



FAIR Data Management and Open Data

Sarah Jones

Digital Curation Centre, Glasgow

sarah.jones@glasgow.ac.uk

Twitter: @sjDCC

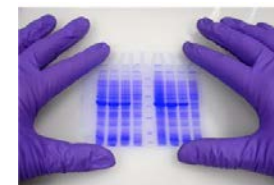
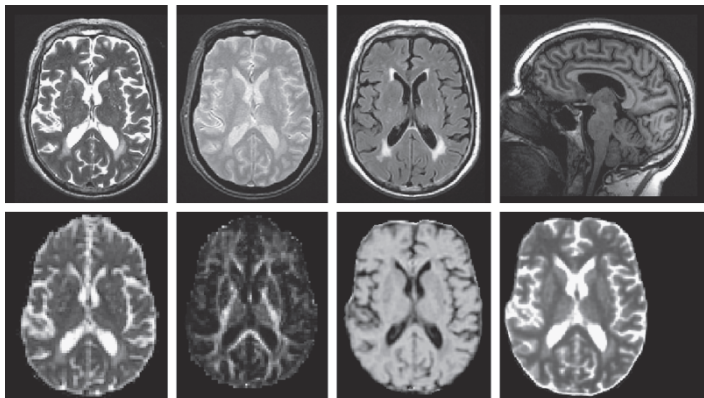
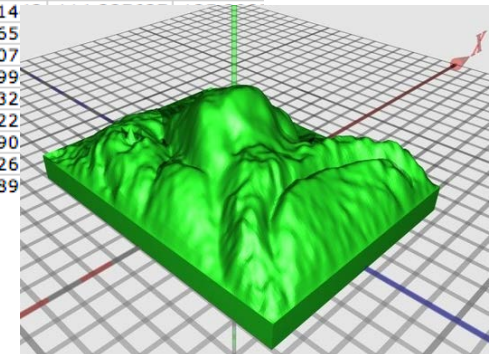
What is research data?

“In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form.”

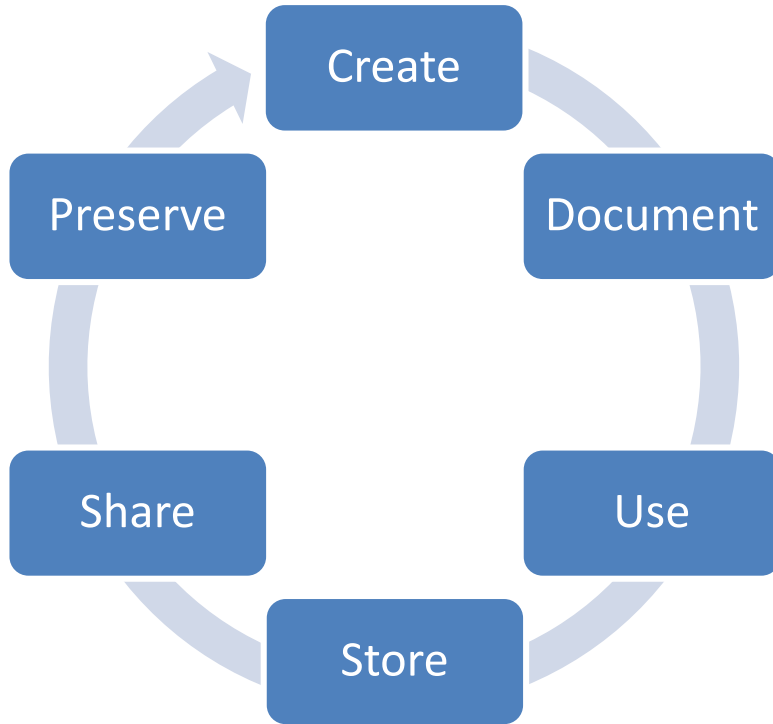
Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020

J: Yes.
JN: And which daughter were you?
J: 3rd.
JN: And do you know how her deliveries went? Did she have easy births?
J: Yes.
JN: And your grandmother delivered them all?
J: Yes, my grandmother.
JN: Was there a hospital close by?
J: No.

10/29/04	124.761606	129.27356	122.260995	209.662
11/1/04	119.977679	129.534731	124.739135	176.316
11/2/04	130.46875	135.839924	130.84732	168.289
11/3/04	135.895502	149.510531	140.795689	120.686
11/4/04	134.127052	140.495868	132.823819	206.138
11/5/04	129.851598	137.880438	124.888856	189.675
11/8/04	123.797241	131.84633	126.146789	202.496
11/9/04	118.435374	130.691651	112.877008	140.366
11/10/04	112.401212	121.5614		
11/11/04	112.388488	128.4965		
11/12/04	129.011813	138.8807		
11/15/04	127.077465	139.2899		
11/16/04	124.9785	135.3632		
11/17/04	124.294035	133.2422		
11/18/04	125.663717	135.1590		
11/19/04	123.704853	127.6126		
11/22/04	118.926697	122.8189		



What is Research Data Management?



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

By managing data effectively you can ensure it is FAIR, and, where appropriate, open



Creating data

Data creation tips

- Choose appropriate formats
- Adopt a file naming convention
- Create metadata and documentation as you go
- Ensure consent forms, licences and agreements don't restrict opportunities to share data

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

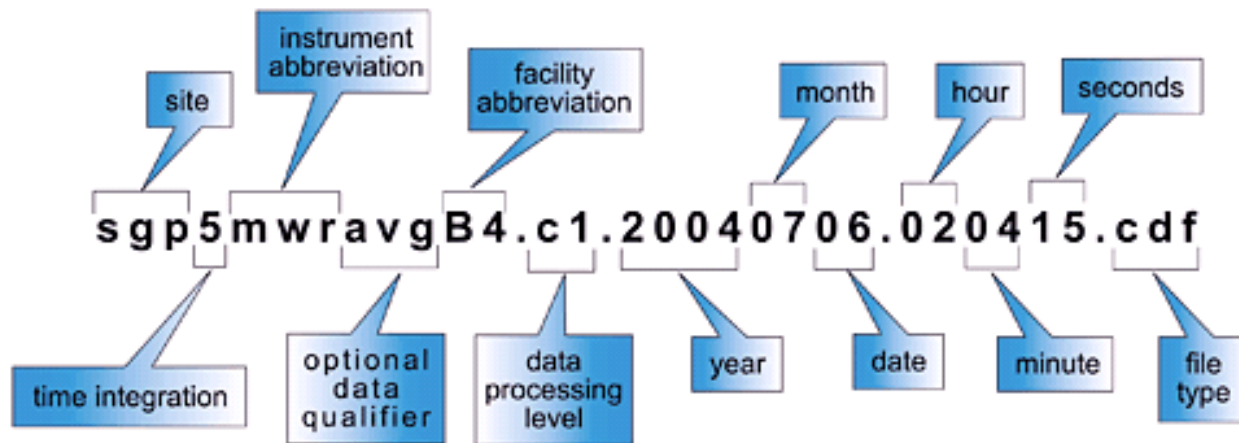
BioformatsConverter batch converts a variety of proprietary microscopy image formats to the Open Microscopy Environment format - OME-TIFF

Data centres may suggest preferred formats for deposit

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

How will you name your files?

An example netCDF data file name is depicted below:



- Keep file and folder names short, but meaningful
- Agree a method for versioning
- Include dates in a set format e.g. YYYYMMDD
- Avoid using non-alphanumeric characters in file names
- Use hyphens or underscores not spaces e.g. day-sheet, day_sheet
- Order the elements in the most appropriate way to retrieve the record

Example from ARM Climate Research Facility

www.arm.gov/data/docs/plan

www.jiscdigitalmedia.ac.uk/guide/choosing-a-file-name

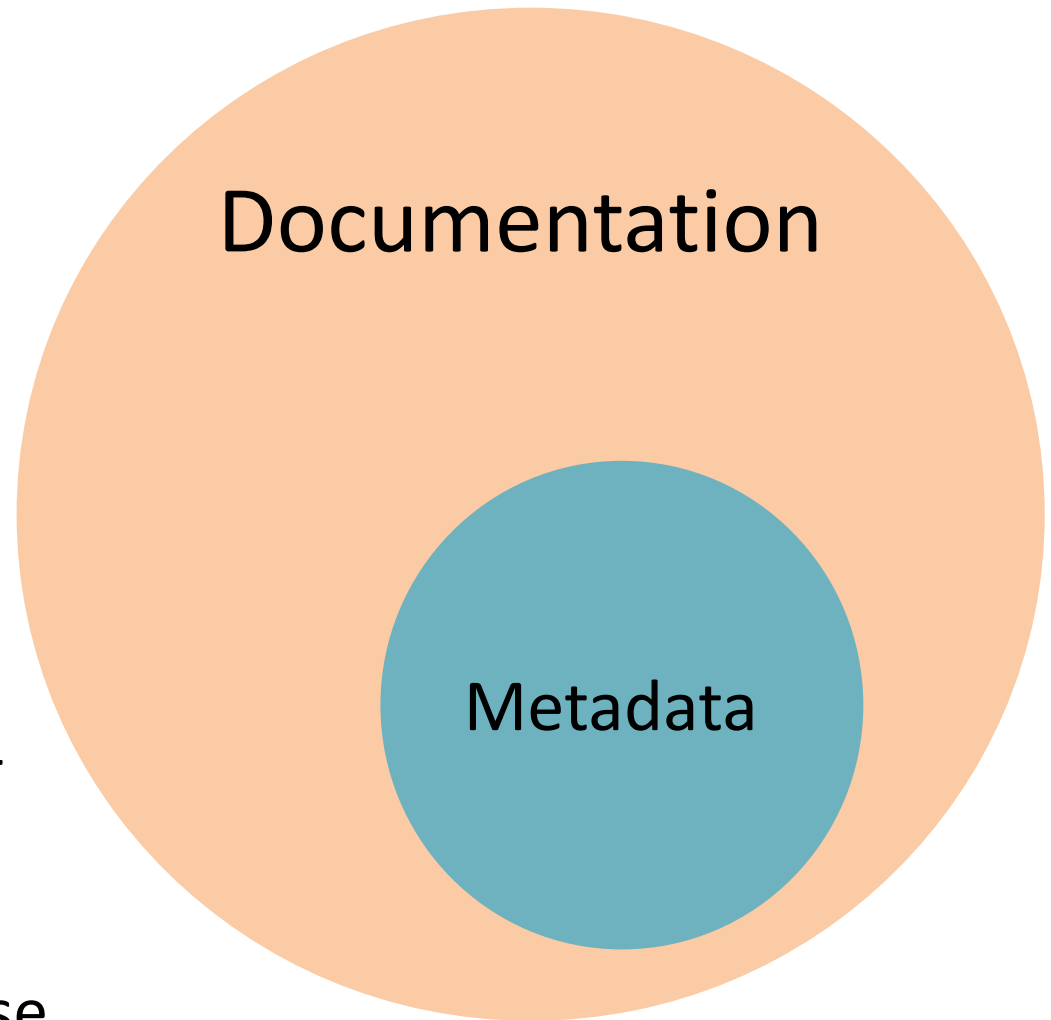
Documentation and metadata

Metadata

- Standardised
- Structured
- Machine and human readable

Metadata helps to cite & disambiguate data

Documentation aids reuse



Metadata standards

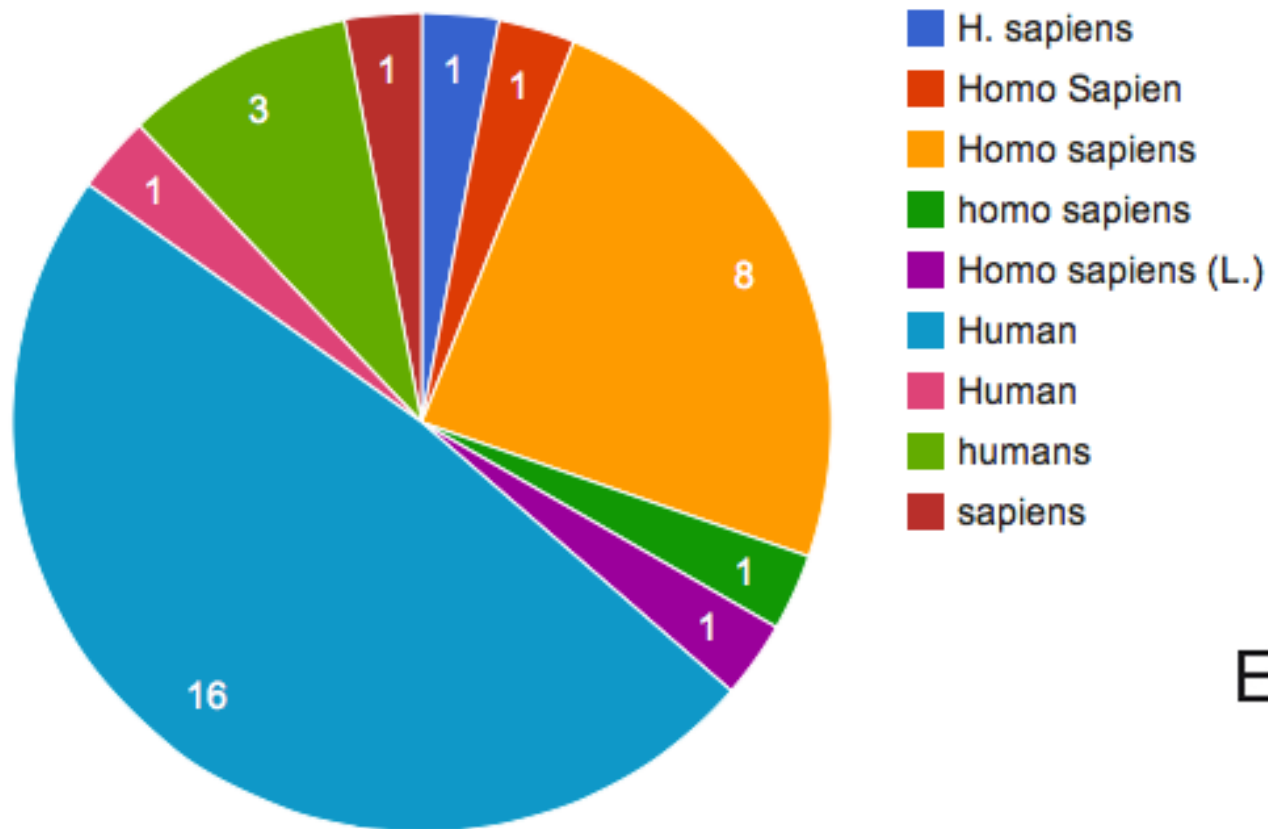
These can be general – such as Dublin Core

Or discipline specific

- Data Documentation Initiative (DDI) – social science
 - Ecological Metadata Language (EML) - ecology
 - Flexible Image Transport System (FITS) – astronomy
-
- Provided in catalogues to aid discoverability
 - Structured so search engines can uncover it
 - Exposed in machine-readable form e.g. XML

Why are ontologies important?

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



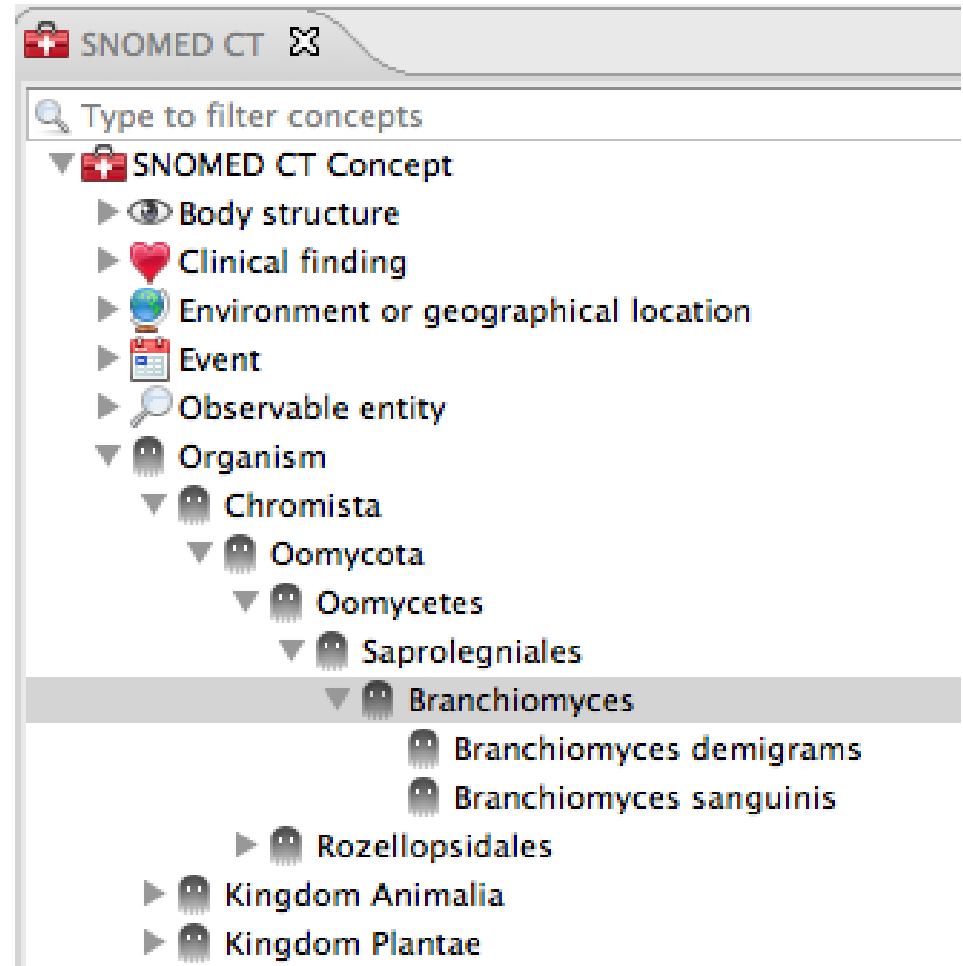
Controlled vocabularies

E.g. SNOMED CT (clinical terms) or MeSH

Include ontologies as well

- Defined terms + taxonomy

Useful for selecting keywords to tag datasets



Documentation to ensure data utility

Is it clear what each bit of your dataset means?

- Data dictionaries
- Columns/rows labelled
- Variable ranges defined





ReadMe files

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

<http://datadryad.org/pages/readme>

Ask for consent for data sharing

If not, data centres won't be able to accept the data
– regardless of any conditions on the original grant.

SAMPLE CONSENT STATEMENT FOR QUANTITATIVE SURVEYS

Thank you very much for agreeing to participate in this survey.

The information provided by you in this questionnaire will be used for research purposes. It will not be used in any manner which would allow identification of your individual responses.

Anonymised research data will be archived at in order to make them available to other researchers in line with current data sharing practices.

How to keep you data secure?

Develop a practical solution that fits your circumstances

- Store your data on managed servers
- Restrict access to certain groups
- Encrypt mobile devices carrying sensitive information
- Keep anti-virus software up-to-date
- Use secure data services for long-term sharing





How to make data open

Degrees of openness

Five star open data



**SECURE
DATA
SERVICE**
enabling the
research community

Unable to share

Open

Restricted

Closed

Content that can be freely used, modified and shared by anyone for any purpose

Limits on who can use the data, when, how or for what purpose

- Embargo periods
- Charges for use
- Restrictive licences
- Data sharing agreements
- Peer-to-peer exchange
- ...

CLASSIFIED



Four steps to make data open



<https://okfn.org>

1. Choose your dataset(s)

- What can you may open? You may need to revisit this step if you encounter problems later.

2. Apply an open license

- Determine what IP exists. Apply a suitable licence e.g. CC-BY

3. Make the data available

- Provide the data in a suitable format. Use repositories.

4. Make it discoverable

- Post on the web, register in catalogues...

License research data openly



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

Horizon 2020 Open Access guidelines point to:



or

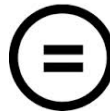


CREATIVE COMMONS LIMITATIONS



NC Non-Commercial

What counts as commercial?



ND No Derivatives

Severely restricts use

These clauses are not open licenses

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.

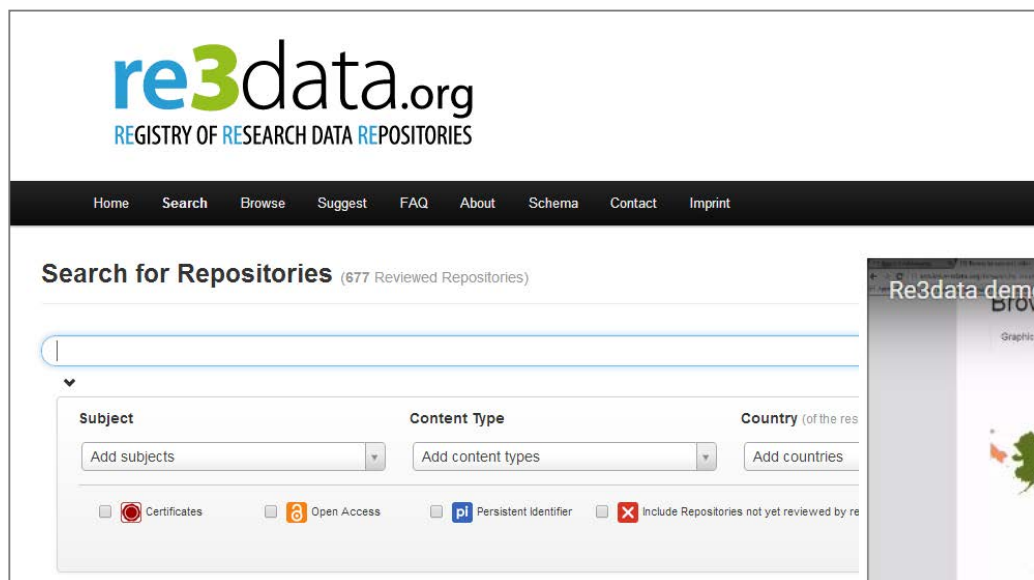
Public Domain Dedication (CC Zero)

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.

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

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



The screenshot shows the re3data.org website. The logo at the top left reads "re3data.org" with "REGISTRY OF RESEARCH DATA REPOSITORIES" underneath. A navigation bar contains links for Home, Search, Browse, Suggest, FAQ, About, Schema, Contact, and Imprint. Below the navigation bar is a search section titled "Search for Repositories (677 Reviewed Repositories)". It features a search input field, a dropdown menu for "Subject", and input fields for "Content Type" and "Country (of the res)". There are also checkboxes for "Certificates", "Open Access", "PI Persistent Identifier", and "Include Repositories not yet reviewed by re".

<http://service.re3data.org/search>



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

How to select a repository?

- Look for provision from your community, university, publisher, funder etc
- Check they match your particular data needs: e.g. formats accepted; mixture of Open and Restricted Access.
- See if they provide guidance on how to cite the deposited data.
- Do they assign a persistent & 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.

www.openaire.eu/opendatapilot-repository



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

Sharing data increases citations

Want evidence?

- Piwowar, Vision – 9% (microarray data)
- Drachen, Dorch, et al – 25-40%, astronomy
- Gleditch, et al – doubling to trebling (international relations)

Open Data Citation Advantage

<http://sparceurope.org/open-data-citation-advantage>

How to cite data

Key citation elements

- Author
- Publication date
- Title
- Location (= identifier)
- Funder (if applicable)

AWARENESS LEVEL

A Digital Curation Centre Briefing Paper
19th July 2011

DCC
JISC

Data Citation and Linking

By Alex Ball and Monica Duke, UKOLN, University of Bath

- Introduction
- Short-term Benefits and Long-term Value
- Perspectives on Data Citation
- Roles and Responsibilities
- Issues to be Considered
- Related Research
- Additional Resources

Introduction

On the surface, citing datasets is a trivially easy thing to do. Style manuals such as the *Publication Manual of the American Psychological Association* and the *Oxford Manual of Style* have provided sample citations for datasets since at least the early 2000s. The process of making datasets citable, however, is rather more difficult. In consequence of this and other factors, a culture of citing datasets has been slow to develop. Nevertheless, it is vital that researchers cite the datasets they use, if datasets are to be regarded as legitimate academic outputs in their own right.

Short-term Benefits and Long-term Value

There are several short-term benefits to making datasets citable, citing them in practice, and linking datasets to papers that make use of the data.

- If the authors of a scientific publication properly cite the data that underlies it, it is much easier for the reader to locate that data. This in turn makes it easier for the reader to validate and build on the publication's findings.

- Data citations ensure that data contributors receive proper credit when their work is reused by other researchers.
- If a dataset links back to the paper that describes its collection, a reader coming to the dataset direct can use that link to put it in context and understand the methodology used.
- If a dataset links to other papers that make use of it, these links can be used by the contributors and data publishers to demonstrate the impact of the data. Potential reusers might use these links to discover critiques of the data or to provide inspiration for how to use them.

Once a culture of data citation has been established, several other benefits are likely to become apparent.

- The publishing infrastructure that makes the data citable will also help to ensure they are available for reference and reuse long into the future.
- There will be less danger of rival researchers 'stealing' results from those who publish their data openly, as failure to give due credit would amount to plagiarism and thus be punishable.
- Services built around data citation will make it easier for researchers to discover relevant datasets.
- Data citations could be used to measure the impact of both individual datasets and their contributors.
- Researchers could gain professional recognition and rewards for published data in the same way as for more traditional publications.

Taking these points together, there would likely be an increase in the quantity and quality of data published, with all the benefits this implies for the transparency and rate of scientific research.

www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

How do you share data effectively?

- Use appropriate repositories, this catalogue is a good place to start

<http://www.re3data.org>



- Document and describe it enough for others to understand, use and cite

<http://www.dcc.ac.uk/resources/how-guides/cite-datasets>



- Licence it so others can reuse

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



FAIR data checklist

Findable

- Persistent ID
- Metadata online

Accessible

- Data online
- Restrictions where needed

Interoperable

- Use standards, controlled vocabs
- 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

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, [EUDAT](#). Image CC-BY-SA by [SanyvaPundir](#)

Thanks for listening

DCC resources on DMPs

www.dcc.ac.uk/resources/data-management-plans

Follow us on twitter:

@digitalcuration and #ukdcc

@DMPonline and #ActiveDMPs



D|C|C

because good research needs good data