A brief overview of metadata for datasets

Sally Rumsey
The Bodleian Libraries
University of Oxford
Structure

- Citation
- Discovery
- Documentation and re-use
- Timing of metadata creation
- WIP at Oxford
1. Citation: Why?

- Attribution
  - Recognition
  - Reward
  - Impact
- Discovery
  - Referencing
  - Avoid repetition
  - Potential collaboration
- Location
  - For access and re-use
  - Includes non-digital data
- Access
  - Verification
- Re-use
  - ‘Standing on the shoulders of giants’
  - Documentation of methods and descriptions of data
- Administration
Citation: Key documents/principles

- **DataCite** organisation [https://www.datacite.org/](https://www.datacite.org/)

- **Force11** Joint Declaration of Data Citation Principles. 8 principles covering purpose, function and attributes of citations. [https://www.force11.org/datacitation](https://www.force11.org/datacitation)

- **DCC** How to Cite Datasets and Link to Publications [http://www.dcc.ac.uk/resources/how-guides/cite-datasets](http://www.dcc.ac.uk/resources/how-guides/cite-datasets)

- **Out of cite, out of mind**: The Current State of Practice, Policy, and Technology for the Citation of Data. CODATA-ICSTI Task Group on Data Citation Standards and Practices, Yvonne M. Socha (Ed). CODATA *Data Science Journal*. 13 September 2013 [https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article](https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article)
1.3 The Metadata Schema

The DataCite Metadata Schema is a list of core metadata properties chosen for the accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. The resource that is being identified can be of any kind, but it is typically a dataset.

Table 1: DataCite Mandatory Properties

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
<th>Obligation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifier (with type sub-property)</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Creator (with name identifier sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>Title (with optional type sub-properties)</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>Publisher</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>PublicationYear</td>
<td>M</td>
</tr>
</tbody>
</table>
Minimum reference

Creator (PublicationYear): Title. Publisher. Identifier
Deluxe reference

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier
2. Discovery

• In addition to 5 mandatory elements, how might an interested person search for a dataset?

• May include (not exhaustive):
  – Subject/keywords
  – Resource type
  – Description text/abstract
  – GeoLocation
  – Methodology
3. Documentation & Re-use

EPSRC policy framework on research data: Principles
http://www.epsrc.ac.uk/about/standards/researchdata/Pages/principles.aspx

“Sufficient metadata should be recorded and made openly available to enable other researchers to understand the potential for further research and re-use of the data. Published results should always include information on how to access the supporting data.”
When are the metadata created?

- DMP
- Active data phase
- Stable version for archiving
  - Third party archive
  - Local institutional archive
- Harvesting
- Editing, enhancing and updating
WIP at Oxford

- Library perspective: data archiving and cataloguing
- ORA-Data branding (Oxford University Research Archive)
- Phase 1: Basic service
- Hydra implementation http://projecthydra.org/
  - Work completed under JISC Damaro project
  - Checking all elements/fields
  - Consistency with other item types
- Metadata creation/harvest
  - Expect mixed model
  - Low entry barrier
  - Training/advocacy on benefits of rich metadata
- Integration with CUD (Core User Directory)
- FAST subject headings
- Work in parallel to service set-up
- Funder compliance as main driver for University