

Managing Scientific Data for Long-term Access and Use¹

Melissa H. Cragin¹

W. John MacMullen²

Jillian Wallis³

Ann Zimmerman⁴

Anna Gold⁵, Moderator

1. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820-6211. e-mail: cragin@uiuc.edu

2. School of Information and Library Science, University of North Carolina, Chapel Hill
CB# 3360, 100 Manning Hall, Chapel Hill NC 27599-3360. e-mail: macmw@ils.unc.edu

3. Center for Embedded Networked Sensing, University of California, Los Angeles
3563 Boelter Hall, Los Angeles, CA 90095-1596. e-mail: wallis@moebiustrip.org

4. School of Information, University of Michigan
1075 Beal Avenue, Ann Arbor, MI 48109-2112. e-mail: asz@umich.edu

5. MIT Engineering & Science Libraries
77 Massachusetts Ave., Cambridge, MA 02139-4307. e-mail: annagold@mit.edu

Abstract

Preservation of data for long-term use will require data management strategies that include curation and preservation planning and implementation. While data management and curatorial activities have been an integral part of some scientific domains for years (see for example, high energy particle physics), these are new concepts in other areas of science. Concepts such as provenance, representation for re-use, and work-flow capture are rarely understood, let alone addressed. By bringing together theories and best practices from archives, museum studies, and library and information science (LIS), it is possible to address these problems.

Based on current research into scientific data management problems, this panel will consider questions about sharing and re-use of data, curation and preservation, and the intersection of scientific production and scholarly communication. Our research explores information work and problems across a range of scientific areas in the life and physical sciences, including genomics, neuroscience, ecology, and earth science. As more scientific work products are shifted to open or shared data collections (including archives, repositories and databases), we will need to understand how these systems are implemented and used to support collaboration and discovery, as well as scholarly and scientific communication.

Introduction

Data-intensive research, which is dependent upon distributed, high-throughput networking infrastructures for the access and analysis of very large data sets, is increasing with the availability of data stores and the creation of new digital instruments and informatics techniques. The need to integrate data from differing scales (e.g. from gene to tissue), or across data collections presents numerous problems that are not solved easily with established data management systems (see for example the National Science Board (2005) report on Long-lived

¹ A panel presented at the American Society for Information Science and Technology Conference in Austin, Texas on November 6, 2006.

digital data). While data management and curatorial activities have been an integral part of some scientific domains for years (see for example, high energy particle physics), these are new concepts in other areas of science, often addressed in only a segmented fashion. Even the database community sees both the absence and opportunity in the development of tools and techniques for the capture and presentation of data provenance (Jagadish & Olken, 2004). One area where this is particularly evident is in researchers' preservation activities, which are often equated with the creation of an "archival back-up;" concepts such as provenance, representation for re-use, and work-flow capture are rarely understood, let alone addressed.

If we are to support scientists in these activities, preservation of data for long-term use requires careful planning, and would benefit from some new approaches. Interestingly, these problems point to the need to combine theories and practices from archives, museum studies, and library and information science (LIS); that is, it may not be possible to solve some of these issues without bringing together knowledge and best practices from the traditions of all three disciplines. For example, if research libraries are to provide data management services or add data sets (more) regularly to their digital libraries, then concepts and best practices concerning provenance will need to become more prominent in collection policies and acquisition procedures.

The members of this panel are involved in current research on the activities and work practices that pertain to the acquisition and curation of shared scientific data. While there are often panels at ASIST to present new techniques or tools for problems, we rarely hear about how new or emerging techniques are integrated into work practice, and ultimately, what this means for scholarly and scientific communication. Our research explores information work and problems across a range of scientific areas in the life and physical sciences, including genomics, neuroscience, ecology and environmental sciences. While using different approaches, each of us is investigating what happens at the intersections of scientific research, informatics, data collection, and data management for shared or collaborative use.

Based on current research into scientific data management problems, we will present recent and summative findings on the work of scientists using informatics tools and shared systems of data and infrastructure. As more scientific work products are shifted to open or shared data collections (including archives, repositories and databases), we will need to understand how these systems are implemented and used to support collaboration and discovery. General questions for the panel and audience include:

- What are some of the roles for the LIS field to fulfill?
- In what ways can research libraries / digital libraries add value to scientific data collections, as suggested by Lagoze et al, (2005)?
- Are common strategies emerging for long-term use solutions across different fields? If so, what sorts of obstacles impede these common solutions?

Panel Members and their contributions:

Melissa Cragin:

Scholarly communication: Scientists' views of a changing landscape

In the biological sciences, there is great hope for understanding and solving problems such as the development and progression of human disease. The availability of publicly accessible data stores opens new possibilities for cross-scale integration of heterogeneous data that would support novel inquiry beyond the original, often singular purposes for which data were generated. However, the actual scientific work practices that must evolve to support the re-

use of data are in conflict with the traditional mode of basic science, where researchers generate their own new data for their specific investigations but then leave that data behind as they move on to new questions.

Building on a study of the development and use of a neuroscience cell image repository, I will present findings from current dissertation research on the use of shared digital data collections and their roles and functions in scholarly communication. Data from interviews with principal investigators and their laboratory research staff indicates a great deal of variation reported on the benefits of depositing data for public use, and how this will affect the conduct of scientific research. The concept of publishing data raises a number of questions and concerns, including when data are deemed “published,” and the lack of standards for citing biological data, which poses problems for both the depositing scientists and anyone seeking to re-use the data. Depositing scientists (authors) are concerned that they will not be credited for their research products, and end users need know how to cite data and collections they use.

Two other significant issues are the lack of any system for designating parties who will be responsible for the long-term maintenance of neuroscience data collections, and the allocation of resources for on-going curation. While at first glance these are policy problems, they affect the stability of access for future verification of claims or re-use of data. In effect, this becomes a concern for scholarly communication. It is evident that shared scientific data collections act as points of intersection between scientific production work and scholarly communication. Discussion will include examples of how data collections have particular roles in neuroscience production and communication.

Panel / Audience Questions:

What sorts of partnerships and collaborations are developing between academic libraries and scientists to support curation activities throughout the data life cycle?

What are the emerging roles for institutional repositories to support scientific communication activities?

How are emerging data curation infrastructures interacting with scholarly communication processes?

W. John MacMullen:

Achieving cross-species knowledge integration with model organism databases and the Gene Ontology

Genetic and genomic data is frequently archived with supporting annotations in vertical repositories that are based on an organism or a disease type. Modern biomedical research is also often performed on a large scale, with multi-disciplinary collaborators. In addition to information overload, a major challenge facing biomedical research is the integration of related data, information and knowledge across these boundaries. The model organism database (MOD) community is addressing this challenge in part through the use of a common annotation vocabulary: the Gene Ontology project.

Through the Annotation of Structured Data project at UNC-SILS, we have been investigating annotation as a strategy for information and knowledge management across a wide spectrum of disciplines, including biomedical research. Our focus in that area is on the quantification and representation of annotation relationships within and across MODs, and

the investigation of human facets of the curatorial process, such as measures of annotation quality in MODs, that could assist in the development of improved systems for data management and discovery support.

Panel / Audience Questions:

What can we learn from the work practices of both curators of biological data repositories and their scientist end-users that might drive the development of information tools and systems to facilitate:

- a) better literature integration across organizational and disciplinary boundaries;
- b) better tools for visualization of multi-disciplinary and multi-modal data; and
- c) better approaches to knowledge discovery support systems that help uncover relationships among existing data?

Jillian Wallis:

CENS: A test bed for the study of scientific data management and use

Sensor networks are a developing technology that requires development in nearly every aspect. The data management team of CENS is devoted to identifying standards and off-the-shelf solutions, as well as developing tools and policy, in order to create a cyberinfrastructure for the use of these sensor networks. Developing cyberinfrastructure in turn requires an understanding of the intended audience communities and how they are going to use their data throughout the data chain. Thus the cyberinfrastructure will be composed of tools to support the practices of participating scientists, including data acquisition, aggregation, cleaning, verification, analysis, management, markup, sharing, publishing, and preservation, while fitting into their existing workflows. Our current research explores the social, architectural, and system needs of the scientists participating in this large research center.

Our project team has access to the ideal environment in which to study the set of research problems outlined above: the Center for Embedded Networked Sensing (CENS), a National Science Foundation Science and Technology Center based at UCLA [<http://www.cens.ucla.edu/>]. Participants in this highly collaborative research center include large numbers of cooperating scientists, technologists, and educators. CENS researchers' goals are to develop, and to implement in diverse contexts, innovative wireless sensor networks. The fundamental properties of these systems are investigated; new technologies are designed and tested; and novel scientific and educational applications are explored.

CENS currently focuses on applying embedded networked sensor technology in four scientific research areas: (i) contaminant transport monitoring: directed toward the recycling of wastewater, and preventing the impact of nitrates on groundwater; (ii) marine microorganism monitoring: detecting harmful algae using immuno-based methods; (iii) biocomplexity/habitat monitoring: developing robust tools that can be operated remotely, both in uncontrolled natural settings and agricultural settings; and (iv) structural/seismic monitoring: continuously recording data from UCLA's Factor Building (the most densely instrumented building in North America), and from a 50-node seismic network. We continue to focus on CENS' activities in these last two areas—habitat biology and seismology, as the former is an example of a data poor discipline with a history of minimal instrumentation, and the latter is data rich and highly instrumented.

In the first year (2002-2003), we sat in on team meetings across CENS scientific activities and we inventoried data standards for each area. In year 2 (2003-4), we conducted open-

ended interviews with scientists and teams, and continued to inventory metadata standards. We used the results of the first two years to conduct an ethnographic study of habitat biologists. In the current year (2005-6), we are interviewing engineers, scientists, statisticians about habitat biology data and participating in meetings of other CENS groups. Our current survey instrument, based off of all our previous observations, explores data management practices and uses, as well as attitudes towards data sharing.

Panel /Audience Questions:

In developing these tools to support data management and use, what kinds of social engineering should we build into a system? In our case, we are trying to encourage data sharing, but is it really our place to encourage instead of support this behavior?

In our own research we have seen a correlation between the amount of effort expended on data collection or verification and the scientist's sense of data ownership. Is this a phenomena being observed elsewhere?

We have also observed a spectrum of what counts as data to a scientist. Our habitat biologists will only consider some measurement as data once it has been verified and cleaned, whereas our computer scientists think that all measurements coming from our sensors are data. Is this also the case on other projects?

Ann Zimmerman:

Many challenges, multiple solutions: Facing the barriers to data sharing and re-use

Data are the building blocks on which scientific knowledge depends, serving as representations of the physical world and as evidence to support scientific claims. Recently, there has been a growing emphasis on using the raw data gathered for one purpose to answer new and different sets of questions (Arzberger et al., 2004). While there have always been issues related to data sharing and re-use, current demands are particularly pressing due to several trends. The scale and complexity of human challenges; the availability of sophisticated technologies to capture, store, and disseminate data; and the creation of policies that require researchers to share data are fueling scientists' efforts to aggregate and analyze existing data to address difficult problems. A number of alterations to scientific culture, practice, and communication are possible as scientists move from assembling and analyzing their own data to relying more on data collected by others (Brown, 2003; Glasner, 2002; Hilgartner, 1995). These changes have implications for institutions, such as digital libraries and archives that are concerned with the need to organize, preserve, and provide access to scientific data and to support their re-use.

The sharing and re-use of scientific data are promoted because the educational, scientific, and socioeconomic benefits are thought to be substantial. Yet, these activities face considerable legal, organizational, social, and technical challenges. Large-scale solutions to these challenges have been slow to develop, although some progress is being made in the technological arena. I have been studying how individual scientists and large collaborative teams of researchers overcome these obstacles. In this presentation, I report on preliminary results from an analysis of the approaches that several large-scale biomedical and environmental science collaborations are employing to address data sharing challenges. These findings are based on data collected by the Science of Collaboratories project (www.scienceofcollaboratories.org), which was a five-year study funded by the National Science Foundation to investigate large, distributed collaborations across many disciplines. Further, I compare results from studies of these collaborations with previous work I

conducted on the re-use of data by ecologists (Zimmerman, 2003). I describe two main types of data sharing solutions and the specific challenges they are intended to address.

One type of data sharing solution allows individual scientists to work as they always have, while the labor necessary to prepare data for sharing and to support their re-use are handled by others. I refer to this as the "backward-compatible" approach to data sharing since considerations for sharing are not injected into the data collection process. An example of this solution type is the work that information technologists, data curators, and others perform to create federated databases by aggregating data from many different sources. For instance, a significant portion of data in WormBase, a resource for information on the genetics, genomics and biology of *C. elegans*, are extracted from the published literature by a team of more than twenty curators (Chen et al., 2005). In contrast, the second approach forces scientists to consider barriers to data sharing and aggregation at the outset of data collection and to develop solutions in advance to deal with these issues. I define this as the "forward-compatible" solution. For example, researchers in one of the multi-institutional collaborations I studied spent almost a year to develop standardized data collection and management protocols. No data were produced at any of the sites until these common methods were in place. Hybrid approaches also exist. For example, in some cases, scientists document and submit data to an archive, and the archive integrates, maintains, and disseminates the data.

My results show that different types of data sharing solutions place different demands on those who produce data and on those who collect and manage data and make them available for others to use. In addition, individuals or small teams of researchers can often conduct their work privately, whereas large-scale collaborations are subject to increased accountability, greater interdependencies, and intensified needs for standardization. These factors affect the production, organization, and sharing of data and have implications for long-term preservation of scientific data.

Panel / Audience Questions:

What factors affect the degree to which data sharing and re-use occurs in various scientific domains? How might better knowledge of these factors be used to guide the collection and long-term preservation of data?

How will increases in the sharing and re-use of data influence the way that data are collected, organized, and managed? What opportunities might such changes present for archives, libraries, and museums?

Digital libraries are places for social interaction as well as spaces for diverse collections of data and information. How can librarians and other information professionals support collaboration around data collections and capture information from these interactions that might be useful to others' use of the data?

References

Arzberger, P., et al. (2004). An international framework to promote access to data. *Science*, 303(5665), 1777-1778.

Brown, C. (2003). The changing face of scientific discourse: Analysis of genomic and proteomic database usage and acceptance. *Journal of the American Society for Information Science and Technology*, 54(10), 926-938.

Chen, N. et al. (2005). WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Research*, 33, D383-D389.

Glasner, P. (2002). Beyond the genome: Reconstituting the new genetics. *New Genetics and Society*, 21(3), 267-277

Hilgartner, S. (1995). Biomolecular databases: new communication regimes for biology? *Science Communication*, 17(2), 240-263.

Jagdish, H.M. & Olken, F. (2004). Database management for life sciences research. *SIGMOD Record*, 33(2), 15-20.

Lagoze, C., Krafft, D. B., Payette, S., & Jesuroga, S. (2005). What Is a digital library anymore, anyway? Beyond search and access in the NSDL. *D-Lib Magazine*, 11(11). Available: <http://www.dlib.org/dlib/november05/lagoze/11lagoze.html> [Accessed 2/7/2006]

National Science Board (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Available: <http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf> [Accessed February 7, 2006]

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists*. Unpublished dissertation, University of Michigan, Ann Arbor.