Just enough metadata

Metadata for research datasets in institutional data repositories

Sally Rumsey
The Bodleian Libraries

Uses for metadata

- Citing (at package level)
- Discovery
- Compliance with funder requirements
- Explanatory: Additional information for potential users (assessment of usefulness)
- Preservation
- Reporting and business intelligence (internal & external)

Bodleian Libraries
UNIVERSITY OF OXFORD

Starting point: DataCite kernel

- Creator
- Title
- Date
- Publisher – auto generated
- ID – auto generated

- Fulfills citation and basic discovery requirements
- Does not fulfill other requirements

Bodleian Libraries
UNIVERSITY OF OXFORD

EPSRC principle vi

- “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”
- How much is ‘sufficient?’
- What should be mandatory in these circumstances?
- Using EPSRC roadmap as a starting point

Bodleian Libraries
UNIVERSITY OF OXFORD

Making things easy?

- Person information inc affiliation
  - Reliant on central and other systems
- Import data from DMPs
- Imported metadata – you get what you're given. Enhancement?
- Import data from equipment

Bodleian Libraries
UNIVERSITY OF OXFORD

Bodleian Libraries
UNIVERSITY OF OXFORD

Populating DataFinder

Sources of metadata

Manual entry

- Generally disliked
- Can lead to inaccuracies
- Can lead to richer metadata

Import existing

- Not much exists
- From data repositories (eg UKDA, Dryad)
- From central systems (eg RIM or DMP systems)
- From other systems (eg ROS)
- From machines that generate the data

Auto generated

- From DMP systems
- From DataStage

Minimum metadata set

- Mandatory
- Contextual
- Optional – including disciplinary

Recommended minimum core metadata set for Oxford [WIP]

Element		Auto Gen	DataCite	Note
Record/digital object ID		UUID	M	
Location of dataset	URL/ DOI	DataBank auto		If no URL: contact details
[Medium]	Default: digital (+ non-digital).			To enable indication of non-digital data. Check box + options. On/offline
Creator (if not depositor)	Repeatable	WebAuth/OxDMP	M	If depositor draw from WebAuth. (see optional)
Creator affiliation (if not depositor)	Repeatable (see optional)	WebAuth/OxDMP		If depositor draw from WebAuth; CUD; Imply subject
Title			M	
Publisher of data	Default University of Oxford	Default	M	
Publication year	Default current	Default	M	If an embargo period has been in effect, use the date when the embargo period ends.
Access terms & conditions	Default + options			
Data owner	Default Department	WebAuth/OxDMP		For curation; ALT Name (Person or role) + Data owner contact. + Qu 'Do you own the rights for this data?Need policy
Access date to data	Default current			To set embargo
Rights for metadata	Default: CC0? ODC?			
[Subject]	FAST + options			Import where possible using available data. Encourage imupt.+ K/w option. See Optional

Contextual mandatory metadata [WIP]

Element		Auto Gen	DataCite	EPSRC
Funding agency	Multiple	OxDMP		M
Grant number	Multiple	OxDMP		M
Project information	Link to project web page/blog			
Last access request date		Automatically determined		M
Source	If imported record	Automatically determined		
Source URL	If imported record	Automatically determined		
Data generation process	Text or link to paper/document			M
Why the data was generated/Abstract/ Brief description	Might be link to project page			M
Date	Repeatable; eg date (range) of data collection; format described in W3CDTF		O	M
Reason for embargo	Repeatable; List options			[M]

All manual depositors to be prompted "Do you have publications associated with this data?" Provide DOI, URL or location plus ORA link. See related identifier (optional)

Optional metadata – a selection

Element		Auto Gen	DataCite
Co-creators/contributor	Repeatable	OxDMP	O
Role	Repeatable		
Affiliation	Repeatable	OxDMP	
Sub-title			
Subject	Default FAST or discipline specific		O
Keywords	Free text		
Date (other)	Repeatable		O
Language	Default English		O
ResourceType			O
AlternateIdentifier	Eg DOI		O
RelatedIdentifier		eg DOI of publication	O
Size		System gen if DataBank	O
Format			O
Version			O
Data generation process			
Abstract/Brief description			
Documentation 1	Text/Link/URI	descriptive or contextual information about the dataset (e.g. machine settings and experimental conditions under which the data were gathered)	
Documentation 2	Text/Link/URI		
Subject specific m.d.	XML		
Subject specific m.d.	XML format		
Subject specific classification	Repeatable		
Subj specific classn scheme	Repeatable Populate with fixed values eg AMS; JEL		
Data complying with known standards eg DDI			

A metadata conundrum