IJDC | *Peer-Reviewed Paper*

Mapping Methods Metadata for Research Data

Tiffany C. Chao
Center for Informatics Research in Science and Scholarship,
Graduate School of Library and Information Science
University of Illinois Urbana-Champaign

Abstract

Understanding the methods and processes implemented by data producers to generate research data is essential for fostering data reuse. Yet, producing the metadata that describes these methods remains a time-intensive activity that data producers do not readily undertake. In particular, researchers in the long tail of science often lack the financial support or tools for metadata generation, thereby limiting future access and reuse of data produced. The present study investigates research journal publications as a potential source for identifying descriptive metadata about methods for research data. Initial results indicate that journal articles provide rich descriptive content that can be sufficiently mapped to existing metadata standards with methods-related elements, resulting in a mapping of the data production process for a study. This research has implications for enhancing the generation of robust metadata to support the curation of research data for new inquiry and innovation.

Received 16 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to Tiffany Chao, 501 E. Daniel, Champaign IL 61820. Email: tchao@illinois.edu

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The generation of metadata by scientists and researchers is recognized as a time and resource-intensive activity yet vital for the curation of research data. Funders have long been urging their grantees to collect and maintain metadata, but this call has been met with minimal adherence or completely disregarded. As explained by Edwards, Jackson, Bowker and Knobel (2007) this "metadata conundrum represents a classic mismatch of incentives: while of clear value to the larger community, metadata offers little to nothing to those tasked with producing it and may prove costly and time intensive to boot." In the past decade, one emerging form of metadata for research datasets is the *data article* or *data paper* published within a *data journal*. The focus of these data articles is on the description of a dataset where details related to data collection, processing, software and file formats are explained and emphasized, rather than reporting formal analysis and findings as presented in conventional journals. More importantly for researchers, publishing the data article and associated dataset results in a citable work not only acknowledges researchers' efforts in producing rich metadata but is also an incentive resonant in the scholarly community similar to the traditional journal research publication. The establishment of data journals and publication of data articles are now available in different areas of research[1] but with nascent adoption within their respective research communities.

As noted, a key component in a data article is the description of those processes used by researchers in generating a dataset. It is anticipated that such descriptive information is detailed in the methods section of a scholarly journal article but few studies have explored the content of this section for metadata purposes. With a focus on research methods, this study examines scholarly journal articles as a potential source for contributing descriptive metadata for datasets, or *methods metadata*, and content for data articles.

Background

Metadata plays a critical 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 and Halkola, 2006). This "friction" of generating metadata can ultimately inhibit the long-term access and use of data across disciplines (Edwards, Mayernik, Batcheller, Bowker and Borgman, 2011). Sources of friction that influence minimal metadata documentation by researchers have been identified in studies of scientific data practice (i.e. Mayernik, Batcheller and 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.

[1] See Ubiquity Press for discipline-specific data journals (http://www.ubiquitypress.com/journals); Earth Systems Sciences Data Journal (http://www.earth-system-science-data.net/home.html); and the International Journal of Robotics Research, "Data Papers Submission" (http://www.uk.sagepub.com/repository/binaries/pdf/Data_Papers_Submissions.pdf).

 doi:10.2218/ijdc.v10i1.347

Metadata Generation Approaches and Long Tail Science

The use of semi-automated approaches to generate metadata can be an effective technique that produces an initial foundation of descriptive metadata for research data without being a large burden on data producers. Automated approaches for harvesting and extracting bibliographic metadata have been applied to text (e.g. Kovacevic, Ivanovic, Milosavljevic, Konjovic and Surla, 2011) but still necessitate a degree of human intervention to ensure quality documentation (Greenberg, 2004). With the growth of computationally-intensive research, technological advancements in tools and systems to record scientific workflows provide a semi-automated alternative to manually documenting the concise step-by-step description of the scientific procedure and protocols enacted (McPhillips, Bowers, Zinn and Ludäscher, 2009) while also creating a record of provenance for data generated during the research process. Although the use of these documenting technologies to ameliorate the process of generating metadata is not widespread across scientific domains (Davis et al., 2012), their adaptation and application would be beneficial to areas such as long tail science. For instance, in field-based sciences such as ecology, rapid environmental changes demand immediate decisions that may alter research methods and protocols in order to properly capture a particular phenomenon (Mayernik, Wallis, Pepe and Borgman, 2008; Karasti, Baker and Halkola, 2006). Presenting an ecological research data set for public consumption should include discussion of missing values, modifications during procedure implementation, or natural disturbances that occur in the ecological environment at the time of collection (Karasti and Baker, 2008). These dynamic changes are often manually recorded and if not well documented may impact the overall integrity of the dataset and limit future reuse.

The long tail data produced from these scientific fields are highly heterogeneous. These data remain a challenge to curate not only due to their diversity but also the limited resources for sustained data management and maintenance, which includes producing metadata (Heidorn, 2008). Adopting a metadata standard can alleviate some of the variation that deters integration and reuse, especially in the realm of long tail science research where data are heterogeneous and there is a tendency to use localized conventions developed with the small lab group environment (Wallis, Rolando and Borgman, 2013). On the whole, long tail data are representative of the vast majority of scientific data and are a rich source for new discoveries and innovations (Ferguson, Nielson, Cragin, Bandrowski and Martone, 2014) which warrant curation consideration.

The Role of "Methods" in Data Reuse

The importance of information on methods and those processes of data production is a prominent theme in studies of research data reuse. The research methods employed can convey the level of professionalism and expertise of the data producer within his or her scientific community (Faniel and Jacobsen, 2010). Scientists in the environmental sciences determine whether to trust the quality of environmental datasets by first evaluating the scientific processes that were employed in creating the data and then assessing the personal and professional reputation of the individual, group, or organization that produced the dataset in order to counteract any biases that the chosen methods for generating data may have (Van House, Butler and Schiff, 1998). Similarly, Zimmerman's (2008) study of ecological research practices uncovered that the documentation on methodologies was instrumental in appraising trust in data and guiding selection of data for reuse. The provision of methods documentation is also

reflective of research best practice in scientific communities. For instance, methods and protocol information for genomics research is often made available through project websites to complement a dataset deposited in GenBank[2], and assessment of methods deployed to produce data is common practice in the peer-review process for publications in astronomy research (Swan and Brown, 2008).

Journal Articles as a Methods Metadata Source

Scientific journal publications remain a primary mechanism of communication among scientists and scholars. 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 and Callaghan, 2011; Parsons, Duerr and Minster, 2010). For research data, the journal publications of data producers are one of the dominant modes for communicating scholarly information within scientific communities and could be a rich source of content for generating metadata for datasets. Moreover, with the rise of the number of open access journals, published articles are more readily available than in the past. The representations of data (i.e. figures, tables, charts, etc.) and narrative content embedded in journal articles, particularly descriptions of methods implemented in the research, can play a vital role for researchers in assessing data for reuse (Faniel and Jacobsen, 2010). Data underlie the results published in journals, and they are increasingly made accessible as supplements to published articles (Borgman, 2012) or deposited in domain repositories in response to journal publisher policies, further emphasizing the role of articles in representing the data but also the significance of linking a research publication with its respective dataset.

Assessing the use of journal publications requires an understanding of the practices of data production and what aspects of these practices need to be represented in the metadata describing a dataset. As stated by Gray, Szalay, Thakar and Stoughton (2002) "(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." Initial research to understand what information contained in soil science journal publications by data producers can be used to inform metadata description for data indicate that articles are a viable source for evidence of methods implemented in generating data (Chao, 2014a). Articles generally encompass particulars of the study site where collection of soil samples occurred, the instruments and techniques applied in collecting and processing the soil samples including units of measurement, and what variables were used for statistical analysis. The journal articles provided description of the processes and practices related to how data emerge and have potential application for imparting descriptive metadata for data that can contribute to curation efforts.

The primary aim of this study is to examine how methods description from journal articles can be utilized to generate metadata content for datasets. Existing metadata schemes for data are used to guide analysis and map journal article content for methods. By mapping to a metadata scheme, there is possibility to automatically generate data articles.[3] There are two questions guiding this research: 1) What metadata elements for methods map to journal article content? and 2) What similarities in methods mapping are visible across the metadata schemes? The mapping process may also reveal potential gaps in metadata schemes in conveying methods information to facilitate data reuse.

2 Genbank: http://www.ncbi.nlm.nih.gov/genbank/

3 See Chavan and Penev (2011) for example in biodiversity science; PREPARDE project for geoscience example: http://www2.le.ac.uk/projects/preparde

Research Design

In this exploratory study, 24 full text research articles were collected from three peer-reviewed journals in soil ecology: Soil Science Society of America Journal[4], Plant and Soil[5], and Applied Soil Ecology[6]. Soil ecology is investigated as a research area representative of long tail science, where data generated are in high need of curation support and are primarily analyzed and used locally within a research group using field-based approaches (Cragin, Palmer, Carlson and Witt, 2010). These top tier journals were selected based on published rankings from Scimago for the year 2012[7], reflecting quality research and different publisher practices for that scientific domain. Due to the high volume of journal papers available, the sample was limited to articles published between 2006-2013 and the thematic research area of "earthworm"-related studies. Using these criteria, the top eight research articles returned for each journal were retrieved for analysis.

The journal articles were analyzed in two phases for methods metadata content. In the first phase, the National Environmental Methods Index (NEMI)[8] metadata scheme was applied by mapping methods content from the journal articles to the relevant metadata fields to create a record. NEMI was developed for documenting analytical and field methods for all phases of environmental monitoring and was the most applicable metadata standard for methods. Such a scheme provides an initial framework for demonstrating how methods content from journal publications would fit.

The second phase involved reviewing the element mapping developed using the NEMI standard with two prominent metadata schemes relevant to soil ecology data, the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (CSDGM) and the Ecological Metadata Language (EML). These schemes are well supported and adopted by national data centres and repositories. Metadata schemes have benefitted from community development (Yarmey and Baker, 2013) in order to promote discovery of data, accommodate reuse by the original investigator and external researchers, and enable human and automated use of data (Michener, 2006). Such standards are often embedded in the scientific practice of the community, drawing on vocabulary familiar to the discipline (Willis, Greenberg and White, 2012). Most notably, each of these schemes contains a section dedicated to description of those processes used to generate data (Chao, 2014b). For CSDGM, this is the "Lineage"[9] module while the EML has the "eml-methods"[10] module. Metadata records using these two schemes were generated as needed. The addition of these discipline-specific schemes for analysis can reveal additional metadata fields for detailing methods. For all phases, the articles retrieved were manually coded and reviewed for metadata mapping.

4 Soil Science Society of America Journal: https://www.soils.org/publications/sssaj
5 Plant and Soil: http://link.springer.com/journal/11104
6 Applied Soil Ecology: http://www.journals.elsevier.com/applied-soil-ecology/
7 Scimago: http://www.scimagojr.com/. Note: 2012 is the year that is the most recently available.
8 NEMI: https://www.nemi.gov/about/
9 CSDGM Data Quality Information – Lineage: http://www.fgdc.gov/metadata/csdgm/02.html
10 EML eml-methods module: https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./eml-methods.html

Findings and Discussion

Mapping Data Production Processes

From the first phase of analysis, the mapping of NEMI metadata elements to journal article content established the importance of linking information on study samples and study variables with related procedures undertaken to gather, process, and analyse data (see Table 1 for NEMI metadata elements). The soil science articles generally contain designated "methods" sections detailing the generation and use of multiple data sources, and these systematic connections between data and processes bring greater meaning to the article text for metadata generation. Relationships between data sources (i.e. physical samples collected) and those processes applied for data production can be established through identifier use. For NEMI, the element *Method Number/Identifier* allows metadata creators to self-assign an identifier to a grouping of activities associated with a specific data source. Within such a grouping, the elements most commonly mapped were *Media Name*, *Method Descriptive Name*, *Brief Method Summary*, and *Instrumentation*. Figure 1 shows a brief example of mapping metadata elements to journal content and how information about the data source (*Media Name*, soil) with the related processing activities (*Method Descriptive Name*, *Brief Method Summary*, and *Instrumentation*) is brought together by the use of the identifier.

> Soil for chemical analysis was air-dried, gently crushed with a mortar and pestle, and passed through a 2 mm sieve prior to analysis. Soil texture was determined by the pipette method (Gee and Or, 2002). Total C and N were measured by dry combustion (Nelson and Sommers, 1982) in an Elementar Vario Max CNS analyzer (Elementar Analysensysteme, Hanau, Germany). (Smetak, Johnson-Maynard, & Lloyd, 2007, p. 163)

Element	Value
Method Number/ Identifier	SJML-2007-02
Media Name	Soil
Method Descriptive Name	Chemical analysis for soil
Brief Method Summary	Soil for chemical analysis was air-dried, gently crushed with a mortar and pestle, and passed through a 2 mm sieve prior to analysis. Soil texture was determined by the pipette method (Gee and Or, 2002). Total C and N were measured by dry combustion (Nelson and Sommers, 1982) in an Elementar Vario Max CNS analyzer (Elementar Analysensysteme, Hanau, Germany).
Instrumentation	Soil preparation: mortar and pestle, 2 mm sieve
Instrumentation	dry combustion: Elementar Vario Max CNS analyzer (Elementar Analysensysteme, Hanau, Germany)

Figure 1. Example of mapping with NEMI metadata elements; the top part of the figure is text from a sample journal article and the lower portion of the figure is one representation of how the text can be used as metadata for methods. The *Method Number/Identifier* value is solely used for illustrative purposes in this study.

While NEMI provides a basic framework to accommodate the uniqueness of the methods process for each study reported in the articles, the CSDGM and EML metadata schemes (see Table 2) examined in the second phase of analysis go a step further to highlight specific details of the processes used for generating data. For CSDGM, the *Source_Citation_Abbreviation* is self-assigned and can be used to link data sources used or produced with associated research activities. The application of the *Process Step* and *Process Description* elements from CSDGM also brings a greater level of granularity for illustrating the sequential nature of some research procedures from the narrative descriptions of methods in journal articles. Mapping of the EML elements to article content also demonstrates this level of granularity for documenting methods metadata through the introduction of *subStep* to show the hierarchical nature of some methods description. However, it is not clear if an identifier system is available for EML methods description to demonstrate relationships between data and the related *subStep*(s). The narrative nature of journal article content seems more amenable to the step-by-step framework presented by the CSDGM and EML schemes, which break down the text to make processes and related data more explicit. In contrast, the NEMI scheme would maintain more of the narrative format but also be accommodating of direct extractions of article content for applicable metadata elements. There needs to be a balance, therefore, in determining when it is appropriate to directly use article text as methods metadata and when the use of a citation to the journal article is better suited to convey details on the method processes.

Methods Metadata Element Similarities

The investigation of similar elements present across the mapped metadata schemes provided insight on what information is most important in describing methods for metadata inclusion. Each scheme is fairly consistent in having a general description field that allows for free-text explication of the procedures used. This type of element offers greater flexibility for what methods content to record and is a possible space for extracting full-text sections from the journal article to generate metadata. Another element that is available in all three metadata schemes is related to the citation of an existing method or data source. The prevalence for citation provision across these metadata schemes further enforces a best practice of documenting the provenance for a dataset in order to understand the context of its creation and use (Whyte and Wilson, 2010). These elements were also among the ones most commonly used when mapping to journal article content. One element that was also frequently mapped was “sampling”, which appears in both the NEMI and EML metadata schemes. Within the soil science journal articles, content specific to describing sampling procedures was consistently identified. The emphasis on a specific activity related to research data production reflects a methods procedure that is meaningful across scientific communities.

Identifying Gaps in Methods Metadata

The analysis of journal article content not only informs what descriptive metadata for methods may need to be added to existing schemes but also reveals what metadata elements cannot be readily extracted from article content. A consistent feature across the soil science journal articles was the description of the study site where physical samples were collected and research techniques employed. Study site description often included the longitude and latitude coordinates of the site, average precipitation and relative humidity, along with soil type identification; these details were typically found at the

beginning of the article's methods section. Based on the available method-related metadata elements, it was not always evident if this detailed contextual information could readily be recorded. Within NEMI, there is potential to record the geographic coordinates but no explicit elements to detail the descriptive aspects of the physical study site where data sampling and collection occurred. Similarly, geographic location coordinates can be entered in CSDGM though in a section of the metadata scheme separate from Lineage.

The EML appears to be the only standard where description of the study location may be documented. The *studyExtent* element enables narrative of "both a specific sampling area and the sampling frequency (temporal boundaries, frequency of occurrence)" and is connected to another element where more extensive information about the study site can be explicated (*studyAreaDescription*), which is found in a different module of the EML. The overall lack of visibility to document contextual study site location details within available methods metadata elements potentially signifies that this descriptive study site information may not be as integral in the capture of methods-related metadata as geographic coordinates are, especially if there is the opportunity to describe this context information in a different part of the metadata record.

Just as study site description is an area not explicitly covered in the examined metadata schemes, there are also methods-related elements from these schemes that are not easily discernible from journal article content. Within NEMI, the majority of the required elements can be accommodated from journal article content; those elements that remain may need to be addressed by the repository (i.e. method number/identifier) or manually inputted (i.e. method type/subcategory) based on the metadata submission process. Both EML and CSDGM have similar required elements dedicated to delineating the procedures engaged by scientists for a research study and as discussed in a previous section, there is rich availability of relevant content from journal articles.

Each of the mapped metadata schemes also has elements that are optional or to be used when applicable for detailing methods. The description of quality control (NEMI *Quality Control Requirements*; EML *qualityControl*) along with *Process Date* from CSDGM can potentially be inferred from textual clues from the article narrative but are not necessarily definitive sections within the soil science articles as with sampling. For quality control, practices in soil science include collecting field replicates, or a second sample from the same field location, to monitor field variability and precision in sampling procedures (Boone, Grigal, Sollins, Ahrens and Armstrong, 1999). The description of replicates is often included within discussion of the sampling process and therefore may not be easily recognized as quality control information for metadata purposes. It may be the case that quality control practices are represented more prominently in journal articles from other disciplines, but for the soil sciences articles these practices appear to be more embedded in the overall methods narrative.

Determinations of *Process Date* can be made for some field-related processes based on article text. In one example, the month and year for sampling are detailed, "earthworm sampling was conducted during a 2-week period beginning at the end of May in 2004 and 2005. Each site was sampled once per year" (Smetak, Johnson-Maynard and Lloyd, 2007). Other common representations of date include the season (i.e. spring, fall) rather than specific months. However, laboratory processes such as chemical analysis of soil samples tend not to have any temporal indicators for when they occurred. It is possible to infer that processes subsequent to field sampling would take place in the coming months (i.e. June/July 2005) but this information would need to be verified before inclusion as metadata. There still remain some required and

optional methods-related elements from all three metadata schemes that are not necessarily accounted for by journal article content but these articles provide a solid foundation to build on in producing metadata about methods for a dataset.

Table 1. Methods-related elements from CSDGM and EML.[11]

	Mandatory	Mandatory If Applicable
CSDGM metadata elements from "Lineage" section	Process_Step	Source_Information
	Process Description	Source_Citation
	Process Date	Type of Source Media
	Process Time	Source_Time_Period_of_Content
	Process Contact	Source_Citation_Abbreviation
	Source Produced/Used Citation Abbreviation	Source_Contribution
EML metadata elements from "eml-methods" module	MethodsType	dataSource
	methodstep	sampling
	ProcedureStepType	studyExtent
	description	studyExtent — coverage
	citation	studyExtent — description
	protocol	samplingDescription
	instrumentation	spatialSamplingUnits
	software	spatialSamplingUnits — coverage
	substep	spatialSamplingUnits — referenceEntityId
		citation
		qualityControl
		description
		description_citation
		description_protocol
		instrumentation
		software
		substep

[11] The " — " have been added to distinguish elements with similar names and are not part of the original scheme. The indentation of certain elements is purposeful to show the scheme hierarchy and related elements.

Table 2. Elements from the NEMI metadata scheme.

NEMI Mandatory Elements	NEMI Optional Elements
Method Descriptive Name	Scope and Application
Method type/subcategory (pre-defined list)	Max Holding Time
Brief Method Summary	Quality Control Requirements
Method Number/Identifier	Precision Descriptor Notes
Method Source	Interferences
Source Citation	Concentration Range Units
Media Name	Applicable Concentration Range
Method Official Name	Detection Limit Note
Instrumentation	Detection Limit Type
	Sample Preparation Methods
	Sample Handling

Conclusions

Research methods are a key part of descriptive metadata for scientific data reuse. Generating this methods metadata from available sources, such as journal articles, presents an initial approach to address the time-intensive process of metadata production. The examination of emerging metadata standards for methods, such NEMI, sheds light on what basic information should be documented about an implemented method. The subsequent review and application of the CSDGM and EML metadata schemes highlighted different descriptive approaches as well as shared similarities in methods-related elements. Journal article content, as a whole, provided a robust source for descriptive information that could readily be extracted as methods metadata to illustrate the basic steps of the data production process.

The findings from this study invite further inquiry and exploration. Analysis of actual metadata records generated for research data from data repositories would provide evidence on how schemes are applied in practice and what description information is actually provided about the methods. Particular attention to metadata records with associated publications listed would be one approach to corroborating journal article text with actual metadata content. In addition, designing a taxonomic approach for identifying and documenting methods information from journal articles can have potential implications for the development and advancement of automated processes to capture and enhance data description in supporting data repositories and curation services. With scientific research data expecting to rise in quantity and diversity, this attention to methods and how research data are generated can inform metadata generation and standards development for the curation of data.

Acknowledgements

This material is based upon research supported by the Thomson Reuters Doctoral Dissertation Proposal award. Many thanks to Dr. Carole Palmer, Dr. Michelle Wander, Dr. Jane Greenberg and Dr. Catherine Blake for the thoughtful discussion on the topics and themes addressed in this research.

References

Boone, R.D., Grigal, D.F., Sollins, P., Ahrens, R.J., & Armstrong, D.E. (1999). Soil sampling, preparation, archiving, and quality control. In G.P. Robertson, D.C. Coleman, C.S. Bledsoe, & P. Sollins (Eds.), *Standard Soil Methods for Long-Term Ecological Research* (pp. 3–28). New York: Oxford University Press.

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

Chao, T.C. (2014a). Identifying indicators of description for research data from scientific journal publications. In *iConference 2014 Proceedings* (1038 - 1042). http://hdl.handle.net/2142/48765

Chao, T.C. (2014b, November). *Enhancing metadata for research methods in data curation.* Poster presented at the 77th ASIS&T Annual Meeting, Seattle, WA. Retrieved from https://www.asis.org/asist2014/proceedings/submissions/posters/249poster.pdf

Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2

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. doi:10.1098/rsta.2010.0165

Davis, L., Qin, H., D'Ignazio, J., Romero Lankao, P., Mayernik, M., & Alston, P. (2012). *Variables as currency: Linking meta-analysis research and data paths in science.* Paper submitted to the 75th ASIS&T Annual Meeting, Baltimore, MD. Retrieved from http://dlsciences.org/research/DataConservancy/Variables%20as%20Currency.pdf

Edwards, P.N., Jackson, S.J., Bowker, G.C., & Knobel, C.P. (2007). Understanding infrastructure: Dynamics, tensions, and design. In *Report of a Workshop on History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures,* University of Michigan, School of Information. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/49353

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., & Borgman, C.L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. doi:10.1177/0306312711413314

Faniel, I.M., & Jacobsen, T.E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW), 19*(3-4), 355–375. doi:10.1007/s10606-010-9117-8

Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E., & Martone, M.E. (2014). Big data from small data: Data-sharing in the "long tail" of neuroscience. *Nature Neuroscience*, *17*(11), 1442–1447. doi:10.1038/nn.3838

Gray, J., Szalay, A.S., Thakar, A.R., & Stoughton, C. (2002). Online scientific data curation, publication, and archiving. In *Proceedings of SPIE: Vol. 4846. Virtual Observatories* (pp. 103–107). doi10.1117/12.461524

Greenberg, J., (2004). Metadata extraction and harvesting: a comparison of two automatic metadata generation applications. *Journal of Internet Cataloging, 6*(4), 59–82. doi:10.1300/J141v06n04_05

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280–299. doi:10.1353/lib.0.0036

Karasti, H., Baker, K.S., & Halkola, E. (2006). Enriching the notion of data curation in escience: Data managing and information infrastructure in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work, 15*(4), 321-358. doi:10.1007/s10606-006-9023-2

Karasti, H., & Baker, K.S. (2008). Digital data practices and the long term ecological research program growing global. *International Journal of Digital Curation*, *2*(3), 42–58. doi:10.2218/ijdc.v3i2.57

Kovacevic, A., Ivanovic, D., Milosavljevic, B., Konjovic, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program: Electronic library and information systems*, *45*(4), 376–396. doi:10.1108/00330331111182094

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation, 6*(2), 4–37. doi:10.2218/ijdc.v6i2.205

Mayernik, M.S., Wallis, J.C., Pepe, A., & Borgman, C.L. (2008). Whose data do you trust? Integrity issues in the preservation of scientific data. In *Proceedings of the 2008 iConference.* Retrieved from http://hdl.handle.net/2142/15119

Mayernik, M.S., Batcheller, A.L., & Borgman, C.L. (2011). How institutional factors influence the creation of scientific metadata. In *Proceedings of the 2011 iConference* (pp. 417–425). doi:10.1145/1940761.1940818

McPhillips, T., Bowers, S., Zinn, D., & Ludäscher, B. (2009). Scientific workflow design for mere mortals. *Future Generation Computer Systems*, *25*(5), 541–551. doi:10.1016/j.future.2008.06.013

Michener, W.K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7. doi:10.1016/j.ecoinf.2005.08.004

Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions American Geophysical Union*, *91*(34), 297–298. doi:10.1029/2010EO340001

Smetak, K.M., Johnson-Maynard, J.L., & Lloyd, J.E. (2007). Earthworm population density and diversity in different-aged urban systems. *Applied Soil Ecology*, *37*(1), 161–168. doi:10.1016/j.apsoil.2007.06.004

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs.* Retrieved from Research Information Network website: http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS One, 6*(6), e21101. doi:10.1371/journal.pone.0021101

Van House, N.A., Butler, M.H., & Schiff, L.R. (1998). Cooperative knowledge work and practices of trust: Sharing environmental planning data sets. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 335–343).

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

Whyte, A., & Wilson, A. (2010). *How to appraise and select research data for curation*. DCC How-to Guides. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology, 63*(8), 1505–1520. doi:10.1002/asi.22683

Yarmey, L., & Baker, K.S. (2013). Towards standardization: A participatory framework for scientific standard-making. *International Journal of Digital Curation*, *8*(1), 157–172. doi:10.2218/ijdc.v8i1.252

Zimmerman, A.S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology and Human Values*. doi:10.1177/0162243907306704