Data Sharing, Small Science, and Institutional Repositories

Melissa H. Cragin[1], Carole L. Palmer[1], Marina Kogan[1], Jacob R. Carlson[2], & Michael Witt[2]
[1]Grad. School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA
[2]Purdue University Libraries, Purdue University, West Lafayette, Indiana, USA

Introduction

If cyberinfrastructure is "principally *about* data: how to get it, how to share it, how to store it, and how to leverage it" for scientific discovery and learning (Edwards, Jackson, Bowker, & Knobel, 2007, p. 31), then advancing cyberinfrastructure is dependent on our understanding of how to support data practices and needs. Sharing is at the heart of success, since collecting, storing, and making use of data can only come after the means for sharing are in place. In sciences served by disciplinary or nationally scoped infrastructure initiatives, sharing research data is considered an inevitable trend. Unlike fields such as physics and astronomy that tend to have standard data practices, data sharing in "small science" is not common or expected. It functions largely as a cottage industry where data is exchanged based on professional relationships and personal communication. Nonetheless, over the long term, small science researchers, who span many fields and produce many different forms of highly valuable data, are expected to produce more data than researchers in big science fields (Carlson, 2006). These are also the scientists who are increasingly turning to their university libraries and institutional repositories (IRs) for assistance with their data problems. In response, IRs at many universities are now providing support for primary research data with varying architecture and service model implementations (Choudhury, 2008; Witt, 2008; Wong, in press).

Supporting small science research requires that IRs provide data curation services that are tailored to the unique qualities of the data and researchers' sharing requirements, but that are also aligned with the growing, global e-Science and curation infrastructure. At present, a number of research and development efforts are focused on solutions for small science data (Borgman, Wallis, Enyedy, 2007; Karasti, Baker, Halkola, 2006; Rice, 2009), but, in practice, management remains ad hoc (NSB, 2005); descriptive standards, when they exist, are often not applied, and most data remain concealed locally and are undiscoverable (Heidorn, 2008).

The Data Curation Profiles project is investigating *what* data researchers are willing to share, *when*, and *with whom*, and researchers' needs and requirements for sharing those data through an IR. Following on recommendations from the Research Information Network (RIN) (2008), our research addresses the need for "full account(s) of the different kinds of data that researchers create and collect,… of the significant variations in … behaviors and needs in different disciplines, sub-disciplines…; and make clear the categories of data that they wish to see…shared with others" (p. 17). We analyze several disciplines not covered in the RIN study, elaborating the specific data "forms and varieties" scientists are willing to share. We identify scientists' self-reported practices and views on "sharable data", the data forms or representations that are requested most often by others or are thought to have the most scholarly or re-use value, and consider the implications for curation services. Our disciplinary practices approach aims to compare how data-related scholarly activities vary among disciplines, specialties, and research areas, as these differences are essential to building both technical capabilities and effective policies for multidisciplinary data repositories (Cragin, 2009; Palmer & Cragin, 2008).

Methods

The study was designed to investigate a range of disciplines and different forms of data that might be covered by curation services within an academic IR context. A sample of twenty-two scientists who conduct small science research and have an interest in sharing data were interviewed. Participants from the University of Illinois at Urbana-Champaign and Purdue University were recruited by librarians who already had professional relationships with the scientists. The sample included researchers whose work is data-intensive and who generate large digital data sets, as well as researchers who generate multiple kinds of data. Twelve disciplines were represented (the number of participants is in parentheses): Agronomy & Soil Science (5); Anthropology (3); Biochemistry (1); Biology (1); Civil Engineering (1); Earth & Atmospheric Sciences (2); Electrical & Computer Engineering (1); Food Science (1); Geology (3); Horticulture & Plant Science (2); Kinesiology (1); Speech and Hearing (1). We report here on the 19 participants who completed the full two-stage interview process.

Following Institutional Review Board approval for human subjects research, interviews were conducted utilizing structured worksheets to help focus attention on data issues. In the first stage, a Pre-interview Worksheet was distributed prior to the interview asking the participant to identify their research area and to describe two recent or on-going projects "from the perspective of the data." The interview sessions that followed were semi-structured and ranged from 60-120 minutes. Second stage interviews included a Requirements Worksheet, designed to gather details about curation needs and requirements for the specific forms of data participants had stated they were willing to share in an IR context, and supplemented with customized follow-up questions based on the results of the initial interview.

All interviews were recorded and fully transcribed. The initial code list was developed through independent manual coding of selected interviews by multiple team members. To optimize intercoder reliability, the team

worked together to develop a shared understanding of broad terms and meanings. Transcripts were then coded using NVivo 8 qualitative analysis software applying the initial broad categories followed by iterative micro-analysis of data related to strong emergent themes. Results from the data generated with the Requirements Worksheet were analyzed to identify patterns and contrasts regarding the data scientists were willing to share prior to publication, followed by further analysis of the interview data to draw out associated motivations and rationales. We were particularly interested in capturing how raw data are transformed throughout the research process, and how these representations relate to "sharability".

Findings

In this paper we report on when and with whom researchers are willing to share data, what kinds of data they are willing to share, and sharing incidents that provide insights into barriers to sharing. In general, in the small science research areas represented by our participants, with the exception of genomic data, there were no field-wide norms for sharing. While sharing with close, trusted collaborators happened with ease, sharing with anyone outside of this inner circle, including other members of a project team, took place through "just in time" negotiations. Views on public sharing of data in repositories were primarily speculative, since most respondents had only shared data within collaborations or by request.

The forms of data identified by the participants as sharable were often also considered to have the most scholarly or re-use value. Spreadsheets were the most common, followed by images and databases. Table 1 excerpts 4 of the 19 cases to illustrate characteristics of a sharable data set from the scientist's perspective.

Table 1: Sample of shared data forms and selected data set characteristics

Field	Specific Research Area	Form to be shared	Formats	Type of data set	Size	Shared when?
Atmospheric science	severe weather modeling	compressed output of the model	Vis5D	1 file / dataset	10-100 Mb	4-6 month embargo,
Agronomy	water quality, drainage, and plant growth	cleaned and reviewed sensor and sample data	.xls	approx. 100 files	~1MB each, up to 20 Mb	After publication
Geology	geobiology and microbes	averaged sensor and sample data; photographs	.xls; jpg	1 file; images	< 1 Mb	After publication
Civil Engineering	traffic movement	cleaned and normalized sensor data	MySQL (postgresql)	1 database	approx. 1000 K/day	1 month to 1 year embargo

There is a high level of variation in what is considered a sharable data set. Image formats (4), databases (4), tabular data (10) were the most common, but many sharable units are constructed either from multiple files of different formats or were composite data, where multiple data sources are integrated and analyzed to produce a new complex data set.

Participants had positive views of data sharing in general, and expressed openness to sharing their own data. Not surprisingly, willingness to share increased as data were cleaned, processed, refined, and analyzed in the course of research. Only one participant was willing to share raw data beyond immediate collaborators, and while five participants stated that they were willing to share "with everyone" once the data were normalized, there are few examples of this actually occurring. Of the twelve participants who stated they were willing to share their data after the findings were published, five would first require an embargo period, and these ranged from 1-3 months to 2-5 years. For at least eight participants, sharing any data before publication or embargo was strictly limited to known and trusted individuals who were either immediate collaborators or known associates.

In cases where data had actually been shared, limitations and conditions were common but the "rules" for sharing were far from systematic. For example, during the course of the two interviews, one participant's views changed dramatically due to two experiences where other scientists used his data for publication without proper attribution. He retreated from enthusiastic advocacy for early data sharing to a highly conservative stance of willingness to share only with immediate collaborators just before publication, and only making data public after publication. Another scientist recounted co-authoring and attribution problems that led to withholding of data among a collaborative group. A number of participants also described incidents of wrong or inappropriate interpretation of data and strategies developed to guard against this kind of misuse. In one case, a participant was only willing to share data if she was allowed to approve the new application and interpretations.

Discussion

The RIN report (2008) makes clear the need to understand sharing behaviors and norms at the sub-discipline level to guide decisions and policy development for curation services that will support shared small science data. The findings from this study advance our understanding of scientists' views on the most "sharable" data set for their research area. However, we also have evidence that suggests that the data sets deemed the most

"presentable" or easily shared may not be the most valuable for preservation over the long term, especially for re-use by researchers in other disciplines. Moreover, as noted in the RIN report, scientists rarely have the skills or resources needed to prepare all their data for public sharing, and we know that the cost of returning to the sharable form of data at the end of a project is too high. Sharable forms ought to be acquired at the research stage when it is produced; as, findings here show that researchers will not want these data made public for some period of time, an embargo option will be a necessary feature of university curation services.

New infrastructures will remain underutilized if they do not account for the range and variety of concerns driving these practices. We have substantiated some of the concerns previously reported about misuse of data (Edwards, et. al, 2008; RIN, 2008). Further research is needed on re-use risks with various data forms, as well as a systematic study of levels of contextual metadata required as part of the curation process and effective policies on use and attribution in the academic environment.

As such, provision of curation services at universities will require thorough understanding of the conditions that drive data practices more generally, and sharing practices in particular. Findings from this investigation show that sharing practices are related not only to disciplinary norms and research-related factors, but also to particular, recent sharing experiences. These will need further documentation to inform development as repository use increases. In addition, the high level of variation and complexity in data forms, even among the selected research areas represented in this study, indicate that allocation of resources for curation services for small science will be intense, particularly at acquisition and ingest stages (Beagrie, Chruszcz, & Lavoie, 2008). It is anticipated that individual institutions will not have the capacity to manage or provide curation services for all locally produced data, therefore service models and collection development policies will need to be aligned with institutional missions, while being scoped to meet organizational capacity.

Acknowledgements: This research is supported by the Institute of Museum and Library Services grant # LG-06-070032-07, D. Scott Brandt, PI.

References

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). Keeping research data safe: A cost model and guidance for UK universities. Final Report to JISC. Available: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf. Accessed July 9, 2009.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal of Digital Libraries, 7*(1-2), 17-30.

Carlson, S. (2006). Lost in a Sea of Science Data. *The Chronicle of Higher Education*, 23/06/2006.

Choudhury, G. S. (2008). Case Study in Data Curation at Johns Hopkins University. *Library Trends, 57*(2), 211-220.

Cragin, M. H. (2009). Shared scientific data collections: Use and functions for scientific production and scholarly communication. Unpublished dissertation, University of Illinois at Urbana-Champaign.

Edwards, P. N., Jackson, S. J., Bowker, G. C., & 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". Available: http://hdl.handle.net/2027.42/49353.

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends, 57*(2), 280-299.

Karasti, H., Baker, K.S., & Halkola, E. (2006). Enriching the notion of data curation in e-Science: Data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work, 15*(4), 321-358.

National Science Board (September 2005): NSB-05-40, Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Available: http://www.nsf.gov/pubs/2005/nsb0540/. Accessed April 7, 2009.

Palmer, C. L. & Cragin, M. H. (2008). Scholarly and disciplinary practices. *Annual Review of Information Science and Technology, 42*, 165-212.

Research Information Network. (2008). To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network. Available: http://www.rin.ac.uk/data-publication. Accessed June 25, 2009.

Rice, R. (2009). DataShare Final Report. Available: http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf . Accessed June 29, 2009.

Witt, M. Institutional Repositories and Research Data Curation in a Distributed Environment. *Library Trends, 57*(2), 191-201.

Wong, G. K. W. (in press). Exploring research data hosting at the HKUST institutional repository. *Serials Review*, doi:10.1016/j.serrev.2009.04.003. Accessed July 7, 2009.