Collaborate, Automate, Prepare, Prioritize: Creating Metadata for Legacy Research Data

Inna Kouper
Indiana University, USA
inkouper@indiana.edu

Stacy R. Konkiel
Indiana University, USA
skonkiel@indiana.edu

Jennifer A. Liss
Indiana University, USA
jaliss@indiana.edu

Juliet L. Hardesty
Indiana University, USA
jlhardes@iu.edu

Abstract

Data curation projects frequently deal with data that were not created for the purposes of long-term preservation and re-use. How can curation of such legacy data be improved by supplying necessary metadata? In this report, we address this and other questions by creating robust metadata for twenty legacy research datasets. We report on the metrics of creating domain-specific metadata and propose a four-prong framework of metadata creation for legacy research data. Our findings indicate that there is a steep learning curve in encoding metadata using the FGDC content standard for digital geospatial metadata. Our project demonstrates that when data curators are handed research data "as is," they may be successful in incorporating such data into a data sharing environment. We found that data curators can be successful in creating descriptive metadata and enhancing discoverability via subject analysis. However, curators must be aware of the limitations in applying structural and administrative metadata for legacy data.

Keywords: metadata; research data; metadata quality; legacy data

1. Introduction

Data curation projects frequently involve data that were not created with long-term preservation and re-use in mind. Curating such data (hereafter called "legacy data") poses several difficulties associated with the considerable resources required to prepare data for digital repositories and provide tools for effective search and retrieval. Lack of metadata is a major barrier to providing rich access and discovery capabilities for data. How can curation of legacy data be improved by supplying necessary metadata? How much time and effort is required?

In this report, we address the questions above by creating robust metadata for twenty legacy research datasets. This effort is part of a larger project called *Sustainable Environment - Actionable Data*, or SEAD (Hedstrom, Alter, Kumar, Kouper, McDonald, Myers et al., 2013). The SEAD project focuses on the development of tools that enable sustainability scientists to curate and share their data at earlier stages of research as well as "downstream," after the data have been collected and stored.

We report on quantitative and qualitative metrics of creating domain-specific metadata and make recommendations for other librarians and researchers. In benchmarking the process of enhancing legacy dataset metadata, we pursue several goals. First, we make datasets available for effective search and re-use within newer data sharing environments. Second, we advance knowledge among researchers and data professionals about the needs, barriers, and requirements of curating legacy research data. Ultimately, our efforts will contribute to the development of efficient metadata creation practices for research data.

2. Methodology

For this project, we used twenty datasets that are publicly available via the National Center on Earth-surface Dynamics (NCED) repository (National Center on Earth-surface Dynamics [NCED], n.d.). Because these datasets originate from the interdisciplinary domain of earth sciences, the choice of a domain-specific metadata standard was not easy. Considering that the

data contained a significant amount of geospatial information, we decided to use the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (Federal Geographic Data Committee [FGDC], 1998).

A team of four librarians and data professionals (called "encoders" henceforth) contributed to metadata creation. Two encoders (Encoder 1 and Encoder 3) have significant expertise with library metadata schemas, such as Dublin Core, while two other encoders (Encoder 2 and Encoder 4) are well acquainted with scientific data and metadata schemas. Each encoder received five datasets of varying sizes (ranging from 0.01 to 664 gigabytes and from 1 to ~140,000 files per set) without regard for the datasets' content.

The encoding was done in two phases. During Phase I, encoders created standalone XML metadata files for each dataset using basic information provided by the NCED repository and information available via quick Internet searches. During Phase II, encoders undertook extensive research to find more information about datasets, particularly concerning the processes by which datasets were created and used. Encoders timed their encoding activities and logged their experiences in a journal.

3. Findings

We began this project by planning the suite of metadata standards and tools we would use to create metadata for legacy datasets. One of the primary sources for working with the FGDC content standard is the FGDC website and its page on metadata (FGDC, 2012). Working through the FGDC website was challenging because it contained a lot of information overall, but not enough information about how to choose the most relevant tool or approach. After an extensive search for suitable aides and tools, encoders decided to use XML editor tools such as Oxygen XML Editor and Notepad++ source code editor.

Encoding the basic metadata during Phase I required nine minutes to four hours per dataset. Providing additional metadata during Phase II required 20 minutes to 1.5 hours per dataset. Presented in the following table are the times required to create metadata during Phase I:

TABLE 1: Time to create metadata (h:mm) during Phase I.

Encoder / Dataset	Encoder 1	Encoder 2	Encoder 3	Encoder 4
Dataset 1	1:38	1:23	1:33	4:00
Dataset 2	0:35	0:26	0:46	0:45
Dataset 3	0:18	0:23	0:43	0:40
Dataset 4	0:26	0:45	0:26	0:45
Dataset 5	0:09	0:15	1:33	1:00

The average time to create Phase I metadata per dataset was 54 minutes. It took four hours to create an XML file "from scratch" using the FGDC content standard publication (FGDC, 1998). The rest of the encoding activities relied on existing templates and examples available on the web. Even excluding the four-hour experience of Encoder 4, the table demonstrates a rather steep learning curve for working with the FGDC content standard. All encoders required more than one hour to complete their first dataset. After encoding the first dataset, metadata creation proceeded more quickly.

The information that encoders entered into metadata during Phase I largely corresponded to the FGDC content standard's mandatory elements requirement (FGDC, 2000). TABLE 2 lists the elements that were encoded in all twenty datasets during Phase I.

TABLE 2: FGDC content standard fields encoded during Phase I. (Non-mandatory elements are italicized.)

Content Standard Section	Content Standard Elements
1 Identification Information	Abstract (1.2.1 abstract), Purpose (1.2.2 purpose), *Currentness (1.3.1 current)*, Progress (1.4.1 progress), Maintenance and Update Frequency (1.4.2 update), Place and Theme Keywords (1.6.1.1 themekt, 1.6.1.2 themekey, 1.6.2.1 placekt, 1.6.2.2 placekey), Access and Use Constraints (1.7 accconst, 1.8 useconst), *Point of Contact (1.9 ptcontac)*
7 Metadata Reference Information	Metadata Date (7.1 metd), Metadata Contact (7.4 metc), Metadata Standard Name (7.5 metstdn), Metadata Standard Version (7.6 metstdv)
8 Citation Information	Originator (8.1 origin), Publication Date (8.2 pubdate), Title (8.4 title), Geospatial Data Presentation Form (8.6 geoform), Publication Place (8.8.1 pubplace), Publisher (8.8.2 publish), *Online Linkage (8.10 onlink)*
10 Contact Information	Contact Person (10.1.1 cntper), *Contact Organization (10.1.2 cntorg)*, Contact Organization Primary (10.2 cntorgp), *Contact Position (10.3 cntpos)*, Address (10.4.1 addrtype, 10.4.2 address, 10.4.3 city, 10.4.4 state, 10.4.5 postal, *10.4.6 country*)

Assigning values to most of the mandatory metadata elements was relatively unproblematic. However, most datasets did not contain spatial domain information in a readily available form; therefore encoders could not complete the mandatory element Spatial Domain (1.5 spdom). Only one dataset contained bounding coordinates as part of its shapefile metadata. Other datasets had large amounts of spatial information in varying forms, including geospatial files (shapefiles), maps in PDF and TIFF formats, binary files, and so on. Identifying bounding coordinates from these files would have required a significant amount of time and access to the originating software. As such, we did not provide data values for spatial domain. Another challenge encountered was finding correct and up-to-date contact information for the data managers for each data set (10.1.1 cntper, 10.3 cntpos). As research groups change, it becomes unclear who maintains the data and is responsible for providing documentation.

One of the areas where encoders' efforts were especially fruitful was the use of thesauri and controlled vocabularies in assigning keywords. We used the NASA Global Change Master Directory (GCMD) Science Keywords (Olsen et al., 2012) and ISO 19115 Topic Categories (FGDC, 2011) to assign thematic keywords and places for all datasets. We were also able to find information about instruments used in data collection for some of the datasets.

Phase I allowed us to collect descriptive metadata, which describes resources for the purposes of discovery and identification (NISO, 2004). During Phase II, we attempted to supply richer metadata, encoding the composition of research objects as well as their technical and preservation information. Below are the elements that were encoded in all datasets during Phase II:

TABLE 3: FGDC content standard fields encoded during Phase II. (Non-mandatory elements are italicized.)

Content Standard Section	Content Standard Elements
1 Identification Information	Theme Keyword (1.6.1.2 themekey), *Point of Contact (1.9 ptcontac)*, *Data Set Credit (1.11 datacred)*
5 Entity and Attribute Information	Entity and Attribute Detail Citation (5.2.2 eadetcit)
6 Distribution Information	Distribution Liability (6.3 distliab), Digital Form (6.4.2 digform), Digital Transfer Information (6.4.2.1 digtinfo), Format Name (6.4.2.1.1 formname)

During Phase II encoding, only one additional Theme Keyword (1.6.1.2 themekey) was found for one of the datasets, which prompted questions about the added value of such a task, in light of the added time investment. The FGDC mandatory elements that encoders were unable to identify for all twenty datasets during Phases I and II are found in TABLE 4.

TABLE 4: Mandatory FGDC content standard fields that were problematic to encode during Phases I and II. (Elements that are 'mandatory if applicable' are underlined.)

Content Standard Section	Content Standard Elements
1 Identification Information	Bounding Coordinates (1.5.1 bounding)
2 Data Quality Information	Attribute Accuracy (2.1 attracc), Logical Consistency Report (2.2 logic), Completeness Report (2.3 complete), Positional Accuracy (2.4 posacc), Lineage (2.5 lineage)
3 Spatial Data Organization Information	Indirect Spatial Reference (3.1 indspref), Direct Spatial Reference Method (3.2 direct)
4 Spatial Reference Information	Horizontal Coordinate System Definition (4.1 horizsys), Vertical Coordinate System Definition (4.2 vertdef)
5 Entity and Attribute Information	Entity Type (5.1.1 enttype)
9 Time Period Information	Single Date/Time (9.1 sngdate), Multiple Dates/Times (9.2 mdattim), Range of Dates/Times (9.3 rngdates)

The efforts of Phase II were significantly less successful due to the difficulties of accessing and processing the datasets. Some of our datasets were several hundred gigabytes and contained tens or hundreds of thousands of files. Some datasets were simply impossible to download and access due to their size. Others were inaccessible due to specialized software requirements. Most of the datasets were heterogeneous, i.e., they contained text, graphics, video, scripts, executable files, geospatial files, and so on. Accessing multiple file types and retrieving information about their native environments and relationships is not a trivial task, both in terms of time and resources.

Despite the difficulties, encoders were able to find additional information not readily available via the NCED repository. Ten metadata fields were enhanced during Phase II, adding such information as references to grants and funding information, distribution conditions, digital access and transfer information, and citations to related datasets and published articles. However, as Table 4 indicates, most of the information for sections 2 "Data Quality Information," 3 "Spatial Data Organization Information," 4 "Spatial Reference Information," and 5 "Entity and Attribute Information" from the content standards was difficult or impossible to obtain. The information that we found can enhance discovery opportunities of legacy research data, but may be insufficient to support the tasks of preservation, reproducibility, and re-use.

4. Discussion and Conclusion

The FGDC Content Standard for Digital Geospatial Metadata is a powerful tool for representing descriptive, structural, and administrative metadata (NISO, 2004). In dealing with legacy research data, however, the capabilities of this tool become seriously limited. Unlike other information resources, such as books or images that remain accessible and relatively transparent for preservation and sharing efforts, research data are complex compound objects. Formats, structure, relationships, and provenance become opaque once the data has been created. Our project demonstrates that data curators who are handed legacy research data "as is" can be very effective in creating descriptive metadata – particularly, in conducting subject analysis and assigning keywords based on controlled vocabularies and thesauri. However, identifying structural and administrative metadata for legacy data is extremely difficult.

The findings from our project support the argument regarding modeling metadata offered by Qin, Ball, and Greenberg (2012). They argued that to meet the requirements for data management, discovery, and use while remaining easy to use and maintain, metadata creation should adhere to the following principles: the least effort principle - utilizing existing databases and tools to populate metadata fields; the infrastructure service principle - incorporating semantic, temporal and geospatial metadata as well as its scientific context into research data infrastructure; and the portability principle - making metadata schemas portable and flexible enough so that multiple and separate metadata creation efforts can be merged together.

Based on our findings and collective experience, we propose a four-prong approach to metadata creation of legacy research (see FIG 1).

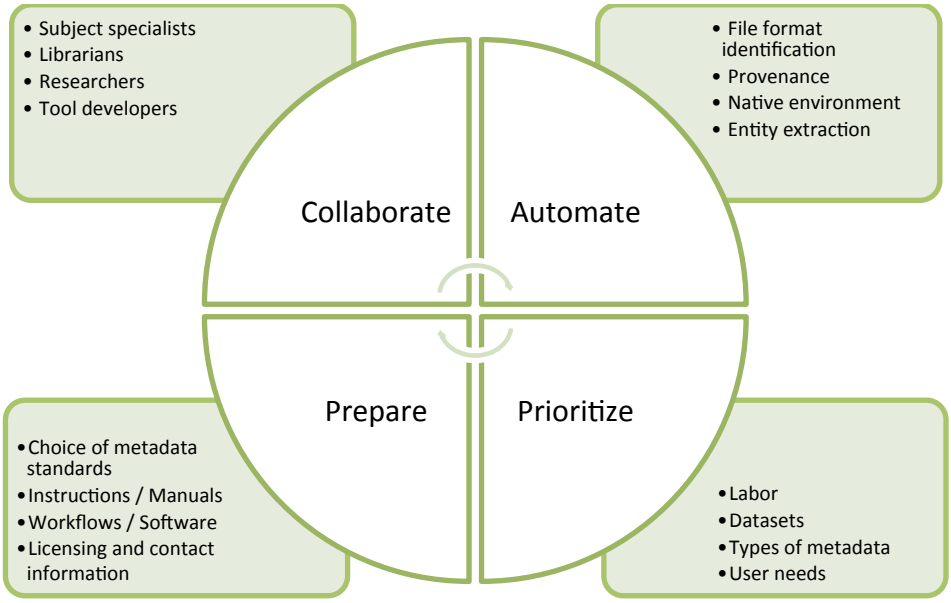

FIG. 1. CAPP Framework: Collaborate, Automate, Prepare, Prioritize.

Our approach, *CAPP: Collaborate, Automate, Prepare, Prioritize* is based on the premise that metadata creation or enhancement projects need to rely on a collaborative effort and on a combination of automated and manual labor. Data managers need to collaborate with subject specialists, researchers, and tool developers to define the requirements of specific data curation projects. Researchers as data producers and consumers can contribute to metadata creation by indicating what elements are valuable and for what purposes. They can also supply additional information that can be used in completing metadata records.

At the beginning of a legacy data curation project, data managers may also want to make the decision-making explicit and prioritize which datasets they need to curate and what user needs will guide the curation. Tasks such as file format identification, provenance capture, and entity extraction need to be automated. Existing tools, such as the JSTOR/Harvard Object Validation Environment (JHOVE, http://jhove.sourceforge.net/), MIME Type Detection Utility (mime-util, http://sourceforge.net/projects/mime-util), or Internet Assigned Number Authority's MIME Media Types (IANA, http://www.iana.org/assignments/media-types) can be used to automate identification of technical metadata, including file formats. Tools such as GeoServer (http://geoserver.org/display/GEOS/Welcome) can provide access to specific metadata within certain formats, such as shapefiles, and automate the extraction of bounding coordinates and other geospatial information. Researcher identification registries such as ORCID (http://orcid.org/) may help mitigate some of the challenges of finding up-to-date information about data set contributors that the encoders encountered in Phase I. Librarians and data managers can contribute to

automation by providing system and user requirements, identifying a minimal set of metadata elements, and encouraging other partners to become involved in data sharing initiatives.

In the future, we plan to enhance the CAPP framework by benchmarking other processes of metadata creation, such as the usability and effectiveness of certain tools and workflows, the impact of collaborations on metadata creation, and the effects of domain orientation or interdisciplinarity on the effectiveness and completeness of metadata. At its current early stage, CAPP framework is a proposition that needs to be developed into a rich research agenda. We hope that our framework will be considered by the Dublin Core community for further development, testing, improvement, and incorporation into the set of best practices for metadata creation.

5. References

Federal Geographic Data Committee [FGDC]. (1998). FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C. Retrieved April 4, 2013, from http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf.

Federal Geographic Data Committee [FGDC]. (2000). Content standard for digital geospatial metadata (For use with FGDC-STD-001-1998). Federal Geographic Data Committee. Washington, D.C. Retrieved April 4, 2013, from http://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf.

Federal Geographic Data Committee [FGDC]. (2011). Preparing for International Metadata. Retrieved April 4, 2013 from http://www.fgdc.gov/metadata/documents/preparing-for-international-metadata-guidance.pdf.

Federal Geographic Data Committee [FGDC]. (2012). Geospatial Metadata. Retrieved from http://www.fgdc.gov/metadata/index_html.

Hedstrom, M., Alter, G., Kumar, P., Kouper, I., McDonald, R. H., Myers, J., & Plale, B. (2013). SEAD: An Integrated Infrastructure to Support Data Stewardship in Sustainability Science. *CASC Research Data Management Implementation Symposium*. Arlington, VA. doi: 10.6084/m9.figshare.651719.

National Center on Earth-surface Dynamics [NCED]. (n.d.). Retrieved April 4, 2013, from https://repository.nced.umn.edu/.

NISO. (2004) Understanding Metadata. Bethesda, MD: NISO Press. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Olsen, L.M., Major, G., Shein, K., Scialdone, J., Ritz, S., Stevens, T., Morahan, M., Aleman, A., Vogel, R., Leicester, S., Weir, H., Meaux, M., Grebas, S., Solomon, C., Holland, M., Northcutt, T., Restrepo, R.A., & Bilodeau, R. (2012). NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 7.0.0.0.0 Retrieved April 4, 2013 from http://gcmd.nasa.gov/learn/keyword_list.html.

Qin, J., Ball, A., & Greenberg, J.. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Available at http://dcpapers.dublincore.org/pubs/article/view/3660.

SEAD: Sustainable Environment – Actionable Data. (n.d.). Retrieved from http://sead-data.net.