Units of Evidence for Analyzing Subdisciplinary Difference in Data Practice Studies

Melissa H. Cragin Tiffany C. Chao Carole L. Palmer

Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
cragin; tchao; clpalmer@illinois.edu

ABSTRACT

Digital libraries (DLs) are adapting to accommodate research data and related services. The complexities of this new content spans the elements of DL development, and there are questions concerning data selection, service development, and how best to align these with local, institutional initiatives for cyberinfrastructure, data-intensive research, and data stewardship. Small science disciplines are of particular relevance due to the prevalence of this mode of research in the academy, and the anticipated magnitude of data production. To support data acquisition into DLs – and subsequent data reuse – there is a need for new knowledge on the range and complexities inherent in practice-data-curation arrangements for small science research. We present a flexible methodological approach crafted to generate data units to analyze these relationships and facilitate cross-disciplinary comparisons.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, systems issues, user issues.*

General Terms

Theory

Keywords

Methods, data practices, small science, data curation

1. INTRODUCTION

As libraries assess how to best align themselves with institutional eResearch initiatives and participate in the emerging global system of data resources, there are a range of questions facing digital libraries (DLs) taking on data stewardship roles, such as which data to acquire and what services to provide. It is anticipated that vast amounts of data will be produced by "small science" research that is generally conducted by a single scientist or small group whose funding is limited and success is based on collecting and analyzing their own data. These data are highly variable in type and format, and additionally complex due to uneven application of standards and idiosyncratic data management practices. Such practices are diverse even within disciplines, and these distinctions have significant implications for data policy and stewardship. To support acquisition into DLs, and subsequent data reuse, there is a need for new knowledge on the range and complexities inherent in practice-data-curation arrangements for small science research [1]. Methodological approaches for studying scientific practices across data communities must be targeted yet flexible, especially during periods of technological change.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

2. SMALL SCIENCE AND DIGITAL LIBRARIES

Previous efforts to support small science data in DLs have generally focused on curation issues or data management by scientists [2], [3]. A long-term study at the Center for Embedded Network Sensors (http://research.cens.ucla.edu/) has given careful consideration to the handling of research data and implications for digital libraries [4], [5]. Together, these studies have illuminated problems for DL development related to data variation in small science. In contrast, the RIN report on data sharing [6] stands out as a comprehensive study across several large disciplines, providing an important framework for comparison, along with discipline-specific analysis that can inform high-level policy. However, we have found the sub-discipline to be a more optimal level of analysis, allowing critical focus on the domain questions and data types that produce the 'science' in a community–the social unit where data sharing practices and reuse can best be explored.

3. PERFECTING THE METHOD

In our studies of data practices over the last several years, we have refined a qualitative approach for investigating data practices that integrates targeted semi-structured interviews, data inventorying, artifact analysis, and purposive sampling. The sequence of data collection, targeted participants, and multiple data sources work together and are all essential in producing dense, high-quality units of evidence. In essence, our units of evidence are case studies; and, as is often true with case study research, there may be considerable difference in the volume or density of each kind of data collected within or across a given case. When carried out systematically, the case study is a rigorous approach for investigating how and why contemporary phenomena occur in the ways they do.

In our current approach [see Table 1], a Pre-Interview Worksheet is used to orient participants to questions about their data and serve as a base of reference for the Research Interview. Verbal validation of the Worksheet's content during the initial session facilitates deep discussion on the participant's research. A subsequent Follow-up Interview is used to clarify or address gaps from the initial Research Interview. Specific data types identified by researchers as having scholarly importance for future generations are targeted for additional assessment to document deposition requirements and essential curation actions.

Table 1. Components of the methodological approach

Participant Contacts	Instrument Objectives	Instrument Benefits
Pre-interview worksheet	- orient the participant to our investigation - capture description of research area and significant research questions - identify data types generated or collected	- initiates a relationship with participants - supplies background literature pertinent to understanding research context - provides a common ground for participants and investigators
Research Interview	- locate the science in a professional and academic context - capture details of data generation, gathering, and use - specify services\needs articulated for data use	- creates mutual understanding between the participant and investigator - facilitates participant awareness of relationship between their research problems and practices applicable to data repository development
Follow-up interview	- clarification of points addressed in Research Interview - further inquiry on lingering questions	- fill in gaps from Research Interview. - opportunity for investigators to realign interview questions
The Data Interview	- capture breadth and depth of data produced to address specific research question(s) - processes to generate, collect, and use the data	- clarifies 'what' the data are, how they are used, uncovers limitations for aggregation and use
Lab visit	- in situ observation of practices and tools employed - validation of earlier discussions of practice	- insight into the social and cultural interactions that shape the research setting - additional system requirements for DLs

The operation of this approach varies in duration across cases. Completion of the process may take as little as two to three months, while data collection for one case has lasted nearly a year.

4. EARLY OUTCOMES

The Pre-interview worksheet has become an essential component for initiating contact with new participants, and in conducting interviews across multiple disciplines. Based on feedback from participants, the process of completing the form grounds them in the scope of questions that follows in the Research Interview. Likewise, it alerts the investigator to vital, but often veiled domain-specific information necessary to support reuse.

Our most current data collection with 18 participants yielded promising results that will contribute to the development of curation and repository services. The analytical units are designed to illuminate how small science data are valued for re-use, which adds to our understanding of appraisal for collection development. Moreover, examination of three distinct scientific sub-disciplines revealed the significance of 'systems research', in which scientists are investigating questions that require integration and analysis of data from multiple disciplines. The 'systems' construct now serves as a basis for comparison of various research communities and their management of data to address complex scientific problems.

5. CONCLUSIONS

Research on the nature of practice-data-curation systems is essential during this period of emergent global data resources to facilitate design of data infrastructures. These new collections and services will be the foundation for innovative research and scholarship in the eResearch landscape. The robust units of analysis and evidence applied in this methodological approach offer a tested strategy for understanding the relationship between real-world work practices and the curation activities supporting data preservation and reuse.

ACKNOWLEDGMENTS

This research was supported by the Institute of Museum and Library Services (LG-06-07-0032-07) and National Science Foundation (OCI-0830976).

6. REFERENCES

[1] Cragin, M.H., Palmer, C.L., Chao, T.C. 2010. Relating Data Practices, Types, and Curation Functions: An Empirically Derived Framework. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, Oct. 22-27, 2010, Pittsburgh, PA.

[2] Karasti, H., & Baker, K. S. 2008. Digital Data Practices and the Long Term Ecological Research Program Growing Global. *Int. J. Dig. Curation*, 3(2).

[3] Zimmerman, A. S. 2008. New Knowledge from Old Data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, 33(5), 631-652.

[4] Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007). Drowning in data: digital library architecture to support scientific use of embedded sensor networks. In *Proceedings of the 7th ACM/IEEE-CS*, JCDL '07 (pp. 269–277). New York, NY: ACM. doi:10.1145/1255175.125522

[5] 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. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 333–340).

[6] 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. (http://www.rin.ac.uk/data-publication).