11th Annual Brick and Click Libraries: Academic Library Symposium.
Maryville, Missouri, November 4, 2011

E-science and Libraries (for Non Science Librarians)

Eric Snajdr
Assistant Librarian
Indiana University - Purdue University Indianapolis, Indianapolis, IN

Abstract

Information Technology is rapidly changing the world of scientific research. We have entered a new era of science. Some call it e-science, while others call it the 4th paradigm of science. Scientists, with the aid of technology, are continually amassing larger and more complex datasets. These data are accumulated are at an ever-accelerating rate. How will this information be organized? What, if any of it should be preserved for future use? How will it be preserved? If it is preserved, how will it be made publically accessible? The NSF and others describe the solving of problems such as these as some of the major challenges of this scientific generation. They also state that tackling these problems will take expertise from many fields, including library and information science.

A recent movement of this new era of science is an increasing requirement for scientists to archive and make their research data public. For example, the National Science Foundation (as of January 18, 2011) is requiring scientists to articulate how they will accomplish these goals within data management plans that must be submitted with each grant proposal.

What role can libraries play in this new realm of science? What role are libraries already playing? Several libraries have taken the lead in initiating efforts in assisting scientists with a variety of data management needs. This presentation will include a brief overview of the current trends as well as possible future directions in librarianship that this new era of science may lead.

Snajdr, E.A. 2011. *E-science and libraries (for non science librarians*). 11th Annual Brick and Click Libraries: Academic Library Symposium. Maryville, Missouri, November 4, 2011.

Introduction

It is important for librarians of any discipline or area of specialization to stay abreast of changes in the world of scientific research, as these changes inevitably will affect libraries and the services that libraries provide. Information Technology is rapidly changing the way in which scientific research is being done. Technology has not only allowed scientists to collect larger and more complex datasets but it has opened the possibility for scientists to share data very quickly across a global scale. Science and technology have become intertwined giving rise to what has been called e-science.

The Association of Research Libraries ("Transforming Research Libraries, e-science") has adopted the following definition of e-science from the UK National e-science Centre (Taylor), "…the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet…". Some have argued that these changes in the world of science are so significant that we have entered a new era of science. Jim Gray (xviii) calls the new era the Fourth Paradigm of science, the Fourth Paradigm being the fourth major phase of how science has been conducted throughout history.

One excellent example of e-science in action can be found in the field of astronomy. Digital images from the night sky are now available online to scientists as well as the public (SDSS.org). With the help of these public data, some professional and amateur astronomers are observing outer space, not by looking out of their own telescopes, but by looking at images created and recorded electronically. With the Internet making these images accessible to a worldwide audience, the number of eyes examining the night sky via these high-powered images has increased enormously. This, in turn, has vastly increased the possibility of new discoveries, and new discoveries have indeed occurred. For example ScienceNews, reported that amateur and professional astronomers working together have discovered a new type of galaxy ("Astronomer Unveils the Mysteries of 'Green Pea' Galaxies"). In this case, science and technology have truly intertwined and as a result, opened new possibilities for scientific discovery.

New trends for preserving/sharing science data

Traditionally, most of the scientific information available for research purposes, as well as that which was saved for perpetuity, was found in the form of published research findings. This came mainly through research articles in scientific publications, which was (and still is) available to anyone with access to expensive subscriptions of peer reviewed scientific publications. However, the traditional primary output of science found in peer reviewed scientific journal articles represents only a small portion of the scientific information that is produced through the process of scientific research.

In the various subfields of science there have therefore, been tremendous amounts of data that have never seen by anyone other than the researcher that conducted the experiment. When the researcher retired or passed away, most or all of their data has been lost forever. And in many cases, the loss of data has been huge. For example, even a single experiment that might have been conducted over a period of several months might have a vast array of data files associated with it. These data were likely summarized into a graph, table, or a few sentences in a scientific publication. Therefore, the scientific community, and certainly the public, never had a chance to see the corresponding scientific data in its raw form.

Recently, however, there has been an overall trend in the scientific community, toward making data more open. Just within this past year, there have been increasing requirements for scientists to archive and make their research data public. For example, as of January 18, 2011, the National Science Foundation (NSF) has required scientists to articulate how they will accomplish these

goals within data management plans that must be submitted with each grant proposal (Grant Proposal Guide).

Another example of the trend of data becoming more open is the Dryad repository. Dryad (*Dryad.com*) is an international repository composed of a group of 84 journals in biology, which include many core journals in the fields of molecular, evolutionary, and conservation biology. As of January 2011, authors submitting papers to journals belonging to Dryad have been required to also submit the data on which the research is based. The submitted data resides in a public online archive and can be accessed and viewed alongside the research publications. In the half of a year since the onset of the Dryad repository, 1,908 data files have been added (*Dryad.com*).

Advantages of preserving/sharing data

But why bother with preserving and sharing data? The answer lies in the many advantages and opportunities that archived and open data bring. One advantage is that the more eyes that are scanning data, looking for trends, and connections, increases the possibilities of new discoveries and new understandings. The sharing of data also opens a wide range of possibilities of combining data sets from multiple studies from multiple scientific disciplines in order to investigate large scale questions.

The NSF (Sustainable Digital Data Preservation and Access Network Partners) points out that data collected in an experiment can often be used to generate new scientific questions. One example that helps to demonstrate how data collected for one purpose can sometimes be used in new ways comes from a famous American author and naturalist. When Henry David Thoreau lived on Walden Pond, he took careful notes of his observations of the natural world that surrounded him. Among other things, Thoreau made notations of the timing of wildflowers as they bloomed in the spring. A group of scientists at Harvard University and Boston University (Wilis et al. 17029) have recently used Thoreau's data and linked it with present day observations in order to investigate the effect of global warming on plant species in the Walden Pond area. The research paper reporting these results appeared in the prestigious science journal, *Proceedings of the National Academy of Sciences of the United States of America.* This research study has resulted in more than an important scientific discovery; it also has served to demonstrate the great value of historical data. It is doubtful that Thoreau could have anticipated how his notes would be used in new ways. Researchers today, similarly could be collecting data for a particular purpose, only to have their data be of value for a completely unforeseen purpose in the future.

Another advantage of preserving data and making it accessible comes from the nature of science itself. In the field of science it is important that experiments can be replicated and that discoveries can be verified. Both replication and validation can be more easily done if one has the data from the original experiment. A researcher wishing to replicate an experiment will likely have a greater understanding of how the original work was conducted if they have the data of the original experiment on hand, since having the opportunity to examine the data opens avenues for a greater understanding of the subtleties of the original research. Also, having widespread access to data more easily enables consensus and fact checking throughout the greater scientific community.

Challenges and possible roles for libraries

With the aid of technology, scientists are continually amassing larger and more complex datasets. Since data are accumulating are at an ever-accelerating rate, the scientific community, in general, is experiencing an over abundance of information. A single research group, for example, could very quickly find themselves inundated with an overwhelming and constantly growing pool of

data. This ever-expanding accumulation of information creates several concerns. For example how is one to go about organizing, securely storing, and backing up all of these data. Additionally, will the researcher be able to easily retrieve the data a week, a month, or even years from now?

In this age of e-science, questions that will need to be answered include: What data should be preserved for future use? How will it be preserved? If it is preserved, how will it be made publically accessible? And if it could be shared and made publically available for the future, would others be able to make sense of the data? This brings up the important point that any data that are going to be preserved need to be sufficiently annotated with descriptive information (metadata, data dictionaries, and other supporting documentation) such that scientists unfamiliar with the particular study can correctly interpret the various categories and codes within the data.

The NSF and others describe the solving of these big picture problems such as these as some of "the major challenges of this scientific generation" ("Sustainable Digital Data Preservation and Access Network Partners"). They also state that tackling these problems will take expertise from many fields. One of the fields that they mention specifically is library and information science.

Tracy Gabridge (15-6) frames the curation of scientific data within research libraries as a "the last mile" of the librarian liaison role. Gabridge points out (15-6) that a major challenge for research libraries is creating the place or infrastructure within the library system to house the data. Another challenge depends on the success of liaison librarians as they work with research faculty with these new data services. Indeed, it will not be an easy task for the library to be seen as "the place" for faculty to turn to for data management needs.

Despite the challenges of forging new ground, several libraries across the country have taken the lead in the e-science arena. One example of this is reflected in the creation of new positions, with titles such as "data librarian" or "e-science librarian", that have been created in some libraries. These relatively new positions center on data management and data preservation responsibilities.

As mentioned earlier, the NSF is now requiring scientists to include data management plans as part of their grant proposals. Librarians at some institutions have become involved assisting researchers with the creation of these documents. For example, librarians at University of Virginia (Scientific Data Consulting) are providing data management plan templates that lead researchers through questions or prompts for categories that the NSF requires as part of the data management plan.

Many academic libraries have created pathfinders or subject guides for data. The content on the guides varies but in general include information and assistance in organizing data, backing up data, storing data, creating metadata, citing data sets.

One of the barriers with this new area of librarianship is that, because it is so new, few librarians are trained in this area. However, some library and information science programs are starting to make strides in this area. For example, the Graduate School of Library and Information Science at University of Illinois is offering a degree in data curation (Master of Science: Specialization in Data Curation) which in turn will help to pave the way for a new breed of librarianship.

Conclusion

Academic libraries have seen vast changes over the last several decades. And the one certainty of academic libraries of the future seems to be that changes will always be on the horizon. Despite this climate of uncertainty, one promising future role for academic libraries is the curation of the unique research products of each institution's research faculty, especially those products that lie

beyond the official research publications. In the sciences this can include the vast array data that is produced throughout the process of scientific research.

Data preservation and the openness of data is a new area for scientists. In fact this represents a major change in the way that science has operated in the past. Data preservation is a new area for libraries as well. However, libraries playing a role in the organization and preservation of information is not new. Working out the details of how this will be accomplished with scientific data collections will undoubtedly take active communication and cooperation between scientists and librarians. This new realm of e-science has opened new avenues for libraries and it is up to librarians to welcome the changes and play a vital role in assisting our research faculty in this arena.

Works Cited

"Astronomer Unveils the Mysteries of 'Green Pea' Galaxies." *Sciencedaily.com.* ScienceDaily, 2010. Web. July 3, 2011.

Dryad.com. Homepage. North Carolina State University, 2011. Web. 3 July 2011.

Gabridge, Tracy "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data." *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*, no. 265 2009: 15–21.

"Grant Proposal Guide." *nsf.gov.* National Science Foundation. 2011. Web. July 3, 2011.

Gray, Jim. Jim Gray on eScience: A Transformed Scientific Method in The Fourth Paradigm: Data-Intensive Scientific Discovery. Eds. Tony Hey, Stewart Tansley, and Kristin Tolle. 2009. xviii-xxxi. Web. 5 July 2011.

"Master of Science: Specialization in Data Curation." *lis.illinois.edu.* The Graduate School of Library and Information Science. 2011. Web. 10 July 2011.

SDSS.org. Homepage. Sloan Digital Sky Survey. 2011. Web. 3 July 2011.

"Scientific Data Consulting." *lib.virginia.edu.* University of Virginia Library. 2011. Web 3 July 2011.

"Sustainable Digital Data Preservation and Access Network Partners.". *NSF.org.* National Science Foundation, 2010. Web. 27 June 2011.

Taylor, John. "Defining e-Science." *nesc.ac.uk.* National e-Science Centre, n.d. Web. 2 July 2011.

"Transforming Research Libraries, E-science". *ARL.com*. Association of Research Libraries, 2011. Web. 10 July 2011.

Willis, Charles, Brad Ruhfel, Richard Primack, Abraham Miller-Rushing, and Charles Davis. "Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change." *Proceedings of the National Academy of Sciences of the United States of America* 105 (2008): 17029–17033. Print.