Exploring the role of 'research methods' in metadata description for data reuse

Tiffany Chao
30 May 2014

Introduction

A key part of the curation process is ensuring that metadata are available to describe datasets for future use. The importance of metadata for research data use has been shown through studies on scientists' expectations for data sharing and reuse (Borgman, Wallis, & Enyedy, 2007; Cragin, et al, 2010; Tenopir et al., 2011). A common theme from theses studies is the significance of research methods description provision for determining reuse of data. In particular, the descriptions of methods and processes used to generate data can convey 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 (Faniel & Jacobsen, 2010). Little is known however about how to document these implemented methods and what information should be included in order to be useful for others. To address this gap, this exploratory study investigates how research methods descriptions are represented in contemporary metadata schema for research data, with a focus on what information about the data production process is required for metadata inclusion. The paper is organized as follows: the next section examines the issue of scientists' attitudes toward metadata provision and their metadata expectations for the sharing and reuse of research data. Next, the research questions of the study are detailed along with the proposed approach on how they are answered. The findings are then discussed and the final section concludes with next steps toward a better understanding of metadata needs for research data and the support of data sharing and reuse.

Background

Generating metadata for research data is a time-intensive process for scientists to undertake during the course of research study. The provision of these 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). Research funders have long been urging their grantees to collect and maintain metadata, but this call has been met with minimal adherence or completely ignored. As explained by Edwards et al., 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" (2007, p. 32).

Yet, metadata is an essential component to making datasets more accessible by others and facilitating meaningful interpretation and use. Providing metadata for long tail science research data is 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). Examples of long tail science include areas of study within the Earth Sciences (Cragin et al., 2010), particularly with field-based work, where scientists often lack the financial support or tools for metadata generation thereby limiting future access and reuse of data produced.

There are renewed 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). This action follows earlier requirements set by federal funding agencies for researchers to produce data management plans as part of funding proposals and the increasing federal investment in the development of cyberinfrastructure and services for research data. As data sharing and open access initiatives are implemented more broadly in universities and other research institutions, many research communities will be in need of curation support for describing their data.

Methods description for data reuse

In order to better support metadata generation, it is critical to understand what information is of value to researchers in the use of data. Studies of scientific data practices are providing an important base of knowledge for the development of curation processes for research data. From these studies, scientists consistently highlight description of "research methods" as key information to include when making data available for others. The methods of a research study, more generally, provide an account of the various procedures, protocols, and algorithms "that can be used to generate scientific knowledge" (Committee on the Conduct of Science, 1989, p. 9060). Researchers may use multiple methods or a particular grouping of methods to address a particular question or understand a specific phenomenon. Methods are also perceived as "socially accepted standards of science" (Institute of Medicine, 1995, p. 4) where the selection and implementation of processes are often mediated through contact with fellow research group members or collaborator interactions.

Methods description has been used in appraising trust in data and guiding selection of data for reuse. 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 & Schiff, 1998). Similarly, Zimmerman's

(2008) study of ecological research practices uncovered that the documentation on methods was more significant in appraising trust in data and guiding selection of data for reuse than the prestige of the investigator who collected the data. Within field-based science, the research methods and protocols are often prone to modification due to rapid environmental change, and documenting these modifications is not only vital to integrity of the dataset but also informative for future users (Mayernik, Wallis, Pepe, & Borgman, 2008; Karasti & Baker, 2008). The importance of information on methods and processes of data production is also a persistent theme in studies of data practices beyond long tail sciences. Methods and protocol information for genomics research is often made available through project websites to complement a dataset deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/), and assessment of methods deployed to produce data is common in the peer-review process for publications in astronomy research (Swan & Brown, 2008).

Other attributes of value indicated for inclusion in descriptions of methods relate to the research activities of "data generation" and "analysis". Wallis (2012) describes the variation in ecology between traditional practices of taking sensor readings by hand and the alternative use of networked embedded sensor technologies that automatically collect the same contextual variables. The adoption, or lack of adoption, of new technologies by a research team is an obvious difference in data gathering within a single research community that needs to adequately documented in the metadata. For soil survey documentation in databases, a detailed description of the analytical methodology used to measure survey parameters and the time frame for field sampling are needed (Lacarce et al., 2009). While not a formal community practice, Staudigel et al. assert that any description of physical geological samples within journal publications should at minimum provide a full account of the sampling process and analytical process including “information on the origin of the data” and “how the data were normalized” (2003, p. 7). This emphasis on specific activities related to research data production further supports the role of methods for inclusion in metadata for long tail science data.

Research Questions

To encapsulate the significance of methods description for the use of data, e-Science pioneer Jim Gray and colleagues state “(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). These studies of scientists’ data practices on sharing and reuse behaviors and expectations provide an introduction to some of the criteria needed in describing research methods for data use; criteria include documenting changes in data production processes, instruments and tools used for data collection, sampling methods, and data analysis procedures. For this study, existing metadata schemas are examined to more comprehensively understand how scientists’ expectations for methods description of research data can be supported in the generation of metadata. The questions guiding this study are: 1) How is description for methods represented in metadata schemas? 2) What information/criteria comprise methods description in metadata schemas? Analysis of existing metadata schema for methods-related elements may reveal potential gaps in description for data or encourage more pro-active use of these standards.

Method

A total of eight metadata schemas were identified for investigation of research method descriptive elements (see Table 1). This sample builds on those schemes reviewed by Willis, Greenberg, and White (2012) in their study of metadata goals for scientific metadata standards and is extended with additional schema that have potential application for long tail data. These selected schemas are in active use for describing data and range in date of issue with the Directory Interchange Format (DIF) released in the early1990s to the more recent the Dublin Core Dryad Application Profile (DCDryad), released in 2009. The schemas are primarily community-driven efforts with repository application and often, international reach, such as the Data Documentation Initiative (DDI) and Macromolecular Crystallographic Information File (mmCIF). Some schemas are more domain-specific (i.e. Ecological Metadata Language (EML) for ecology data) while others can be more broadly applied (i.e. Content Standard for Digital Geospatial Metadata (CSDGM) for geospatial data).

Each metadata schema and related documentation were retrieved and analyzed using a set of observation categories (see below). The findings from this analysis are detailed in Table 2.

Observation Categories:

- Presence: this first category examines whether the term "methods", "methodology", or similar derivation is an explicit element within the schema. For this study, "methods" is used to encompass these element name derivations. If there is no explicit "methods" element, additional schema elements will be examined in relation to research methods definitions and the data production process.
- Detail: the second category explores how the "methods" element is defined and what information is required to complete the field. Findings can be used to better understand what information on methods researchers need to contribute in generating metadata.
- Status: the third category of interest reports on whether the "methods" element is mandatory for a completed metadata record. This category provides insight to how the metadata schema supports methods description for the reuse of data.

Table 1. Sample of metadata schemas for analysis.

Schema name	Overview	Examples of Repository Application
Content Standard for Digital Geospatial Metadata (CSDGM) *https://www.fgdc.gov/metadata/csdgm*	Metadata for digital geospatial data; originally applied to data from US federal agencies and governments but now transitioning to ISO 19115; maintained by Federal Geographic Data Committee (1998 first published)	National Climatic Data Center
Darwin Core (DwC) *http://rs.tdwg.org/dwc/terms/simple/index.htm*	Content specification designed for data about the geographical occurrences of species; maintained by Biodiversity Information Standards-TDWG (1998 first published)	GBIF Data Portal
Data Documentation Initiative (DDI) *http://www.ddialliance.org/specification*	Specification for describing data from the social, behavioral, and economic sciences through the life cycle; maintained by the DDI Alliance (2000 first published)	ICPSR, CESSDA
Directory Interchange Format (DIF) *http://gcmd.gsfc.nasa.gov/add/difguide/index.html*	Metadata format to describe scientific datasets; maintained by the Global Change Master Directory, NASA (1990 first published)	Global Change Master Directory (GCMD)
Dublin Core – Dryad Application Profile (DCDryad) *http://wiki.datadryad.org/Metadata_Profile*	An application profile based on the Dublin Core Metadata Initiative Abstract Model, used to describe multi-disciplinary data underlying peer-reviewed scientific and medical literature; maintained by Dryad (2009 first published)	Dryad
Ecological Metadata Language (EML) *https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1/./index.html*	Specification developed by the ecology discipline and for the ecology data; developed by National Center for Ecological Analysis and Synthesis (1997 first proposed)	ESA Data Registry, LTER Network Data Repository
Macromolecular Crystallographic Information File (mmCIF) *http://www.rcsb.org/pdb/static.do?p=file_formats/mmcif/index.html*	Describes small molecule organic structures and from crystallographic experiments; developed by the International Union of Crystallography (IUCr) (mid 1990s first published)	Protein Data Bank (PDB)
ThermoML (TML) *http://trc.nist.gov/ThermoML.html*	An XML-Based IUPAC Standard for Storage and Exchange of Experimental Thermophysical and Thermochemical Property Data; maintained by the TRC Group, NIST (2002 first published)	ThermoML Archives

Table 2. Summary of "methods" representation observations across metadata schema sample.

	Presence & Detail		Status
Schema name	"Methods" element in schema	Schema elements related to "methods"	
CSDGM	(none)	**2.5 Lineage** -- information about the events, parameters, and source data which constructed the data set, and information about the responsible parties.	Mandatory with optional elements
DwC	(none)	<samplingProtocol> The name of, reference to, or description of the method or protocol used during an Event. Refer to <Event> for additional terms. <Location> A spatial region or place, named or not.	Recommended
DDI	<Methodology> concerns data collection, determining the timing and repetition patterns for data collection, and sampling procedures. Includes following content elements: DataCollectionMethodology, DataCollectionSoftware, DeviationFromSampleDesign, SamplingProcedure, TimeMethod	Related: <SamplingProcedureType> Describes the type of sample, sample design and provides details on drawing the sample. In addition to the descriptive narrative supports the use of a brief term or controlled vocabulary to classify the type of sampling procedure described.	Mandatory
DIF	(none)	<summary> The "Summary" field provides a brief description of the data set along with the purpose of the data. <instrument (sensor name)> name of the instrument used to acquire the data <temporal coverage> specifies the start and stop dates during which the data were collected.	<summary> is mandatory; others are highly recommended
DCDyrad	(none)	<dcterms:description> Human-readable description of the resource; an abstract or summary. <Spatial Coverage> description of the data specified by a geographic description and/or geographic coordinates. <Temporal Coverage> description of the data, as geologic timespan.	Mandatory
EML	**eml-methods module** - Methodological information for resources, describes the methods followed in the creation of the dataset, including description of field, laboratory and processing steps, sampling methods and units, quality control procedures.	**eml-protocol module** specific to prescribed procedures whereas eml-method describes the procedures that actually occurred.	Not clear; suggests use of EML modules as needed
mmCIF	<method_list> part of Data category with related Data items; method description: stores the experimental method used to create the structure	(not clear)	Not mandatory
TML	<MethodName>, describes method of property generation; distinguished by **eMethodName** [enumeration] or **sMethodName** [string] for methods that are not enumerated	<MethodName> part of <BioProperties> element and other "Property" elements	Mandatory

Findings & Discussion

Explicit and implicit representations of methods

The examination of eight metadata schemas for the inclusion of methods provided insight to how descriptions of data production processes are supported as well as what components comprise methods description. Of the investigated schemas, half (4) have an explicit element for “method” or “methodology” while the remaining ones had potential elements that would support the documentation of methods. The DDI and EML, both of which contain a <methods> element, have additional sub-elements that provide opportunity for comprehensive structured description of method procedures, temporal and/or spatial coverage related to an implemented procedure, and the tools and instruments used in these processes. In contrast, the <methods> element for mmCIF had minimal definition on what information to include while specific attention was given to the formatting of methods description in TML. Despite the variation in the level of detail for this element, the definitions and content from these schemas with explicit <methods> elements were used to inform the identification of potential elements for methods description from the remaining schema.

For the remaining schemas that did not have an explicit <method> element (CSDGM, DIF, DCDryad, DwC), each element of the schema along with definition were reviewed to determine implicit representation of methods. The <lineage> element in CSDGM has the closest association with definitions observed for <method> in the DDI and EML, containing multiple sub-elements to further describe the processes, data collection sources, and temporal factors related to the creation of a dataset. In this respect, the use of <lineage> and <method> in these schemas are nearly synonymous given the element definitions. While the CSDGM presents a more concrete representation of methods, both the DIF and DCDryad offer only potential metadata elements where methods used in data production may be described. The elements <summary> (DIF) and <description> (DCDryad) have broader application in explaining what the data are, where details on how data were produced may or may not be included as part of this explanation. Other implicit elements related to methods observed across this group of schemas include spatial and temporal characterization along with instrument details, all of which are comprised in the <method> elements of EML and DDI.

Schema element names and discipline influence

Metadata standards have benefitted from community development (Yarmey & 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, Green, & White, 2012). These disciplinary distinctions also seem to be apparent in the formation of element names used in schema affiliated with methods. For instance, there is a distinction made between the use of "methods" and "protocol" within the EML modules which accommodate how data are managed and documented in ecology research (EML Contents, n.d.). The EML denotes that "protocols" are often established procedures known in the community with particular steps in place while "methods" describe what actual procedures were implemented during the course of research. The element name of <lineage> from CSDGM may also have significance within the geospatial data community despite the similarities in definitions and structure with the <method> element from DDI and EML. The <lineage> element is actually part of the Data Quality Information section of CSDGM, corroborating findings from Faniel and Jacobsen's (2010) study of data reuse by earthquake engineers on the importance of data production process documentation as a measure of quality for data.

An element that seems to span multiple disciplines and schemas is "sampling", a research technique generally applied for capturing a quantity or portion for analysis that is representative of a greater phenomenon. Within DwC, the <samplingprotocol> element is associated with an overarching Event (characterized by a time and place) and is used more generally to detail any methods or procedures used in the Event time frame. The broader DwC definition, which is typically applied to biodiversity related data, differs from the DDI use of "sampling" for social science-oriented data. The element <sampProc> or "sampling procedure" from DDI is connected with the process of data collection and more narrowly defined as "The type of sample and sample design used to select the survey respondents to represent the population. May include reference to the target sample size and the sampling fraction." Although different meanings are associated with the term" sampling" by these distinctive scientific communities, the inclusion of "sampling" as a metadata element provides a more granular level of detail for describing methods and the process for how data are produced.

Use of methods elements

Findings from the "status" category indicate that metadata elements related to methods description vary in how they are used in the resulting metadata record. Among those metadata schemas with explicit <methods> elements, only the DDI and TML clearly state that this element should be completed as part of the record. From the remaining schemas, CSDGM, DCDryad, and DIF have some mandatory elements that may include methods information but related sub-elements are often optional or recommended to complete. The variation in use of these elements related to methods description reflects variation in support for capturing metadata about methods to be made accessible for the discovery of datasets. This also suggests different aims or priorities in metadata generation for datasets, with other elements prioritized to satisfy a record for data discovery and access.

Returning to the research questions, each of the sample metadata schemas have some element(s) where description for methods could be included. The DDI, EML, and CSDGM, are the more robust schemas relative to the others for articulating definitions and criteria specific to methods and the processes of producing data. The composition of sub-elements and conciseness of definitions observed in these three schemas can potentially be adapted to other metadata standards. Identifying the element(s) to include methods description in the other sample schemas was less explicit where any information pertaining to the methods may not be provided at all as many of these metadata fields are optional or recommended to use. In this regard, scientists' limited utilization of metadata standards (Tenopir et al., 2011; Mayernik, Batcheller, & Borgman, 2011) may be warranted if basic information about the method of data production cannot be readily documented or shared. On the other hand, the number of associated metadata elements specific to methods as seen in the more robust schemas may equally overwhelm a researcher and also result in minimal compliance.

Conclusions

This exploratory study is only the start to a more extensive discussion on metadata standards and the generation of research methods information for data reuse. Based on the sample analysis of existing metadata schemas, there appears to be some reflection of researcher expectations for methods description but it is not widely supported across disciplines in these current frameworks. The metadata records generated from those schemas with more comprehensive elements for methods can be a source of future investigation on how such a schema is used in practice and what metadata is actually provided about the methods. The application of controlled vocabularies specific to methods within these schemas is another facet to investigate that can shed light on producing useful metadata for data reuse. Metadata standards for documenting methods, such as the National Environmental Methods Index (NEMI) (https://www.nemi.gov/about/), can be leveraged to address potential gaps in what basic information should be documented about an implemented method.

It is evident that curation support for metadata generation must go beyond the integration of a more explicit "methods" element in existing standards. Further inquiry into the scientific practices of researchers is needed for determining viable solutions to alleviate some of the pressures of producing metadata. The analysis of journal articles produced by scientists, for instance, can be a rich data source for extracting methods information for metadata (Chao, 2014). With scientific research data anticipated to increase in quantity and heterogeneity, there is considerable need for deeper understanding of scientific data practices and expectations in different research communities, as well as techniques for optimizing data description as demand for curation services continues to grow.

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 on Digital Libraries*, 7(1), 17–30. doi:10.1007/s00799-007-0022-9

Chao, T. (2014, February). *Exploring description for research data in soil science journal publications*. Poster presented at the 9th International Digital Curation Conference, San Francisco, CA.

Committee on the Conduct of Science, National Academy of Sciences of the United States of America. (1989). On being a scientist. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(23), 9053–9074.

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

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 & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures,* University of Michigan, School of Information,(January). Retrieved from *http://deepblue.lib.umich.edu/handle/2027.42/49353*

EML Contents. (n.d.). *Module Documentation: eml-methods*. Retrieved from https://nis.lternet.edu/nis/schemas/eml/eml-2.1.0/docs/eml-2.1.0/eml-methods.html

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

Gray, J., Szalay, A. S., Thakar, A. R., & Stoughton, C. (2002). Online scientific data curation, publication, and archiving. In *Astronomical Telescopes and Instrumentation* (pp. 103–107). International Society for Optics and Photonics. Retrieved from http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=874929

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

Holdren, J. P. (2013, February 22). *Expanding public access to the results of federally funded research.* Washington, DC: The White House. U.S. Executive Office of the President. Office of Science and Technology Policy. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Institute of Medicine, National Academy of Sciences, and National Academy of Engineering. (1995). *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition*. Washington, DC: The National Academies Press.

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 (CSCW)*, *15*(4), 321-358. Springer. 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. *The International Journal of Digital Curation*, *2*(3), 42-58. Retrieved from http://ijdc.net/index.php/ijdc/article/view/86

Lacarce, E., Le Bas, C., Cousin, J.-L., Pesty, B., Toutain, B., Houston Durrant, T., & Montanarella, L. (2009). Data management for monitoring forest soils in Europe for the Biosoil project. *Soil Use and Management*, *25*(1), 57–65.

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. *iConference 2008*. 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). Retrieved from http://dl.acm.org/citation.cfm?id=1940818

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological informatics*, *1*(1), 3–7.

Staudigel, H., Helly, J., Koppers, A. A., Shaw, H. F., McDonough, W. F., Hofmann, A. W., … Derry, L. A. (2003). Electronic data publication in geochemistry. *Geochemistry, Geophysics, Geosystems*, 4(3). DOI: 10.1029/2002GC000314

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs: Main report*. London, UK: Research Information Network, (JISC) Committee for the Support of Research, and the National Environment Research Council UK. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report

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.

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. (2012). The distribution of data management responsibility within scientific research groups. Unpublished dissertation, Information Studies, University of California, Los Angeles.

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.

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 & Human Values*. doi:10.1177/0162243907306704