2012

Beyond The Low Hanging Fruit: Archving Complex Data and Data Services at University of New Mexico

Robert Olendorf
University of New Mexico, olendorf@unm.edu

Steve Koch
University of New Mexico, stevekochscience@gmail.com

Journal of Digital Information, Vol 13, No 1 (2012)

Beyond The Low Hanging Fruit: Archving Complex Data and Data Services at University of New Mexico

Robert Olendorf
University Libraries, University of New Mexico
olendorf@unm.edu

Steve Koch
Department of Physics and Astrononmy, University of New Mexico
stevekochscience@gmail.com

Abstract

Open data is becoming increasingly important in research. While individual researchers are slowly becoming aware of the value, funding agencies are taking the lead by requiring data be made available, and also by requiring data management plans to ensure the data is available in a useable form. Some journals also require that data be made available. However, in most cases, "available upon request" is considered sufficient. We describe a number of historical examples of data use and discovery, then describe two current test cases at the University of New Mexico. The lessons learned suggest that an instituional data services program needs to not only facilitate fulfilling the mandates of granting agencies but to realize the true value of open data. Librarians and institutional archives should actively collaborate with their researchers. We should also work to find ways to make open data enhance a researchers career. In the long run, better quality data and metadata will result if researchers are engaged and willing participants in the dissemination of their data.

1. Introduction

The role of data in science has been garnering considerable attention in recent years. Scientists have always used a combination of data, theory and experimentation to increase our understanding of the natural world. The computer age, however, has ushered in the possibility of sharing data and ideas almost instantaneously any where in the world. It has allowed the collection of huge amounts of data, often in born digital formats. The data can now often be aggregated and linked in a variety of ways. As a result, sharing and archiving of data has been an increasingly important issue in the sciences. Funding agencies are starting to require data management plans and that data be made available to others.

This is a substantial change of culture in the sciences. Even though most researchers would agree that science would benefit from increased sharing, data is still often a closely guarded secret. Additionally researchers have paid little attention to the curation and preservation of their data, and have little training or time to spend on data management and curation. Many faculty at the University of New Mexico feel there is little reward to working to make their data open and accessable. Even when researchers say they will make data available, requests to receive data are only successful about 25% of the time even

though the authors had signed certified that the data would be available.(Wicherts et al 2006). Additionally, 28% of researchers in a survey report not being able to duplicate or confirm previously publishd findings due to data withholding (Campbell et al. 2002). Despite this reluctance, science has always benefited from the flow of ideas, and there are many historical examples of discoveries made because of the fortuitous rediscovering, intersection or rethinking of ideas and data.

At the University of New Mexico we are working to archive all academic output from the University if the researchers desire it. As part of our development process, we are working with a number of test cases from researchers with large and complex data sets. Our experiences with these test cases have shown us that merely bit level preservation of data is inadequate. Several of our test cases generate large amounts of highly complex data that would not be understandable to anyone without a considerable amount of curation. Additionally, these researchers are interested in having their data used in a variety of ways, such as for education. They have found our current institutional repository to be inadequate, and are experimenting with other ways to make their data available. Finally, a recent experience with a serendipitous discovery through shared data has highlighted the need for not only making data open and available, but for creating a variety of mechanisms for finding the data.

In this paper, we outline some historical cases of data use and scientific discovery that highlight important aspects of discovery and data use. We then describe two test cases we are working on that show how faculty are currently trying to archive and disseminate their data. We then synthesize the lessons from these scenarios and describe our current progress and plans for creating an institutional archive for data that helps to fulfill our researchers needs.

2. Historical Uses of Data and Discovery

Mendel's Peas - Genetics

An early example of data archiving and reuse in biology is the case of Mendel's pea experiments. Between 1856 and 1863, grew about 29,000 pea plants (*Pisum sativum*) at his monastery. This is one of the first cases of data driven and statistical science in biology. While he publish his paper (Mendel 1866), but it was widely misinterpreted at the time, seen as an example of hybridization rather than a radically new and ultimately correct theory of inheritance. While it is unclear if Darwin was aware of Mendel's work, we do know that Darwin and others clung to the concept of blended inheritance throughout his life rather than Mendel's theory of particulate inheritance even though Mendel's theories and data would have solved several long standing criticisms of Darwin's theory of evolution (Bowler 1984).

Darwin's theory of evolution continued to suffer from an inadequate theory of inheritance, while Mendel's work, which would have solved this issue, was misunderstood and forgotten, only being cited 3 times over the next 35 years or so. By the early 20th century, researchers had begun testing other theories of inheritance, it wasn't until the rediscovery of Mendel's work that the modern age of genetics began. Perhaps most important among those who championed Mendel's work were Thomas Hunt Morgan, who while initially skeptical of Mendel's work, came to accept it and expand it, developing the Mendelian-chromosomal theory of inheritance and developing the method linkage mapping. T.H Morgan won the Nobel Prize in Medicine for his work in 1933. Mendel's work was also instrumental in the development of quantitative genetics by R.A. Fisher and Sewall Wright, whose work laid the foundation for modern evolutionary as well as animal and plant breeding (Bowler 1984).

Mendel's work is still analyzed today, by science historians. R.A. Fisher was the first to point out that Mendel's data was too close to the predicte ratios. Fisher (1936) argued that Mendel had altered his data to make his results cleaner, an accusation

that is still debated today (e.g. Hartl & Fairbanks 2007).

Watson & Crick - The Structure of DNA

With the rediscovery of Mendel's work, the fields of evolutionary biology, genetics as well as plant and animal breeding progressed rapidly, despite the fact that the structure of DNA and the physical mechanics of inheritance were unknown. By the mid-20th century the big unanswered question was the structure of DNA, where most people thought the genetic code resided. Watson and Crick entered the race, and quickly decided to use only other people's data (Watson 1968). They deduced the correct structure for DNA in 1953 (Watson & Crick 1953). They won the Nobel Prize in Medicine for their work in 1962.

The discovery of the structure of DNA is an excellent example of the importance of data sharing and reuse. Watson and Crick created no data of their own, relying entirely on the data of others. They were clear in their later writings that their discovery could not have happened without them seeing Rosalind Franklin's x-ray diffraction data, even though she was trying to keep it secret (Watson 1968).

Watson and Crick's methods are similar to current researchers who rely primarily on various molecular databases to conduct their research. While Watson and Crick are sometimes criticized for using Franklin's data unfairly and not giving her enough credit, a case could be made that her group should have shared the data. Other groups, including Franklin's were all privy to most of the same data. However, they were using the wrong x-diffraction images due to some erroneous assumptions. It is clearly important for everyone to have access to data, and also important to give credit for data. Rosalind Franklin died before the Nobel prize was awarded (Nobel prizes are not awarded posthumously), she was acknowledged in the seminal Nature paper, but not directly credited for the data (Watson & Crick 1953).

Lauterber - Magnetic Resonance Imaging (MRI)

During most of his career, Paul Lauterbur worked in the field of Nuclear Magnetic Resonance, working primarily in the field of chemistry. According to Lauterbur, in 1971 he attended a lecture on applying NMR technology to cells. After the lecture, he went to a local Big Boy restaurant and imagined applying the technology to the burger. He noted down his initial thoughts on a napkin, and the idea eventually led to then invention of Magnetic Resonance Imaging (Lauterbur 1973) and Nobel Prize in 2003. While not directly pertaining to data, this discovery illustrates the often unpredictable and sometimes quirky connections required to make a discovery.

NCBI Genbank - DNA Sequence Database

The successful DNA sequence repository Genbank began in 1979 in Los Alamos National Laboratory as the Los Alamos Sequence Database and became Genbank in 1982. Since then, its use has grown exponentially, with the number of deposited base pairs doubling every 18 months (Denton et al. 2009, Figure 1). Most biological journals now require sequences published in them to have a Genbank accession number associated with them. This growth is coupled with an associated growth in use of Genbank for discovery. It is now common for researchers to obtain nearly all their data from genbank and similar resources and make new discoveries entirely from this data.

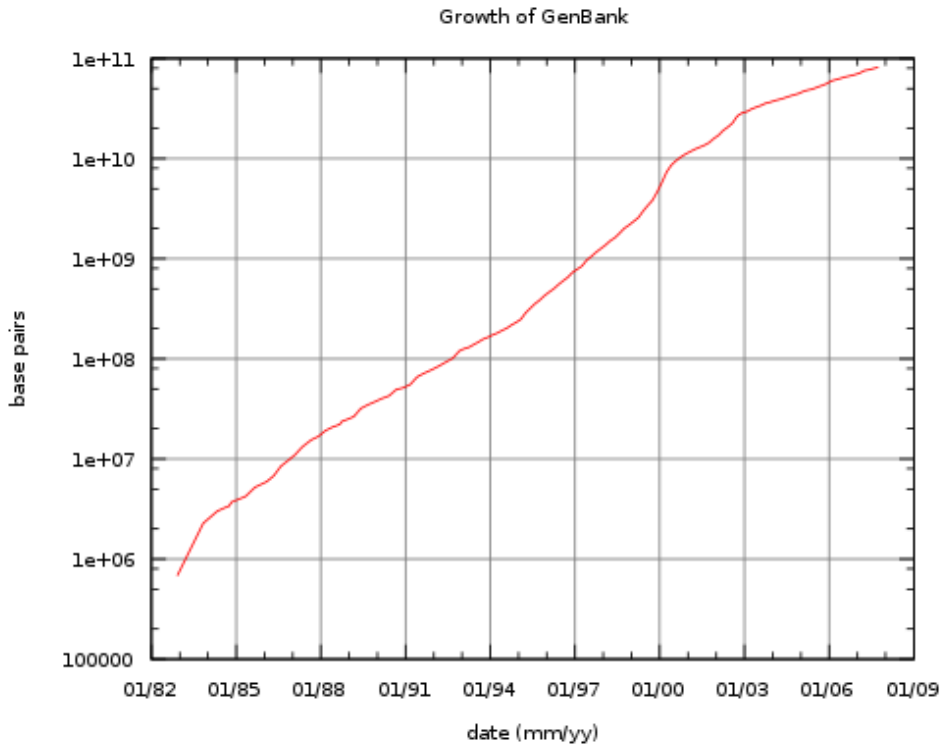

Figure 1. The number of basepair in Genbank is shown over time. Note that the graph is in a semi-log scale so that a linear increase on the graph denotes exponential growth of the database.

Genbank is an early and clear example of the potential of shared data in the digital era. Despite the success, however, there are still some limitations. Perhaps the primary problem with Genbank is the lack of curation. Depositors are responsible for providing accurate metadata with their submissions. However, most researchers are not trained in data curation and additionally there is little incentive for them to spend a great deal of time doing it. Therefore, users of the data must often spend a great deal of time cleaning up the data. Also, due to errors in metadata, many sequences are probably overlooked or misinterpreted. Fortunately, despite all the complexity of life, the structure of DNA is very simple, a linear chain of 4 different bases. Therefore the lack of curation is to some extent not an issue.

3. Test Cases for Data Services

While we can in principle accept data into our current institutional repository at The University of New Mexico, a DSpace instance branded as Lobo Vault, several researchers have found this to be insufficient for their needs. We have identified a number of faculty with both a variety of special needs for their data and also an interest in Open Data. Most of these projects involve large, complex data sets. Here we describe two test cases, chosen because they are the furthest along and because these two researchers have already been working on their own to disseminate their data in a variety of ways and in one case we have a documented instance of a serendipitous discovery being made. We have learned a number of important lessons already, and are currently working to incorporate those lessons into our services.

Kinesin Motility - Steve Koch and Andy Maloney

Steve Koch and his lab are very dedicated to the concept of open data. Typically, when they collect data, they place the data on their web server and make it freely available public domain (CCO). They also publish their protocols and other lab notes on an open source wiki type system called "OpenWetWare" (Koch et al. 2010a). Additionally, because much of their data is in the form of video, they upload examples of their data to YouTube, with some descriptive metadata, and link back to their website. This has already yielded results in the form of a recent discovery and publication. An excerpt from Steve Koch's blog perhaps describes it best.

I think it's a great little success story for open data and data reuse. In a nutshell (and I can answer questions): Some people found Andy's microtubule gliding assay data on youtube and emailed us to say it was very interesting to their theoretical work and could they use our data in a pre-print. We replied "of course!" "woo hoo!" and we told them that it's all public domain data so they are free to do whatever. As a courtesy, we said we'd like a shout-out. They went further and offered co-authorship, but Andy and I decided an acknowledgment was more appropriate at this time. Andy suggested they acknowledge open notebook science, etc. and they did in their pre-print. You can find the pre-print here: http://arxiv.org/PS_cache/arxiv/pdf/1101/1101.2225v1.pdf see Figure 3A for Andy's data and the acknowledgments section.

I think it's a great success story because (A) they never would have known about our data if it weren't open. It didn't necessarily have to have an open license, but it needed to be discoverable. (B) we never would have thought to use our data for this purpose. So obviously value was created via openness. (Koch 2011)

One very important aspect that alluded to in (B) should be highlighted. They did not think the part of the data that got reused would be useful, nor did they imagine researchers from this field using their data.

Despite their successes, they still suffer from a few difficulties. First, they often generate large amounts of complex data. The data alluded to above consists of approximately 500GB of data, in 500, 000 files. While their file structure is logical, and they do add README files in the data structure, the ability of those outside the lab to understand the data is highly limited leaving explanation of the data to the researchers and the vagaries of their memory and lifetime. Second, hosting their data on their own servers takes from their own time and resources to do research. They must pay for and maintain their own storage and arrange for incremental backup. Also, any time spent formatting and preparing their data for display is also time away from their research. Finally, they do not currently have expertise in web design, interface design or digital preservation.

Cave Microbiology - Diana Northup

A second case study involves the study of the micro-fauna of caves. Diana Northrup's research involves collecting samples of cave micro-fauna and studying them using scanning electron microscopy (SEM), spectroscopy and other techniques to study unique aspects of cave microbes. The resulting data set are groups of data, macroscopic images of collection areas, SEM images of those samples at low magnification, higher magnification images of areas inside those lower magnification images and often spectroscopy and other forms of data. Additionally there are physical specimens to be handled as well. Clearly this results in complex relationships between the various pieces of data (Figure 2). Preserving these relationships is critical to understanding the data and keeping it understandable.

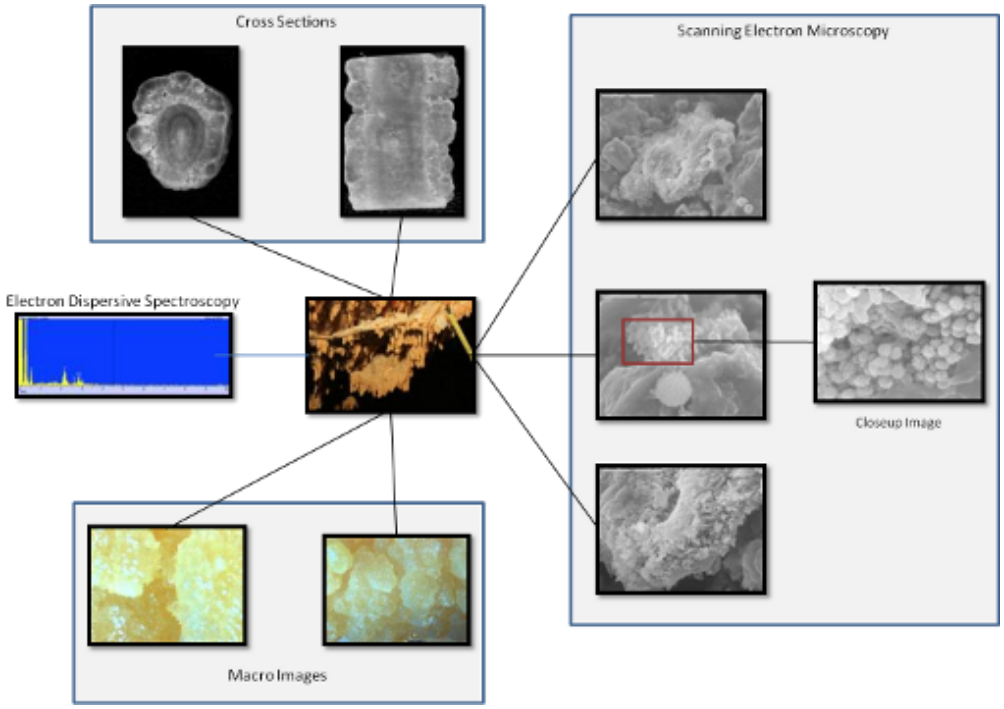

Figure 2. The relationships between some of the image data from a microbial sample taken from a cave. Note the spectroscopy data, while shown as an image here, is in fact a complex piece of data itself. Images were taken from the IDEC website (Northup 2011)

Perhaps more interesting is Northrup's vision for other uses of her data. Cave bacteria often take unusual forms that are hard to recognize. In fact, it is often difficult to know for sure if a structure in an SEM is bacteria or even a produced by some inorganic process. Exobiologists (researchers who study life beyond Earth) are often faced with similar problems. She sees her data as potentially being useful to exobiologists and is working to establish relationships between the fields with shared data. Similarly, she also has an interest in K-12 education and is interested in having her data, as well as other researcher's data in her field being used for educational purposes. She has been commended in NSF grants for her efforts as well. To accomplish these goals, she has created a web site to showcase hers as well as other data from both cave microbiology and exobiology (Northup 2011). In addition, she has attempted to incorporate lesson plans using this data into her website. However, just as in the Koch lab, her expertise and time are limited.

4. Synthesis

Using these historical examples, case studies as well as current practices in data management, curation and archiving we can derive a few lessons. First, most obviously, we must preserve the data. However, curating and archiving data is not free, and financial limitations may often dictate limiting what data can be accepted into a repository. Such decisions should be taken with great care. As shown by the case of Mendel's data as well as the Koch data, often times ideas and data are misunderstood, over looked or just lost for periods of time. Also, what might be useless in one context could be crucially important in another. The value of failed studies and flawed data is also often overlooked. Making such failures available in

some form can often prevent the repeating the same mistakes, and may also allow other researchers to correct the mistakes and succeed where others have failed.

Second, the data must be made open and available. This idea is generally accepted amongst archivists and data curators, less so among researchers themselves, although many do adhere to this principle. One could make a valid argument that Rosalind Franklin and her group should have made their data available to others, rather than forcing Watson and Crick to access it on the sly. Other groups with access to the same critical x-ray diffraction images were failing to arrive at the correct solution. Had they made their data available from the beginning, while stipulating the terms of use, they might well have found themselves being named as contributors to a great discovery.

A third lesson we can take away from these examples is the unpredictability of discovery. It would be difficult to say if MRI would have been invented if Paul Lauterber hadn't seen that lecture, if he hadn't gone to Big Boy or had a hamburger. Likewise, the connection between the Drosophila researchers work and the Koch data may not have been made if the researcher had not rather serendipitously run across it on YouTube. This suggests that when designing data repositories and when curating data, we must try not to create data models, assign metadata, or otherwise impose a model of how we think the data should be used. Rather we should actively try to create agnostic metadata that crosses disciplinary boundaries. Additionally, we should realize that all repositories are silos, no matter how well advertised or large they are. In fact, as of 2009 there were at least 1,170 moleuclar biology databases alone (Havukkala 2010) suggesting that not only are we facing a data deluge, but we are also creating a repository deluge. We should embrace the use of social media such as YouTube, Flickr, Twitter and any others and work to create metadata and well crafted text that results in successful searches by search engines rather than just relying on the searches provided by the repository. We must also consider that our repositories be indexable by search engines and also try to facilitate search engine optimpization. In doing so we can create a wider net that is more likely to cause the intersection of ideas and data.

Fourth, we cannot abandon curation. This is especially true of complex data sets. While very large data sets may tax our storage, complexity limits our understanding. If the data is not understandable, there is little reason to keep it in storage. Well curated data will be understandable to researchers in wide array fields..

Lastly, we should listen to and collaborate with our researchers, not only in designing our repositories, but actively collaborating them in projects dealing directly with their data. Working to make data and other products of research useful and available not just to scientists, but to the world at large. These collaborations between librarians and researchers would likely result in new ideas and products that could be broadly applied to data archiving as well.

5. e-Science and Data Servies at the University of New Mexico

At the University of New Mexico (UNM) we have a team of Librarians working closely with our research support staff to provide data and e-Science services to faculty. We try to begin working with researchers at the planning stages of their projects, working with them to write data management plans for their grants and consult with them to improve data management in their research facilities at their request. We also consult with researchers on appropriate repositories for their data. Our reposiotry is committed to accepting all academic output from our institution in the state its given to us by the researcher and ingest it into our repository with as little modification as possible where no other domain specific repositories exists. We will curate the data as required and consult with the researchers on other mechanisms for advertising and disseminating their data.

Current and Future Resources

At UNM our strongest resource at UNM is our library faculty and our research support staff. We currently have 4 data librarians who are specialists in various types of data, natural sciences, engineering, geospatial, business and economic, social sciences and the humanities. Additionally we have two librarians managing our DSpace instance. Our technical and physical resources are lacking but quickly catching up. As we just noted, our repository is currently a DSpace instance, with 500GB of storage. Given that the Koch data described above takes approximately 500GB storage, we are currently in the process of installing additional storage that we expect will last for 5-10 years in the future. We also benefit from having the DataONE project centered at UNM and expect collaborations with DataONE to provide an important synergy.

Current Data Archiving Strategies

Until we are able to implement a repository that is better able to handle complex data sets, we are working with DSpace as the primary entry point for data deposited in or repository. Our current strategy is to use DSpace to generate and maintain a persistent URL to the data and also to keep the high level Dublin Core metadata for the data set. The actual data is housed in its own file and storage system. We are currently annotating the data using XFDU as a structural metadata schema. XFDU is similar to METS, but with a slightly better focus on data rather than documents. Within the XFDU document we can include any metadata schema required to adequately describe the data just as in METS. We currently commonly work with PREMIS, Dublin Core, Open Microscopy Environment (OME) and Darwin Core, although we can include any schema that is appropriate.

While this system has several benefits it has some significant drawbacks. It is not scalable to very complex datasets. The actual size in bytes is not the critical issue, rather its the number of files and the complexity of the directory system that causes problems with scalability. In our largest most complex case, the Koch data described above, 500GB of data are contained in 500,000 files. The Java based code we developed for automatically processing the files and creating just the XFDU and some automatically generated Dublin Core metadata took about 5 hours to complete and resulted in approximately 5 million lines of XML code. Additionally we must still hand write much domain specific and techinical metadata. The large XFDU files were difficult or impossible to open with most XML editors, so it was necessary to segment the data into 5 pieces to handle it. Also, there are few adopters of the XFDU schema making sharing of objects difficult. Given the scalability and sharing issues were already encountering, we are now looking to another, more scalable repository system to house our data. Additionally, since DSpace is not handling the data directly, creating a user interface to the data and metadata is still required. This negates one of the biggest advantages of DSpace.

Planned Archival Strategies

Given the weakness of our current system, a better approach might be a data repository based on the perhaps familiar technology stack of Fedora repository software, a Drupal front end and Apache Solr search. Fedora has many advantages over our current system. Perhaps the greatest advantage of using this technology stack is the large number of users and developers using it already. Drupal has a large number of modules that are useful to us and we will borrow Drupal modules as needed. We also expect develop data curation specific modules as needed.

The modular nature of Fedora's content model alleviates the scalability issues we were encountering. Rather than creating large, unwieldy XML files, each directory or file is associated with a modestly sized FOXML file (Fedora's native XML digital object schema). Our current data model allows us to link the FOXML files using standard RDF to reflect the file structure of the data, as well as linking to internal and external metadata files in any schema.

A third important advantage of this system is the ability to easily share Fedora objects. In addition to the repository, we are currently collaborating with faculty in several departments where we will developing platforms for e-Research that address specific needs for their research domains or projects. These platforms can incorporate some of the modules from the archive to facilitate such tasks as metadata creation, documentation of the data's revision history and provenance as well as other aspects of data curation while the data is created. The Fedora objects created here can then just be exported with the data, easing the curation and ingest process as well as increasing the accuracy and reliability of the metadata.

Collaborative Projects

In addition to the data repository, we are collaborating with a number of faculty on e-Research projects with synergistic applications with the repository but with specific functionalities to certain domains or departments. Additionally they can provide further funding and development resources for developing data services at UNM. All the projects provide the opportunity to help researchers create metadata as the data is created, rather than doing it as an afterthought. We can automate much of the provenance and also perhaps create automated submission into our institutional repositories as well as other domain specific repositories.

These projects include a consortium of molecular motor researchers that draws people from empircal theoretical and informatic points of view, allowing them to collaborate freely, using primarily an online collaborative platform. Researchers will be able to create research groups, and control access to their data, allowing access to only themselve, named individuals, a group or even open access. Additionally we plan to provide version control, allowing better collaboration and better documenting provenance. We are also working improve upon the current IDEC Cave Microbiology platform, to allow greater and easier collaboration among researchers and educators. The site will allow the formation of groups, submission of both data and lesson plans. Also, we will provide mechanisms for enriching the data with metadata and linking the objects so that the data is easily understandable to those even outside the field. Lastly we are working to create regional online herbarium, digitiziing plant specimens for online access. All of the data on this project is already intended for general consumption. Probably the harder aspects will be normalizing metadata and data across institutions.

Many of these projects, will likely use some for of the Fedora/Drupal/Solr technology stack. By doing so we can ensure that the content objects generated among them can be shared easily wiht the institutional repositiory, and hopefully with other similar projects worldwide.

6. Concluding Remarks

Science has alway functioned on the exchange of ideas and sometimes data. With computer age and e-Science has made the exchange of ideas and data so efficient, that we are now swimming in data. Up till now, most of the data that is easily available has been the product of big science, or easily archived and understood data such as DNA sequences. However, granting agencies and a small number of pioneering researchers are starting to make a case of open data. As shown from our

examples above, sharing of data and ideas obviously aids scientific progress. In addition, there are hints that shared data can benefit's one career, as in the case of the Kinesin data, additionly publications or citatoins are possible.

Despite these benefits, many researchers are hesitant to make their data available. They often cite that they are afraid to be scooped. In addition, researches are not trained in managing their data and may be slightly embarrassed to have others see their mess. The current strategy of funding agencies to require data management plans and requiring that data be made available takes a stick approach. While this may be necessary, our experience is that if sharing data benefitted a researchers career, they would be far more willing to share. By show casing success stories such as our test cases described above, we feel we can increase researchers willingness to share. By building a robust data services suite, we will also alleviate much of the perceived cost in time. Librarians and curators especially must work to do more than place data in a repository. We need to develop value added services that researchers see as adding to their data.

Much of the changes need to be made in the culture of science and academia as well. As researchers become more comfortable sharing data, its likely a culture of citation will develop just as with the citing of journal articles. Changes in how tenure and promotion decisions must be made as well. If cited and reused data counted toward tenure, just as having your publications cited counts, researchers would seek to make their data available knowing that it directly benefits them. This of course requires methods of citing data which are currently in their infancy. Despite these difficulties, examples such as the ones above and others suggest that data is a public resource and should be shared for the common good. It is likely inevitable that the culture of research will evolve to fulfill that promise.

4. Acknowledgements

Andy Maloney generated the data for the Kinesen motility experiments, and his enthusiasm and understanding of open data greatly aided our work. We would like to thank Diana Northup for her work on IDEC. Also the hard work of all the data services librarians, Amy Jackson, Lori Townsend, Kevin Comerford and Jeff Dickey for their work in developing the services at UNM and their conversations about data.

5. References

- Benton, D. et al. (2009). "GenBank". Nucleic Acids Research 37 (Database): D26–D31.
- Bowler, P. J. (1984). Evolution, the history of an idea. Berkeley: University of California Press.
- Campbell, E. G., Clarridge, B. R., Gokhale, N. N., Birenbaum, L., Hilgartner, S., Holtzman, N. A., Blumenthal, D. (2002). Data withholding in academic genetics - Evidence from a national survey. Journal of the American Medical Association, 287(4), 473-480. doi:10.1001/jama.287.4.473
- Fisher, R. A. (1936). Has Mendel's work been rediscovered? Annals of Science 1:115–137.
- Hartl, D. L.; Fairbanks, D. J. (2007). Mud Sticks: On the Alleged Falsification of Mendel's Data. Genetics 175 (3): 975–979
- Havukkala, I. (2010). Biodata Mining and Visualization: Novel Approaches (Singapore: World Scientific Publishing)
- Koch S. (2011) An open data success story. posted February 5, 2011. http://stevekochscience.blogspot.com/2011/02/open-data-success-story.html
- Koch, S. et al. (2010a). Koch Lab - OpenWetWare. last updated June 2, 2010. http://openwetware.org/wiki/Koch_Lab
- Koch, S. et al. (2010b). KochLab's Channel, - YouTube. accessed September 1, 2011. http://www.youtube.com/user/KochLab

- Lauterbur, P. C. (1973) Image Formation Induced Local Interactions: Examples Employing Nuclear Magnestic Resonance. Nature 242 190-191.
- Mendel, G., 1866, Versuche über Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brünn 4: 3–47.
- Northup, D. (2011), IDEC - Imagery Data Extraction Collaborative | Exploring two little-known worlds: subterranean and extraterrestrial. accessed September 1, 2011. http://idec.aisti.org/
- Spice, Byron (2003-10-07). Nobel Prize for MRI began with a burger in New Kensington. Pittsburgh Post-Gazette.
- Watson, James D. (1968). The Double Helix: A Personal Account of the Discovery of the Structure of DNA. Atheneum.
- Watson J. D., Crick F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171 (4356): 737–738.
- Wicherts, J. M., Borsboom, D., Kats, J., Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61(7), 726-728. doi:10.1037/0003-066X.61.7.726