Metadata Tensions: A Case Study of Library Principles vs. Everyday Scientific Data Practices

Matthew S. Mayernik
Department of Information Studies
Graduate School of Education & Information Studies, UCLA
mattmayernik@ucla.edu

ABSTRACT
Data sharing requirements and mandates are becoming more common, and many institutions are investigating data curation methods. Metadata are a critical component to any institutional scientific data curation initiative. As libraries and other information institutions become more active in this area, a number of metadata challenges will arise. In this paper we discuss four important issues: 1) the ambiguous responsibility for metadata creation between information professionals, working scientists, and hardware/software tools, 2) the tension between the highly principled library metadata approach and the ad hoc everyday practices of working researchers, 3) the ways that metadata creation and knowledge are distributed socially in research settings, and 4) the role of metadata at different stages of the data life cycle. We illustrate how they are manifested in a case study of data and metadata management in a large science and technology research center.

Keywords
Metadata, scientific data, cataloging, data management.

INTRODUCTION
Research data are becoming increasingly important as a scholarly product (Borgman, 2007). The National Science Foundation announced in May of 2010 that research proposals would be required to include data management plans, but how this requirement is to be fulfilled is still an open question (Mervis, 2010). With such data management policy formalizations becoming more prominent, researchers in many situations will find themselves facing challenges related to managing, sharing, and preserving data that they have never had to address before.

Research libraries are becoming increasingly interested in "curating" digital research data as a product of scholarly research. In this context, "data curation" refers to the activities involved in managing, storing, and preserving digital data, and, just as importantly, ensuring that it is discoverable, accessible, and usable to interested users. Long-term data curation must be an institutional commitment, as most individual researchers have neither the inclination or the expertise to curate data in perpetuity. Libraries are the obvious potential curatorial institutions, and jobs specific to data repository work are increasingly being developed within libraries for people with library training (Delserone, 2008; Choudhury, 2008).

Few established practices exist, however, for managing and curating research data. Many different issues still impede increased institutional support for such curation, and are being actively researched by members of the ASIS&T community (see for example Hodge, Furlani, Greenberg, Hourcle, and Jones, 2009). Data cannot be integrated into library collections without standardized metadata. Cataloging and metadata creation are core to library services, and are based on decades to centuries of established practices and standards. Metadata practices for describing digital data, however, are far less standardized.

In this poster, we discuss four metadata-related issues that information professionals will face as data curation initiatives within libraries pick up steam: 1) Who is responsible for metadata?, 2) Library principles vs. metadata practices, 3) The social nature of metadata, and 4) Where does metadata fit within the data lifecycle? We provide the conceptual framework for these four issues and illustrate how they are manifested in an ongoing case study of data and metadata management within the Center for Embedded Networked Sensing, a science and technology research center based at UCLA.

CASE STUDY: THE CENTER FOR EMBEDDED NETWORKED SENSING
We are performing an ongoing study of data management within the Center for Embedded Networked Sensing (CENS). This poster illustrates our initial findings of a new thrust to specifically study metadata practices in CENS. CENS is a National Science Foundation Science and Technology Center based at UCLA with four partnering institutions in central and southern California and over 200 faculty members, students, and research staff. The main focus of CENS is to develop sensing systems for real-world scientific and social applications through interdisciplinary collaborations between seismologists, terrestrial ecologists,

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

aquatic biologists, and computer scientists and engineers. Other members of the center come from urban planning, design and media arts, and information studies. CENS was founded in 2002 for an initial five years, and received renewal funding in 2007 for an additional five years.

CENS is producing data that have potential value to synthesis and longitudinal studies of large-scale phenomena such as climate change, species shifts, and global seismic activity. Sensor networks are regularly cited as contributors to the "data deluge" that researchers in many disciplines are currently facing (Borgman, 2007). CENS researchers collect many different kinds of data, including environmental parameters like temperature and relative humidity, physical parameters like chemical concentrations and pH, as well as many parameters related to the equipment themselves, such as battery voltages, disk space, and wireless link quality. CENS data vary widely in type and formats, including numerical time-series data, images, audio recordings, hand collected physical samples, among many others.

METHOD

This research uses a combined approach of ethnographic fieldwork and information system design and evaluation to characterize the creation, use, and management of metadata for research data within CENS. Our working research questions are the following:

- RQ1 - How and where are metadata created, by whom, and for what purpose?
- RQ2 - How are metadata creation tasks learned and parceled out in research groups?
- RQ3- How do local metadata practices translate to the creation of metadata for shared community repositories?

Our ethnographic data sources include four sources of data that are important to qualitative field studies: direct experience, social action, talk, and supplementary data (archival records, physical traces, and photographic data). Since 2007, we have participated in more than 10 CENS sensor deployments, encompassing more than 30 days of participant observation. Also, as members of CENS, our study is informed by regular interaction with CENS researchers, at formal gatherings, such as research reviews and retreats, weekly research seminars, as well as informal gatherings and discussions in labs and offices.

Second, we have designed a Dublin Core based metadata repository for CENS data. The repository is in its initial stages of use, and is designed to enable potential data users to discover what CENS data exist, to determine whether those data may be useful, and to learn how to acquire data of interest. In the first stages of development, we are collecting metadata from CENS researchers. As the repository matures, the metadata descriptions will be put on the CENS web site to make them accessible to web search engines (Wallis, et al., 2010). This combination of ethnographic study and information system design/evaluation provides a good test-bed from which we can develop a better understanding of the issues that institutional data management initiatives encounter.

CONCLUSION

Metadata are a critical component to any institutional scientific data curation initiative. As libraries and other information institutions become more active in this area, a number of metadata challenges are likely to complicate the process. We have identified four such important issues: 1) the ambiguous responsibility for metadata creation between information professionals, working scientists, and hardware/software means, 2) the tension between the highly principled library metadata approach and the ad hoc everyday practices of working researchers, 3) the ways that metadata creation and knowledge are distributed socially in research settings, and 4) the need to be involved at the beginning of the life cycle. These possibilities for these tensions are manifested strongly in our case study of data and metadata management in a large science and technology research center.

We are still in the early stages of investigating these issues. Further implications will arise as we continue with our ethnographic work and system design and analysis. We will describe new issues and approaches at the conference.

REFERENCES

Borgman, C.L. (2007). *Scholarship in the Digital Age: information, infrastructure, and the internet.* Cambridge, MA: MIT Press.

Choudhury, G.S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, 57(2): 211-220.

Delserone, L.M. (2008). At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries. *Library Trends*, 57(2): 202-210.

Hodge, G., Furlani, C., Greenberg, J., Hourcle, J., and Jones, E. (2009). Standards and Best Practices in Scientific Data Management: Promoting Interoperability and Re-use (Parts 1&2). In A. Grove (ed.) *Proceedings of the American Society for Information Science and Technology*. US: Richard B. Hill.

Mervis, J. (2010). NSF to Ask Every Grant Applicant for Data Management Plan. *ScienceInsider*, May 5, 2010. Retrieved May 26, 2010 from http://news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html

Wallis, J.C., Mayernik, M.S., Borgman, C.L., & Pepe, A. (2010). Digital libraries for scientific data discovery and reuse: from vision to practical reality. *2010 Joint Conference on Digital Libraries.*