Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators

Matthew S. Mayernik
Department of Information Studies
Graduate School of Education & Information Studies, UCLA
00+1+3102060029
mattmayernik@ucla.edu

ABSTRACT

Cyberinfrastructure systems for digital data will depend on effective ways of creating and sharing metadata. In distributed scientific collaborations, creating and collecting metadata is a significant challenge. Metadata creation is often an unfunded mandate. We present a preliminary study of metadata creation by data authors in a large science and technology research center. We asked researchers to create metadata using the Dublin Core-based metadata fields for inclusion in a center-wide metadata repository. The results of our pilot test indicate that data authors face a number of challenges in creating metadata, including organizing group vs. individual knowledge, adapting an unfamiliar metadata scheme to the specifics of their project, and drawing boundaries between data sets.

Topics

Information management
Information technology and services
Nature and scope of *i*Schools and *i*Research
Preserving digital information

Keywords

Cyberinfrastructure, metadata, scientific data practices

1. INTRODUCTION

Metadata is a key component of data storage, sharing and preservation systems. Quality metadata helps to facilitate the management, discovery, access, and use of data resources. Cyberinfrastructure systems for digital data will depend on effective ways of creating and sharing metadata [5, 11]. In distributed scientific collaborations, however, creating and collecting metadata is a significant challenge. Metadata creation is often an unfunded mandate. Information or data specialist positions are not yet common in cyberinfrastructure projects [14]. Many cyberinfrastructure projects therefore rely on data authors to create metadata that can be discovered and used by others outside the projects. Little research has examined the experiences of researchers in distributed research projects in creating metadata to be shared in public or community data repositories. Understanding how data authors approach the task of metadata creation – what their understandings of metadata are, what problems they encounter, and their work practices in performing the task – will provide guidance in developing metadata collection policies, processes, and technological systems for future cyberinfrastructure projects.

In this paper, I outline a study of metadata creation by data authors in a large science and technology research center. Researchers within the center are being asked to create metadata for a new center-wide metadata repository as a means to make their research products more visible to the scientific community. We are studying the challenges data authors face when creating new metadata for potential users outside the center using an unfamiliar schema. Preliminary findings indicate that researchers face a number of challenges, including organizing group vs. individual knowledge in creating metadata, adapting an unfamiliar metadata scheme to the specifics of their project, and drawing boundaries between data sets.

2. BACKGROUND

Metadata can be defined in various ways, from “data about data” to “descriptive information about data that explains the measured attributes, their names, units, precision, accuracy, data layout and ideally a great deal more. Most importantly, metadata includes the data lineage that describes how the data was measured, acquired or computed” [9]. Metadata can be created through both automated and manual processes. Both of these methods present challenges. Automated metadata creation techniques exist for text-based documents, but these techniques do not extend to creating metadata for scientific data, as a significant proportion of scientific data is not text-based. Further, automatic techniques require customization for every new type of data creation instrument, as the particulars of the data creation instrumentation and processes are a critical component of metadata descriptions. Much metadata creation thus depends on manual effort.

Additionally, the responsibility for creating metadata falls on different individuals depending on the institutional setting. The National Science Board *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century* report [12], outlines four main actors who play important roles in the data collection and curation process:

- *Data creators*: the scientists, educators, students, and others involved in research that produces digital data.
- *Data managers*: the organizations and data scientists responsible for database operation and maintenance.

Copyright and Disclaimer Information

- *Data scientists*: the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection.

- *Data users*: the larger scientific and education communities, including their representative professional and scientific communities. (pg. 25-28)

Swan and Brown [16] describes how the "data scientist" role, as presented in the Long-Lived Digital Data report, did not accurately portray the roles they observed in a study of data management practices in the United Kingdom research community. They instead identify the important data management roles as the following:

- *Data creators or data authors:* researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data, gained through experience and as a result of need or personal interest…

- *Data scientists:* people who work where the research is carried out – or, in the case of data centre personnel, in close collaboration with the creators of the data – and conduct all or a number of the [data author, data manager, and data user] functions... In origin and training they may be domain experts, computer scientists or information technologists and their career development may have required them to assimilate skills from a discipline from which they did not originate…

- *Data managers:* people who are computer scientists, information technologists or information scientists and who take responsibility for computing facilities, storage, continuing access and preservation of data…

- *Data librarians:* people originating from the library community, trained and specialising in the curation, preservation and archiving of data. Originally, the term data librarian seemed to be confined to librarians dealing with social science data, but the title now encompasses people with data skills in all disciplines… (pg. 8)

As Swan and Brown note, the boundaries between these roles may overlap, with certain individuals taking on more than one role. The responsibility for metadata creation can be equally as fuzzy. Data scientists, managers, or librarians are typically tasked with the job of creating metadata for shared data repositories, but these positions are far from ubiquitous in research settings. In practice, data creators are often expected to create metadata for their data without training or help.

As Edwards, et al. note, "data are the product of 'working epistemologies' that are very often particular to disciplinary, geographic, or institutional locations" [7]. Metadata representations created by data authors are likewise products of 'working epistemologies', and are enacted in different ways in different situations. The standardization of metadata practices varies on a discipline-to-discipline basis. For example, astronomers have made substantial progress in developing community data and metadata standards [10], while habitat ecology has been less successful in this endeavor [2, 13]

Part of the challenge in developing data and metadata systems for research data is that data authors often have little experience in creating structured metadata. Effective information system design can mitigate some of the difficulties resource authors may encounter while creating metadata [4], but research data challenge the metadata creation process in ways that other digital resources, such as web pages and digital documents, do not. The next section introduces our work in designing a metadata repository for data collected by researchers in a large academic science and technology collaboratory.

3. RESEARCH CONTEXT

The research reported here take place within the Center for Embedded Networked Sensing (CENS). CENS is an ideal setting in which to study the emergent changes in scientific research that are being brought about by advanced technology. CENS is a distributed research center [3] based at UCLA with five partnering institutions in central and southern California. Over 200 faculty members, students, and research staff from a number of disciplines are associated with CENS at any given time. The main focus of CENS is to develop sensing systems for real-world scientific and social applications through collaborations between seismologists, terrestrial ecologists, aquatic biologists, and computer scientists and engineers. CENS was founded in 2002 for an initial five years, and received renewal funding in 2007 for an additional five years. Other members of the center come from such disparate disciplines as urban planning, design and media arts, and information studies.

As the center has matured, CENS has been more proactive in making research products available. This has stemmed from internal needs, including the administrative need to keep better track of the center's growth, as well as from external pressure from the NSF to increase the visibility of the center's research output. The first effort in this direction was to make the center's research publications available on the web through the University of California eScholarship repository [15]. This process is still ongoing, but has been largely successful in increasing the visibility and utilization of CENS publications.

The second thrust in our work focuses on research data. Data are taking on growing importance as a product of institutionalized research [1]. We are currently working with the CENS administration to develop a "data sharing" system. The system is being designed to enable re-use and re-purposing of CENS research data by potential outside users. The system will not be collecting the data themselves, due to the marked heterogeneity of data resources and data collection practices of the CENS community [2]. Rather it will focus on making CENS data visible and discoverable to the public by collecting metadata descriptions. CENS researchers will be asked to "register" data using a set of descriptive fields. The data descriptions will then be posted on the CENS website, allowing them to be discovered through web searches or by visitors to the CENS site by provide descriptive information about the data, as well as ways for interested users to get in contact with the appropriate person at CENS for more information.

Our goals in developing a data sharing system for CENS are:

1. *Make CENS data discoverable.* We focus on data "discoverability" because individual/lab data collection and storage practices vary widely within the center. CENS researchers collect a large variety of data resources, including images, audio files, physical samples, and numeric data in both digital and analog form. These resources are spread around many different community, lab, and individual computer systems. Some CENS data is available online through lab websites, but large amounts of data are not currently available online. Because of this variability, collecting and integrating all of the center's data into a single system would be prohibitively expensive and time consuming. Instead, we are designing a metadata repository that allows potential data users to find what data exists and whether the data might be useful to them, as well as providing details about how to get access to data if desired, through links to data or through contact information for the relevant researchers. We are investigating possible policies regarding the timeline for contribution to the metadata system, such as one year from collection or one year from initial funding.

2. *Help CENS researchers keep track of data resources.* In addition to providing a tool that makes CENS data more visible to individuals outside of the center, the metadata repository is intended to help individual research teams within the center to keep track of data created by their own group. Similarly, the metadata repository will provide the center's administrators with a new tool to illustrate the research output of the center.

3. *Sustainability.* The metadata collection system should be sustainable beyond the funding of the center, which will end in 2012. Thus, we are focusing on designing the system to be lightweight, in that it should be easy to use with minimal assistance. Additionally, we are using open source software tools for the back end database and web display.

As an initial step in the design process of this system, we created a preliminary metadata schema that could be used to test possible data description fields. The next section describes a pilot test of these metadata fields, including the fields tested and the test method.

4. TEST METHOD

Four CENS researchers have taken part in the pilot test of the metadata fields, two computer scientists, an engineer, and a domain scientist. The domain scientist and one of the computer scientists are part of the same research team. The participants in this test were chosen through targeted sampling of individuals who were known to have participated in original data collection, and as well as to sample from multiple disciplines and projects within the center. We asked these researchers to create metadata using the below fields for the main data that they were using in their primary day-to-day research. We used a "talk-aloud" protocol, asking the testers to describe what they were thinking and writing as they completed the metadata descriptions. During the test, we observed and took notes of the researchers' activities and comments as they completed the task.

After the testers completed the form, we asked targeted questions about their experience in performing the task. Post-test questions included asking the researchers which fields they felt were the most and least useful in describing their data, what additional fields might be necessary, and what benefits (if any) they feel that they receive from creating this metadata, among others.

The metadata fields used in this test are based on the Dublin Core metadata set [6], as shown in Table 1. The Dublin Core metadata set was chosen for it's flexibility and simplicity in providing descriptive fields for resource discovery. Discipline-specific metadata schemas were considered, such as the Ecological Markup Language and SensorML, but these were deemed to be too inflexible for the diversity of research and data types found in the center.

Table 1. Metadata Fields Used Preliminary Test

Data description fields	Dublin Core elements*
1. CENS project name	title
2. CENS research group	publisher
3. dates (of data collection)	date
4. place	coverage
5. people	-
- contact person	creator
- other participating researchers	contributor
6. data type	type
7. data description	-
- research question (why collected)	description
- what collected (variables)	description
- data collection process and equipment	description
- size, format	format
8. related publications (eScholarship URL)	relation
9. related deployment info. (CENSDC URL)	relation
10. keywords	subject
11. location of the data (URL)	identifier
12. permissions	rights
13. funding source	source

*the Dublin Core element "language" is not used

All of these fields were presented to the user as free-text entries, except for two fields, the "CENS research group" field and the "data type" field, for which the testers were asked to choose from a pre-defined list. The option list for the "CENS research group" field were taken from an established set of categories that exists within the center, and the option list for the "data type" field were taken from the list of data types given in the DCMI type vocabulary. The tests and follow-up questions took between 20 and 30 minutes per tester. The next section describes the main points of interest that arose during these pilot tests and our subsequent analysis.

5. PRELIMINARY RESULTS

The preliminary findings of the pilot test identify a number of issues that complicate the metadata creation task. Due to the limited scope of this pilot test, these are not meant to be definitive results; rather, they outline important issues that we will use as points for further investigation as our project matures.

- *Item in hand vs. distributed objects:* In many CENS projects data are not individual self-contained items. They may have many constitutive pieces, such as multiple files and database tables, and they may be spread around multiple locations, such as lab servers and personal computers, or for one pilot tester, even located in multiple institutions. In creating metadata, researchers have to decide what is to be described as part of a single project or data set, and where to draw boundaries between data sets.

- *Non-self-describing resources:* Much of the data collected by CENS researchers are not textual, thus researchers must either create textual descriptions from scratch to describe image, audio, or numeric data, or they else adapt existing text from research publications or technical reports to the data description task.

- *Sense making:* Metadata fields may not make sense to a researcher who has not seen them before. In all four pilot tests, the researchers asked for clarification of what they were expected to include in particular fields, requesting examples or further explanation. Some fields, such as "permissions", were problematic to all testers, while other fields, such as "size and format", were only confusing to individual testers.

- *Projected/reverse sense making:* The potential users and uses of research data are often not obvious, even to the researchers who collected them [2]. Researchers must try to project what ambiguous potential future users will need to make sense of their data. In the pilot tests, one strategy people used was to imagine why potential users might be interested using their data. In contrast, one tester took the strategy of thinking about it from the other direction: his own use of outside data. As he said, "If I was going to use somebody else's data, I would want to know…"

- *Talking vs. writing:* In describing their approach to filling out a specific field, particularly fields that they were less sure about, the pilot testers would "talk through" a field until they were surer about what to include. These verbal discussions about what should or should not go in a given field were not always reflected in what was written down. Often a rich verbal discussion resulted in a brief written statement.

- *Individual knowledge vs. group knowledge:* CENS research takes place in group settings. Individual researchers may not know what to include in certain fields, but do know who in the group to ask. For example, a couple of the pilot testers said that they would need to ask their principle investigator about how to fill out the "funding" and "permissions" fields. Another related issue is that different individuals in the same project may have different perspectives on what the boundaries of the data set are, and what descriptive information should be included. For example, the domain scientist and the computer scientists who are part of the same research team emphasized different parts of the same data. As part of the data description, the domain scientist emphasized the physical work involved in installing research equipment in the field and did not provide many technical details. In contrast, the computer scientist emphasized technical features of the data and the way it was collected, and gave no reference to the field work.

- *State of a project:* Different CENS projects are in different states of completion. Metadata has different importance at different stages of a project. One of our pilot testers is involved in a project that has been collecting the same data for over a year and a half, while another pilot tester has been involved in his current project for about six months, with data collection taking place for less than half of that period. In the latter case, the tester described how creating metadata for our system does not benefit him much currently, because his data is so limited in scope that it would not be useful to anyone else. He went on further to say that if they were to expand their data collection considerably, which they hoped to do in the future, then our metadata system would be very useful to him as a means to make his data more accessible to outside users. Additionally, at this early stage of the project, they had not produced any publications or reports on the project, which meant that he did not have any existing text on which to draw in creating metadata descriptions.

Reflecting back on our initial goals – to make CENS data more discoverable, to help research groups keep track of their own data, and to develop a sustainable system – a couple of key challenges require further study. First, many of the issues identified above illustrate the lack of expertise that data authors have in metadata creation. This points to the development of training programs and more explanatory metadata creation systems, including examples and fuller descriptions of the metadata fields. Second, the ambiguity of boundaries around data sets and the fluidity of prospective users and uses of data suggest that training material and activities will need guidelines regarding the focus of the metadata creation process. Third, the tensions between individual and group knowledge suggest that we investigate metadata creation methods that include both individual and group contributions. And fourth, the varied states of project maturity suggest that we investigate the ways that metadata are, or can be, created piece-by-piece during the lifetime of a project.

6. FUTURE DIRECTIONS

Our work on this project is ongoing. We hope to have the initial metadata collection system online by February. At the conference we will present results from this pilot test, as well as results of further tests that take place in the interim period. Over the longer term, we plan to extend this study of metadata creation by data authors in a number of ways. We plan to perform targeted interviews with data creators focusing on understanding their current metadata practices in their own work. We will ask researchers what “metadata” means to them, what their current data description practices are, and what is involved in sharing their data with people both inside and outside their research group (including the role of metadata in that process). As part of this, we will ask to see the data organization schemes (folder structures, naming conventions, database layouts, etc) currently used by research teams and perform content analysis on these schemes. This will help to characterize the typical state of personal data archives in distributed research environments.

Second, to continue the present study, we are forming plans to use video-taping as a research method. Video-taping will enable more careful study of researchers as they use the metadata creation system introduced in this paper. This will allow us to perform more grounded analysis as our study increases in scale beyond the handful of testers included in our pilot study. Additionally, we plan to investigate new methods of group-oriented metadata creation in our community. Video-taping will be useful in collecting and analyzing the complex interactions that take place in group settings.

These preliminary findings point to the potential contributions of a larger study of the metadata creation process for data creators, including an understanding of:

- the practical difficulties in situations where data managers do not exist for data creators when creating metadata for contribution to a data sharing system
- how metadata creation tasks are parceled out in research groups
- how the setting for metadata creation activities (i.e. individual vs. group) impacts how those activities are conducted

7. CONCLUSION – IMPLICATIONS FOR iSCHOOL RESEARCH

The implications of our research on metadata creation in cyberinfrastructure projects are multi-fold for the iSchool research community, and we hope to promote discussion of these issues. Data are a growing component of the scholarly information infrastructure and must be integrated into larger discussions of technology, institutions, practices, and policy [1]. iSchool research has focused much more on documents than on data. Techniques that have been effective in promoting access and interoperability of documents may not be applicable to data and other digital scientific resources. Research relating to scientific data practices and data preservation and curation are small but growing areas of iSchool expertise. The development of a larger research base in these areas is critical to enhance our understanding of the cyberinfrastructure "blank canvas" [8], and to facilitate the development of a trained workforce of individuals with expertise in data and metadata management [12].

8. ACKNOWLEDGEMENTS

CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator; Christine L. Borgman is a co-Principal Investigator.

9. REFERENCES

[1] Borgman, C.L. 2007. Scholarship in the Digital Age. Cambridge, MA: MIT Press.

[2] Borgman, C.L., Wallis, J.C., Mayernik, M.S., and Pepe, A. 2007. Drowning in Data: Digital Library Architecture to Support Scientists' Use of Embedded Sensor Networks. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (Vancouver, BC). ACM. http://repositories.cdlib.org/cens/wps/216/

[3] Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., & Olson, G. 2007. From Shared Databases to Communities of Practice: A Taxonomy of Collaboratories. Journal of Computer-Mediated Communication. 12(2): 652–672. doi:10.1111/j.1083-6101.2007.00343.x

[4] Crystal, A., and Greenberg, J. 2005. Usability of a metadata creation application for resource authors. Library & Information Science Research, 27(2): 177-189.

[5] Cyberinfrastructure Vision for 21st Century Discovery. 2007. Washington, D.C.: National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

[6] Dublin Core Metadata Initiative. 2009. Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/

[7] Edwards, P.N., Jackson, S.J., Bowker, G.C. and Knobel, C.P. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Final report of the workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" [pg. 32]. http://hdl.handle.net/2027.42/49353.

[8] Freeman, P. A., Crawford, D. L., Kim, S., and Munoz, J. L. 2005. Cyberinfrastructure for Science and Engineering: Promises and Challenges. Proceedings of the IEEE, 93(3): 682-691.

[9] Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D., and Heber, G. 2005. Scientific Data Management in the Coming Decade. CTWatch Quarterly, 1(1). http://www.ctwatch.org/quarterly/articles/2005/02/scientific-data-management/

[10] Hanisch, R.J. 2006. Data standards for the international virtual observatory. Data Science Journal, 5: 168-173. http://www.jstage.jst.go.jp/article/dsj/5/0/168/_pdf

[11] Lynch, C. 2008. Big data: How do your data grow? Nature, 455(7209): 28-29. http://dx.doi.org/10.1038/455028a

[12] Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. 2005. Washington, D.C.: National Science Foundation, National Science Board. http://www.nsf.gov/pubs/2005/nsb0540/

[13] Millerand, F. and Bowker, G.C. 2009. Metadata standards: trajectories and enactment in the life of an ontology. in Martha Lampland and Susan Leigh Star (eds) Standards and Their Stories [pp. 149-165], Ithaca, NY: Cornell University Press.

[14] Palmer, C.L., Heidorn, P.B., Wright, D., and Cragin, M.H. 2007. Graduate Curriculum for Biological Information Specialists: A Key to Integration of Scale in Biology. The International Journal of Digital Curation, Volume 2, Issue 2. http://www.ijdc.net/index.php/ijdc/article/viewFile/42/27

[15] Pepe, Alberto, C.L. Borgman, J.C. Wallis, and M.S. Mayernik. 2007. Knitting a fabric of sensor data and literature. in Information Processing in Sensor Networks. 2007. Cambridge, MA: Association for Computing Machinery/IEEE.

[16] Swan, A. and Brown, S. 2008. The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Report to the JISC, School of Electronics & Computer Science, University of Southampton. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf