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

Melissa H. Cragin ◆ Tiffany C. Chao ◆ Carole L. Palmer ◆ Center for Informatics Research in Science and Scholarship

Graduate School of Library and Information Science

INTRODUCTION

The relationships among data characteristics, scientific work practices, and curation activities are central to how data infrastructures develop and services operate, yet they are not well understood. 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). These variations 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. How we shape our collective data structures will, ultimately, reshape how data assets contribute to science. We present a conceptual framework being developed for data curation; its application here illustrates potential relationships and dependencies among scientific data practices, types of data produced and used, and associated curation activities.

The framework is being elaborated through empirical studies of data practices in the earth and life sciences, and is a part of the information science research on the Data Conservancy. The Data Conservancy, led by the Sheridan Libraries at Johns Hopkins University, is one of two awards through the U.S. National Science Foundation's DataNet program, an initiative to develop long-term preservation and access services for scientific and engineering data (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0830976). This effort draws upon the expertise of infrastructure designers, domain scientists, and information professionals to form sustainable organizations intended to advance complex scientific endeavors and promotes data use.

METHODS

Grounded in on-going empirical research on disciplinary differences in data practices, the framework is a work in progress and intended to adapt to the evolving curation environment. On a quarterly basis, candidate terms from the literature are analyzed for fit. We employ a feedback loop for testing and external review, and validate against actual curatorial processes being applied to initial acquisitions into the Data Conservancy.

Initial terminology was generated from:
- The Purdue-Illinois Data Curation Profiles Project (Witt, Carlson, Brandt, & Cragin, 2009)
- Systematic integration of categories and terms from the literature, (e.g. Cole's (2005) work on describing data-related concepts associated with geographic information systems)
- Terms uncovered during coding and analysis of qualitative data collected in our studies

Development and internal testing:
- Internal review based on analysis of qualitative data collection
- Consultation with the Data Conservancy's Data Concepts team for critical semantic assessment and validation

External review and validation:
- Review and comments by the Data Conservancy's distributed research and development team
- Domain and data scientists will review and make recommendations on key releases of the updated graphical expressions of practice-data-curation relations

NEXT STEPS

The next iteration of the framework will focus on extending the Data component and addressing compiled "orphan" terms through systematic testing for fit and function. Ongoing analysis on practice-based differences across disciplines served by the Data Conservancy will also be used to build the Research 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. The framework is intended for broad application; as new curation systems are developed, the refined framework can be applied to describe full accounts of deposited data sets, associated data products, and workflows required to maintain the coherence and context of complex data collections. In addition to providing a foundation for a taxonomic view of curation, the framework is intended to be a flexible tool for identifying and representing curation requirements, facilitate description and assessment of existing or planned curation infrastructure and services.

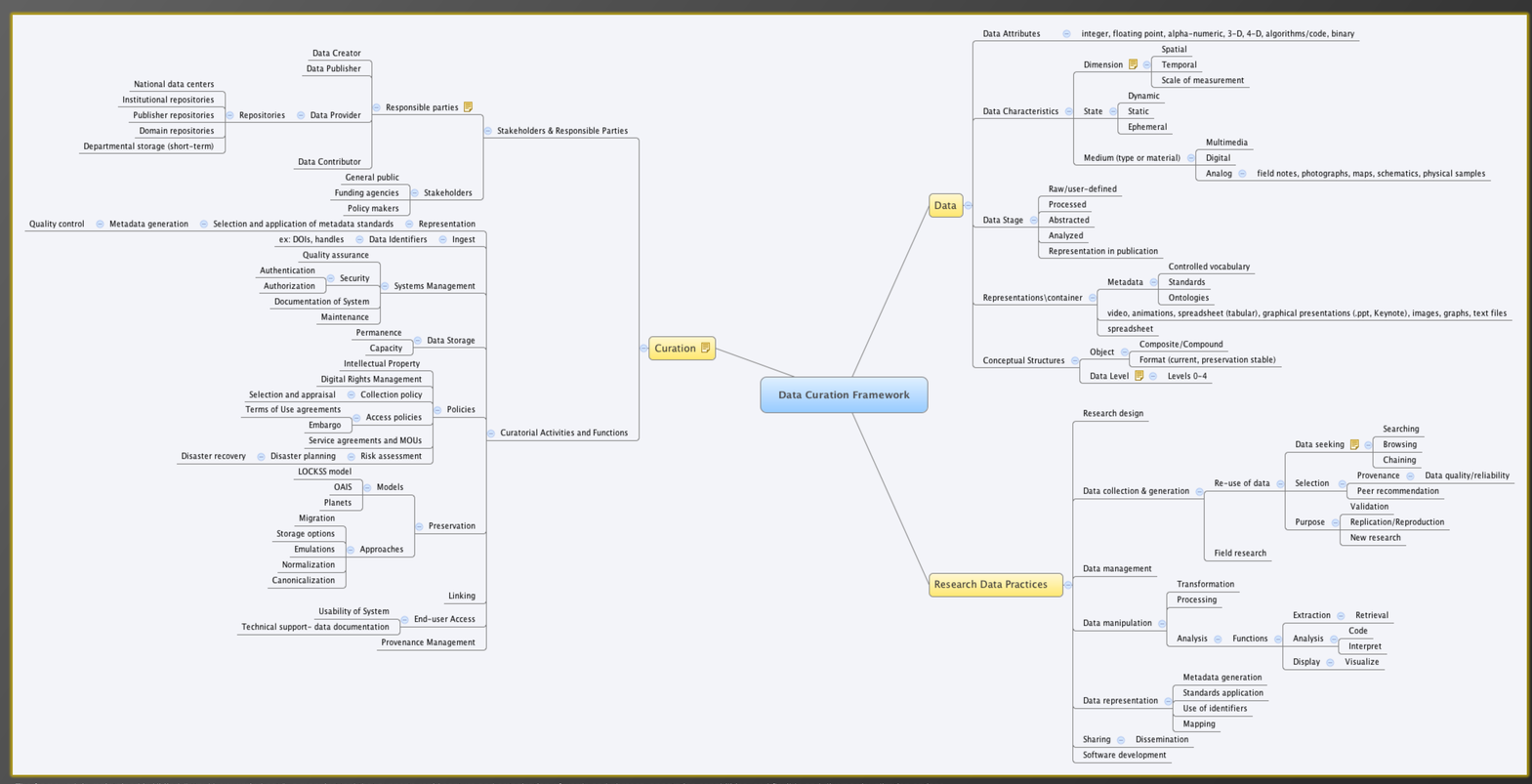
CURATION FRAMEWORK

FRAMEWORK COMPONENTS

Data: types and characteristics of data

Curation: services, functions and associated activities

Research Data Practices: processes and activities related to scientists' work with data

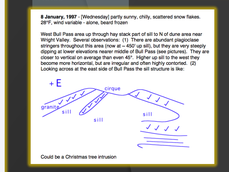


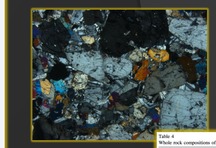
The framework is maintained in XMind (http://www.xmind.net/), a mapping tool that supports rapid capture and organization of words and phrases, annotation capabilities, and flexible and diverse visualization options.

Geology in the McMurdo Dry Valleys: Data types in this special collection

(Right) Rock sample

(Bottom) digitally transcribed Field Notes

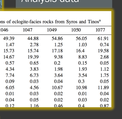




(Top) Thin Section slide

(Left) Polarized image of Thin Section

(Bottom) Chemical Analysis data

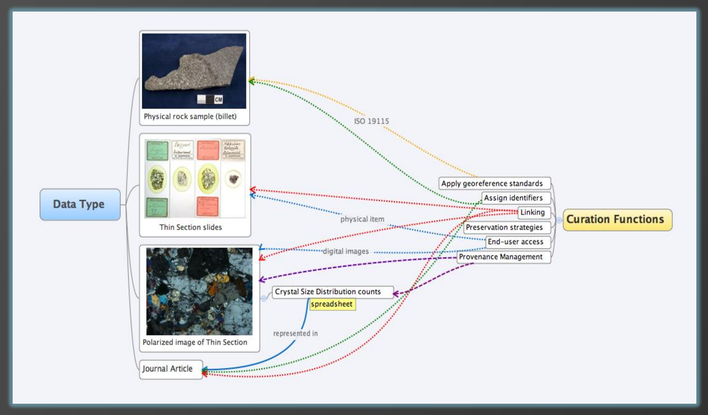


(Top) Digitized 35mm slide image

(Right) Metadata record in XML (for field photo)

```
<?xml version="1.0" encoding="UTF-8"?>
<McMurdoDryValleys>
<FieldPhotos>
<FieldPhoto>
<fileName>McDV_FieldPhoto_1993_00002.tif
</fileName>
<fieldSeason>1993</fieldSeason>
<sampleProfile>Solitary Rocks</sampleProfile>
<dcTitle>McDV_FieldPhoto_1993_00002</dcTitle>

<dcSubject>Geology</dcSubject>
<dcSubject>Petrology</dcSubject>
```



Mini-case study: Thin section data are integral to answering fundamental research questions about the nature of geologic structures. A key research area for this lab is the study of magmatic flow dynamics as observed through crystal size distribution. The presence of crystal structures in rock samples collected from different elevations of a vertical "rock profile" show the paths of magma movement within the volcanic system, an indicator of their developmental process.

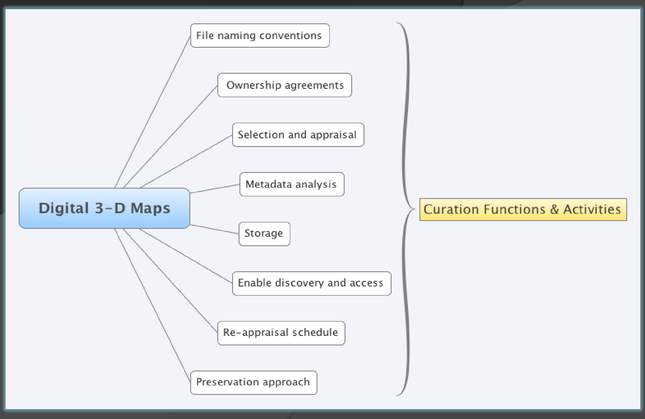
APPLICATION OF THE FRAMEWORK

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 the use of publically shared collections (Cragin, Palmer, Carlson, & Witt, 2010). As noted in the introduction, one contribution of the framework is to support mapping relationships and dependencies among types of data and associated curation activities. Here we have introduced a large special collection of data associated with geological research conducted in Antarctica; the mini case studies that accompany the two images beside this text provide a brief context for the data types represented.

LEFT: The categories of curation processes identified for the thin section data are those currently known to apply to these data sets and related publications. Some curation activity categories, such as "systems management," apply to the entire collection in a repository or archive; these are not included in this illustration. Others, such as "storage," will be assessed as part of the preservation strategy designed for these data types once we have determined the requirements for particular item sets.

RIGHT: The case on digital 3-D maps describes several higher-level curation processes, many of which must occur before data ingest; these present a different set of priorities for managing acquisition. For example, establishing ownership will be a significant step for these data products, and this is necessary for generating depositor agreements.

Mini-case study: A collaborator from a research center at another university is creating digital maps of the geographic features and surface areas for the field sites. Map materials will be stored at two separate locations; data "behind" the 3-D maps will be part of the whole collection in the Data Conservancy repository.



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; Images on Poster: 'Thin Section slides' http://www.microscopy-uk.org.uk/mag/artsep09/bs-coppock.html, 'Chemical Analysis' from Bröcker, M., & Enders, M. (2001). Unusual bulk-rock compositions in eclogite-facies rocks from Syros and Tinos (Cyclades, Greece): implications for U-Pb zircon geochronology. *Chemical Geology, 175*(3-4), 581–603, All other images used with permission from Bruce Marsh, Keith Kaneda, and Mark Cyzyk.
Acknowledgements: This research was supported by the Institute of Museum and Library Services (LG-06-07-0032-07) and National Science Foundation (OCI0830976).

NSF

Institute of Museum and Library Services