Basics in good research data management (RDM) for reviewing DMPs

S. Venkataraman

Digital Curation Centre, Edinburgh

s.venkataraman@ed.ac.uk

https://doi.org/10.5281/zenodo.1461601

OpenAIRE

FOSTER

FOSTER & OpenAIRE webinar, 22nd October 2018
https://www.openaire.eu/open-access-week-2018

WHAT IS RESEARCH DATA MANAGEMENT?

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 here we will help you review them

Choose a appropriate file formats

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

Data centres may suggest preferred formats for deposit

https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

Data Formats

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Image data	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format GIF (.gif) TIFF other versions (.tif, .tiff) RAW image format (.raw) Photoshop files (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)
Audio data	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)
Video data	MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)	AVCHD video (.avchd)
Documentation and scripts	Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)	plain text (.txt) widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0

Document your data as fully as possible

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

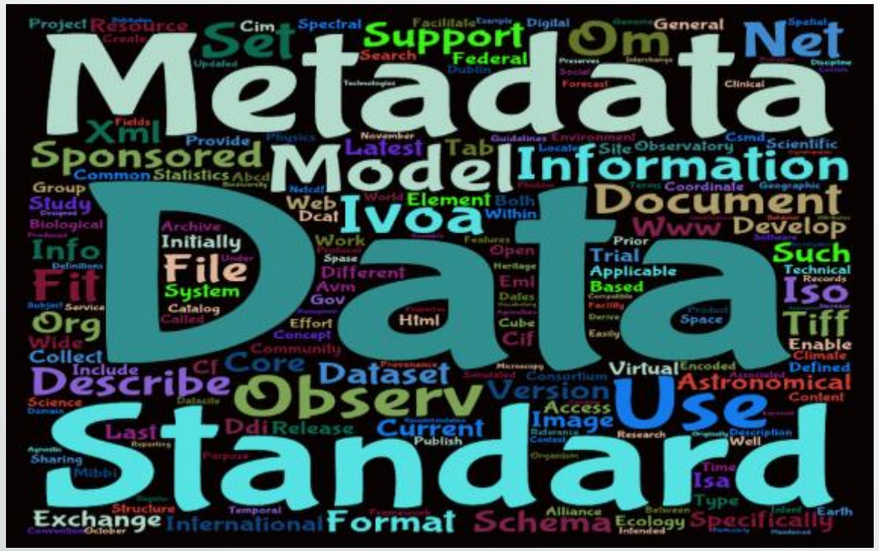
Ecological Metadata Language (EML) - ecology

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

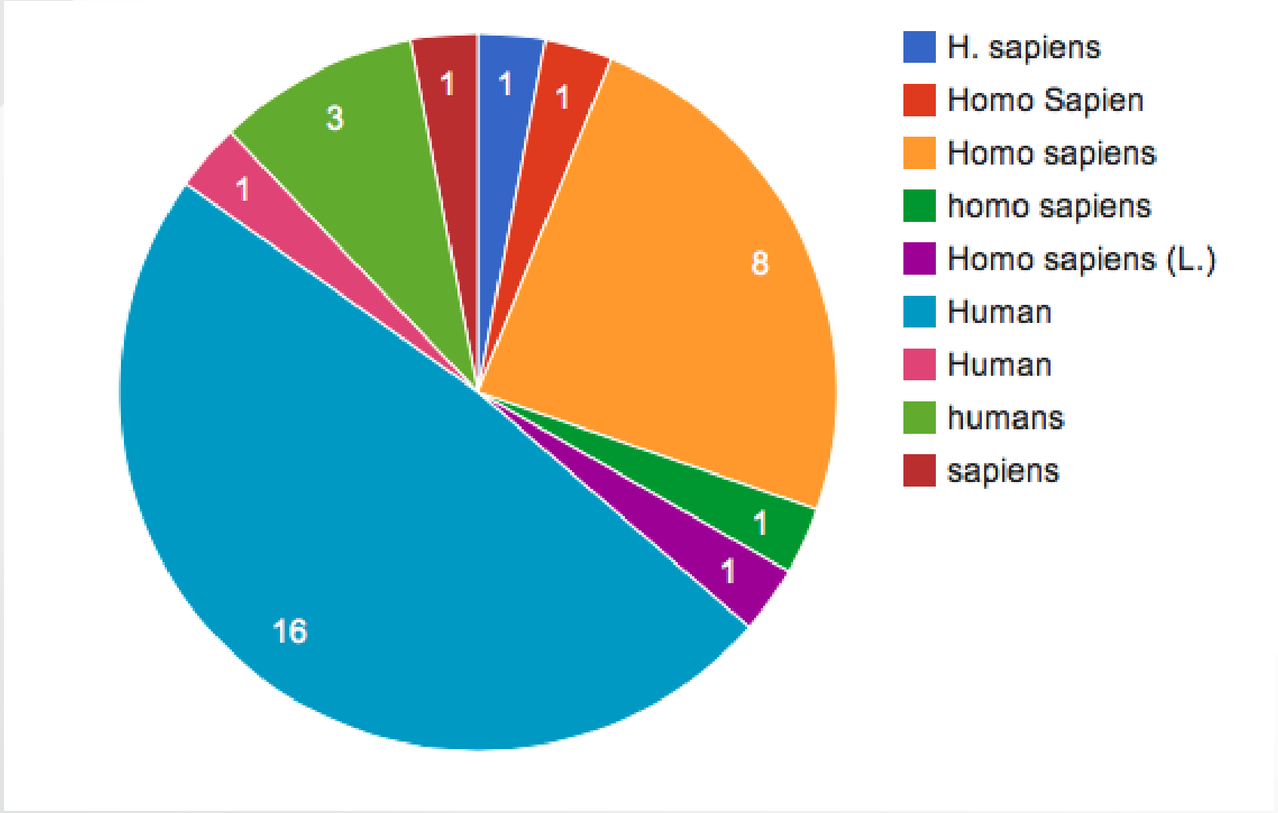
FAIRsharing

- 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
- Example: compare anatomical components in two distinct species of organism…

- **Organism A**
  - Term A1
  - Term A2
  - Term A3
    - Term B1
    - Term B2
  - Term C4
  - .
  - .
  - .
  - Term *n*

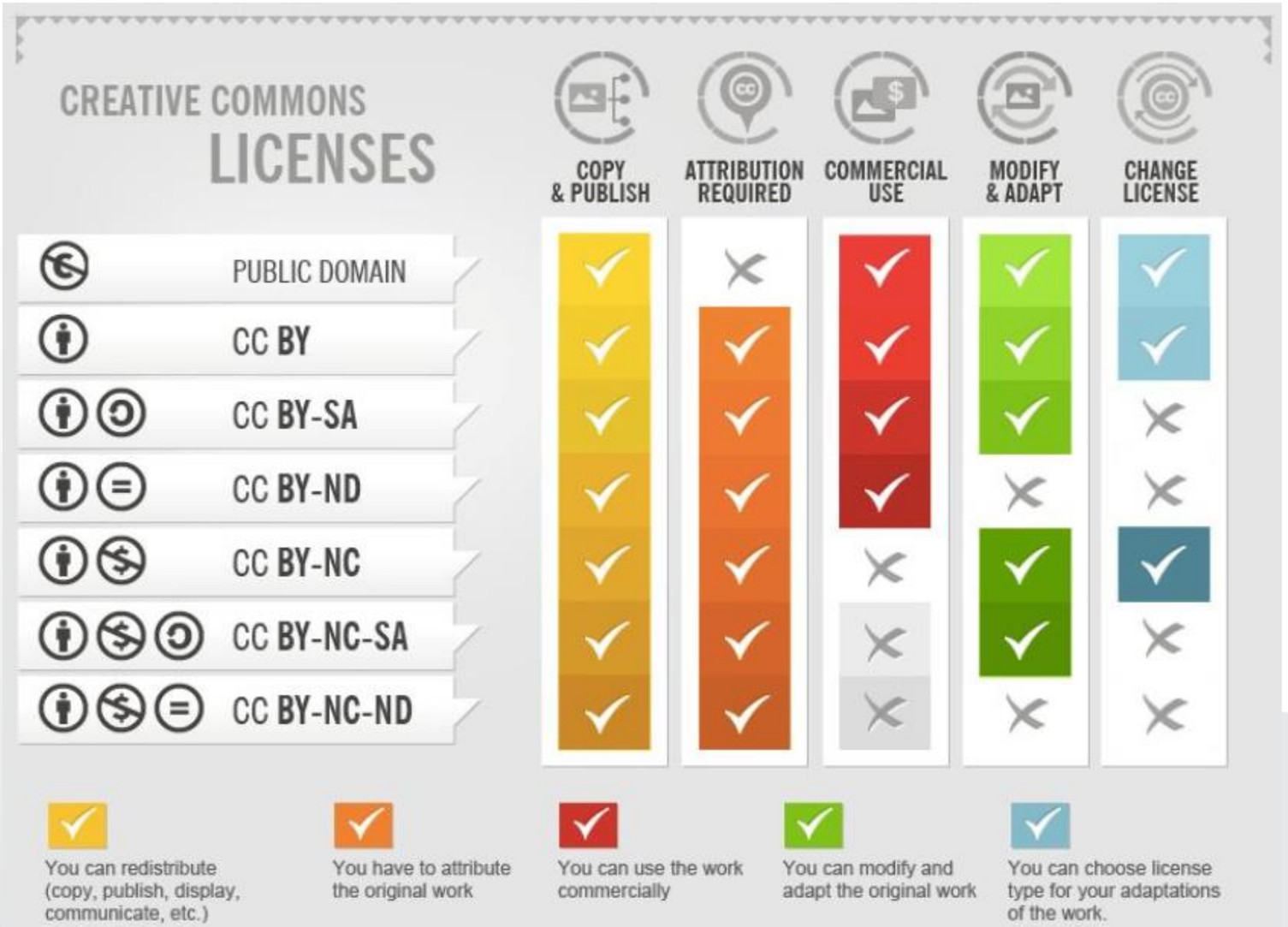
- **Organism B**
  - Term A1
  - Term A2
  - Term A3
    - Term B1
    - Term B2
  - Term C4
  - .
  - .
  - .
  - Term *n*

Ensure your data is as visible as possible

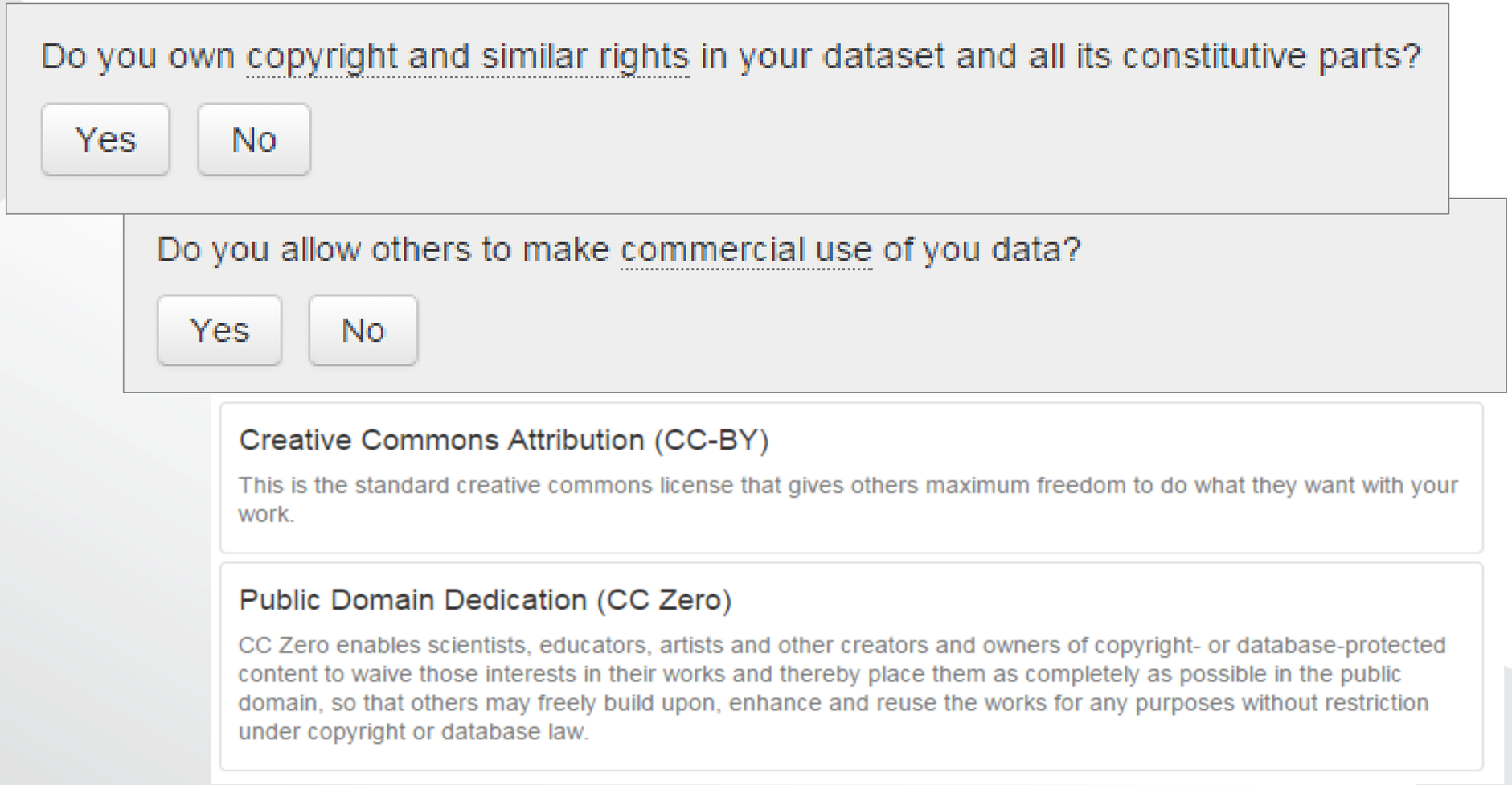
Dataset licensing

Horizon 2020 guidelines point to:

or



EUDAT licensing tool

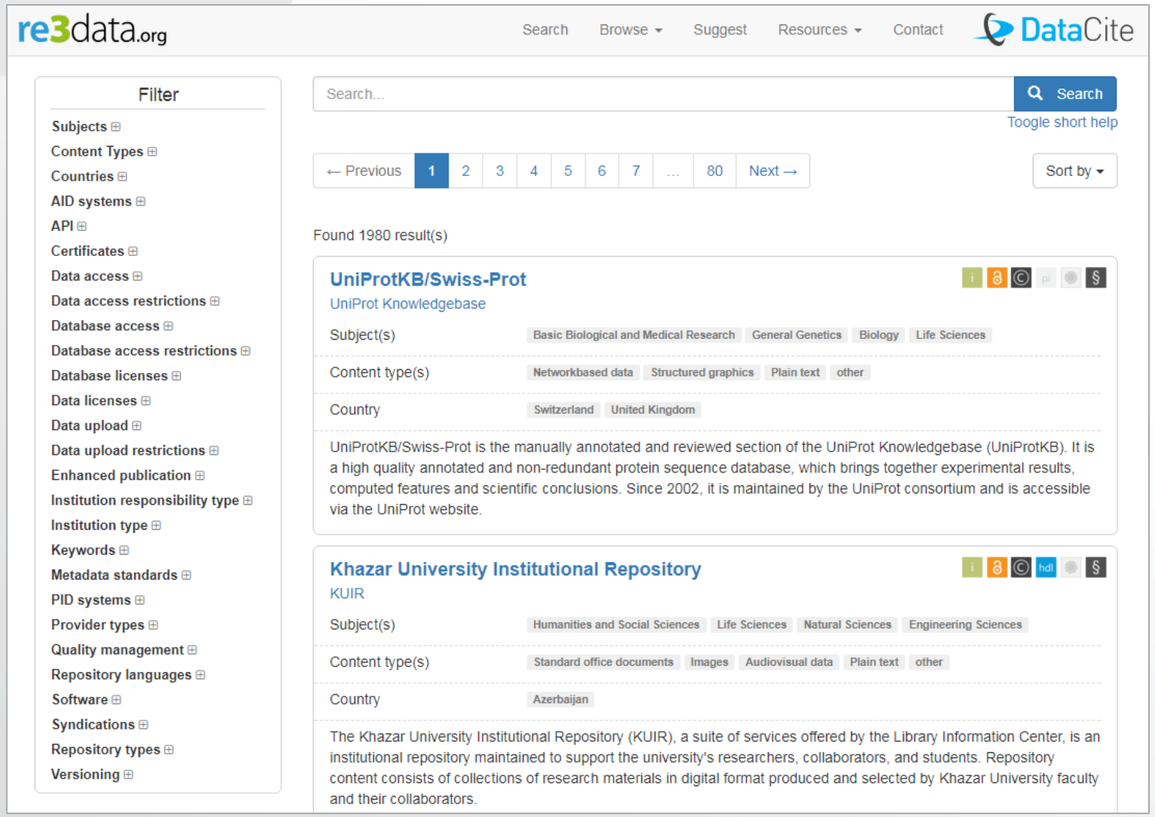


https://ufal.github.io/public-license-selector

Choose a suitable repository

Data repositories

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

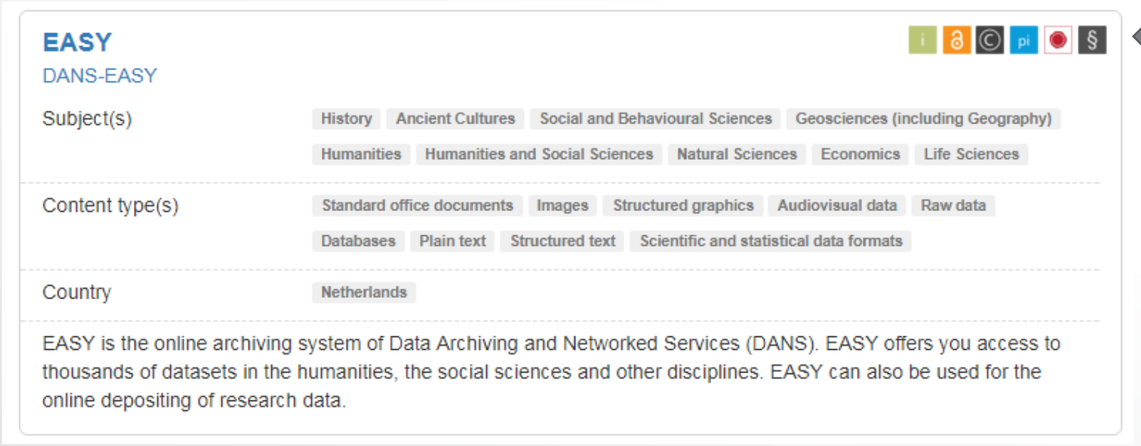


www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

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

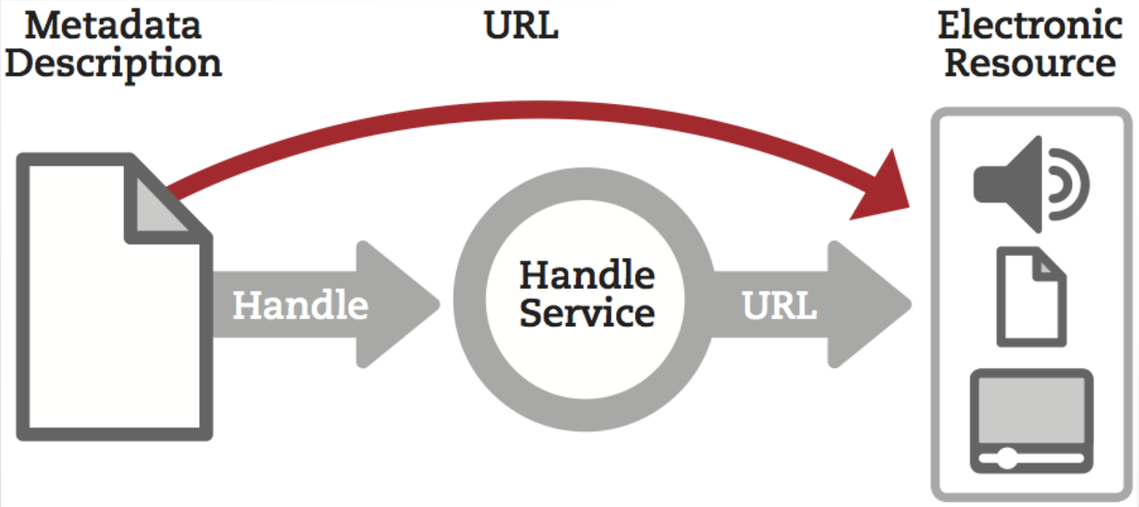


Icons to note open access, licenses, PIDs, certificates…

Make sure that data can be accessed in perpetuity

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

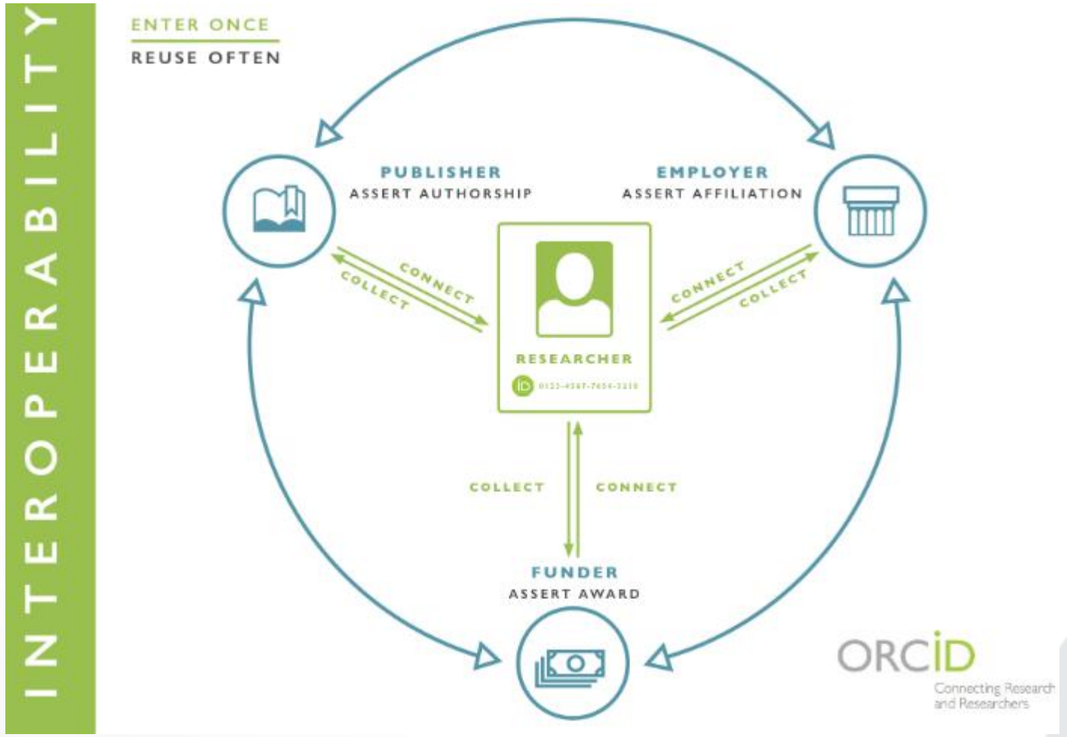
Keyword(s):
FAIR, FAIRness, checklist, research data, Findable, Accessible, Interopeable, Reusable, PID, repository, DOI, metadata, licence, data sharing, research data management,

Grants:
European Commission:
- EUDAT2020 - EUDAT2020 (654065)

License (for files):
Creative Commons Attribution 4.0

Persistent Identifiers

A specific example: ORCID



https://orcid.org/blog/2017/10/04/building-information-infrastructure-research-institutions

https://orcid.org/blog/2016/10/31/organization-identifier-project-way-forward

Thanks for watching!

More info at:

www.dcc.ac.uk/resources/

https://www.fosteropenscience.eu/

https://www.openaire.eu/