LA-UR-12-23984
Approved for public release; distribution is unlimited.

Title: Quarterly Report - May through July 2012

Author(s): Miller, Laniece E.

Intended for: Quarterly Report to be submitted to Internship Committee

Quarterly Progress Report
May through July 2012
Laniece Miller
Mentor: James Powell

The first quarter of my postgraduate internship has been an extremely varied one, and one which I have tackled several different aspects of the project. Because this is the beginning of a new investigation for the Research Library, I think it is appropriate that I explore data management at LANL from multiple perspectives.

I have spent a considerable amount of time doing a literature search and taking notes on what I've been reading in preparation for potential writing activities later. The Research Library is not the only research library exploring the possibility of providing services to their user base. The Joint Information Systems Committee (JISC) and the Digital Curation Centre (DCC) in the UK are actively pursuing possibilities to preserve the scientific record. DataOne is a U.S. National Science Foundation (NSF) initiative aimed at helping to curate bioscience data. This is just a tiny sample of the organizations actively looking into the issues surrounding data management on an organizational, cultural, or technical level. I have included a partial bibliography of some papers I have read (see Appendix A).

Based on what I read, various discussions, and previous library training, I have begun to document the services I feel I could provide researchers in the context of my internship. This is still very much a work in progress as I learn more about the landscape in libraries and at the Laboratory. I have detailed this process and my thoughts on the issue below.

As data management is such a complex and interconnected activity, it is impossible to investigate the organizational and cultural needs of the researchers without familiarizing myself with technologies that could facilitate the local cataloging and preservation of data sets. I have spent some time investigating the repository software DSpace. The library has long maintained the digital object repository aDORe, but the differences in features and lack of a user interface compared to DSpace have made DSpace a good test bed for this project. However my internship is not about repository software and DSpace is just one potential tool for supporting researchers and their data. More details my repository investigation .

The most exciting aspect of the project thus far has been meeting with researchers, some of which are potential collaborators. Some people I have talked with have been very interested and enthusiastic about the possibility of collaborating, while others have not wanted to discuss the issue at all. I have had discussions with individual researchers managing their own lab as well as with researchers who are part of much larger collaborations. Three of the research groups whom I feel are of particular interest are detailed below. I have added an appendix below which goes into more detail about the protein crystallography community which has addressed the complete data life cycle within their field end to end (see Appendix B).

The issue of data management is much bigger than just my internship and there are several people and organizations exploring the issues at the Laboratory. I am making every effort to stay focused on small science data sets and ensure that my activities use standards-based approaches and are sustainable.

Website

In Fall 2011, Mike Wenman, another GRA, and I were tasked with putting together content for a set of data management webpages similar to MIT's set of webpage, http://libraries.mit.edu/guides/subjects/data-management/index.html. We were told these webpages are supposed to mostly informational and compiled the content as such. Our internal website can be found at http://int.lanl.gov/library/knowledge/data/index.shtml (see figure 1). We included general information and a checklist of different questions researchers need to ask themselves and defined 'Data' in this case, which is basically any electronic output results from scientific research activities. We discussed data management during the planning and collection stages, documentation and metadata, sharing, storage and backup, and ethical issues. We also included links to various funder requirements and a breakdown of the NSF requirement. Since the start of my internship, I edited this content, weeded out what was not ready to post, and cleaned it up enough to give to the web team. Matthew Hopkins uploaded it to our website and I presented the webpages to library staff July 30th, 2012.

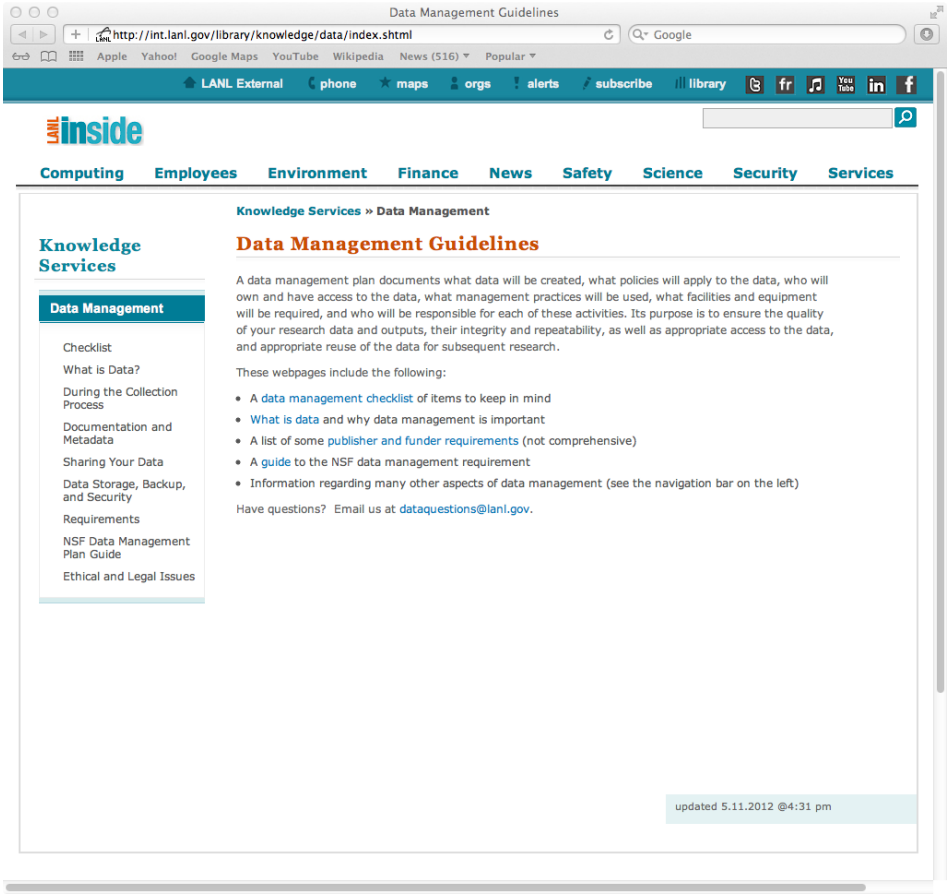


Figure 1 – The main page for the Data Management Guidelines webpage

Potential Services

From my readings in the literature, discussions with various people, and my library training, I have started to compile details about potential services I could provide as part of my internship. This

initial draft contains items such as data management planning, controlled vocabulary, documentation, information organization, and providing general information for data management issues that are more technical. When creating this list, I attempted to focus on transferring traditional library skills to the digital data domain, such as categorizing and cataloging information and controlled vocabulary. Ultimately James and I decided to table this work for now and let this list of services librarians are trained to provide be an outcome of the project as I learn more about what help researchers need and want.

DSpace Repository Software

DSpace is an open source repository system which ingests files and corresponding metadata. James has set up an instance of DSpace on his local computer to allow me a chance to explore the software and evaluate its potential usefulness. As a set of test data for me to learn about DSpace, James gave me data from scientists studying wind turbines. The library had previously worked with this data and thus it was a good test case. I was able to upload the files in a couple of different organizational structures by breaking the data into different groups. I was also able to define a new metadata schema within the user interface, however I have not been able to actually map the metadata collected during input onto these fields because that requires access to the behind the scenes configuration files. Access to the different groups of data files and individual files can be restricted as needed and can easily be set differently depending on the collection.

I am finding that how to organize files into items, collections, sub-communities, and communities takes some thought, as there is different metadata associated with each level. While the metadata fields with each level can be customized, this process is not a straight forward task in the web interface. For example, with the wind data, the data was collected across three different variables, creating two files for each combination of variables, giving a total of 54 files for 27 runs. The disk I was given has a set of 9 file folders, with 3 folders in each. Several questions quickly appeared regarding what was considered an item and how to organize the files. I tried packaging the files two different ways. One was to create 9 items and putting the 6 files in one item. This issue with this method is the documentation files associated with the project were not easily included in these items, or would need to be included 9 times, instead of just once. A separate item could have been created but it still broke up associated files. The other method I tried was to put all 54 files plus the documentation files into one item, which packaged the entire set of associated files together. However, with the poor naming schema of the files made creating enough metadata to differential the individual files difficult due to the limitations of DSpace.

Working with the ISR data I have been given is also posing challenges in figuring out the best way to package the files together as items. What I have been given at this point in time does not have good documentation readme files included, so trying to package those appropriately is not the issue. However, I have been given a couple of published papers which fill in some of the gaps. The data is satellite data for two different satellites, spanning several years, of which I have one year. I have two folders, one for each satellite, with one file per day. At this point I am still attempting to figure out the best way to organize the data so I can begin to work with the researchers to determine if DSpace would be useful. We have been able to point AutoPlot (a visualization tool described more during the ISR

discussion) at files already loaded in DSpace and received an appropriate plot (figure 2), demonstrating compatibility between a digital object repository and a data visualization tool.

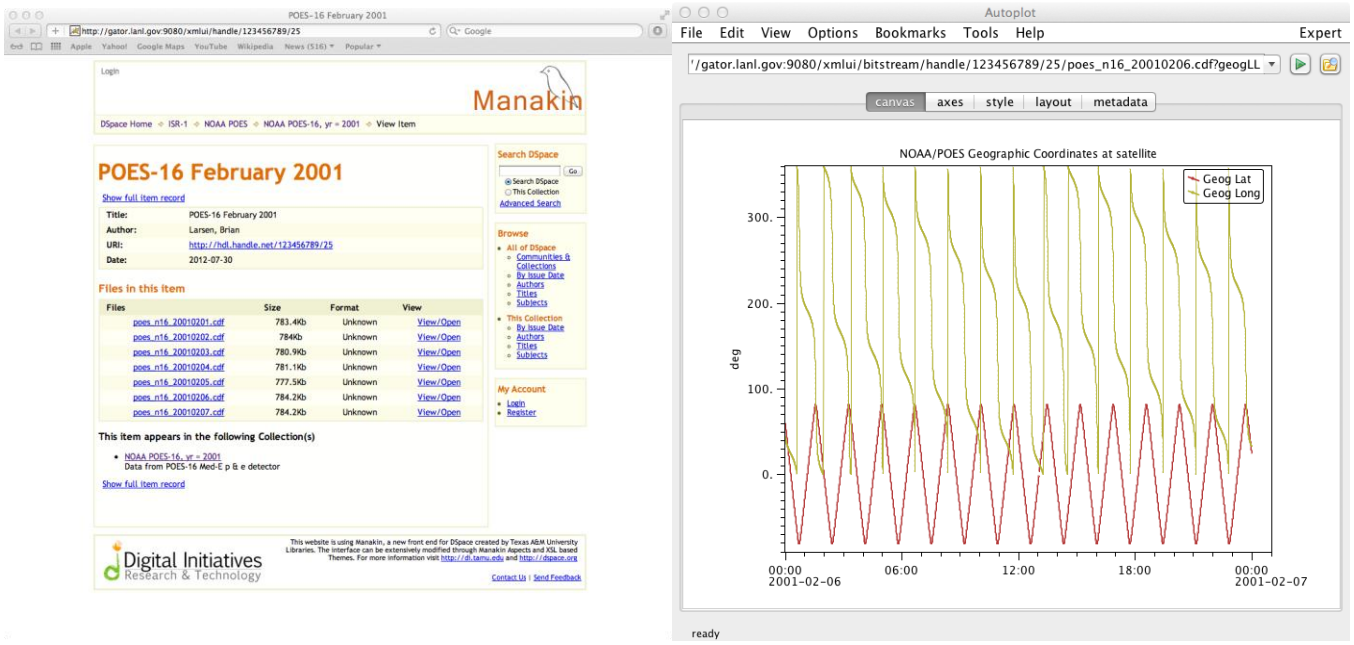


Figure 2 – A screenshot of a DSpace record and Autoplot plotting one of the listed files.

I still need to investigate:

- Batch loading files
- Automatically loading files
- Security of the environment
- Changing the metadata input screens
- Statistic options

Genomics

LANL participates in the Joint Genomic Institute (JGI) which at a national level specializes in bioenergy, the carbon cycle, and biogeochemistry, mainly via DNA sequencing. The JGC-LANL group has several different machines to sequence DNA, each with its own advantages and draw backs. Once the various samples of an organsim's genome are sequenced, the scientists then reassemble the data into a complete genome for the organism. This sequence is then used for analysis and well is deposited in GenBank. GenBank is a NIH funded database / repository for DNA sequences. Both the sequence itself and the final report are deposited into GenBank.

The sequences from DNA samples, complete sequences, and analyzed files are managed by a Laboratory Management System (LIMS) and stored on network storage space purchased from NIE. Much of this data needs to be kept for verification because sometimes they receive requests for it so other scientists can verify the analysis, results, and conclusions of the work. The LIMS system allows at least some level of management of the data itself. The researchers use command line tools to do a lot of the analysis of the complete genes or genomes. Sequences of the tools are often stacked together

into workflows, which are managed by workflow management systems. The individual tools are not well managed, especially being what is used at LANL tends to be slight modifications of open source software. Creating a software repository for these codes potentially could be very useful to help the group keep track of what changes they have made over time. I have done a little preliminary looking at the software tool MyExperiment.

Another data management problem which the group deals with is their archives, a series of magnetic tapes stored in a cabinet. These tapes are organized by date and they have very little documentation describing what they represent. No one has verified that these tapes, especially the older ones are still viable media. Once a month on average, data from these archives is requested, not just the sequence and final report which were submitted to GenBank.

As a first step in attempting to help with their data management issues, I was permitted to attend a series of tours and talks which were provided for an incoming student in the group. The morning sessions were about the different sequencing techniques used. The afternoon session was about the overall workflow of the group, but did not really go into details about specifics and how data really moves from one step to the next. I was unable to attend a session about the primary analysis performed to reassemble to complete gene or genome. At this time, I have not heard more from the Genomic group about working with them, so it is uncertain if I will actually collaborate with them.

Protein Crystallography at LANSCE

Some meetings have resulted in clear paths forward for potential collaboration and some have been more learning experiences. Meeting with Dr. Zoe Fisher, who is an instrument scientist at LANSCE was definitely more of a learning experience. Dr Fisher is a protein crystallographer. She introduced us to a process used by her and her fellow collaborators and resarchers, which essentially takes into account data management from the instrument all the way to the publication of findings. At the core of this process is the Protein Data Bank (PDB), which is a database / repository of bio-molecular structures including proteins and nucleic acids. The PDB contains structural data, which is what scientists refer back to; the raw data is only useful during an investigation for fraud. This means that the raw data from the experiment is minimally preserved, but the major funders and journals require deposition of data in the PDB which actively preserves the data they hold. Appendix B covers what we learned from Dr Fisher in more detail.

Intelligence and Space Research (ISR)

ISR-1 works with NASA and satellite missions, many of which stream some data back to the LANL facility. The researchers I have been working directly with deal with the radiation belts surrounding Earth. The solar wind brings a particle plasma from the Sun towards Earth, and this plasma carries a magnetic field with it, which interacts with the Earth's magnetic field. These fields can cause magnetic storms which can disrupt electric and electronic systems on Earth such as communication and navigation systems as well as the power grid. Soon to be launching are the Radiation Belt Storm Probes which will fly electron and proton detectors.

Much of the data that researchers use is stored in NASA's .cdf format. The common data format (.cdf) was initially developed in 1985 to store multi-dimensional arrays. Beyond just handling multi-dimensional data, the files themselves can store metadata which aid researchers in manipulating and analyzing the data. Associated with this general data format are the CDF libraries which translate into multiple different programming languages such as C and FORTRAN. These libraries can then be used to create programs which allow access to the data stored in the file formats. The other main file type used is ASCII .txt files which span a broad range of completeness. The better organized files include a header which supplies metadata, often in a machine readable format like the JSON format which software can parse. Others have been totally separated from their metadata, thus the columns of numbers have become utterly meaningless.

I am still learning about the exact metadata schemas which are encoded in the .cdf files and JSON headers. One metadata schema used extensive in this field is the Space Physical Archive Search and Extract (SPASE) metadata. My initial reaction to what I have seen of this metadata is that it does not completely describe the cataloged items, but I also have not seen full SPASE records. I am not sure if this is the only metadata schema they normally use and what other schemas and standards exist for this field. This is an area I need to do more research in.

One visualization tool researchers in ISR use is Autoplot, which takes a data file properly formatted and plots it. Researchers can then play with the data to zoom in on areas, re-plot it using other variables in the data, re-plot it in a different view, etc. Autoplot is a java based package which normally requires little setup for use. Because it is low barrier and gives a way to quickly visualize the data, many of the websites for this type of data will actually link the data with Autoplot so researchers can more easily decide what data is useful before downloading numerous files. Before ingesting many files into our test DSpace instance, I first uploaded a single file. I sent the link to our DSpace instance to one of our contacts in ISR and asked him to verify that he could see DSpace and could plot the file I uploaded. From there I then determined what could be a useful organizational structure and partially populated it, just enough to get a good feel for what the structure would look like without spending too much time on a instance that has so many restrictions. We are now evaluating what I have already done to determine suitability and how to proceed.

The ISR data sets seem like they may be one of the best data sets I have encountered thus far for this internship. The data is valuable because, as one of the ISR scientists noted, NASA only typically flies a mission and instrument combination once, so interest in this irreplacable data is high. The bandwidth between a probe or satellite and earth is typically rather low, so the data sets are small. And once a mission is complete, the principle investigator becomes an important resource for would-be future users of the data, and yet they are typically not readily available to answer questions. Thus documenting, cataloging, and preserving this data and the information that can be extracted from the PI is especially important. The researchers from ISR are interested in a LANL and library branded presence for this data and are eager to collaborate with the library. The combination of well-structured self-describing data sets, librarian crafted metadata for context and discovery, and cross-platform visualization tools like AutoPlot which can retrieve the data from a digital object repository is a model which addresses a typical usage scenario for data. Today this data is presented through ad-hoc websites that differ from mission to mission and are not well maintained or predictably structured. A centralized

repository, or data hub for these data sets would be a high profile and yet low effort win for the library, and seems to be a good fit for my internship.

Appendix A: Selected Bibliography

Angevaare, I. (2009). Taking Care of Digital Collections and Data : " Curation " and Organisational Choices for Research Libraries Digital Publications : from Storage to Access, *19*(1).

Beagrie, N. (2006). The International Journal of Digital Curation Individuals Definition and History of Digital Curation, *1*(1), 3-16.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? *Working Paper Series of the German Data Forum*, (161).

Carpenter, L. (2005). Supporting Digital Preservation and Asset Management in Institutions. Retrieved from http://www.ariadne.ac.uk/issue43/carpenter/intro.html

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *368*(1926), 4023-38. doi:10.1098/rsta.2010.0165

Day, M. (2008). The International Journal of Digital Curation Toward Distributed Infrastructures for Digital Preservation : The Roles of Collaboration and Trust, *3*(1), 15-28.

Henty, M. (2008). Investigating data management practices in Australian universities. Retrieved from http://en.scientificcommons.org/36795434

Heslop, H., & Wilson, A. (2002). An Approach to the Preservation of Digital Records, (December).

Jessop, M., Weeks, M., & Austin, J. (2010). CARMEN: a practical approach to metadata management. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *368*(1926), 4147-59. doi:10.1098/rsta.2010.0147

Kowalczyk, S. (2011). e-Science Data Environments: A View from the Lab Floor, (December). Retrieved from http://ivl.slis.indiana.edu/km/pub/2011-kowalczyk-phd-thesis.pdf

Lord, P., & Macdonald, A. (2004). From data deluge to data curation. *Proc 3th UK e-Science All ...*. Retrieved from http://www.allhands.org.uk/2004/proceedings/papers/150.pdf

Lyon, L. (2007). Dealing with Data : Roles , Rights , Responsibilities and Relationships Consultancy Report, (June), 1-65.

MacDonald, S. (2008). Libraries in the converging worlds of open data, e-research, and Web 2.0. *Online*. Retrieved from http://ie-repository.jisc.ac.uk/227/1/Online_mar08.pdf

Marcus, Ball, S., Delserone, L., & Loftus, W. (2007). Understanding Research Behaviors , Information Resources , and Service Needs of Scientists and Graduate Students : A Study by the University of Minnesota Libraries, (June), 1-38.

Organization for Economic Co-operation and Development. (n.d.). OECD principles and guidelines for access to research data from public funding. OECD. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:OECD+Principles+and+and+Guidelines+for+Access+to+Research+Data+from+Public+Funding#0

Piwowar, H. a. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one*, *6*(7), e18657. doi:10.1371/journal.pone.0018657

Pritchard, S. M., Libraries, U., & Carver, L. (2005). Informatics and Knowledge Management for Faculty Research Data, *2005*(2).

Pryor, G. (2007). Project StORe: making the connections for research. *OCLC Systems & Services*, *23*(1), 70-78. doi:10.1108/10650750710720775

Witt, M., Carlson, J., & Brandt, D. (2009). Constructing data curation profiles. *Journal of Digital Curation*, *4*(3), 93-103. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/137

del Pozo, N., Long, A. S., & Pearson, D. (2010). "Land of the lost": a discussion of what can be preserved through digital preservation. *Library Hi Tech*, *28*(2), 290-300. doi:10.1108/07378831011047686

Appendix B

E-Science and Protein Crystallography at LANL

L. E. Miller, J. E. Powell

Dr. Zoë Fisher is the instrument scientist for the Protein Crystallography Station (PCS) at the Los Alamos Neutron Science Center's (LANSC) Lujan Neutron Scattering Center. She helps schedule researchers who intend to use the instrument to collect data, and provides in depth support for their activities. Users submit proposals for beam/instrument time via LANSCE proposal review system. In 2012, there were about 20 proposals submitted for this instrument. The instrument scientists review the proposals online. Accepted proposals are scheduled via an aggregate calendar which takes into account staff and resource availability, and the scientist is notified via email when their proposal is accepted and their requested time is scheduled. The entire PCS data acquisition and processing workflow is streamlined through various locally developed and commercial software packages. One 24 hour period produces one 200 Mb file, giving a total of maybe 2-5 Gb of data for the entire run. This data is then transferred to a hard disk in Dr. Fisher's office where she views the data with the customer and compresses the data to a text format which she sends them. This compression translates the data from an electron density to structural coordinates, which are the products submitted to a protein structure database.

As noted above, the raw experimental data is stored onsite at LANSCE on workstations maintained by the instrument scientist. It is extraordinarily rare for anyone to request this data, although the remote possibility of an audit by a funding organization motivates its limited preservation. The raw data is not rigorously backed up, but only stored on a single hard drive. Interestingly, only about 50% of the experimental data actually ends up deposited and described in peer reviewed publications; the data that is not published tends to either not be viable structures or is calibration data.

Dr. Fisher does protein crystallography research using both neutron and x-ray scattering techniques. Many of the major funders as well as the major journals dealing with protein crystallography require deposition of the structural data in the Protein Data Bank (PDB). Files formatted for the PDB are automatically generated when the data is compressed. The header files in the PDB included experimental conditions of the experiment as well as experimental methods. Depending on the completeness and how 'hot' of a topic, it may not be needed to contact the original experimenter about using the data. Having said that, not all of the data is accurate and does requires some back and forth with the creators of the data. The RCSB PDB staff at Rutgers University goes through all submissions and works with the submitters to verify that the data meets their minimum standards of completeness and robustness.

The Protein Data Bank (PDB) was initially created by Walter Hamilton at Brookhaven National Laboratory in 1971 after discussions about the value of scientists having access to structural biology data. Originally a partnership between Brookhaven and the Cambridge Crystallographic Data Center, the idea was conceived as a global initiative, which is certainly has become with partner sites in the US, Europe, and Japan. The PDB now contains structures determined from many different experimental techniques (Berman et al. 2012). Deposited structures are assigned a unique ID, and the structures are

embargoed until the publication that references and describes them is published. The PDB staff often monitors these publications and takes the initiative to release protein structures when papers describing them are published.

Dr. Fisher records setup and experimental details in word documents and inserts printed copies into paper lab notebooks. These details appear in the final published papers and the header files for structures in the PDB. Analysis of data collected at the PCS is performed with a combination of locally developed tools and commercial products which are capable of outputting data suitable for importing into the PDB. While the original output data from the LANL instrument is stored indefinitely on a hard disk, the analysis results in a text file that, as described above, which represents the structure of the protein, which can be modeled and explored via tools that scientists in this domain have access to and are familiar with.

The entire process is well understood and well-supported by software used by researchers in this field. The incorporation of the PDB into research->analysis->publication is embraced by the international community of researchers in this field. There are mirror depository sites for the PDB in several countries. Curation of the submitted protein structures is rigorous, although Dr. Fisher noted that some structures are rushed to publication with what she termed “bogus filler”, which is possible since protein structures are 50-70% water. This is one of the things other scientists who look at deposited data will look for.

A case study on HIV protease was conducted on the data in the PDB. There was no further data collection beyond the existing published data. The main conclusion regarding the data was the need for some level of quality control before using the data, but that useful conclusions could be made strictly using PDB data (Venkatakrishnan et al. 2012).

Figure 1: A flowchart indicating the general workflow of data from the PCS to curation in the PDB.

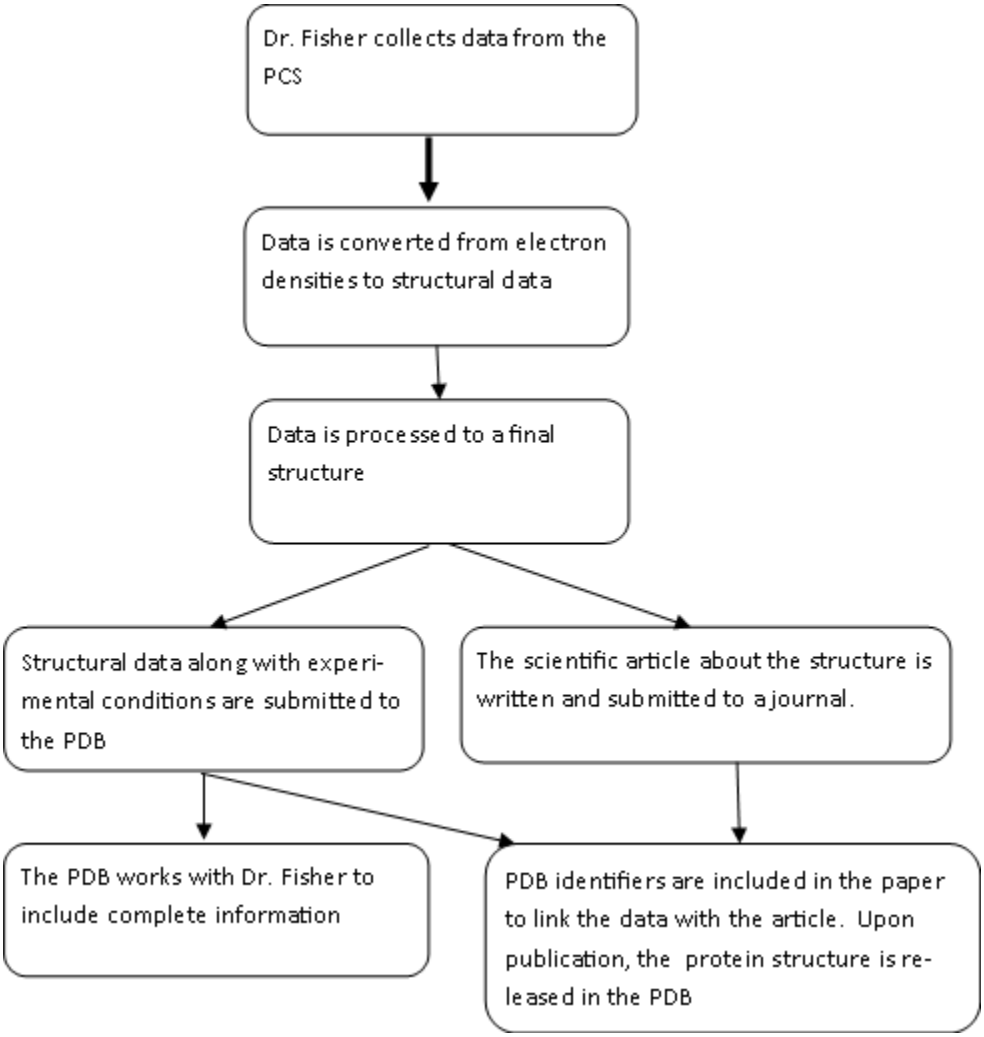
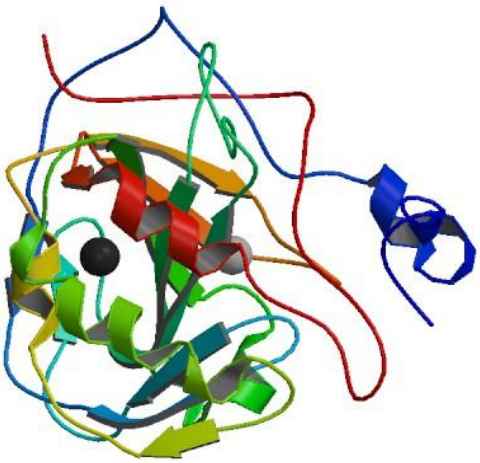


Figure 2: An excerpt of a PDB header, featuring journal "metadata"

```
HEADER    HYDROLASE (SERINE PROTEINASE)           13-APR-88   1SGT       1SGT   3
COMPND    TRYPSIN (/SGT$) (E.C.3.4.21.4)                                 1SGT   4
SOURCE    (STREPTOMYCES $GRISEUS, STRAIN K1)                             1SGT   5
AUTHOR    R.J.READ,M.N.G.JAMES                                           1SGT   6
REVDAT   1   16-JUL-88 1SGT     0                                        1SGT   7
JRNL        AUTH   R.J.READ,M.N.G.JAMES                                  1SGT   8
JRNL        TITL   REFINED CRYSTAL STRUCTURE OF STREPTOMYCES $GRISEUS     1SGT   9
JRNL        TITL 2 TRYPSIN AT 1.7 ANGSTROMS RESOLUTION                   1SGT  10
JRNL        REF    J.MOL.BIOL.                  V. 200   523 1988       1SGT  11
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070   1SGT  12
```

Figure 3: A link to an entry for a protein analyzed by the researcher we interviewed, along with a visual model of this protein

http://www.pdb.org/pdb/explore/explore.do?structureId=3v3f



We would like to acknowledge Dr. Zoë Fisher for taking the time out to show up their process and her workflow. Most of this write up is based off our conversation with her.

Berman, H. M., G. J. Kleywegt, H. Nakamura & J. L. Markley (2012) The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure,* 20**,** 391-396.

Venkatakrishnan, B., M. L. Palii, M. Agbandje-McKenna & R. McKenna (2012) Mining the Protein Data Bank to Differentiate Error from Structural Variation in Clustered Static Structures: An Examination of HIV Protease. *Viruses-Basel,* 4**,** 348-362.