but currently are not made widely available on the web.

Value of the data

The protein data structures have high value as they are in a common format and language, and can be used by researchers outside of the field. This data are also currently linked to the researcher’s publications and should be preserved to retain the link between the research outputs. The researcher stressed that all of the data have high value for internal use, since preservation makes re-analysis possible for the lab group. While the researcher said that all of the data should be preserved indefinitely for internal use, there are some instances where it might be preferable to re-optimize an experiment and retake the data.

Contextual narrative

The researcher stressed that all data produced in the stages prior to publication were highly contextual and are likely to be difficult to interpret by others outside the lab. Each experiment has its own set of unique conditions and parameters that are difficult to communicate thoroughly so that others can properly analyze it.

During processing, there should be no changes made to the parameters in the dataset. If a change needs to be made in the middle of the analysis, processing starts again at the beginning step, starting from the raw data files. The researcher notes that the overall process is one of starting from the original files and then continually refining the data until the end of the analysis process. In this sense, provenance of the data focuses more on the details of one person’s analysis, rather than changes made by multiple people over time.

Intellectual property context and information

Data owner(s)

The researcher stated that it seemed logical that the lab group owned the data, but added that there was some uncertainty about ownership in light of funder requirements.

Stakeholders

Primary stakeholders in this data are the researcher, the lab group – consisting of post-docs and graduate students – and any other collaborators.

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

The researcher has concerns about broad access to data generated prior to publication, as there is a high potential for misinterpretation of the raw data. The target data for sharing and making widely available, namely the crystal structures that validated their hypothesis, are in a common understood format, and should be made available on the web after publication in a repository known to researchers in the field.

Attribution

The researcher indicated that the ability to cite the publicly available protein crystal structure is a high priority, as it will be referenced in the published paper and others should be able to find it. The ability of others to cite this data if they make use of it in their research is also a high priority for the researcher.

Organization and description of data (incl. metadata)

Overview of data organization and description (metadata)

Data are stored as image files off the detector (TIF is the closest approximation), Matlab matrices (text), and ATSAS files. The initial data files generated from the detector are typically named by the detector software sequentially (with an option to set a pre-factor to the filename). Once the image data have been processed and converted to Matlab matrices, the resulting text files are organized by filename. There is no agreed-upon convention in the lab for naming files, though filenames are generally interpretable by those in the lab group. (See **Locally developed standards** for further explanation of file naming.)

Formal standards used

There are no formal standards applied to describe or organize the data. The researcher notes that this is due to the highly specialized and unique nature of the lab's research. (Note that instrument design is also within the purview of the group.)

Locally developed standards

Filenames given to the data produced are intended to reveal the most relevant information – for instance, the protein sample used, the conditions of the experiment, timing information, and any other qualifiers or parameters. Every person in the lab approaches naming files slightly differently, so it might not always be obvious how to interpret a file name. The researcher states that it should not be impossible to decipher a filename, as there is a shared understanding of what is important about each experiment within the group. In the ideal situation, there would be an agreed-upon naming scheme in the lab, but this may be impractical given the uniqueness of each experiment.

Crosswalks

Not discussed.

Documentation of data organization/description

The researcher did not talk in depth about a formal system for documenting data organization, though team members keep individual lab notebooks containing accounts of all processing steps during the data acquisition and processing stages. The information in a lab notebook should be sufficient for the researcher to trace a lab member's steps and confirm the validity of a particular conclusion. The team uses common software for processing data (Matlab, ATSAS) so processing descriptions should be understandable by others in the group.

Ingest

The researcher indicated that the ability to personally submit data to a repository was dependent on the interface of the repository. If submission was simple and straightforward, this was not a high priority. The ability to control the release of the data was of high priority for the researcher – data should only be added and/or made available after publication of results.

Sharing & Access

Willingness / Motivations to share

The researcher is willing to make the protein crystal structures publicly available on the web after the results have been published because it allows others to build off the group's work.

For data produced and analyzed prior to publication, the researcher is willing to share selectively, depending on the stage in the research process. In collaborative projects (that is, the lab group is working with researchers from another lab or department) the researcher is willing to share raw data with the collaborators. (This would include data that have not yet been filtered for quality control.) The researcher is willing to share quality control filtered data with others in the field in order to try to clarify any uncertainties in the data at that point. At this point, the researcher is likely to only share a small portion of the data, or share a description of the data (rather than sending the actual data) in order to gather information from peers. As the data progresses through the processing stages, the researcher notes an increased reluctance to share data as widely because the team is formulating their hypotheses. While there may be exceptions to this, such as an elevated level of interest in the field of the problem (which may necessitate a more guarded approach), or conference timing (where data may be discussed earlier than usual), these are more rare.

Once the results have been validated through peer-review and publication, the researcher is more willing to share data. It is important to note, however, that even after publication, the researcher expressed strong concerns about making raw data from the earlier stages available, as there is a high likelihood of misinterpretation. Published results may represent a small fraction of the original raw data that could possibly be shared, though the researcher noted that it would require a great deal of effort to make this available to others.

Embargo

Embargo functionality for a data repository is best tied to the release of the related publication, rather than a fixed amount of time after deposit. The researcher noted that either data deposit would be delayed until after publication of results or the repository should support a feature where the data link “goes live” after the associated paper is officially published. This seemed especially important for the protein crystal structures, since there is great value in having a functional link to that data as soon as the paper is made available.

Access control

The researcher indicated that the ability to restrict access to the protein crystal structures was not a priority. For all data prior to the final data stage, access should be restricted to lab members (see **Security / Back-ups** for further information on internal access to data). It is important to note that the researcher was disinclined to put processed (not final) data in an external repository.

Secondary (Mirror) site

The ability to access the published data at a secondary (mirror) site if the repository goes off-line is a low priority for the researcher.

Discovery

The ability for the general public to easily find the data is not a priority for the researcher. Understanding the data requires such a great deal of specialized scientific knowledge that it would be exceptionally challenging to provide it in such a way that the general public could make informed use of it.

The researcher stated that the ability for those within the discipline to easily find the dataset was a high priority for the protein crystal structures, but not a priority for all of the other data. The protein crystal structures are published in a standard format and are in a “common language” so it is especially important that this data are easily searchable and findable by researchers interested in the problem.

Researchers from outside of the discipline should be directed to the data through the publication. The researcher stated that the ability for people to easily discover the data using Internet search engines, such as Google, was a medium priority. The researcher imagines that people who are specifically interested in the problem, and have the scientific and technical vocabulary to describe it accurately, should be able to find the publications that describe the data through Web of Science (this was specifically mentioned as a critical resource by the researcher), Google or Google Scholar, or the lab group's website.

Tools

The data are generated by a detector that outputs image files plus text readouts representing any monitoring done during the experiment. The lab uses a Matlab-based package to analyze the data. The "nuts and bolts" of this package are available on the web, though the version the group uses has been developed in-house over many years. The researcher mentioned that it would be possible to analyze the data using other freely available packages on the web.

The ability to connect the data to visualization tools is a high priority for the researcher, especially connecting the protein crystal structures to externally built tools as these would be too complicated to develop internally.

Linking / Interoperability

The researcher indicated that support for the use of web services APIs was not applicable for this data. The ability to connect or merge data with other datasets was a high priority for the publicly available data, the protein crystal structures, but not a priority for the rest of the data as the data are time-resolved and it is difficult to make accurate comparisons with other datasets. The researcher was uncertain about the priority for connecting the data with publications or other outputs.

Measuring Impact

Usage statistics

The researcher indicated that the ability to see the number of times the dataset had been accessed was a not a priority, and the ability to track data citations was a medium priority.

Gathering information about users

Gathering information about people who have looked at or made use of the data was a medium priority for the researcher. The ability to have others comment on or annotate the dataset was not a priority; the researcher expressed concern about unmoderated user comments on data (and publications), citing the issues that can arise when misinformed comments on data (or publications) are publicly viewable. It is reasonable to infer from the researcher's statements that the ability to moderate any comments on the data would be worth consideration.

Data Management

Security / Back-ups

All data for ongoing projects are stored in the lab. During the raw data acquisition stage, there are multiple copies made of data in case there is an issue with one of the machines. After data acquisition, a third copy is made and transferred to the lab and the first copy is deleted. Data are only available to lab members and is protected by firewalls. Ideally, lab members back up their data once a week. When a project has finished, or an individual leaves the lab group (for example, after graduation), the data are transferred to an external hard drive and given to the researcher.

Secondary storage sites

While there is no secondary storage site, some data are backed up and the media is stored in a fireproof box on site.

Version control

The researcher noted that version control was a high priority for the data as it was being processed in the lab by the team. Specifically, the team uses Matlab workspaces to take “snapshots” of the data allowing for minor corrections during processing. It is important to note that this data are not part of the candidate data for sharing, and so the high priority for version control is for internal use.

Preservation

Duration of preservation

The researcher stated that the dataset should be preserved indefinitely, clarifying that this preservation was critical for internal use to allow for review or re-processing of the data if needed, <u>not</u> as part of the scientific record. (See **Data provenance** for further explanation of internal re-processing.) The data have potential to retain research value for the lab group indefinitely, as the researcher could imagine a scenario where measurements taken during data collection and not originally used might become useful later. Data made publicly available after publication of results, such as the protein crystal structure, should be preserved indefinitely as part of the scientific record.

Data provenance

The researcher noted that, in the context of long-term preservation, documentation of any changes made to the data over time was not applicable for this data. All processing or re-processing of the data starts from the initial files (“Raw data acquisition” or “First pass analysis”) and progresses linearly through the data stages, with documentation recorded in lab notebooks. (See **Documentation of data organization/description** for further explanation.)

Data audits

The ability to audit the data to ensure its structural integrity over time is a low priority for the researcher.

Format migration

The researcher noted that the ability to migrate the data to new formats over time was a low priority because most of the data files (in all stages) were in stable and recognizable formats (TIF, Matlab, text). There may be a need to migrate data to new media over time, but this is likely to not be a high priority as the researcher noted, in some cases, it might be preferable to retake data than recover it from obsolete media.

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.