Approaches to Building Metadata for Data Curation

Hsueh-hua Chen  陳雪華
University Librarian
Professor of Library and Information Science
National Taiwan University
sherry@ntu.edu.tw
Location of TAIWAN
Location of TAIWAN
Portuguese sailors, passing Taiwan in 1544, first jotted in a ship's log the name of the island Ilha Formosa, meaning "Beautiful Island". In 1582 the survivors of a Portuguese shipwreck spent ten weeks battling malaria and aborigines before returning to Macau on a raft.
TAIWAN:

- Land: 36K km$^2$ (1/200 of Australia)
- Population: 23M similar to Australia
- Land: 7,702K km$^2$
TAIWAN: A small island with multiple cultures

- A melting pot of multiple cultures, including:
  - Chinese (Min-nan, Hakka…)
  - Indigenous Peoples
  - Immigrants from Vietnam, Indonesia, Thailand…
Because of the different ethnic groups well integrated in Taiwan, the rich historical background has provided Taiwan with a multifaceted culture.
Surrounded by the sea and dominated by high mountains, Taiwan has been generously endowed by nature.
Welcome to visit Taiwan!
Approaches to Building Metadata for Data Curation

Hsueh-hua Chen
University Librarian
Professor of Library and Information Science
National Taiwan University

2013/09/05
Introduction

- Need for data curation in National Taiwan University (NTU).
- NTU Library aims to develop and provide data curation services.
- Users: for university faculty and researchers.
- Data collected: research data of studies from different disciplines, especially “small sciences.”
- Based on DC metadata standards.
- Goal: develop one application profile that will suit different needs.
Steps to creating metadata

1. Investigate user needs
2. Set out a metadata plan
3. Find and choose appropriate metadata fields
4. Test and create metadata entries
5. Revise and propose metadata fields
Data Curation can...

- Help researchers maintain, manage, preserve and add value to data throughout its lifecycle.
- Relies on metadata, which allows researchers to identify, retrieve, and access research data.
- Goal: the discovery and reuse of data over time.
- For this study, we focus on the aspects of data preservation, access and reuse.
Preliminary design of data curation services in NTU Library

Role of the Library:
- Develop metadata profile
- Provide repository framework

Researchers from any field may:
- Register with the repository
- Submit primary data of concluded studies
- Provide background information of their research datasets
Concerns when designing metadata

1. Metadata must be applicable across different research domains.
2. Maintain high interoperability with existing and future metadata standards.
3. Control over intellectual property rights and access permissions.
4. Focus on “collection-level” instead of “item-level” descriptions.
Existing Metadata Schemas

1. DataShare Profile
2. DataStaR minimum metadata
3. DataCite Metadata Schema

Three metadata schemas selected due to:
1. Their use in data curation.
2. Their application across diverse scientific fields.
## Project Comparisons

<table>
<thead>
<tr>
<th></th>
<th>DataShare</th>
<th>DataStaR</th>
<th>DataCite</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Institution</strong></td>
<td>University of Edinburgh</td>
<td>Cornell University</td>
<td>DataCite</td>
</tr>
<tr>
<td><strong>Funding</strong></td>
<td>Joint Information Systems Committee, JISC</td>
<td>National Science Foundation, NSF</td>
<td>DataCite</td>
</tr>
<tr>
<td><strong>Number of participating organizations</strong></td>
<td>5</td>
<td>1</td>
<td>20 organizations from 10 countries</td>
</tr>
<tr>
<td><strong>Number of entries</strong></td>
<td>Not available</td>
<td>55 datasets</td>
<td>800,000 entries</td>
</tr>
<tr>
<td><strong>Subjects</strong></td>
<td>Sciences (multi-discipline)</td>
<td>Sciences (multi-discipline)</td>
<td>Sciences (multi-discipline)</td>
</tr>
<tr>
<td><strong>Interoperability</strong></td>
<td>DCMI Metadata terms</td>
<td>DC, FGDC-CSDGM, EML, OLAC</td>
<td>DC Application Profile</td>
</tr>
</tbody>
</table>
Compare and integrate fields

- Map and combine fields from the three metadata schemas.
- Results: 22 metadata fields
- 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.
Interviews

• To assess usefulness of the 22 metadata fields.
• Interviews conducted with 12 NTU professors.
• Research backgrounds: anthropology, social work, biochemistry, applied physics, atmospheric sciences, geology, geography, etc.
• 13 metadata entries are created beforehand: each describing a dataset from the professor’s field of research.
Interview Process

1. Create metadata for research datasets
2. Explain the concept of metadata and purpose of study
3. Ask interviewees to introduce their work, how data is used and produced, status of data repositories
4. Data repositories exist in their fields
5. Data repositories do not exist in their fields
6. Discuss and confirm the pre-created metadata entries
7. Ask for further suggestions
8. Revise original fields based on interview comments
9. End interview
Discussion (1)

• Discovery: the level of detail required for each metadata element varied greatly between different disciplines.
• Example: atmospheric sciences, geography and geology require detailed geographic locations and temporal records.
• Not needed in physics and engineering.
Revision (1)

• Revised usage guidelines for fields Coverage-geographic and Coverage-temporal.
• Researchers may enter N/A in these fields if such data is not applicable.
• More examples given based on usual practices in different scientific domains.
• Format field originally designed to be automatically generated by the system.
• Some research datasets contain very complicated file formats, and might not be successfully machine-harvested.
Revision (2)

- Revised and opened format field to be both automatically generated and manually input by researchers.
Discussion (3)

- Some difficulties communicating the differences between collection-level descriptions and item-level descriptions.
- The concept and advantages of collection-level metadata would have to be promoted and understood by researchers before fully implementing data curation services.
# Results (1/3)

<table>
<thead>
<tr>
<th>Label</th>
<th>Property</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title*</td>
<td>dc: title</td>
<td>A name given to the dataset.</td>
</tr>
<tr>
<td>Alternative Title</td>
<td>dcterms: alternative</td>
<td>An alternative name for the dataset.</td>
</tr>
<tr>
<td>Creator*</td>
<td>dc: creator</td>
<td>A person primarily responsible for making the research data. A person who conducted new research based on previously collected data. Listed according to priority.</td>
</tr>
<tr>
<td>Contributor</td>
<td>dc: contributor</td>
<td>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.</td>
</tr>
<tr>
<td>Publisher</td>
<td>dc: publisher</td>
<td>A person, organization, or service responsible for making the dataset publicly available.</td>
</tr>
<tr>
<td>Dataset Description</td>
<td>dc: description</td>
<td>An abstract describing the research the dataset belongs to, or other information that cannot be described in other fields.</td>
</tr>
<tr>
<td>Item Description</td>
<td>dc: description</td>
<td>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.</td>
</tr>
<tr>
<td>Label</td>
<td>Property</td>
<td>Definition</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Type</td>
<td>dc: type</td>
<td>The types of the data included in this dataset, using DCMI terms.</td>
</tr>
<tr>
<td>Format</td>
<td>dc: format</td>
<td>Automatically generated. The file format of the data included in the dataset. This field can also be input manually.</td>
</tr>
<tr>
<td>Size</td>
<td>dcterms: extent</td>
<td>Automatically generated. The file size of the dataset.</td>
</tr>
<tr>
<td>Subject*</td>
<td>dc: subject</td>
<td>Keywords describing the topic of the data.</td>
</tr>
<tr>
<td>Coverage-geographic</td>
<td>dcterms: spatial</td>
<td>The location and country that best describes where the included data belongs to.</td>
</tr>
<tr>
<td>Coverage-temporal</td>
<td>dcterms: temporal</td>
<td>The time range of the included data. Examples include the start and end date of data creation, or a single time and date.</td>
</tr>
<tr>
<td>Available Date*</td>
<td>dc: date</td>
<td>The date when the data becomes available to the public.</td>
</tr>
<tr>
<td>Date</td>
<td>dc: date</td>
<td>System generated dates related to the usage of metadata. Submitted Date Accepted Date Updated Date</td>
</tr>
<tr>
<td>Label</td>
<td>Property</td>
<td>Definition</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Language*</td>
<td>dc: language</td>
<td>The language used in the primary data.</td>
</tr>
<tr>
<td>Source</td>
<td>dc: source</td>
<td>Name of the source of the data.</td>
</tr>
<tr>
<td>Relation</td>
<td>dc: relation</td>
<td>Describe relations to other resources.</td>
</tr>
<tr>
<td>Rights*</td>
<td>dc: rights</td>
<td>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.</td>
</tr>
<tr>
<td>Access permissions: metadata*</td>
<td>dc: rights</td>
<td>People or organizations that are permitted access to the metadata.</td>
</tr>
<tr>
<td>Access permissions: download item*</td>
<td>dc: rights</td>
<td>People or organizations that are permitted to download item-level data.</td>
</tr>
<tr>
<td>Identifier</td>
<td>dc: identifier</td>
<td>Automatically generated. Independent sequence of the data.</td>
</tr>
</tbody>
</table>
Future Actions

• A repository platform for data curation currently under development.

• Studies made of data curation frameworks, such as DSpace and Fedora, used by Edinburgh DataShare and DataStaR at Cornell University.

• The new data curation platform to implement the metadata developed in this paper.

• The metadata to be further revised according to user feedback and new developments in data curation.
Thank You!

sherry@ntu.edu.tw