Relating Data Practices, Types, and Curation Functions: An Empirically Derived Framework

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

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

ABSTRACT

We present a general conceptual framework that maps relationships and dependencies among scientific data practices, types of data produced and used, and associated curation activities. As part of the Data Conservancy initiative, the framework is being elaborated through empirical studies of data practices in the earth sciences and life science and validated against use cases as curatorial services are developed around data being prepared for ingest into the repository. The framework can be applied more broadly for identifying and representing curation requirements and to support description and assessment of existing or planned curation infrastructure and services. It will support full accounts of the data products and workflows required to maintain the coherence and context of complex data collections.

KEYWORDS

Data curation, scientific data, scientific work practice

INTRODUCTION

The relationships among data characteristics, scientific work practices, and curation activities are not well understood, yet they are central to how data infrastructures develop and services operate. Moreover, across fields, and even within research specialties, data practices are far from uniform (RIN, 2008). Variations may be rooted in disciplinary conventions or more local structures and processes, such as collaborations that share scientific questions and resources (Palmer & Cragin, 2008). They are manifest in how methods are applied (Pritchard, Anand, & Carver, 2005), what units are measured and how (Borgman, Wallace, & Enyedy, 2007), and the nuance of analysis and interpretation. Lifecycles of different types of data also vary in interesting ways, and have tremendous implications for the many goals that have been set for how data will be shared and re-used. How we shape our collective data structures will, ultimately, reshape how data assets contribute to science.

There is a growing base of evidence from which we can begin to specify the range and combination of data practices and data types, which will configure and constrain data management, and use of publically shared collections (Cragin, Palmer, Carlson, & Witt, 2010). These investigations, and our current work with the Data Conservancy to develop a cross-disciplinary data curation strategy, are the basis for the framework we have developed for mapping relationships and dependencies among arrays of scientific data practices, types of data produced and used, and associated curation activities. The framework can be applied to identify and represent curation requirements and to support description and assessment of existing or planned curation infrastructure and services.

METHOD

The framework is grounded in our on-going empirical research on disciplinary differences in data practices, with initial terminology generated from analyses of the literature and results from the Purdue-Illinois Data Curation Profiles Project (Witt, Carlson, Brandt, & Cragin, 2009). Systematic integration of categories and terms from the literature, such as those developed in Cole's (2005) work on GIS, continue to be evaluated and applied. However, evidence for testing and validating the framework is now primarily derived from our studies of data practices in the earth sciences and life sciences, as part of the Data Conservancy project (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0830976).

We are employing a feedback loop design for testing and external review. On a quarterly basis, candidate terms from the literature are analyzed for fit, and either added to the structure, relegated to the "orphan" category, or discarded. Terms uncovered during coding and analysis of qualitative data collected in our studies are regularly tested and added. We consult with the Data Conservancy's Data Concepts team for critical semantic assessment and validation. Finally, domain and data scientists will review and make recommendations on documented and hypothesized graphic expressions of practice-data-curation relations on key releases of the framework.

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

The framework is maintained in XMind (http://www.xmind.net/), a mapping tool that supports rapid capture and organization of words and phrases, annotation capabilities, and flexible and diverse visualization options.

DATA CURATION FRAMEWORK

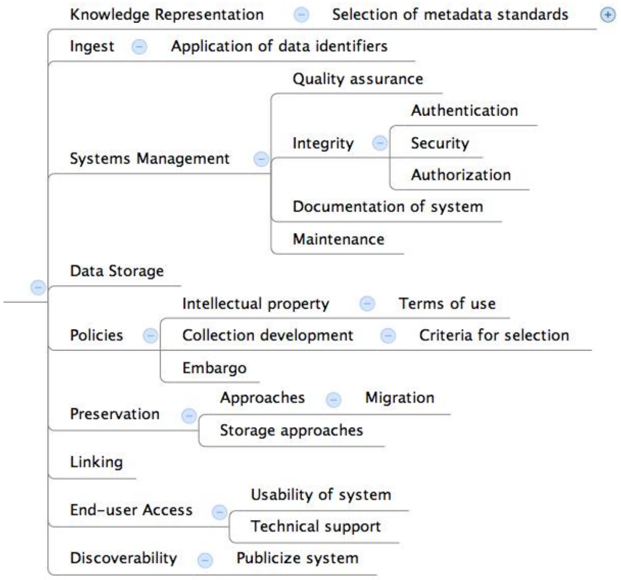

Figure 1. Segment of the Curation Category

The framework consists of three major categories: Data, Data Practices, and Curation. The Data category contains types and characteristics of data; Data Practices contains research processes and activities related to scientists' work with data; and Curation contains services, functions and associated activities. Figure 1 is an extracted segment of the Curation category and illustrates primary curation functions on the left, with nested sub-functions appearing to the right.

The framework is currently being tested on Data Conservancy use cases for curatorial service. For example, Figure 2 maps the curation functions for a subset of research products from an extensive set of volcanology data being processed for ingest. It includes physical rock samples, sliced "thin sections" of the rocks, digital microscopy images of the sections, as well field notes, digital photographs, and a related journal article.

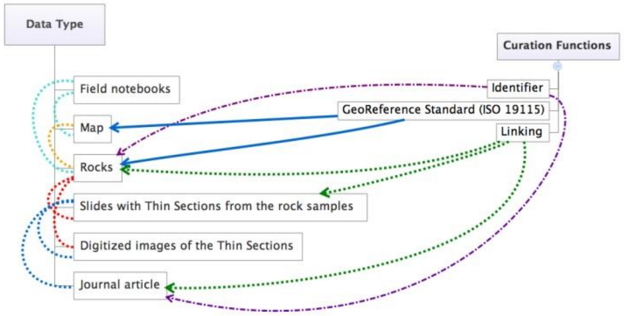
Data Type
Curation Functions
Field notebooks
Identifier
GeoReference Standard (ISO 19115)
Map
Linking
Rocks
Slides with Thin Sections from the rock samples
Digitized images of the Thin Sections
Journal article

Figure 2. Mapping curation functions to data objects

Due to space constraints, this example does not capture the 35mm slides and digital photos, chemical analyses, 3D maps, video, and the metadata records for the collection, which will be required to maintain the value of the collection, and to meet the range of user requirements.

CONCLUSION

The next iteration of the framework, slated for the coming months, will address "orphan" terms through systematic testing for fit and function in the framework. Ongoing analysis on differences across disciplines served by the Data Conservancy will be used to elaborate the Data Practices and Data categories, and data sets submitted for ingest will be used to validate the Curation category and to identify and confirm relationships among all three categories. As new curation systems are developed, the refined framework can be applied more broadly to support assessments of curation services and full accounts of deposited data sets and the associated data products required to maintain the coherence and context of complex data collections.

ACKNOWLEDGMENTS

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

REFERENCES

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*(2-1), 17-30.

Cole, F. (2005). The discourse of data: exploring data-related vocabularies in geographic information systems description. *Journal of information science*, *31*(1), 44.

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*, *368*(1926), 4023-4038.

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

Pritchard, S.M., Anand, S., & Carver, L. (2005). Informatics and knowledge management for faculty research data. *EDUCAUSE Research Bulletin* 2005. Retrieved July 16, 2010, from http://net.educause.edu/ir/library/pdf/ERB0502.pdf.

Research Information Network. (2008). To share or not to share: Publication and quality assurance of research data outputs. Retrieved July 16, 2010, from http://www.rin.ac.uk/data-publication.

Witt, M., Carlson, J.R., Cragin, M.H., & Brandt, D.S. (2009). Constructing Data Curation Profiles International *Journal of Digital Curation*, *4*(3). Retrieved July 16, 2010, from http://www.ijdc.net/index.php/ijdc/article/view/137.