Approaches to Building Metadata for Data Curation

Hsueh-hua Chen
NTU Library, Taiwan
sherry@ntu.edu.tw

Yu Lin
NTU Library, Taiwan
b94106038@gmail.com

Cynthia Chen
NTU Library, Taiwan
cynthiachen@ntu.edu.tw

Abstract

In National Taiwan University (NTU), the Library aims to provide data curation services for university researchers from different research fields, particularly focusing on those from small sciences. In this paper, we will first investigate existing metadata schemas used for data curation services in North America and Europe. Next, we will attempt to develop an application profile, proposing metadata fields to be applied to data curation services in NTU. Finally, we will discuss our findings in this study, and take further action to develop a repository platform.

Keywords: data curation; metadata; academic library

1. Introduction

Data curation can help researchers maintain, manage, preserve, and add value to data throughout its lifecycle, with the goal of providing for its re-use over time (Digital Curation Centre, 2010). While the importance of data curation has been much established in North America and Europe, such services are not yet fully implemented in Taiwan.

In National Taiwan University (NTU), the Library aims to provide data curation services for university researchers across different research fields. While data needs for some of the big sciences are met outside the library, there are other fields, especially small sciences, in which data curation is badly needed (Cragin, Palmer, Carlson & Witt, 2010). Responding to the need for dataset management, NTU Library has formed a team to explore possible actions in order to provide data curation services in the future.

For the purpose of this paper, we focus on the metadata aspects of data curation. Data curation services rely on good metadata practices, with which researchers would be able to retrieve, identify, access and re-use data for new research (Walters & Skinner, 2011). Therefore, in this paper, we will attempt to develop an application profile for collection-level metadata describing datasets that contain primary research data. Our goal is to develop one application profile that will meet the needs of researchers from various different disciplines.

2. Methods

Ideally, data curation would manage data throughout its lifecycle, starting from concept, to data collection, processing, preservation and eventually re-purposing. For this study, we will focus on the aspects of data preservation, access and re-use.

The preliminary design of our data curation service appears as follows: With the repository framework provided by NTU Library, researchers from any field may register with the repository and submit the primary data of their concluded studies. They will be asked to provide collection-level information, in which they describe the backgrounds and purposes of their research.

With data coming from various fields of research, the metadata would have to be applicable across different disciplines. We also wish to maintain high interoperability with existing and future metadata standards, and choose to build on Dublin Core metadata.

In consideration of the sensitive, unpublished nature of some research data, we understand that researchers are particularly concerned with how data is accessed, and by whom. Therefore, the

design of metadata for data curation would also focus on intellectual property rights and access permissions (Buneman, Müller & Rusbridge, 2009).

2.1. Existing Metadata Schemas

Taking such concerns into account, we begin to create an application profile that is cross-disciplinary, highly interoperable, and flexible in rights and access control (ANDS, 2011, pp. 9-10). To do this, we first look to three existing metadata schemas used for data curation, including DataShare Profile (Rice, Macdonald & Hamilton, 2008), DataStaR minimum metadata (Dietrich, 2010) and DataCite Metadata Schema (Starr, 2011). These are selected due to their use in data curation and application across different scientific fields (Greenberg, 2010, pp. 75-78; Ball, 2011).

We make comparisons to the three projects based on several aspects, including their scope, scale, source of funding, number of participating organizations, types of data collected, number of entries, user interface, rights and access properties, and many others.

Next, we begin map and compare fields from these three metadata schemas. Even though the three data curation projects differ in scale and scope, they are similar in many ways, such as supporting Dublin Core elements and focusing on data discovery. After analyzing and integrating fields from the three existing standards, we end up with 22 fields that are appropriate for our project.

Our list of metadata elements needed for the project include: Title, Alternative Title, Creator, Contributor, Publisher, Dataset Description, Item Description, Type, Format, Size, Subject, Coverage-geographic, Coverage-temporal, Available Date, Date, Language, Source, Relation, Rights, Access permissions: metadata, Access permissions: download item, and Identifier.

2.2. Interviews and Revisions

In order to assess the usefulness of the proposed 22 metadata fields, we conducted interviews with 12 professors from NTU, who are from various different backgrounds including anthropology, social work, biochemistry, applied physics, atmospheric sciences, geology, geography, etc. As part of the preparations for the interview sessions, we created 13 metadata entries, each describing a dataset from the professor's field of research.

The interview process is outlined as follows: First, we explain the concept of metadata for data curation and the purpose of this study. The interviewees are asked to briefly introduce their work, how data is produced and used, and whether data repositories already exist in their fields. If yes, data repositories already exist in their fields, then the interview process ends at this point. If no, we continue on to discuss the aforementioned 13 metadata entries, during which the interviewees are asked to confirm whether information is filled in accurately, using correct terminology, and to see whether the usage guidelines are clear and comprehensible. Finally, we ask the interviewees for other suggestions on metadata for data curation.

Based on their comments, we revise and make several changes to the original 22 fields. Some examples of these revisions include: indicating required fields for manual input, adding more examples for geographic and temporal descriptions, etc. The results are shown in the next section of this paper.

3. Results

Following is a list of the revised metadata elements. Due to page limits, detailed usage guidelines are omitted from this table. The gray-shaded rows indicate system-generated information, while required elements are marked with an asterisk (*).

TABLE 1: Proposed metadata elements for research datasets

Label	Property	Definition
Title*	dc: title	A name given to the dataset.
Alternative Title	dcterms: alternative	An alternative name for the dataset.
Creator*	dc: creator	A person primarily responsible for making the research data. A person who conducted new research based on previously collected data. Listed according to priority.
Contributor	dc: contributor	An entity responsible for making contributions to the dataset. Examples of a Contributor include creators of the metadata, the funding organization, a person involved in the collection of research data.
Publisher	dc: publisher	A person, organization, or service responsible for making the dataset publicly available.
Dataset Description	dc: description	An abstract describing the research the dataset belongs to, or other information that cannot be described in other fields.
Item Description	dc: description	Names, descriptions, version number of the items included in the dataset. If this information is written in another text file, the name of the file has to be included.
Type	dc: type	The types of the data included in this dataset, using DCMI terms.
Format	dc: format	Automatically generated. The file format of the data included in the dataset. This field can also be input manually.
Size	dcterms: extent	Automatically generated. The file size of the dataset.
Subject*	dc: subject	Keywords describing the topic of the data.
Coverage-geographic	dcterms: spatial	The location and country that best describes where the included data belongs to.
Coverage-temporal	dcterms: temporal	The time range of the included data. Examples include the start and end date of data creation, or a single time and date.
Available Date*	dc: date	The date when the data becomes available to the public.
Date	dc: date	System generated dates related to the usage of metadata. Submitted Date Accepted Date Updated Date
Language*	dc: language	The language used in the primary data.
Source	dc: source	Name of the source of the data.
Relation	dc: relation	Describe relations to other resources.
Rights*	dc: rights	Statement of intellectual property rights. Links to online copyright statements can be put here, or any other information related to rights, including information about rights held in and over the resource.
Access permissions: metadata*	dc: rights	People or organizations that are permitted access to the metadata.
Access permissions: download item*	dc: rights	People or organizations that are permitted to download item-level data.
Identifier	dc: identifier	Automatically generated. Independent sequence of the data.

4. Conclusions and Future Actions

During the process of building metadata for data curation, we discovered that the level of detail required for each metadata element varied greatly between different disciplines. For example, researchers in atmospheric sciences, geography and geology often conduct field surveys, and would require detailed descriptions of geographic locations and temporal records. However, such

information is sometimes unavailable or unnecessary in other sciences, such as physics and engineering.

Therefore, we revised our usage guidelines for the Coverage-geographic and Coverage-temporal fields to reflect this disparity. Researchers may enter N/A in these fields if such data is not applicable. We also put in more examples based on usual practices in different fields, which would be easier for researchers to follow.

Second, we originally designed our Format field to be automatically generated by the system. However, some research datasets contain very complicated file formats, and might not be successfully machine-harvested. Therefore, we open this metadata field to be both automatically generated and manually input by researchers.

In addition to the technical aspects, on a concept level, we had some difficulties communicating the differences between collection-level descriptions and item-level descriptions. In the interviews, researchers often confused our metadata with item-level metadata. Therefore, the concept and advantages of collection-level metadata would have to be promoted and understood by researchers before fully implementing data curation services.

On a broader level, data repository tools would have to be developed to successfully achieve data curation. The application profile developed in this paper would also have to be revised according to developments of the repository platform.

Currently, we are developing a repository platform for data curation in NTU. We have studied data curation frameworks such as DSpace and Fedora, which are used by Edinburgh DataShare and DataStaR at Cornell University. As there is no all-in-one solution for data curation services, we intend to develop a repository platform according to our needs. The new data curation platform will implement the metadata developed in this paper, and the metadata will be further revised according to user feedback and new developments in data curation.

References

Australian National Data Service (2011). *ANDS Guides –Metadata (Working level).* ANDS. Retrieved from ANDS Website: http://www.ands.org.au/guides/metadata-working.pdf

Ball, A. (2011, September). *Overview of scientific metadata for data publishing, citation, and curation.* Paper presented at the Eleventh International Conference on Dublin Core and Metadata Applications (DC-2011). Den Haag, Netherlands. Retrieved from http://wiki.dublincore.org/images/7/72/Scientific-metadata-ajb.pdf

Buneman, Peter, Heiko Müller and Chris Rusbridge (2009). Curating the CIA world factbook. *The International Journal of Digital Curation, 3*(4), 29-43.

Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical. Physical and Engineering Sciences, 368*(1926), 4023-4038.

Dietrich, Dianne (2010). Metadata Management in a Data Staging Repository. *Journal of Library Metadata, 10*(2). doi: 10.1080/19386389.2010.506376

Digital Curation Centre (2010). *Digital Curation Centre.* Retrieved from http://www.dcc.ac.uk/

Greenberg, Jane (2010). Metadata for Scientific Data: Historical Considerations, Current Practice, and Prospects. *Journal of Library Metadata, 10*(2-3), 75-78.

Rice, Robin, Stuart Macdonald and George Hamilton (2008). *Applying DC to institutional data repositories: DSpace metadata for Edinburgh DataShare*. Retrieved from http://dc2008.de/wp-content/uploads/2008/10/12_rice_poster.pdf

Starr, Joan, et al. (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data*, Retrieved from http://schema.datacite.org/meta/kernel-2.1/doc/DataCite-MetadataKernel_v2.1.pdf

Walters, Tyler and Katherine Skinner. (2011) *New Roles for New Times: Digital Curation for Preservation.* Retrieved from http://www.arl.org/bm~doc/nrnt_digital_curation17mar11.pdf