Data description and Repository Selection

Nanyang Technological University

Jonathan Rans

Digital Curation Centre

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License.
Data about data?

• Citation
• Discovery
• Reuse
Metadata vs documentation

 Gestures

 Metadata
  » Standardised
  » Structured
  » Machine and human readable

Documentation

Metadata
What is the minimum required?

- Repository requirements
- Could be lead by DataCite
- Citation/disambiguation
  - Identifier
  - Creator
  - Title
  - Publisher
  - Publication Year
- Licencing/access conditions
What are persistent identifiers?

- They are an alphanumeric code identifying a resource, organisation or individual
- They must be
  - Unique
  - Persistent
- Ideally they should be actionable too
How do persistent identifiers work
Licencing and access

It is good practice for datasets to carry information on access and any restrictions or conditions that apply.
Aiding discoverability

- Catalogue or discovery metadata
- Structured so that search engines can uncover it.
- Must be exposed in machine-readable form eg XML
- OAI-PMH?
Controlled vocabularies

- E.g. SNOMED CT (clinical terms) or MeSH
- Include ontologies as well
  - Defined terms + taxonomy
- Useful for selecting keywords to tag datasets
Ensuring the utility of the data

-The what, why and how data creation must be understood
-Data dictionaries
-Columns/rows labelled
-Variable ranges defined
Metadata standards

- These can be general – such as Dublin Core
- Or discipline specific
  - Data Documentation Initiative (DDI)
    - Social Sciences
  - Ecological Metadata Language (EML)
    - Ecology
  - Flexible Image Transport System (FITS)
    - Astronomy
Example extract from a UK Data Service DDI catalogue record

This project aims to develop knowledge and understanding of the contemporary globalization of the headhunting industry in Europe and its implications for new forms and geographies of executive search and selection. Europe has become the most complex and sophisticated pan-regional market for executive search, fuelled by free labour mobility within the EU, thereby offering a unique environment in which to study the changing practices of the headhunting industry.

https://www.ukdataservice.ac.uk/manage-data/document/metadata
DCC metadata catalogue

The catalogue lists:

- Metadata standards
- Profiles
- Use cases
- Tools

http://www.dcc.ac.uk/drupal/resources/metadata-standards
Metadata can have a variety of sources

- Publication underlying the data
- Manually added by researcher
  - Folder structure
  - File names
- Grant information from CRIS
- Contextual information generated by equipment
We recommend that a ReadMe be a plain text file containing the following:

• for each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication
• for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units
• any data processing steps, especially if not described in the publication, that may affect interpretation of results
• a description of what associated datasets are stored elsewhere, if applicable
• whom to contact with questions
DATA DOCUMENTATION AND METADATA

Getting the basics down

- Who? Who contributed to the project (authors, research assistants, etc.)?
- What? What kind(s) of data and analysis were used?
- When? When was the data collected? When was analysis performed? Any other pertinent dates?
- Where? Does the project involve a particular geographic area, such as the state of Minnesota, or the Twin Cities, or Antarctica?
- Why? What is the impetus for the project? What questions are you trying to answer?

Getting a little more in-depth

Imagine that you have to leave the project as is for a couple months and then come back to it. What are the most important aspects of the project you’d need help remembering? Some examples:

- file handling (how are they named, how are they divided)
- processing steps (how to get from point A to B)
- field abbreviation/name glossary (now what does ABC130 stand for again?)

Now imagine if you had to leave the project and come back after six months or a year. What else would you add to the list?

Need Help? Download an Example Readme.txt (plain text file)
A metadata record can be the accessible, online representation of a dataset that is:

» Physical
  » Work of art
  » Biological sample
  » Geological sample

» In digital form but too large to access via internet
  » EPSRC datasets

» Too sensitive to make accessible online
  » Patient data
An example from UAL


Type of Research: Performance

Creators: Lazell, Richard

Description: A performance art piece shown in conjunction with the installation at the Hardwick Gallery and as part of Acts Re-Acts 2 at Wimbledon College of Arts. This work was independently conceived of the exhibition and is stand-alone.

Additional Information (Publicly available): WP (Willow Pattern)

Your affiliations with UAL: Colleges > Wimbledon College of Arts

Date: 23 April 2015

Related Exhibitions: WP - liu hua

Locations / Venues:

<table>
<thead>
<tr>
<th>Location</th>
<th>From Date</th>
<th>To Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardwick Gallery, University of Gloucestershire, Cheltenham, UK</td>
<td>23 April 2015</td>
<td>UNSPECIFIED</td>
</tr>
<tr>
<td>The Wimbledon Space, Wimbledon College of Arts, London, UK</td>
<td>13 May 2015</td>
<td>UNSPECIFIED</td>
</tr>
</tbody>
</table>

Measurements of Duration of item: 60 minutes

Date Deposited: 03 Jul 2015 17:02

Last Modified: 04 Sep 2015 22:16

Item ID: 8062

URI: http://ualresearchonline.arts.ac.uk/id/eprint/8062
Supporting researchers

LSHTM’s six questions to consider
1. What legislative requirements must be met?
2. What contractual requirements must be met?
3. What is considered good practice in the subject domain?
4. What type of data are you working with?
5. What do you want to do with it?
6. Who are your target audience?

https://www.lshtm.ac.uk/research/researchdataman/describe/describing_data.html
The level of documentation required is likely to be proportional to the complexity of the data.

Ensuring values and terms are correctly defined is important.

Tools exist to simplify the creation of metadata.

Data which can’t be published in digital form can still be made visible.
Now the where...
Options for closed data

- Institutional data archive/vault
- Safe havens – (e.g. secure patient data)
- 3rd party data archiving
- Cloud storage
- Institutional servers – the ‘do nothing’ option

Apart from exceptional cases it should have a digital presence
Options for open data

- Domain repository
- General repository – Figshare, Zenodo, Dryad
- Institutional repository
- Journal supplementary material
- Departmental web page
Repository selection
General directories
Re3data.org

Domain specific directories
e.g. life sciences – Biosharing.org

Data journal recommendations
Edinburgh research data blog: Sources of dataset peer review

Funding body recommendations
E.g. Wellcome Trust Data repositories and database sources
A conversation with the researcher

- There may be an accepted repository used by peers or required by funders
- Multidisciplinary studies may not have an obvious home
- Data types and volumes will impact on decision
Journal of Open Psychology Data

Recommended Repositories

The following repositories meet our peer-review requirements and are recommended for the archiving of JOPD datasets. Please contact us if you would like to use another repository or recommend that we add it to our list.

<table>
<thead>
<tr>
<th>International repositories</th>
<th>National repositories</th>
<th>Institutional repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOPD Dataverse</td>
<td>DANS (Netherlands)</td>
<td>UCL Discovery</td>
</tr>
<tr>
<td>Dryad</td>
<td>Gesis (Germany)</td>
<td></td>
</tr>
<tr>
<td>Figshare</td>
<td>FORS (Switzerland)</td>
<td></td>
</tr>
<tr>
<td>OpenfMRI</td>
<td>Odum (United States)</td>
<td></td>
</tr>
<tr>
<td>Zenodo</td>
<td>SND (Sweden)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TARKI (Hungary)</td>
<td></td>
</tr>
</tbody>
</table>

Location: http://thedata.harvard.edu/dvn/dv/JOPD

Focus and suitability: Data can be uploaded to the JOPD Dataverse Repository designed specifically for papers in the Journal of Open Psychology Data. We recommended this repository to authors because it is managed entirely by the JOPD editorial staff and ensures maximum interoperability between datasets and data papers.

Cost: Free for all JOPD authors.

Licenses: CCO

Identifiers used: DOI
Finding a repository

- Lists over 1500 data repositories
- Icons for ease of assessment
- Supported by DataCite
Criteria for assessing repositories

- Is the repository reputable?
- Will it accept the data you want to deposit?
- Will data be safe in legal terms?
- Will the repository sustain data value?
- Will the repository support analysis and track data usage?
- Is it accredited?

Data Seal of Approval
ICSU World Data System
nestor seal
ISO 16363
An example of assessment

Zenodo
Thank-you – any questions?

Jonathan Rans
J.Rans@ed.ac.uk
@JNRans

Image credits:
Handles: https://www.clarin.eu/sites/default/files/handles.png
Snomed taxonomy: https://b2i.sg
Magnifying Glass: http://www.bestofmicro.com