Thomson Reuters Doctoral Dissertation Proposal Scholarship
Title: METHODS METADATA: CURATING SCIENTIFIC RESEARCH DATA FOR REUSE

Tiffany C. Chao
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Research Proposal
(a) Description for the research, including significance and methodology
(b) Schedule of completion
(c) Budget and budget justification
(d) Other support (including scholarships, assistantships, and employment)
(e) Name of the dissertation advisor endorsing the proposal

Additional submission materials
1. A cover letter from the dissertation advisor endorsing the proposal
2. An up-to-date curriculum vitae
A. RESEARCH PROPOSAL

Metadata plays a key role in the curation and management of scientific research data for multidisciplinary sharing and reuse. However, the provision of metadata by the data producer, who best understands how and why data are gathered, is not always a common practice or cultural norm (Karasti, Baker, & Halkola, 2006). This “friction” of metadata generation can ultimately inhibit the long-term access and use of data across disciplines (Edwards et al., 2011). Sources of friction that influence minimal metadata documentation by researchers have been identified in studies of scientific data practice (i.e. Mayernik, Batcheller, & Borgman, 2011; Tenopir, et al., 2011) but there is a need to understand what approaches can be taken by information professionals and scientists to overcome this friction and ensure that research data are adequately described to enable future use.

The primary aim of my research is to inform metadata generation processes for data curation services by addressing the role of research “methods” description and its significance for data reuse. Metadata that describes methods is essential for conveying the level of professionalism and expertise of the data producer within his or her scientific community and are used by scientists to assess the quality of data for reuse (Faniel & Jacobsen, 2010). More generally, explanations of methods offer insight into the practices of research communities and their standards for generating new knowledge (Zimmerman, 2007). I encapsulate this critical information component into the term methods metadata, or the type of information needed for basic comprehension of how data were produced in the scientific research context.

This qualitative study will examine the data production practices of three Earth Science subdisciplines representative of long tail science research: soil ecology, volcanology, and stratigraphy. Long tail science research data are a particular concern for curation services since scientists in these fields tend to produce highly heterogeneous data while having few repository options and limited resources for sustained data management and maintenance (Heidorn, 2008).
Two sources of evidence will be used to develop the study: 1) semi-structured interviews covering practices of data production in the three subdisciplines, and 2) content analysis of journal articles published in these subdiscipline areas.

Research questions

I will investigate the following set of interrelated questions:

A) What methods metadata related to how data are gathered, processed, and analyzed for research, can be derived from:
   - qualitative interview techniques?
   - content analysis of research publications?

B) How does methods metadata differ across subdisciplines in Earth Science?

C) How do data sharing expectations and practices differ across subdisciplines in Earth Science and what are the implications for generating methods metadata?

These research questions are not aimed at producing an exhaustive account of methods metadata but rather at understanding what sources are best for developing metadata for research data, as framed in the first question, with the emphasis on the research methods scientists employ. From the literature, data producers cite journal articles as an information resource to understand the study context and processes implemented in the generation of data (Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011; Parsons, Duerr, & Minster, 2010). Journal articles are therefore chosen as a potential source for methods metadata in this study. In formulating the second research question, the level of analysis at the subdiscipline is a change from most previous data practice studies which have tended toward comparative analysis at the broader, discipline level. Examining methods metadata differences exhibited at the subdiscipline level can contribute a more refined perspective on the nuances of subdiscipline practices in relation to data. As methods description is an important component of data sharing and reuse, the third research question draws from the first
two to look at practices and expectations for data sharing in relation to methods metadata generation. The literature on sharing data, particularly in the long tail sciences, provide insight on what, when, how, and with whom data may be shared (i.e. Borgman, Wallis, & Enyedy, 2007; Cragin, Palmer, Carlson, & Witt, 2010; Zimmerman, 2008) and will help to frame what methods metadata can be identified within the research process and the conditions involved with public release.

RESEARCH DESIGN

This study utilizes a case study approach to understand how methods implemented in producing data can be identified and represented in metadata for scientific research datasets. The three Earth Science cases will be developed at the subdiscipline level. Analysis at this level has been successful in distinguishing data practice differences, 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” (Cragin, Chao, & Palmer, 2011, p.441). Soil ecology, volcanology, and stratigraphy, in particular, are useful for comparison of practices since scientists across all three apply similar empirical approaches of collecting observational data from field sites for analysis and experimentation. There are (3) phases in this study design (see Figure 1 for overview) with each phase discussed in detail in the following section.
1) **Phase 1: Methods Metadata Identification and Intra-case analysis**

Phase 1 focuses on identifying how well methods metadata can be determined from two different sources, qualitative interviews with scientists and journal publications authored by scientists.

- **Semi-structured Interview technique**

  The interview data for the three subdisciplines have been collected as part of the Data Conservancy Data Practices research. The Data Conservancy ([http://dataconservancy.org](http://dataconservancy.org)) as a whole performed research and developed data infrastructure and services to foster preservation, sharing, and discovery of research data across institutions and disciplines.

  Products from the interviewing process include transcripts of the qualitative interview sessions, written responses to questions from a Pre-interview worksheet, personal observation notes, and photographs and artifacts of the research process from participants.
A total of (16) interviews and related interview products from (10) participants will be analyzed in this research. The interviews are valuable as models of the kind of narrative that can be generated through discussion of practices and data with scientists. In particular, the interviews show the detailed description of how data were generated and the processes involved, which will provide a foundation for analysis in this research.

- **Qualitative content analysis**

A pilot study was conducted to understand the viability of using journal publications to determine methods metadata for data. Based on the successful pilot, a sample of peer-reviewed journal articles authored by the interview participants will be analyzed, supplemented by analysis of additional articles from respective discipline journals. These additional publications beyond those of the participants will be added for each case to not only validate findings from participant publications but also how the larger research community discusses data production and sharing, and to what extent variations exist within a subdiscipline community. Each subdiscipline case has at least (15) articles published by interviews participants and will be initially supplemented with (8) articles from respective journals until saturation is reached, meaning that no new findings are revealed. A purposive sampling procedure will be implemented to develop the sample of additional journal articles for content analysis. The qualitative approach for content analysis will be used to enable the range of meanings of the data production process to be revealed from the articles, and is more conducive for this research than a quantitative approach, where content is viewed objectively to determine frequencies of instances in the text (White & Marsh, 2006).
• **Intra-case analysis**

This systematic analysis brings together evidence from direct engagement with scientists through interviewing and the relatively indirect approach of examining scientists’ research publications to assess the strengths of the different sources for documenting the methods used by scientists for producing data within each subdiscipline. The analytical framework builds on concepts and themes from the research questions (i.e. “gathering”, “processing”, “analyzing”, “sharing”) and will be applied to the collected data sources. A product of this analysis is the compilation of a Methods Metadata Profile for each subdiscipline (described below).

2) **Phase 2: Methods Metadata Verification**

The aim of Phase 2 is to corroborate the synthesized findings of the Methods Metadata Profiles from Phase 1 with existing metadata records from data repositories.

• **Methods Metadata Profile (Profile)**

The Profile systematically describes the research methods and data production processes for a specific data parameter or variable based on findings synthesized from the interviews and journal papers. For instance, a common parameter observed within soil ecology is “organic carbon,” therefore an Organic Carbon profile will be created to detail the techniques implemented in gathering, processing, and analyzing this data parameter using findings from the different data sources. At least one Profile will be developed for each subdiscipline with the potential for additional ones based on the intra-case analysis.

• **Corroboration with data repository metadata records**

Existing metadata records from established data repositories will be used to determine how robust the derived methods metadata are and what gaps may still exist. I have identified
comparable data repositories for each subdiscipline based on the repository's data collection scope. The sampling of metadata records for analysis will draw on the data parameter or variable described in the Profile. An initial sample of (8) records will be selected with additional records included until saturation is reached. The structured form of the metadata records will provide a baseline for assessing the elements of methods metadata that fit into current schemes and for determining what elements are missing in existing schemes and records related to the data production process and sharing.

3) *Phase 3: Cross-case comparison and Inter-case analysis*

The case comparison of Phase 3 will focus on discerning similarities and differences in methods metadata as well as what research practices for data production and sharing are described by scientists across the subdisciplines. This cross-case analysis incorporates the data repository metadata record assessment with the inter-case analysis for comparison of the findings from the subdiscipline cases within the broader context of the Earth Sciences. Based on this analysis, I will also consider how the Profiles and sources for methods identification can extend to other areas of field-based research in the Earth Sciences.

Across the three phases, this research will produce a case file for each subdiscipline containing a Methods Metadata Profile, comparison with data repository metadata records based on schema dimensions, and observations on data sharing. In addition, there will also be a comparative analysis across the three subdiscipline case files generated from this study. The strength of the case study approach is the ability to analyze and synthesize evidence from multiple sources for each subdiscipline, which in this research, builds on a complex qualitative interviewing process and content analysis of journal publications. Alternative approaches such as surveys or ethnography, both of which have also been used for studies of scientific practices (i.e. Tenopir et al., 2011,
Borgman, Wallis, & Enyedy, 2007, respectively), would not be as effective. Surveys constrain the level of detail that can be generated on practices at the subdiscipline level, and because of the extensive time investment needed for ethnography, it is not a viable approach for practical application in data curation services.

**RESEARCH CONTRIBUTIONS**

*Intellectual contribution*

My research brings attention to the significance of methods description for data reuse and addresses how scientific practices, in particular data production processes, contribute to metadata development. The identification and description of methods metadata can be a starting point in overcoming the friction of metadata generation experienced by scientists through alignment with the practices of gathering, processing, and analyzing data in scientific research. It is anticipated that the methods used by scientists will differ, and this study can contribute an understanding of what information components are crucial for methods metadata description for research datasets.

Past studies of data practices across multiple disciplines (i.e. Swan & Brown, 2008; Lyon, Rusbridge, Neilson, & Whyte, 2010) have conducted analysis within broad disciplinary groupings that can mask important distinctions within research or subject areas. This is especially problematic for the long-tail sciences where research processes and techniques used tend to be localized and unique to the small research team (Wallis, Rolando, & Borgman, 2013). By investigating methods and data production practices at the subdiscipline level, the potential range of similarities and differences will be revealed to not only enhance how metadata schemes are used but also how they can be improved to better support access and reuse of research data across disciplines (Willis, Greenberg, & White, 2012). Methodologically, my proposed study can also further understanding on the benefits and limitations of conducting analysis via interviews and content analysis of journal articles at the subdisciplinary level.
**Practical applications**

A more effective approach for generating metadata for research data is important in developing tools for facilitating data deposit. The combined technique of direct interviewing of scientists in addition to the unobtrusive approach of identifying metadata through journal articles proposed in this study could potentially be adopted by data curation professionals to secure description about methods for datasets deposited by scientists. Currently, “data interviews” are used by research library data services to gather information about practices and expectations surrounding data production, long-term access, and other information needed to describe data (Witt, Carlson, Brandt, & Cragin, 2009; Carlson, 2012), but they are time consuming to coordinate and conduct with busy scientists. The examination of methods description in journal articles, in particular, can contribute to more rigorous automated approaches for metadata extraction of datasets (Greenberg, White, Carrier, & Scherle, 2009). This is especially valuable for enhancing metadata generation in research fields such as in the long-tail sciences that do not use instrumentation that automatically documents the processes undertaken when collecting and processing data (to contrast see Goble et al., 2008 for automated documentation in bioinformatics).

This research is particularly timely as demand for curation services is expected to escalate due to new expectations for public access to digital data produced by federally funded research, as outlined in the February 2013 Office of Science and Technology Policy memo (Holdren, 2013). Research techniques such as meta-analysis are also gaining traction in the long tail science research areas, including Earth Science (i.e. Koricheva, Gurevitch, & Mengersen, 2013), which increases the need for discovery and access of data from multiple studies with common parameters. Moreover, metadata about the research methods used to generate the data for each study will be critical for determining the applicability and compatibility of data for integration and re-analysis. As stated by e-Science pioneer Jim Gray and colleagues: “(d)ata is incomprehensible and hence useless unless
there is a detailed and clear description of how and when it was gathered, and how the derived data was produced” (2002, p. 4).

Summary

Reuse of data in the Earth Sciences requires methods metadata that documents the data production processes and methods used by scientists in data collection and processing. The proposed study investigates how to generate methods metadata through qualitative semi-structured interviews and content analysis of journal publications. It will also identify and describe the similarities and differences in how methods metadata needs to be described across subdisciplines and provide a better understanding of expectations and current practices in data sharing. The results of this study will have implications for professionals working in libraries, repositories, and archives who will be responsible for or expected to aid in the curation of research data. With scientific research data anticipated to increase in quantity and heterogeneity, there is considerable need for deeper understanding of data practices and needs in different research communities, as well as techniques for optimizing data description as demand for curation services.
B. SCHEDULE OF COMPLETION
Dissertation proposal defended and approved March 2014, expected completion: March 2015.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data analysis and drafting chapters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revising Chapters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defend dissertation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. BUDGET JUSTIFICATION
The following areas of support are requested in order to conduct the proposed research:

- **Atlast.ti 7.0 qualitative analysis software, student license: $99.00**
  - [http://www.atlasti.com/licenses.html](http://www.atlasti.com/licenses.html)
- **Dell Latitude E6440 14" laptop: $1036.00**
  - University of Illinois Union Tech Zone: [http://union.illinois.edu/~media/Sites/Union/Files/Tech%20Zone/TZ_Dell.ashx](http://union.illinois.edu/~media/Sites/Union/Files/Tech%20Zone/TZ_Dell.ashx)

Total amount requested: $1135

I am conducting extensive qualitative analysis on a diverse array and large quantity of documents from multiple sources. Access to qualitative analysis software such as Atlas.ti (with which I have prior training) will support me in organizing and managing these documents. I also leverage the useful functionalities of the software to identify, code, and analyze these data sources. Due to Microsoft’s end of support for Windows XP, my current laptop is no longer secure to use for this data analysis. The availability of computing services with Atlas.ti or any other qualitative analysis software on the university campus is minimal where the limited hours and file storage configurations are not conducive to the iterative process of qualitative coding needed for this study. Therefore, in addition to the qualitative analysis software, funding for a computing device is also requested. The laptop selected adheres to the recommended technical specifications to run Atlas.ti and is available for purchase through the University.

D. OTHER SUPPORT (including scholarships, assistantships, and employment)
I have applied for the Graduate School of Library and Information Science Research Fellowship.

E. ENDORSEMENT Name of the dissertation advisor endorsing the proposal.
Professor Carole L. Palmer
RESEARCH PROPOSAL REFERENCES


Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One, 8(7), e67332. doi:10.1371/journal.pone.0067332


