Data Curation Profile – Movement of Proteins

Profile Author	J. Carlson
Institution Name	Purdue University
Contact	J. Carlson, jrcarlso@purdue.edu
Date of Creation	July 14, 2010
Date of Last Update	July 14, 2010
Version	1.0
Discipline / Sub-Discipline	(Withheld)
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 April 2008. • A second interview with the scientist conducted in March 2009. • A questionnaire completed by the scientist as a part of the second interview.
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 Movement of Proteins 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 may be 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/10.5703/1288284315009

Brief summary of data curation needs

The scientist generates hundreds of TB of data but lacks the means of managing, curating and sharing this data with others as effectively as he would like. The data are composed of images, videos and spreadsheets, with some associated metadata captured in MS Word files or lab notebooks. Data are currently stored on "disposable" hard drives and are at risk if the drives crash. Although the scientist is very interested in making his data available to others once he and his collaborators have published their results, currently the data are not organized or described sufficiently for external use. The scientist is exploring how his data may be interoperated with similar data sets produced by other researchers. The scientist has applied for additional funding to address these issues with his data.

Overview of the research

Research area focus

The scientist is collaborating with a physicist to study the positions and movement of protein molecules within a cell. This research is done through taking a colored fluorescent protein molecule derived from a jellyfish inserting them into E. coli cells as they do not impact the background fluorescence of the protein. The proteins are studied for their stationing, localization, rate of movement and direction of movement within the E. coli cell in order to identify the pathways taken by the protein and cell mechanics. Genetic modifications and mutations are then introduced into the cell to study their effects on the movement of the protein.

Intended audiences

The scientist indicated that other researchers in his field or in physics may be interested in analyzing or mining this data for other purposes. Students were named as a particular group who would benefit from having access to the data. The scientist has not given much consideration to how the data might be used by researchers outside of his field.

Funding sources

The National Institutes of Health (NIH) is the primary sponsor of this research. The scientist has also submitted a proposal to the National Science Foundation (NSF) to support this project. The award from the NIH was beneath the threshold that would require the Scientist to share his data with others. However the NSF grant, if awarded, would provide for the development of a publicly accessible database to disseminate this data to others.

Data kinds and stages

Data narrative

The positioning and movement of the proteins within the E. coli cell are captured using a specialized charge-coupled device (CCD) camera that takes a picture every millisecond. The images produced are captured as JPEGs.

These images are then processed through the use of a software package developed for this particular purpose by a 3rd party in Japan. This software performs two functions. First, the software strings the images together to generate video files (most likely in .avi format). Second, it performs some calculations on the data including averaging the data, drawing vectors, making comparisons and other statistical functions. These calculations are captured in very large spreadsheets (MS Excel). It is unclear if all of the images are strung together as video files and calculations are performed on all files, or if video files are made and calculations are performed only for the images identified as being of interest to the project team.

The project team (the collaborating scientists and their graduate students) then review the video files to mine the data and interpret the results. The data are also reviewed at this point to

determine how the cells should be modified or mutated for future study on the effects these changes may have on protein movement. Students may generate some metadata about the data file at this point, including the size of the cell, the date the data was generated, the temperature, etc. Currently metadata are not added regularly or systematically; if metadata are added it is done manually by the student to capture "things that may be important" which may change depending on the specific research question. Metadata are not directly associated with the data file itself; instead they are captured in lab notebooks or MS Word documents.

Analysis of the data are then written up as manuscripts and submitted to publishers. Elements of the very large spreadsheets generated are often incorporated into publications as tables and figures. The data as presented in these tables and figures are insufficient in and of themselves for reuse or repurposing by others.

Data Stage	Output	Typical File Size	Format	Other / Notes
"Raw"	Photos of proteins	Actual file size is small, but the sheer number of files aggregates to TB of data.	.JPEG	Pictures are taken with a CCD camera which can take pictures every millisecond.
"Processed-1"	Video file consisting of strung together photos		.avi (not 100% sure of format)	Pictures are strung together to make videos.
"Processed-2"	Calculations about the Data	Files are very large, though it's unclear as to their specific avg. size	MS Excel	In addition to generating videos of the images, calculations are performed on the data as a part of the processing stage. It's unclear how these calculations are associated with the data, whether they are a part of the video file or not.
"Analyzed"	Metadata		MS Word, or handwritten in lab notebook	Students generate some descriptive metadata during analysis, though it is not uniform or standardized. Metadata are stored in MS Word or are handwritten.
"Published"	Tables or figures within an article		(part of the published article)	Relevant data are extracted, interpreted and represented in a limited fashion through tables and figures in published articles.

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

The scientist is most inclined to make the pictures, video, and spreadsheets available to others, preferably through an online database of some kind. The scientist recognizes the need to provide metadata along with the data to make it useable for others; however the metadata currently being developed by the graduate students in the "analyzed" stage of the lifecycle is of insufficient depth and quality for this purpose. Ideally, data would be disseminated through an online database that incorporates a standardized set of metadata enabling others to find, understand, and make use of the data (this database is planned but does not exist currently).

Value of the data

The data being gathered have the potential to be mined for additional measurements or other aspects of protein behavior beyond what the scientist and his collaborators intends to exploit. The scientist notes that the sharing of this dataset in a repository would be helpful to other researchers in the field who do not have access to the equipment, facilities, or resources needed to collect this type of data themselves. The data could also potentially be correlated with similar data sets that are being generated elsewhere for further analysis.

Contextual narrative

The data are considered to be dynamic as they are still being collected. The scientist was unable to answer specific questions on the size of the data or the eventual size of the collection as the day-to-day data management for this project is more the responsibility of his research partner and the partner's graduate students, than the scientist himself. However, the size of the aggregated files is already hundreds of terabytes and continues to grow rapidly. An individual student in this project generates a terabyte of data every few months.

The types of data comprising this dataset include photos (.jpg), videos (most likely .avi), and statistical calculations (MS Excel) with metadata about the video files stored as word processing files (MS Word), or in physical lab notebooks. The scientist would like to develop a database that would enable the data to be organized in more of a logical fashion and to have the metadata associated with the data directly. The scientist sees this database as the primary mechanism for sharing his data with others.

The software used to process the data was custom designed for this research project by a lab in Japan with whom the scientist is collaborating. Access to the software does not appear to be a necessary prerequisite to accessing or making use of the video or spreadsheet data files. It is unclear if this software would also need to be made publically accessible in order for others to make use of the photographs, (the "raw" data files).

Intellectual property context and information

Data owner(s)

The scientist had not thought much about or discussed ownership of the data with his collaborator. He stated that he was largely unconcerned with intellectual property issues regarding the data outside of attribution.

Stakeholders

Primary stakeholders in this data are the scientist's collaborators and the graduate students who have been contributing to the research.

The NIH as the funding agency supporting this project might also be considered a stakeholder in the data, although according to the scientist they have not made any demands about the data as his level of support is beneath the amount that would trigger the NIH's data sharing requirement. If the proposal made to the NSF is awarded, the NSF would become a stakeholder in the disposition of the data.

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

The scientist has two major requirements for sharing this data publicly. First, he would not share the data before the results of his research are published. Second, he would require that he receive attribution for the data if it were to be used by others (see "attribution" below).

Attribution

The scientist would require someone using his data to give him attribution for providing the data. The nature of the desired attribution would depend upon the usage of the data by others. If another researcher were to conduct an analysis of the data that is similar to the analyses done by

the scientist and his collaborators then co-authorship on subsequent publications might be expected. If the analysis consisted of new areas of exploration beyond the work that the scientist or his collaborators had done, then citing the source of the data in subsequent publications would be sufficient attribution.

The scientist is interested in applying some form of permanent identifiers (handles or DOIs, for example) to his data to enable citation.

Organization and description of data for ingest (incl. metadata)

Overview of data organization and description

The scientist stated that his data is much less organized than he would like it to be, and recognizes the need to develop better systems and practices in this area. He indicated that the amount of organization and description is not sufficient at present for other researchers in his field to understand and make use of the data themselves. He has applied for an NSF grant in part to develop a database to better organize and disseminate the data.

Most of the description, organization, and analysis of this dataset are currently performed by the graduate students associated with the project. Metadata are developed manually by students with some guidance from the scientist and his collaborators and are kept in lab notebooks or MS Word documents. The scientist reports that once a student graduates and leaves the project it can be difficult to interpret the metadata or other documentation that the student generated as it is not standardized or sufficiently detailed. The scientist also implied that file names may change over the course of the research which may contribute to difficulties in associating metadata to the appropriate file.

The scientist also reported that he and his colleagues are still learning what information they themselves need to capture in order to compare lab experiments and that occasionally they will need to go back and generate additional metadata about their data.

Formal standards used (Metadata, Ontologies, Controlled vocabularies)

Currently the scientist is not employing any formal standards for his data, but he recognizes the need to apply formal standards to his data as he is exploring opportunities to interoperate his data with similar data sets being generated by other research groups. He has listed the ability to apply standardized metadata from his discipline or field to the metadata as a high priority.

The scientist is not aware of any existing metadata standards that would be suitable to apply to the data. The NSF grant he and his partners submitted would provide the funds for a workshop for the scientist and other researchers generating similar data kinds to get together and begin to develop a standardized set of descriptive metadata for these data.

Locally developed standards

The types of metadata currently generated for this data include information on the size of the cells, when the experiment was conducted, the temperature, information about any cell mutations, the growth media used, and other information to track the movement of one or more proteins within the cell.

Metadata has not been uniformly applied or standardized by the graduate students on this project.

Crosswalks

The scientist expressed a need to interoperate his data with similar data generated by other researchers. This could potentially require the development of crosswalks, although this was not discussed specifically by the scientist.

Documentation of data organization/description

The scientist did not discuss any other aspect of organizing or describing the data.

Ingest

The scientist indicated that both the ability to automate the submission process and the ability to submit data manually into a repository are both high priorities for him.

The scientist indicated that he wants to share all of his data and that the value of making this data available would be for others to mine the entirety of the data. Therefore, a selection and appraisal process for this data would likely be a low priority, if it were even needed at all.

The data sets are composed of photographs, videos, and tabular data in spreadsheet files. The means of associating these data files together were not identified by the scientist, so it is unclear if associations would have to be identified prior to ingest.

The scientist indicated data files are not described sufficiently for use outside of the scientist's lab. In addition to needing richer metadata, documentation about the methods and procedures used to generate the data may have to be codified or created to make the data understandable and usable by others before it can be ingested into a repository.

Access

Willingness / motivations to share

The scientist expressed a great deal of interest in making the data publicly available online for others to use, provided the data are not released before the results of his work are published and that he receives attribution for providing the data (see "attribution"). He recognizes that the data has potential research value beyond what he and his collaborators will be able to extract and feels strongly that the data should be more fully mined by others; "otherwise it is wasted."

The scientist reported occasions in which he shared data outside of his immediate collaborators before publication with colleagues who he trusts and through presentations at conferences. His motivation in sharing in these instances is to obtain comments and feedback. He is selective about who he shares his data with and will refrain from presenting data if he feels he cannot trust the audience. It is unclear if he has previously shared the protein movement data described in this profile with others or not.

Embargo

The need for an embargo is predicated on whether or not the scientist has published the results of his research, rather than on a specific period of time. Once the research has been published the data can be released for anyone to access.

According to the scientist, he and his colleagues are just beginning to submit their work for publication at this time.

Access control

Beyond the embargo period, further access controls over the data were not discussed by the scientist for this data.

Secondary (mirror) site

The scientist indicated that the ability to access the data at a mirror site if the main site is offline is a high priority.

Discovery

The ability for discovery through internet search engines is a high priority for the scientist, as is the ability for researchers within and outside of his discipline to find the dataset easily.

Tools

MS Excel (or a csv reader), a media player capable of playing .avi video files, and an image viewer for .jpg files would be needed to use this data as it is currently structured. The scientist stated his desire to make the data and metadata available through an online database, though he did not specify the type of database or details about its functionality in the interviews.

As much of the data are visual in nature, the scientist identified the ability to connect the data to visualization tools as a high priority.

Interoperability

Enabling connections between the data and any publications that have resulted from the data are a high priority for the scientist. The scientist is interested in assigning some form of permanent identifiers to his data to enable these connections as well as for the data to be cited.

The scientist indicated that support for the use of web services APIs is a high priority for him. He is particularly interested in connecting his data with similar data sets that are being generated by other laboratories through the use of web services. A major obstacle in interoperating his data with these similar data sets is the lack of identified standards for ontologies and metadata schemes in this particular community. A grant proposal that the scientist has recently submitted to the NSF would provide funds for this community to meet and begin work on identifying or developing these standards.

Measuring impact

Usage statistics

The ability to view usage statistics on how many people have accessed the dataset is a high priority for the scientist.

Gathering information about users

Gathering information about the users of his data was not discussed by the scientist.

Data management

Data files are currently stored on “disposable” 5 TB hard disks, stacked one upon another. Although the scientist has not yet encountered a hard disk crashing, it is a very real concern for him. The grant proposal submitted to the NSF would provide him with resources needed to develop a computer system that would be capable of storing and backing up hundreds of TB of data files for this project.

Security / Back-ups

It’s unclear if the data files are backed up currently, and if they are with what frequency backups occur. Hosting the data in a secure environment that is backed up on a regular basis is important to the scientist; however the details of what this secure environment would consist of (frequency of back-ups, etc.) were not discussed.

Secondary storage site

A secondary data storage site is a high priority for the scientist. A secondary data storage site at a different geographic location is also a high priority.

Preservation

Duration

The scientist estimated that this data set should be preserved between 10 and 20 years. However, he also expressed some uncertainty about the length of time in which his data would have research value. Technological advances in gathering this type of data may reduce the value of his data sets. The scientist was uncertain about the need to preserve this data as a part of the scientific record.

Data provenance

Documentation of any and all changes made to the data over time is a high priority for the scientist.

Data audits

The ability to audit the dataset is a high priority for the scientist.

Version control

The ability of the repository to provide version control for this data set is a high priority for the scientist.

Format migration

The ability to migrate the dataset into new formats over time is a high priority for the scientist.

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

The scientist is collaborating with a physicist who oversees a significant portion of the data management activities on this project. The graduate students associated with the project, although trained by the scientist and his collaborator, are heavily involved in the analysis of the data and devising the metadata. The scientist has also recruited a faculty member from the computer science department as the principal investigator on the grant proposal that was submitted to the NSF, to develop the technological infrastructure to storage, manage and disseminate the data.