



Introduction to RDM concepts and tools

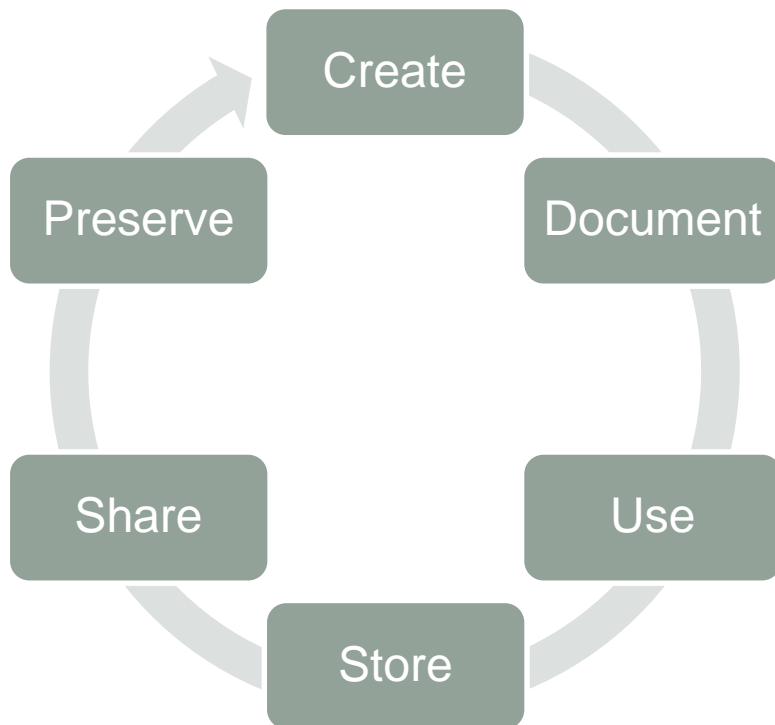
S. Venkataraman

DCC, University of Edinburgh

S.Venkataraman@ed.ac.uk

#fosteropenscience

What is Research Data Management?



“the active management and appraisal of data over the lifecycle of scholarly and scientific interest”

Data management is part of good research practice

RESEARCH DATA - OPEN BY DEFAULT



Concepts to cover

- Data formats
- Metadata
- Licensing
- Data repositories
- Persistent identifiers

These aspects are addressed specifically in Data Management Plans so will help you review them



Data formats

Different formats are good for different things

- open, lossless formats are more sustainable e.g. rtf, xml, tif, wav
- proprietary and/or compressed formats are less preservable but are often in widespread use e.g. doc, jpg, mp3

One format for analysis then
convert to a standard format

BioformatsConverter batch
converts a variety of proprietary
microscopy image formats to the
Open Microscopy Environment
format - OME-TIFF

Data centres may suggest preferred formats for deposit

www.data-archive.ac.uk/create-manage/format/formats-table

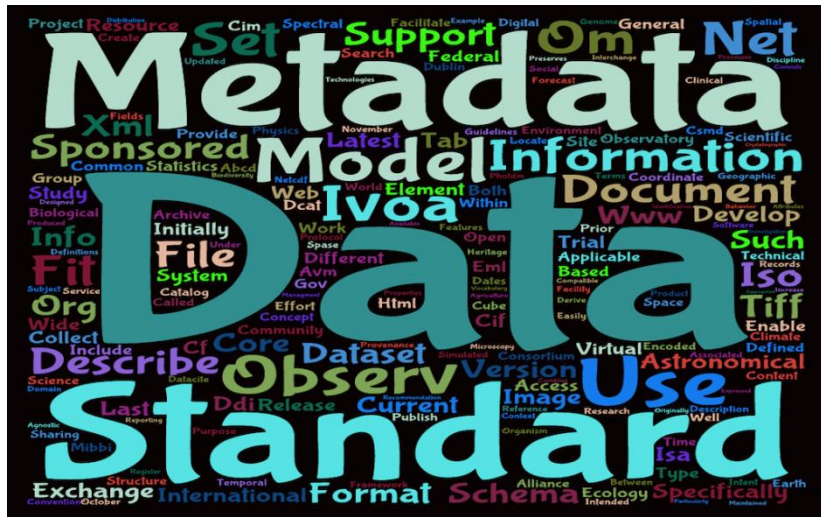
Metadata and documentation

- At a basic level, metadata supports data discovery, disambiguation and citation
- Rich metadata and documentation will support interoperability & reuse
- Standards should be used. These can be general - such as Dublin Core, or discipline specific
 - Data Documentation Initiative (DDI) - social science
 - Ecological Metadata Language (EML) - ecology
 - Flexible Image Transport System (FITS) - astronomy

Where to find relevant standards?

Metadata Standards Directory

Broad, disciplinary listing of standards and tools. Maintained by RDA group



<https://rdamsc.dcc.ac.uk>

FAIRsharing

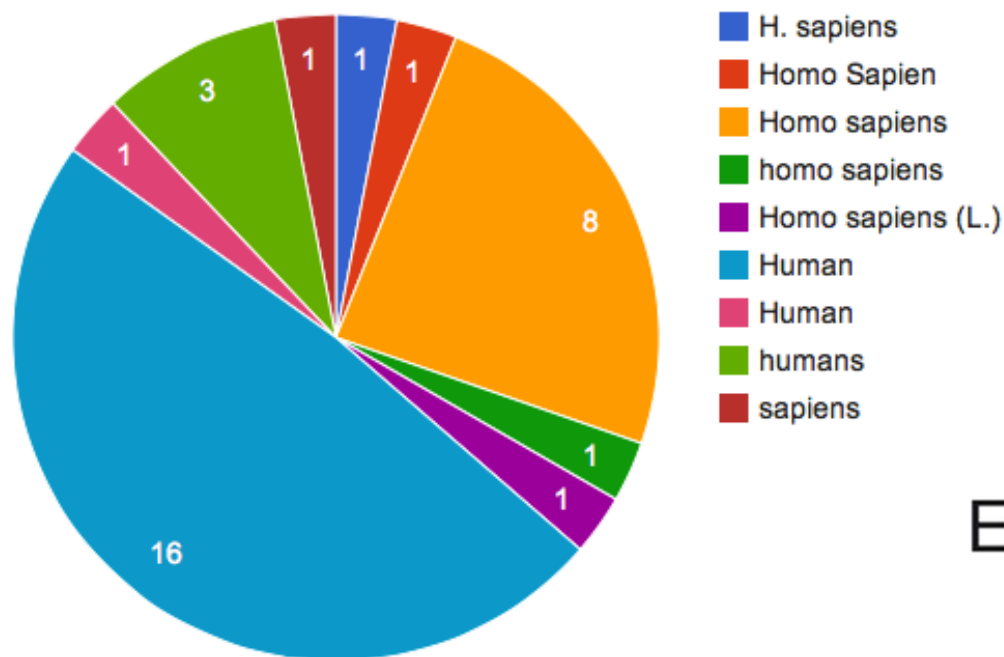
- A portal of data standards, databases, and policies
- Focused on life, environmental and biomedical sciences, but expanding to other disciplines



<https://fairsharing.org>

Value of controlled vocabularies

“MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in.....”



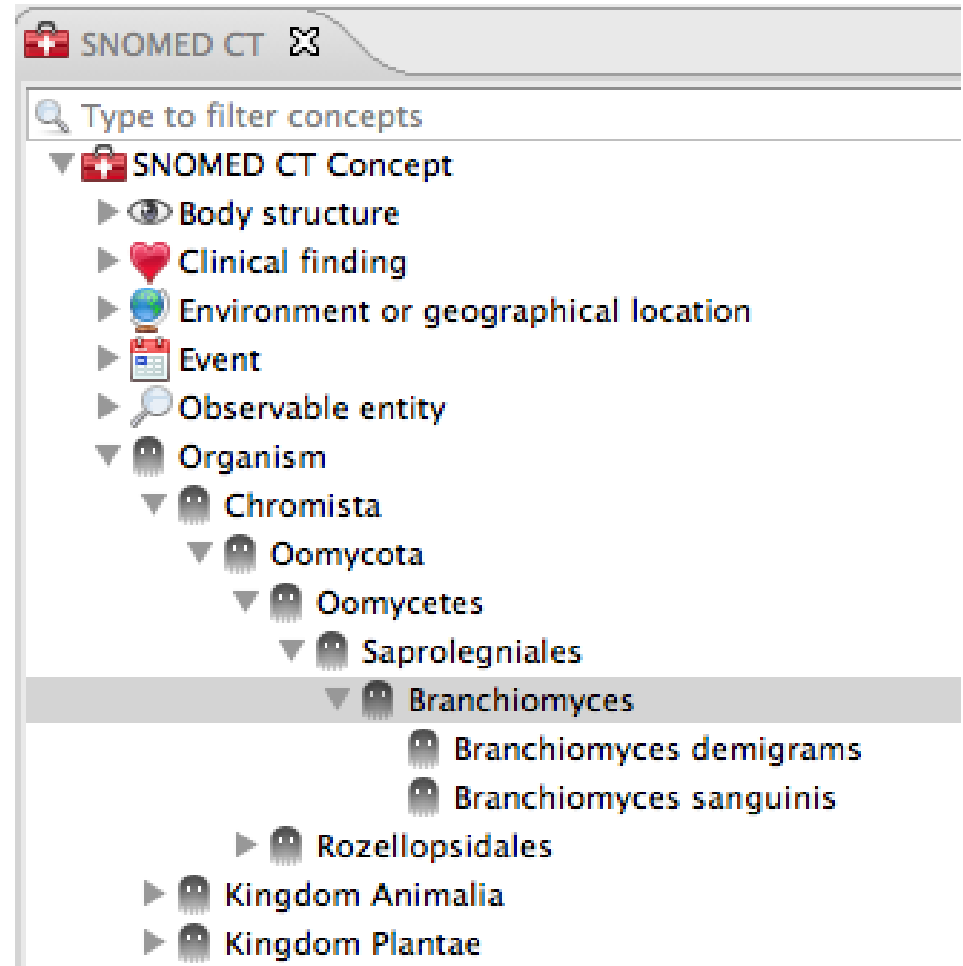
EMBL-EBI



Example courtesy of Ken Haug, European Bioinformatics Institute (EMBL-EBI)

Controlled vocabularies

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



Dataset licensing

Horizon 2020 guidelines point to:



or



CREATIVE COMMONS LICENSES		COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
	PUBLIC DOMAIN	✓	✗	✓	✓	✓
	CC BY	✓	✓	✓	✓	✓
	CC BY-SA	✓	✓	✓	✓	✗
	CC BY-ND	✓	✓	✓	✗	✗
	CC BY-NC	✓	✓	✗	✓	✓
	CC BY-NC-SA	✓	✓	✗	✓	✗
	CC BY-NC-ND	✓	✓	✗	✗	✗

You can redistribute (copy, publish, display, communicate, etc.)	You have to attribute the original work	You can use the work commercially	You can modify and adapt the original work	You can choose license type for your adaptations of the work.

EUDAT licensing tool

Answer questions to determine which licence(s) are appropriate to use

Do you own copyright and similar rights in your dataset and all its constitutive parts?

Yes

No

Do you allow others to make commercial use of you data?

Yes

No

Creative Commons Attribution (CC-BY)

This is the standard creative commons license that gives others maximum freedom to do what they want with your work.

Public Domain Dedication (CC Zero)

CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Data repositories

The EC guidelines point to Re3data as one of the registries that can be searched to find a home for data

The screenshot shows the re3data.org website interface. At the top, there is a search bar and navigation links for Search, Browse, Suggest, Resources, and Contact. A DataCite logo is also present. On the left, a 'Filter' sidebar lists various categories like Subjects, Content Types, Countries, etc. The main content area displays search results for 'UniProtKB/Swiss-Prot' and 'Khazar University Institutional Repository'. The UniProtKB/Swiss-Prot entry shows it is a manually annotated and reviewed section of the UniProt Knowledgebase, with subject(s) in Basic Biological and Medical Research and General Genetics, and content type(s) in Networkbased data, Structured graphics, and Plain text. The Khazar University Institutional Repository (KUIR) entry shows it is a suite of services offered by Khazar University in Azerbaijan, with subject(s) in Humanities and Social Sciences, Life Sciences, and Natural, and content type(s) in Standard office documents, Images, and Audiovisual data.

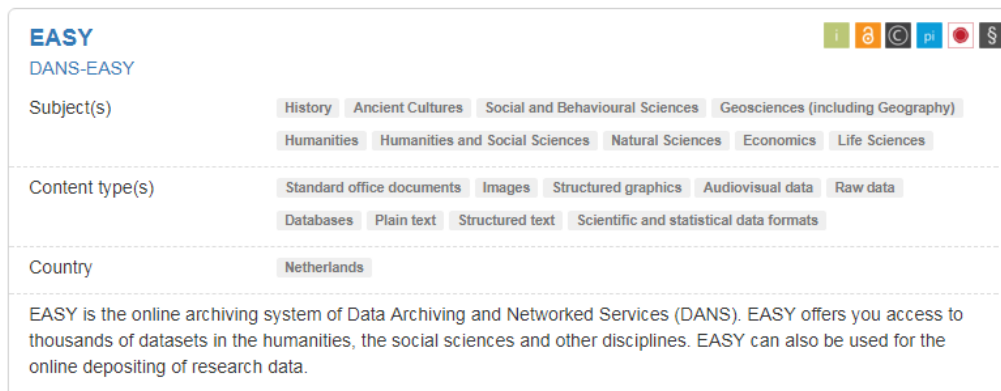


[www.fosteropenscience.eu/
content/re3data-demo](http://www.fosteropenscience.eu/content/re3data-demo)

www.re3data.org

Considerations when selecting repositories

- Often preferable to use a subject specific repository if available
- Useful if repositories assign a persistent identifier
- Look for certification as a '*Trustworthy Digital Repository*' with an explicit ambition to keep the data available in long term.
- Generic repositories are also available e.g. Zenodo or institutional repositories



EASY
DANS-EASY

Subject(s) History Ancient Cultures Social and Behavioural Sciences Geosciences (including Geography)
Humanities Humanities and Social Sciences Natural Sciences Economics Life Sciences

Content type(s) Standard office documents Images Structured graphics Audiovisual data Raw data
Databases Plain text Structured text Scientific and statistical data formats

Country Netherlands

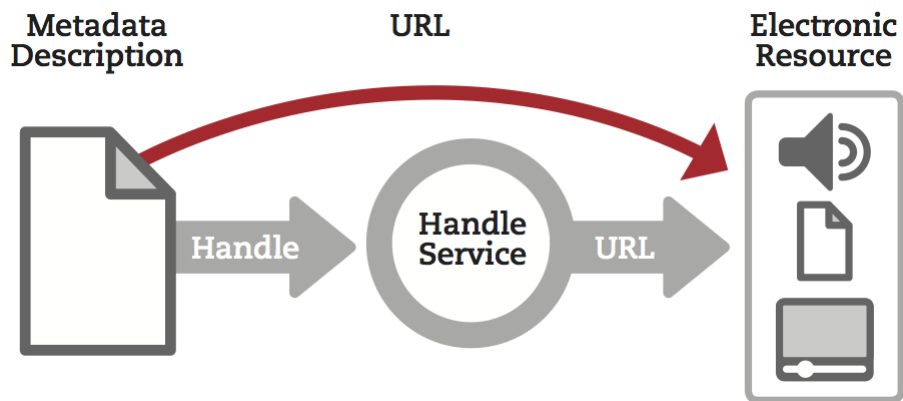
EASY is the online archiving system of Data Archiving and Networked Services (DANS). EASY offers you access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data.

Icons to note
open access,
licenses, PIDs,
certificates...

Persistent Identifiers

a long-lasting reference to a document, file or other object

- PIDs come in various forms e.g. ARK, DOI, URN, PURL, Handles...
- Typically they're actionable i.e. type it into web browser to access
- Many repositories will assign them on deposit



Publication date:
November 24, 2017

DOI:
DOI [10.5281/zenodo.1065991](https://doi.org/10.5281/zenodo.1065991)

Keyword(s):
FAIR, FAIRness, checklist, research data, Findable, Accessible, Interoperable, Reusable, PID, repository, DOI, metadata, licence, data sharing, research data management,

Grants:
European Commission:

- EUDAT2020 - EUDAT2020 (654065)

License (for files):
[Creative Commons Attribution 4.0](#)

Thanks - any questions?

Follow us on Twitter:

[@fosterscience](https://twitter.com/fosterscience) and [#fosteropenscience](https://twitter.com/#fosteropenscience)

FOSTER training events and materials:

www.fosteropenscience.eu/events

