Research data requirements in Horizon 2020

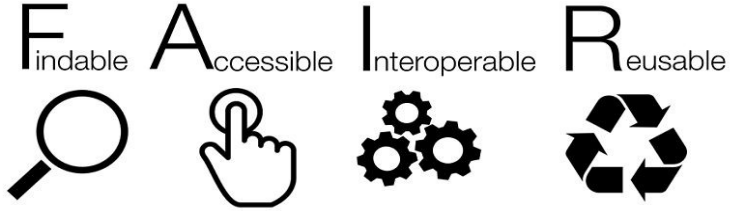
- what you need to know to assess DMPs

Sarah Jones
DCC, University of Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC
#fosteropenscience

RESEARCH DATA – OPEN BY DEFAULT

FAIR vs Open

- FAIR data does not have to be open
- Making data FAIR ensures it can be found, understood and reused – by the creator as well as others
- Data can be shared under restrictions & still be FAIR
- Open data is a subset of all the data shared
- *As open as possible, as closed as necessary*



FOSTER

FAIR data checklist

Findable

- Persistent ID
- Metadata online

Accessible

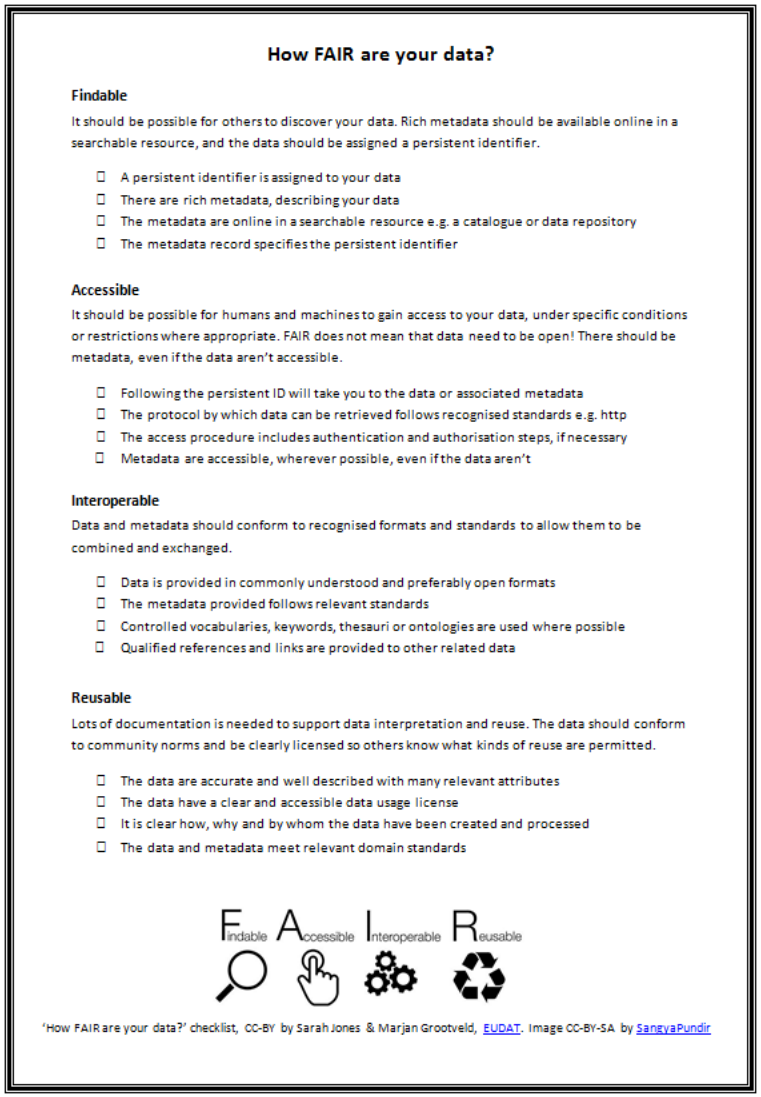
- Data online
- Restrictions where needed

Interoperable

- Use standards, controlled vocal
- Common (open) formats

Reusable

- Rich documentation
- Clear usage licence



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

Findable Accessible Interoperable Reusable

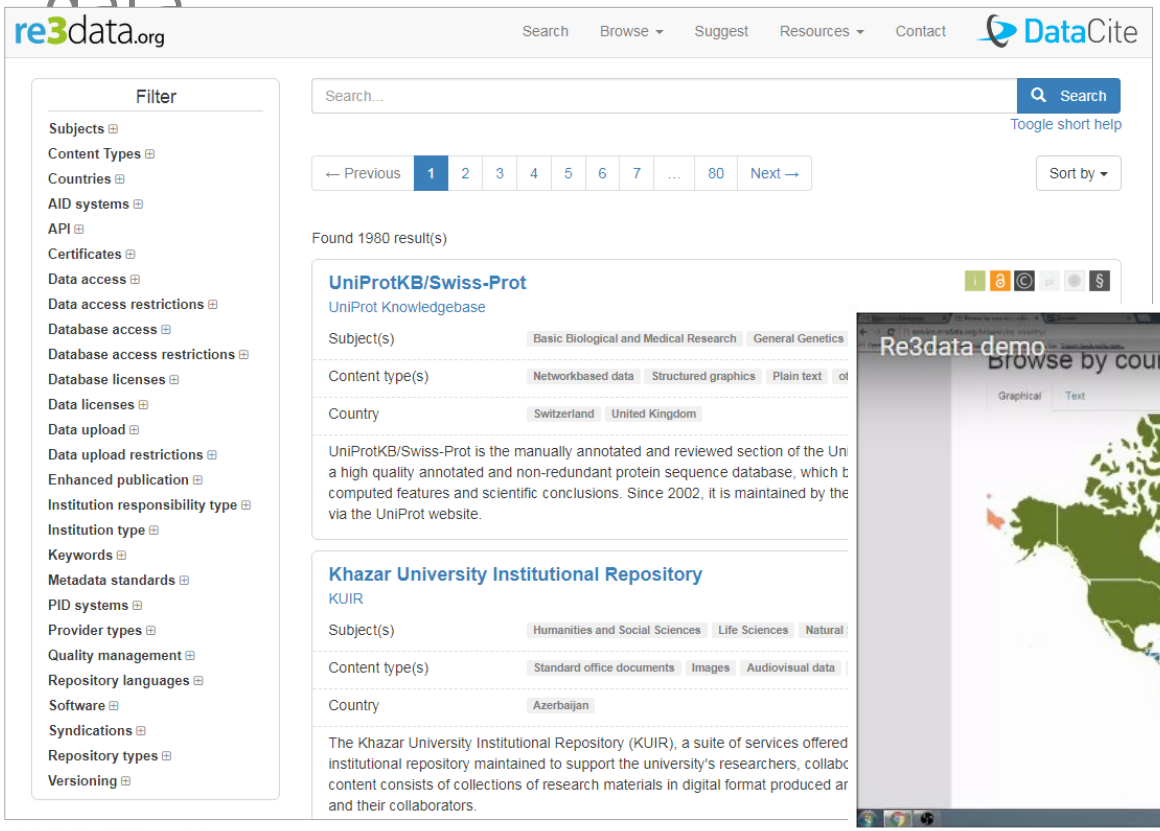
'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT. Image CC-BY-SA by SangyaPundir

https://zenodo.org/record/1065991

FOSTER

Deposit in a data repository

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

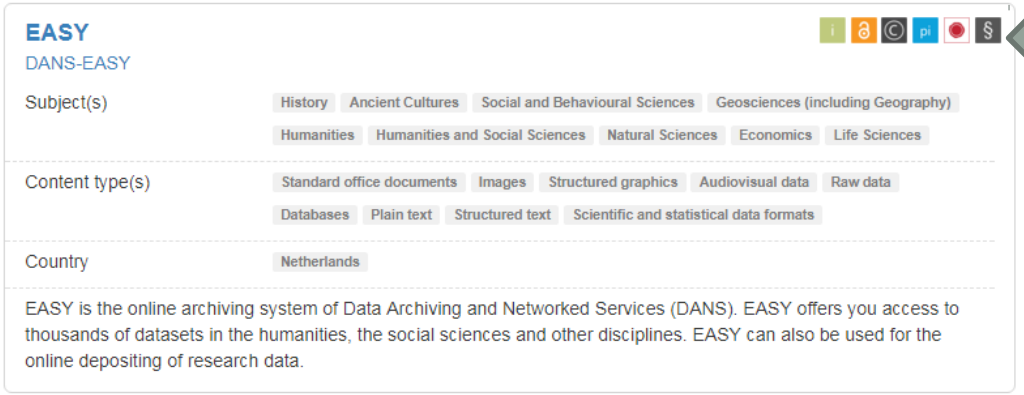


www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

How to select a repository?

- Better to use a subject specific repository if available
- Check they match particular data needs e.g. formats accepted, mixture of Open and Restricted Access.
- Do they assign a persistent and globally unique identifier for sustainable citations and to links back to particular researchers and grants?
- Look for certification as a '*Trustworthy Digital Repository*' with an explicit ambition to keep the data available in long term.



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

FOSTER

Zenodo

Zenodo is a multi-disciplinary repository that can be used for the long-tail of research data

- An OpenAIRE-CERN joint effort
- Multidisciplinary repository accepting
  - Multiple data types
  - Publications
  - Software
- Assigns a Digital Object Identifier (DOI)
- Links funding, publications, data & software

www.zenodo.org

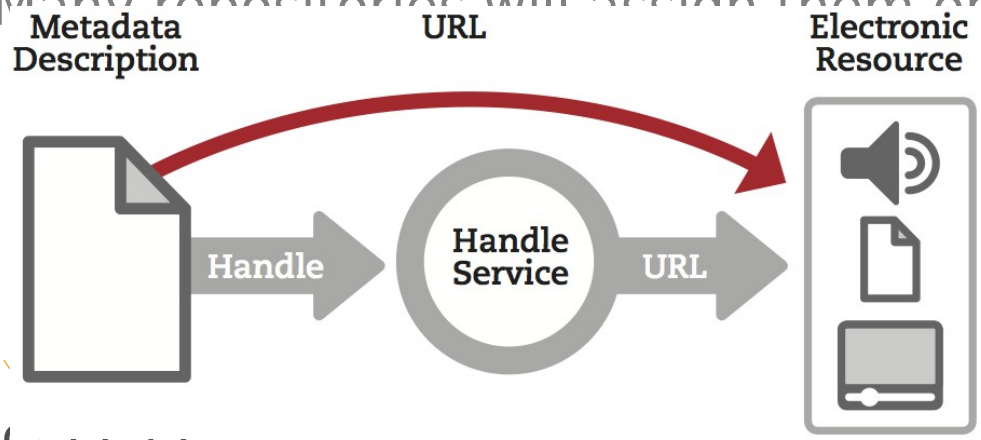
FOSTER

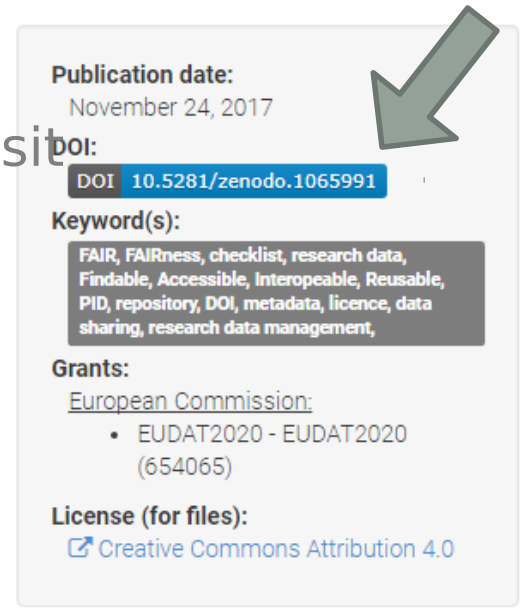
zenodo

What is a Persistent Identifier?

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





FOSTER

Create metadata

- At a basic level, metadata supports findability, disambiguation and citation
- Rich, specific metadata 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

FOSTER

Dublin Core metadata example

Creator: Donald Cooper
Role=Photographer

Subject: Shakespeare, William, 1564-1616, Antony and Cleopatra [LC]

Description: Vanessa Redgrave as Cleopatra

Date: 1973-08-09

Type: Image

Format: JPEG

Identifier:4150 [catalogue no]

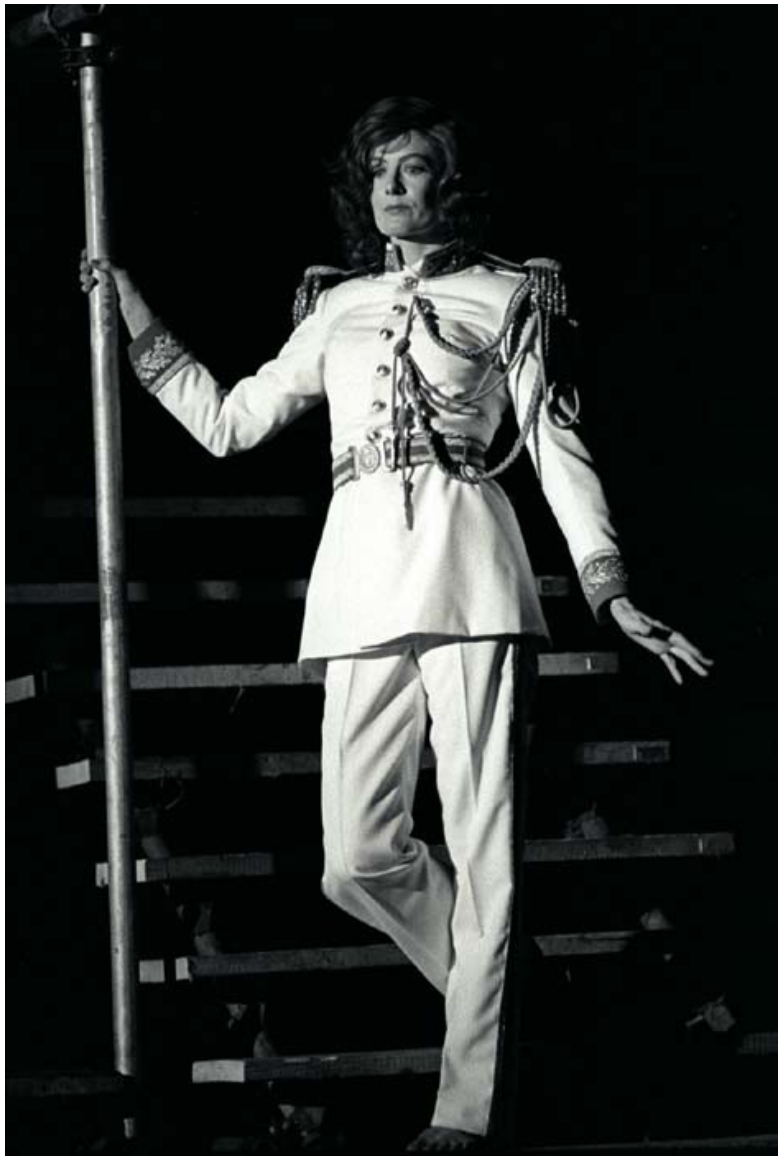
Source: negative no 235

Relation: Antony and Cleopatra: Thompson/73-8
IsPartOf

Coverage: Bankside Globe
Role=Spatial

Rights: Donald Cooper

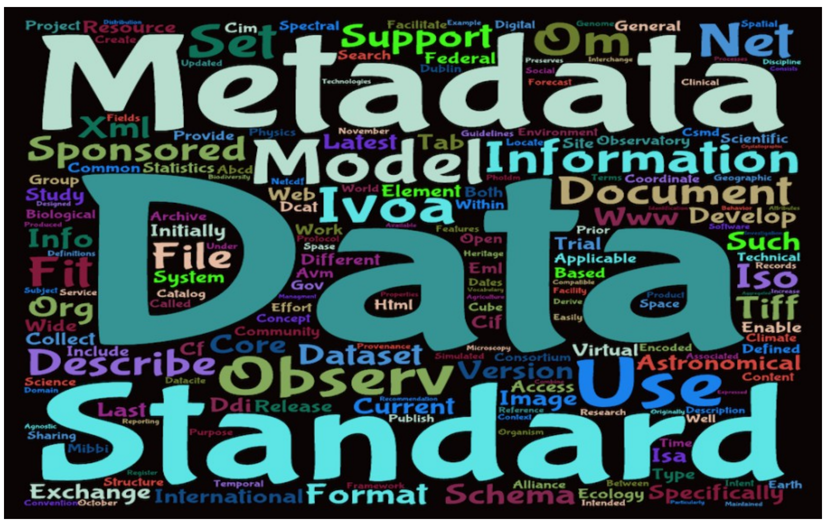
www.ahds.ac.uk/performingarts



Where to find relevant standards?

Metadata Standards Directory

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



http://rd-alliance.github.io/metadata-directory

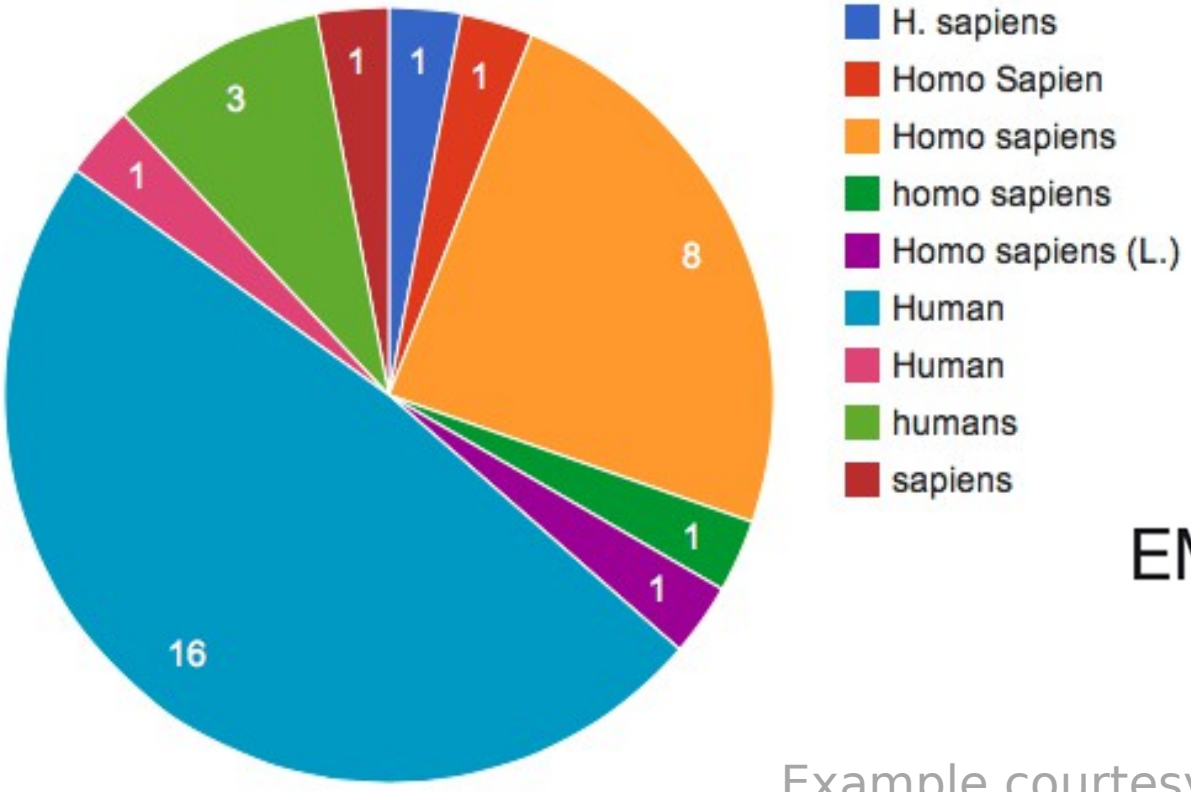
FAIRsharing

- 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……"

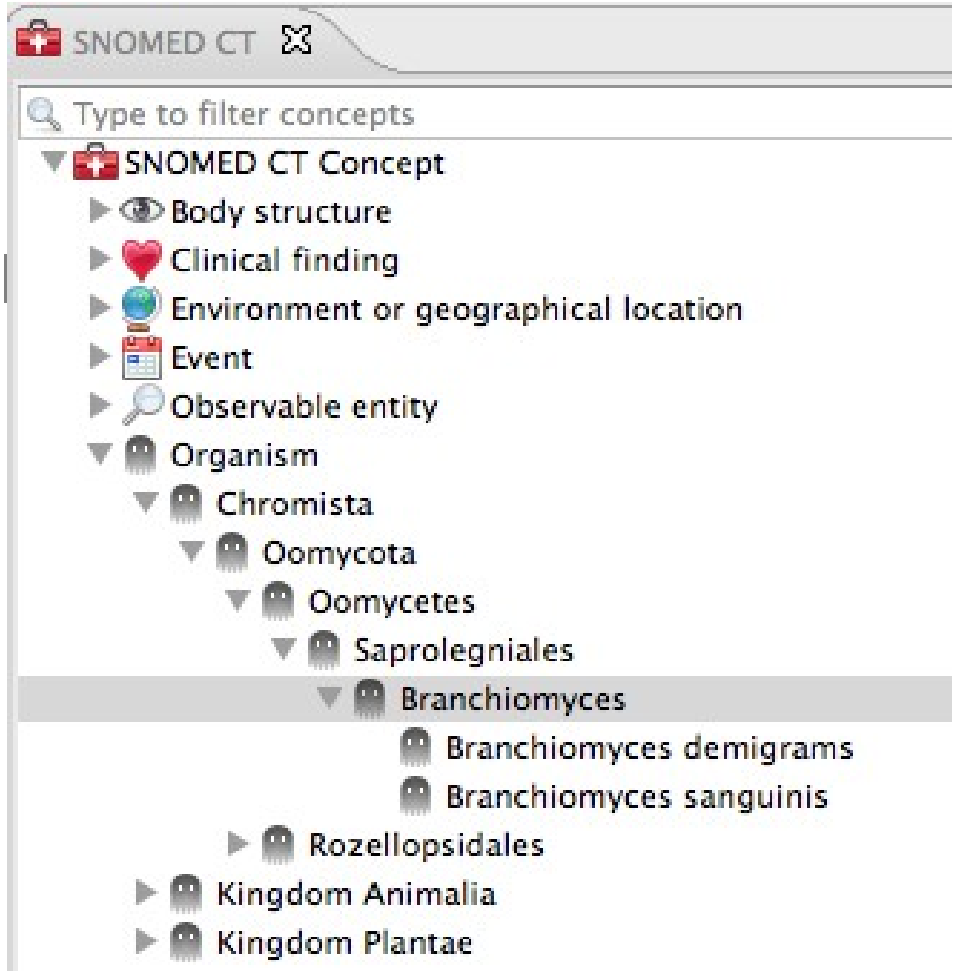


EMBL-EBI

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



FOSTER

Create documentation

We recommend that a ReadMe be a plain text file containing:

- 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

FOSTER

DRYAD

http://datadryad.org/pages/readme

Choose appropriate file 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

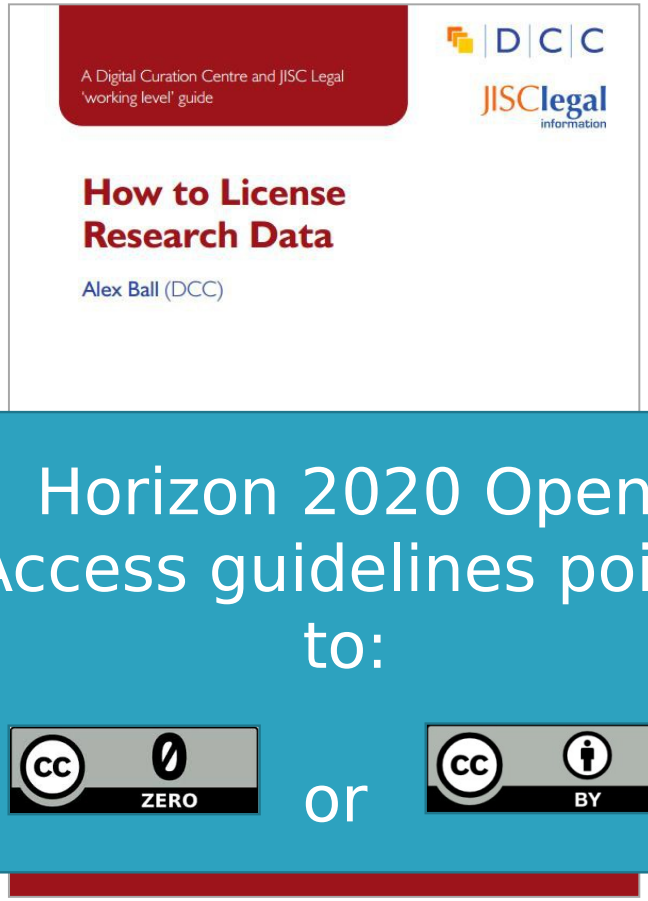
One format for analysis then convert to a standard format

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

License research data openly



Horizon 2020 Open Access guidelines point to:

CC0 ZERO or CC BY

This DCC guide outlines the pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

ND No Derivatives
Severely restricts use

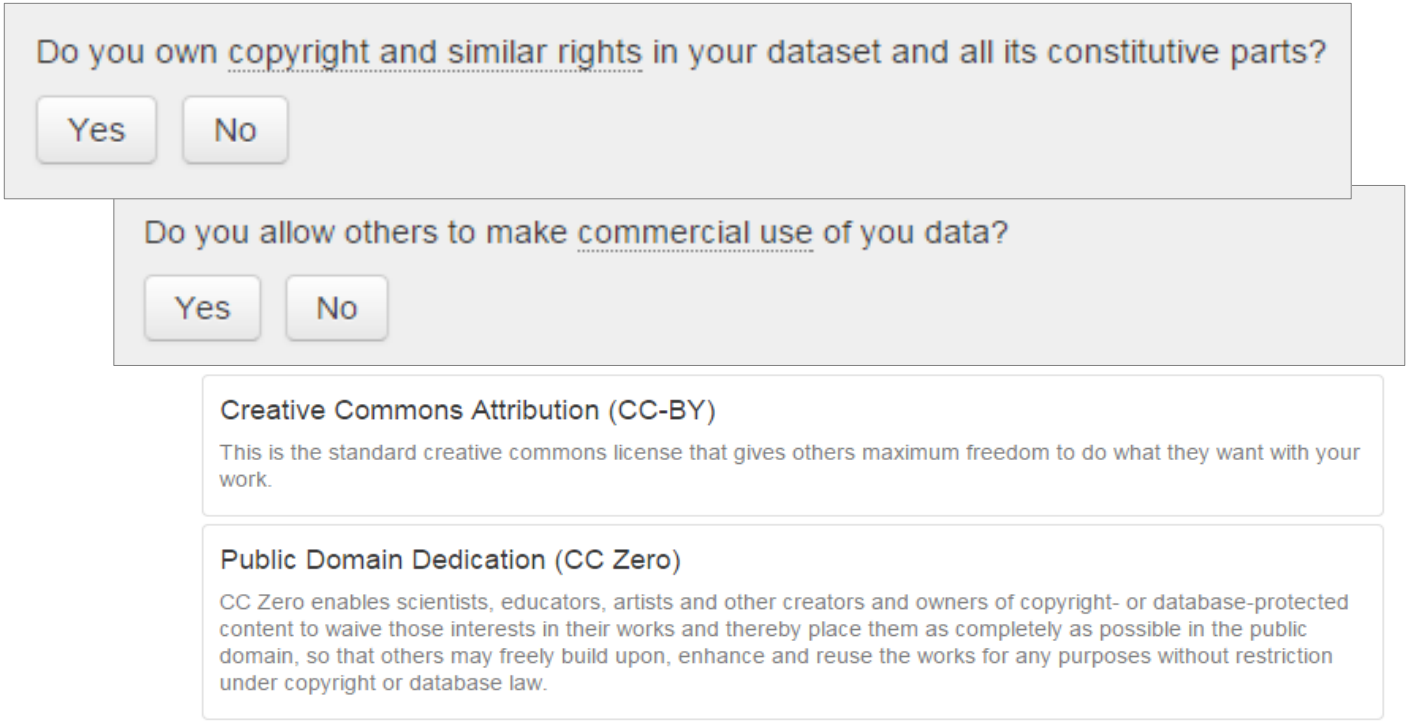
These clauses are not open licenses

FOSTER

www.dcc.ac.uk/resources/how-guides/license-research-data

EUDAT licensing tool

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



FOSTER

http://ufal.github.io/lindat-license-selector

Thanks – any questions?

Follow us on Twitter:
@fosterscience and #fosteropenscience

FOSTER training events and materials:
www.fosteropenscience.eu/events

FOSTER

Reviewing DMPs

Using the EC assessment grid, review one or more DMPs against the main criteria covered here:

- 1.c Are data types and **formats** accurately listed?
- 2.1.a Will the data be assigned a unique and **persistent identifier**?
- 2.1.d Will the data be described with **rich metadata**?
- 2.2.c is it specified where the **data** and associated metadata, documentation and code are **deposited**?
- 2.3.a Is it described how data interoperability will be facilitated e.g. through use of data and metadata **vocabularies, standards** or methodologies?
- 2.4.a Is **data licensing** and its role in facilitating re-use described?

FOSTER