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

Introduction

Digital libraries (DLs) are adapting to include research materials generated upstream in the research-publication cycle. Managing these new content types – and creating services to support their use – spans the elements of DL development, revealing complicated technical requirements (e.g. exposing complex relationships amongst objects) and the need for additional human infrastructure. Building collections of scientific data also raises questions concerning data selection, policy development, collaboration, and outreach efforts and how to best align these with local, institutional initiatives for cyberinfrastructure, data-intensive research, and data stewardship. To facilitate data acquisition and purposeful user services, we require increased understanding of data-practice-curation service arrangements across small science research [1]. We present a flexible methodological approach crafted to generate units of evidence to analyze these relationships and facilitate cross-disciplinary comparisons.

Small Science and Digital Libraries

Small science disciplines are of particular relevance due to the prevalence of this mode of research in the academy, the anticipated magnitude of data production, and potential value for data re-use. Previous efforts to support small science data in DLs have generally focused on:

- curation issues or data management by scientists [2], [3]
- handling of research data and implications for digital libraries (e.g. long-term study at the Center for Embedded Network Sensors; http://research.cens.ucla.edu/) [4], [5].

Together, these studies have illuminated problems for DLs working to accommodate the variety and range of data types and practices 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.

We have found the sub-discipline to be a more optimal level of analysis, allowing critical focus on the research questions of interest to the domain and data types that produce the 'science' in a community–the social unit where data sharing practices and re-use can best be explored.

Method Overview

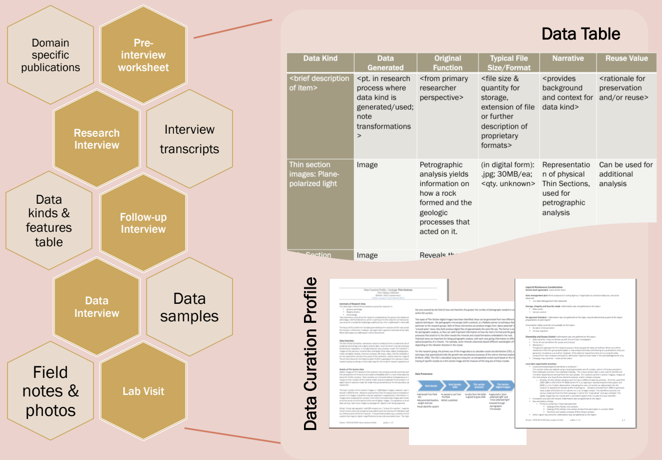
Our units of evidence are case studies; there may be considerable difference in the volume or density of each kind of data collected within or across a given case but they are a rigorous approach for investigating how and why contemporary phenomena occur in the ways they do. Critical to our cases, the combination of data collection approach, targeted participants, and complimentary data sources are all essential in producing dense, high-quality units of evidence. Featured in Table 1 is our current approach and we highlight the process below:

- The **Pre-Interview Worksheet** is used to orient participants to questions about their data and serve as a base of reference for the initial **Research Interview**.
- Verbal validation of the Worksheet's content during that initial session facilitates deep discussion on the participant's research. A subsequent **Follow-up Interview** is used to clarify or address gaps from the Research Interview; multiple follow-ups may occur.
- Data types are investigated and characterized in detail; those identified as having scholarly importance or re-use value are targeted for additional assessment in the **Data Interview**. This interview also addresses deposition requirements and essential curation actions.
- **Lab Visits** provide an opportunity to observe real-world practice and a lens into socio-cultural and technological interactions.

Table 1. Components of the methodological approach for data collection

Participant Contacts	Instrument Objectives	Instrument or Product (material) 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 • alerts the investigator to vital, but often masked domain-specific information necessary to support re-use
Research Interview	• locate the science in a professional and academic context including identification of data communities • capture details of data generation or gathering, processing, use, and sharing • specify services\needs articulated for data use	• creates mutual understanding between the participant and investigator • facilitates participant awareness of relationship between research problems and practices applicable to data repository development
The Data Interview	• capture breadth and depth of data produced to address specific research question(s) • ascertain the processes to generate, collect, and use the data	• clarifies 'what' the data are and how they are used • uncovers limitations for aggregation and re-use
Follow-up interview(s)	• clarification of points addressed in Research +/or Data Interview • further inquiry on lingering questions	• fills in gaps from Research +/or Data Interview. • offers opportunity for investigators to realign interview questions
Lab visit	• in situ observation of practices and tools employed • gather or photograph data samples • serves to validate earlier discussions of practice	• reveals 'workarounds' in local data gathering and use • presents insight into the social and cultural interactions that shape the research setting • provides additional system requirements for DLs

Building a 'unit of evidence': process and outputs



Data Kind	Data Generated	Original Function	Typical File Size/Format	Narrative	Reuse Value
<brief description of item>	<pt. in research process where data kind is generated/used; note transformations >	<from primary researcher perspective>	<file size & quantity for storage, extension of file or further description of proprietary formats>	<provides background and context for data kind>	<rationale for preservation and/or reuse>
Thin section images: Plane-polarized light	Image	Petrographic analysis yields information on how a rock formed and the geologic processes that acted on it.	(in digital form): .jpg; 30MB/ea; <qty. unknown>	Representation of physical Thin Sections, used for petrographic analysis	Can be used for additional analysis
[illegible] Section	Image	Reveals th			

The figure above details the materials generated during analysis of the components that form the 'unit of evidence.' From each component, refined qualitative outputs contribute to further analysis of researchers' data practices. These analyses are synthesized through the Data Curation Profile, a document that abstracts data characteristics and related curatorial requirements likely to be directly applicable to curation and repository services. The refined Data Table is incorporated into the Data Curation Profile, but may be used as a standalone product in our work.

Preliminary Outcomes

The current implementation of the method has been used to collect data from 18 participants; this has yielded promising results that are contributing to the development of curation and repository services at the Data Conservancy (http://dataconservancy.org). In addition to elucidating practices at the data community level [7], 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.

1) *Emergence of an analytical construct:* 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' model now serves as a basis for comparison of various research communities and their management of data to address scientific problems that require composite, or multiplex data sets.

2) *Role of Pre-interview worksheet*: essential in our initial contact with new participants, and has been the basis for prolonged engagement with each research group or lab; it also provides a consistent, structured instrument for conducting interviews across multiple disciplines which facilitates cross-disciplinary comparisons.

3) *Iterative interactions with participants*: necessary for explicating information at the right levels of granularity to support re-use; builds upon the information collected from the pre-interview worksheet which alerts us to vital, and often veiled domain-specific details.

Implications for Digital Libraries

Research on the nature of relationships among sub-disciplinary practice, data types and curation activities is essential during this period of emergent data resources that will be part of a growing global infrastructure. New and interconnected collections will be the foundation for innovative science and scholarship in 21st century research and learning. 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 re-use.

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. *International Journal of Digital Curation*, 3(2). [3] Zimmerman, A. S. 2008. New Knowledge from Old Data: The role of standards in the sharing and re-use 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, Joint Conference on Digital Libraries* (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 re-use: 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). [7] Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. 2010. Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926), 4023-4038.

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