


Data Curation Profile – Biochemistry / Histone Protein Modification

Profile Author	Jake Carlson Katherine Beavis	
Author's Institution	Purdue University	
Contact	Jake Carlson, jrcarls@purdue.edu	
Researcher(s) Interviewed	[Withheld]	
Researcher's Institution	Purdue University	
Date of Creation	August 22, 2013	
Date of Last Update		
Version of the Tool	1.0	
Version of the Content	1.0	
Discipline / Sub-Discipline	Biochemistry / Histone Protein Modification	
Sources of Information	<ul style="list-style-type: none"> • An interview conducted on January 8, 2013. • A worksheet completed by the scientist as a part of the interview. • Lab policy on working with data 	
Notes		
URL		
Licensing	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.	

Section 1 - Brief summary of data curation needs

The researcher is conducting research on the glial and neuronal cells of drosophila larvae, a genus of flies. She is examining changes in gene expression when nuclei lose histone-modifying complexes. The stages of her research involve a long period of methodology development, a shorter data collection period, and then data analysis. Various aspects of the data analysis are outsourced to facilities at the university. The lab notebook contains written documentation of all experiments and trials and links to data, versions of analysis and images, databases, and versions of the developing journal article. Therefore, keeping an organized electronic lab notebook is of high importance for the researcher. The centrality of the lab notebook to the function of the lab means that it is very important for her students to be trained to properly document their experiments using her lab notebook structures and protocols. There is a potential issue of being able to link some of the data files to the notebook because they are very large and must be stored at the analysis facility on campus.

The researcher simultaneously submits some of her data to a database or repository as a part of the publication process. This data is publicly available once the article is published. The researcher has no limitations on access to the data, except that she requires that those using or referencing her data cite the paper from which it came rather than the data itself.

Section 2 - Overview of the research

2.1 - Research area focus

The researcher is conducting research on the brain cells of drosophila larvae. Specifically, the goal is to sort nuclei from glial and neuronal cells from the brains of mutant and wild type drosophila larvae. She will do transcriptome and high throughput ChIP-seq analysis on the cell nuclei to examine the distribution of complexes that modify histones and how loss of those complexes affects transcription. In other words, she wants to know what happens to gene expression when she makes mutants in a specific complex and particular tissue or cell type. She also wants to understand what mechanistically causes those changes in gene expression. By looking at the effect of on other histone modifications or on some of the enzymes that regulate the process, she can understand mechanistically what is occurring.

2.2 - Intended audiences

The researcher believes that the people most interested in her data would be those doing the similar types of research. She believes that there would be interest in both her early stage and her later stage data. Her early stage data involves developing the methodology for her research. Other researchers may want to reproduce her research methodology in their own areas of research, such as with mice. Some people may also want to do the same analysis with flies, but in a different context, such as using a different tissue or using the same tissue but with a different complex to see if they get the same results. Those researchers could then compare their results with hers to identify differences. The later stage data would be used by others to better understand her experiment and findings. Additionally, the researcher stated that people outside the field might be interested in the raw data if it were a high-profile study, though this would be a rare occurrence.

2.3 - Funding sources

The specific sources of the researcher's funding were not discussed. However, she suspects that most funding agencies in her field would require a data management plan. The researcher also indicated that she may need some help with constructing a data management plan in the future, as the policies and procedures she has set up to manage her data are directed to address her localized research needs rather than the broader elements required by funding agencies.

Data sharing through deposit of the data into a community repository is expected practice in the biochemistry field regardless of funding.

Section 3 - Data kinds and stages

3.1 - Data narrative

The researcher's initial data stage is defined as methodology development during which protocols are established and tested to ensure that they work (e.g., are the right genes being expressed in the purified nuclei?). During this stage the researcher's lab develops and saves versions of the protocols as they improve. An important part of this stage is being able to link back to their databases with the genotypes of the flies so that they can identify which flies they're using. DNA gels are generated as reference data at this point to confirm that the flies being used are the correct genotype. Each of these is kept in the OneNote lab notebook as a record. The protocols contain two sections, one for solutions and another for method. At this stage they also maintain PDF files for the reagents and antibodies that they are using as a record of the components involved in the protocol. Additionally, for lab equipment used, PDF's containing descriptions from the company from which they came are included when relevant. Also, included in the lab notebook is documentation of the setup of and what was done in the experiment, so that the results can be interpreted correctly.

At this stage they run quantitative real time polymerase chain reaction (qPCR) on RNA to measure gene expression. These experiments produce QPCR files, a readout from the experiment which are in a proprietary format, and excel spreadsheets that capture the analysis. Information about the fly stocks and reagents used, including their specific genotypes and oligonucleotides, are entered into MS Access databases for reference purposes.

Another component of this stage involves microscopy images that are included in the notebook as links. The images are in ND2 format, which is produced by the microscope software and contains the technical details of the image (e.g., fluorescence, intensity, lasers). ND2 is a proprietary format and so the images are converted to TIFF for export into papers or other purposes. However, the ND2 format contains information about the data represented in the image that is lost upon conversion to TIFF and so the ND2 files are kept and used for reference. ND2 images are linked in the MS OneNote lab notebook to connect them to the other relevant data files and documentation. It is important that the lab notebook contain documentation of what the images are because one cannot determine what the image is just by viewing it. The information in the notebook describes the context of the experiments run by members of the lab. Overall, the results produced in this stage demonstrate whether the experiment worked and allow them to move forward to the next stage.

The second data stage involves the further development of the methodology. During this stage the protocol and flow cytometry data are optimized. Additional data files are generated using flow cytometry software. These FCS files are generated outside of the lab by a bioinformatics facility with the necessary equipment. This facility is a part of the same institution as the researcher. In this stage, annotating the generated files is particularly important to keep track of how the methodology has evolved. This includes clearly identifying and documenting all of the various files, as well as understanding the relationships between them.

The next stage of the project is collecting the data. Once the methodology has been finalized, the samples are again sent out to an on-campus bioinformatics facility (external from the researcher's lab) for sequencing analysis. The result is a series of microscopy images that will comprise the researcher's raw data set. The researcher did not know the specific format of the data files, but did know that the format was proprietary in nature. They are also extremely large and therefore they are stored off site. The researcher will make a note as to where the raw data is located in the lab notebook for reference purposes, though the raw data itself is rarely accessed.

Once the data have been delivered to the research, the next stage of the data lifecycle is for the researcher to conduct some filtering and preliminary analyses on the data. Data from this stage are typically the data that are deposited in a data repository and associated with a paper upon its publication. The filtering and preliminary analysis done by the researcher is enough to give the data some meaning to other researchers and will enable other researchers to perform more specific analyses of their own on the data. Most of the journals in which the researcher would publish her research require the deposit of data into a data repository.

The data are then analyzed further, which is also done outside of the lab at a bioinformatics facility. For the analysis, the researcher will provide the protocols and possibly supply some scripts to be applied to the data. The researcher has not saved the scripts used to analyze her data in the past but recognizes that she will need to start doing so in the future as they are an important part of the methodology and would be needed for others to repeat the experiments. Analyzed data is delivered in an open text format that can be further analyzed by the researcher through running scripts on the data. The analyzed data may still be quite large and therefore may also necessitate storage outside of the lab or the department's servers.

The researcher will then generate visualizations of the data through using software programs including: Origin, Adobe Illustrator, Adobe Photoshop, or MS Excel. The researcher will then write up a text document to accompany the figures, which will incorporate the texts generated during the development of the methodology (Stages 1 & 2). This text will become the basis of an article

for publication in a scientific journal and will ultimately include the write up of the experiment, interpretation, and explanation of the data. The lab notebook would contain copies or links to images at various stages of analysis. The final figures would also be included the lab notebook.

Access to images and other data developed through the various stages of the data lifecycle is important in case she or her students want to return to a different version and try a different analysis. The researcher noted that in her research a particular experiment will be only one of a series of experiments that will contribute to one paper. As a result it is especially important to keep track of the analysis in the lab notebook.

The final stage of her data lifecycle involves refining and finalizing the images and experimental write up and then producing the publication. Upon submission of the article to a journal, the researcher will also submit her analyzed data to a data repository. This is a typical step in the publication process and is required by most journals in the field.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Methodology Development Part 1: Discovery	<p>Microscopy images and descriptions of these images.</p> <p>Documentation of experiments and evolving methodologies</p> <p>Records of the components involved in the protocol (pdfs)</p> <p>Tabular information and photos of DNA gels about the fly stocks used in the experiments</p>	<p>Many small files are generated at this stage. They include:</p> <p>QPCR files: 5-6 mb each</p> <p>Excel files: 38 kb or so each</p> <p>TIFF files: 2-3 mb each</p>	<p>ND2 (proprietary) TIFFs</p> <p>MS Word PDFs</p> <p>QPCR files (proprietary) MS Excel</p> <p>MS Access</p> <p>MS OneNote</p>	MS Word files are for documenting the development of the protocols and are versioned. Data in proprietary formats are kept as reference as they contain information that is lost in the transfer to open formats.
Methodology Development Part 2: Refinement	<p>An optimized research method</p> <p>Annotated files</p>	Similar to the previous stage: multiple small files.	In addition to the file formats listed in the previous stage, FCS files are generated (proprietary format)	Some of the flow cytometrics techniques are performed by a bioinformatics facility within the institution.
Data Collection	<p>“Raw” sequence data</p> <p>Scripts for sequencing the data</p>	Not discussed in detail, but referred to by the researcher as “very, very big”	Unknown but proprietary in nature	These data files are large enough so that they must be stored offsite. The sequencing scripts would be needed to

				replicate the experiment.
First Stage Data Analysis	Sequence data with some preliminary analysis - results	Not discussed in detail		Data from this stage are the ones typically deposited in a public repository.
Late Stage Data Analysis	Sequence data analyzed by the researcher for a specific purpose Figures generated from the data	Not discussed in detail. Not as large as the raw data files from the previous stage, but still quite large.	Text Excel (rarely) Adobe Illustrator Adobe Photoshop Origin	These data files are more likely to be stored on the researcher's or the department's network. "Origin" is software used to graph the data.
Publication	Visualizations of the Data (Figures) Data prepared for ingestion into a repository		Adobe Illustrator Adobe Photoshop Origin MS Word	Data are deposited into a relevant data repository as a part of the publication process. Data are typically prepared for deposit by the bioinformatics facility that generated the data.

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.

3.3. - Target data for sharing

The raw data produced are not shared as they are very large and would not have any meaning to others until they are filtered and put through some preliminary analysis (represented by the "First Stage Data Analysis" stage in the data table above).

Sharing the data is an accepted practice in biochemistry and reviewing data is a part of the publication process, though it is unclear how much scrutiny the data receive from reviewers. Data are submitted to a public repository when the related article is submitted. The repository will produce an identifier of some kind that can be listed in the article to connect people to the data set under discussion. This connection is critical in both directions as access to the data provides a measure of validity to the findings contained in the paper, and the paper provides the necessary context for people to be able to understand the data.

Repositories have strict formatting and other requirements in submitting data which vary by repository and the type of data. These formatting requirements are usually addressed by the outside facility that generated the raw data for the researcher.

3.4 - Value of the data

In the biochemistry field, data are deposited into a repository and made available publicly so that others can better understand and verify the findings described in publications. The images placed into publications can easily be faked using Adobe Photoshop or similar programs and so providing the underlying data is seen as a deterrent to improper manipulation. In addition to the verification aspect, sharing data allows researchers in the same field to conduct analysis of their own. The researcher would consider sharing her data through another venue to increase its

visibility and accessibility to researchers in other fields, however she is not sure who else might make use of the data or what value it would have for others outside of her field.

The researcher does see value in sharing her methodology as well. Researchers could potentially apply the methodology she developed for flies and apply it to a different species, mice for example, to see if they generated similar results. Others may simply want to compare results between different methodologies.

3.5 - Contextual narrative

There have been a number of high profile retractions in biochemistry recently. As a result, the researcher believes that in the future, the field may move toward making researchers' raw data available to further discourage improper data manipulation and fabrication.

Data repositories have a long history in biochemistry research. The data repositories used by researchers are largely supported by grants. The researcher believes the removal of these grants would be a disaster for the field, but does not anticipate that happening.

Section 4 - Intellectual property context and information

4.1 - Data owner(s)

Data ownership was not discussed.

4.2 - Stakeholders

Stakeholders for her data were not discussed directly. Potential stakeholders include the bioinformatics facility. Graduate students working in her lab may also be potential stakeholders.

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

The only condition for others to use the researcher's data is that the article that describes the data should be cited.

4.4 - Attribution

The researcher would want a citation of the paper from which the data came. She prefers this over citation of the actual data because in the promotion and tenure process the number citations to the paper matter, but the number of citations to the data do not. Citing the paper also helps with increasing the perceived impact of paper. The P&T process would have to recognize the value of citing data in order for this situation to change. Thus citing the paper is a high priority and citing the data is not a priority. The researcher does not see this situation changing in biochemistry in the near future.

Additionally, the researcher stated that the bioinformatics group that does the analysis could potentially be listed as a co-author on an article if they have made a significant intellectual contribution to it.

Section 5 - Organization and description of data

5.1 - Overview of data organization and description (including metadata)

The researcher is still developing and revising how she organizes, documents, and describes her data. She described it as somewhat of a trial and error process to find what works and believes that it may take some time to really figure out what works for herself and her lab.

The organization of the data is primarily centered on the lab notebook that the researcher maintains in Microsoft OneNote. The researcher finds that an electronic notebook is

advantageous because it is a searchable document, makes it possible to easily annotate images and allows her to create linkages between files.

The researcher has created Access Databases for the lab reagents that they make, such as fly stocks. These are referenced throughout the duration of the experiment. When a certain fly stock is used, a table of information with the ID number from the database is copied and pasted into the lab notebook in order to be clear about which fly stock has been used. As a safeguard, she also includes the genotype and a descriptor along with the ID number.

Developing a long term process of organizing and describing her data is somewhat difficult because the number of the people in the lab is relatively small. The researcher may need to adopt different practices as the number of people in the lab grows and as time goes on. One area mentioned by the researcher was writing up the analyzed data in her OneNote lab notebook, specifically how to ensure components are linked together in meaningful ways and how to track the development of the images and figures that her lab will be generating.

The researcher has experienced a few challenges so far in organizing her data. The main issue she is experiencing is the large size of some of the files and the ability to link them to the lab notebook. Specifically, she is not entirely sure of how she will connect the analysis files to the lab notebook; she thinks this part will be difficult, but does not know for certain because they are not yet at that stage of the experiment. She would like to be able to link to different stages of making the figures because linking it allows one to see the progression of the experiment. However, she is concerned that the OneNote notebooks may become very large with the number of files linked to or embedded in it, which could lead the program to crash. Lastly, the researcher finds that she must educate her students about organization of the information in OneNote because they do not know the “big picture” of the experiment. She added that it can also be confusing for students to know that a long series of experiments need to be kept together so that they can be easily found.

OneNote has worked well for her thus far as a means of organizing her data and tracking her students work on the data, especially in comparison to using paper notebooks where it is difficult to search and retrieve information in a useful manner. She has considered investigating other e-lab notebook systems but is concerned about the cost and the long term support that would be provided.

She has considered creating templates for experiments, but she has not done it yet for a couple of reasons. First, experiments may change over time with early results potentially leading to a change in approach. Having a template could inhibit the flexibility of adjusting approaches. Second, the researcher believes that students are better trained when they have to develop their own documentation strategies themselves. On the other hand, the researcher also believes that students need to be taught on how to keep their notebook properly, as a part of their lab training, which includes learning the format that is suitable for her lab.

The researcher does believe that the way she has been organizing and describing data so far would be sufficient for another person with similar expertise to understand and make use of the data.

5.2 - Formal standards used

In terms of keeping the data organized and properly described, the researcher stated that there are no established conventions within her department or in the biochemistry field as a whole. Each lab in the department is relatively independent and tends to develop its own method of organizing data and metadata. Practices in data management, organization and description are not topics that are generally discussed amongst the faculty.

The process of submitting data into a repository is highly structured. Data repositories do require the data to be submitted in particular formats. The researcher will be relying on the bioinformatics

facility to handle formatting the data and other structural requirements for submitting it into a repository.

5.3 - Locally developed standards

The researcher has created her own practices for organizing and describing data mostly through experience and trial and error.

The researcher requires her students to save documents, such as figures, with the student's initials and date in the file name. When students document their experiments in their lab notebooks, they need to include a title or aim, information about the genotype used, protocols used, and the results. The results must also have a link to the sample.

The researcher has considered making and using templates for annotating documents for ease of use. She has not done it yet because she has some concerns on how such an aid might affect student learning on documenting experiments.

5.4 - Crosswalks

Crosswalks were not discussed.

5.5 - Documentation of data organization/description

The researcher has drafted policy documents for her lab. She does not plan to expand them for the current experiment and if changes are made they will be minimal.

Section 6 - Ingest / Transfer

The choice of repository is determined by the particular type of data being generated. The data would need to be in the correct format requested by the repository before it could be ingested. The specific format type depends on the guidelines of the repository. The bioinformatics center she works with usually takes care of putting the data in the proper format. The repository also typically provides a template to ensure that all of the appropriate information is included.

If any questions or concerns about the data were to come up an editor may request to see the raw data to be sure an image has not been edited or photoshopped inappropriately. However, this is a rare occurrence.

Section 7 – Sharing & Access

7.1 - Willingness / Motivations to share

The researcher indicated some reluctance to share her methodology (stages one and two) beyond her immediate collaborators. Generating the methodology is the most difficult part of the research and the researcher would want to be confident in the work before sharing it with others in her department. The researcher is uncertain about sharing the methodology beyond the department at this time as the research is still in process.

The researcher would be willing to share her data or methodology especially after publication. However, beyond those in her field, she does not believe that others would have great interest in her data. Also, sharing her data allows for people to see that she has done the analysis properly and whether there are any errors or criticism. The data goes in the repository before submitting the paper, but it is embargoed until it is published. The editor and reviewers are the only ones with access to it.

7.2 - Embargo

There is no embargo on releasing the data publicly once the article is published. However, when the researcher submits her article and data to a publisher, the data has an embargo until the article is published.

7.3 - Access control

Data are typically embargoed by the repository until the paper is published and then they are publically accessible to all. The researcher does not see a need for restrictions on access to her data once they are published.

7.4 Secondary (Mirror) site

The need for a mirror site was not discussed.

Section 8 - Discovery

The researcher feels that her data are sufficiently discoverable when it is kept in the repositories specified by the journals in which she publishes. She believes the data would be more discoverable to a broader range of people in an institutional repository. However, at this point in time, she does not think that putting her data in such a repository would be very beneficial as researchers in her field are the primary audience for her data.

Section 9 - Tools

The following tools are needed for the researcher's experiments:

- QPCR machine (to measure gene expression)
- Access Database (created by the researcher to keep track of reagents, such as fly stocks, which have specified genotypes or oligonucleotides)
- OneNote (for the lab notebook)
- Flow cytometry software (used by bioinformatics)
- Origin (a graphing program to create images)
- Fluorescence Activated Cell Sorting (FACS)
- Typhoon (images western blots)
- Gel Doc (takes photos)
- A scanner (for scanning images)
- Microsoft products (Word, Excel, etc.)
- Adobe products (Illustrator, Photoshop etc.)
- Protocols for experiments (created in Stage 1 and 2)

The tools that produce files in a proprietary format require specialized software to access the data. The files may be exported into open formats however information about the calibrations, settings, protocols followed and other important elements would be lost. Thus, converting the format of the files would negatively affect its utility.

Section 10 – Linking / Interoperability

Linking the experiment's data and various accompanying documents to the OneNote notebook are very important to the researcher. She noted that it would be helpful to link the papers, which her students are writing, to the electronic notebook so that she can see the progress of the paper. She also puts links in the notebook to the appropriate genes in Fly Base, so it easy to reference to read more about the specific gene in the paper. Also, linking to large data files that are not stored on the departmental drive could pose some difficulties.

Outside of the laboratory experiment, the researcher stated that is a priority for the published article to have a link to the repository containing her data. She would also want the data to link back to the originating article. This is important to the researcher because the article is needed to fully understand the data.

Section 11 - Measuring Impact

11.1 - Usage statistics & other identified metrics

Usage statistics were not discussed with the researcher.

11.2 - Gathering information about users

Gathering information about users was not discussed with the researcher.

Section 12 – Data Management

The experiment files are stored in the departmental folders that are dedicated to her lab. Larger files, such as the raw data, are stored with the bioinformatics center because there is not enough space on the departmental drive.

12.1 - Security / Back-ups

The data is stored on the server in the researcher's lab specific folder within the department's drive. The files are backed up daily. The researcher also personally backs up the files manually on a monthly basis. She has had to retrieve files from a back-up in a previous position and so she sees the value of backing up her work on a regular basis.

The accessibility of the files is limited to her lab and some folders are limited to specific people for write access.

12.2 - Secondary storage sites

Secondary storage sites are a low priority for the researcher.

12.3 - Version control

Version control is a high priority for the researcher. Version control is especially important when she or her students are going through a long series of trials in developing the experiment.

Section 13 - Preservation

13.1 - Duration of preservation

The researcher stated she is too early in her research project to know what particular parts will need to be preserved. In general, she believes that the preserved parts will need to be kept at least 20 years, but more likely for at least the duration of her career which could be around 30 to 40 years. The researcher considers preservation to be a high priority.

In terms of beginning the preservation process, the data of students who briefly worked in her lab but did not join would need to be archived earlier on than other data. She did not specify the duration of preservation for this data.

13.2 - Data provenance

Not discussed.

13.3 - Data audits

Not discussed.

13.4 - Format migration

The researcher has considered format migration in terms of the OneNote lab notebooks. She knows that there might be better electronic notebook software, but she feels most comfortable with using OneNote because she has confidence that it will be supported for a longer period of time as compared to other software. Format migration was not discussed regarding the other types of files.

Section 14 – Personnel

Information withheld.

Section 15 – Local Questions

15.1 - Education for Graduate Students (General)

The researcher sees graduate student training as an important aspect of her lab. She believes that training is important not only because it ensures experiments are being done properly in her lab, but also because she believes training in data management is an important for the graduate students' education.

The researcher does train her students on keeping a proper lab notebook. She noted that most students come into her lab not knowing how to write up a laboratory experiment in a lab notebook. She teaches her students how to use OneNote, which is employed as an e-lab notebook by the researcher, and how to format their entries in their lab notebooks appropriately. Elements that students must include in their notebooks include the purpose of the experiment, the genotype used, the protocols followed, and linkages to the samples. The researcher will invest a significant amount of time in training students in how to document their work, reviewing their work daily when a student joins her lab and then less frequently as the student picks it up.

The researcher has considered developing templates for writing up experiments, but so far has decided against doing so even though using templates may be easier on her and her students. One of the factors in her decision was her sense that it might be too easy for students to write up and document their work in a rote fashion without thinking about or really understanding what they are doing and why. Students tend not to see the “big picture” of their experiments and how they may all fit together.

15.2 – Specific Knowledge and Skills Needed by Graduate Students

In addition to documenting their work in lab notebooks, the researcher identified important areas of knowledge and skills for her graduate students to learn. She believed that students should develop an understanding of data formats and databases, in particular that students should know the differences between the proprietary data formats and the formats used for analysis. Students often do not grasp that data loss occurs in transferring data from proprietary formats, such as ND2, to open formats, such as TIFF, nor do they always understand the significance of the loss. Students also need to know what is and is not an acceptable form of image manipulation in the biochemistry field. Some modifications may be required to demonstrate a finding more clearly, but research fraud is a big concern and students must tread carefully when using Photoshop or other visualization software programs.

Areas such as data curation and preservation were less important for students in the opinion of the researcher as many students will not continue on into academia, but instead will go into industry position. Each industry will have its own policies and procedures for working with the

data generated. The researcher did see value in learning how biochemistry based companies perform data management and organization related tasks and developing training for students that would align with these practices.