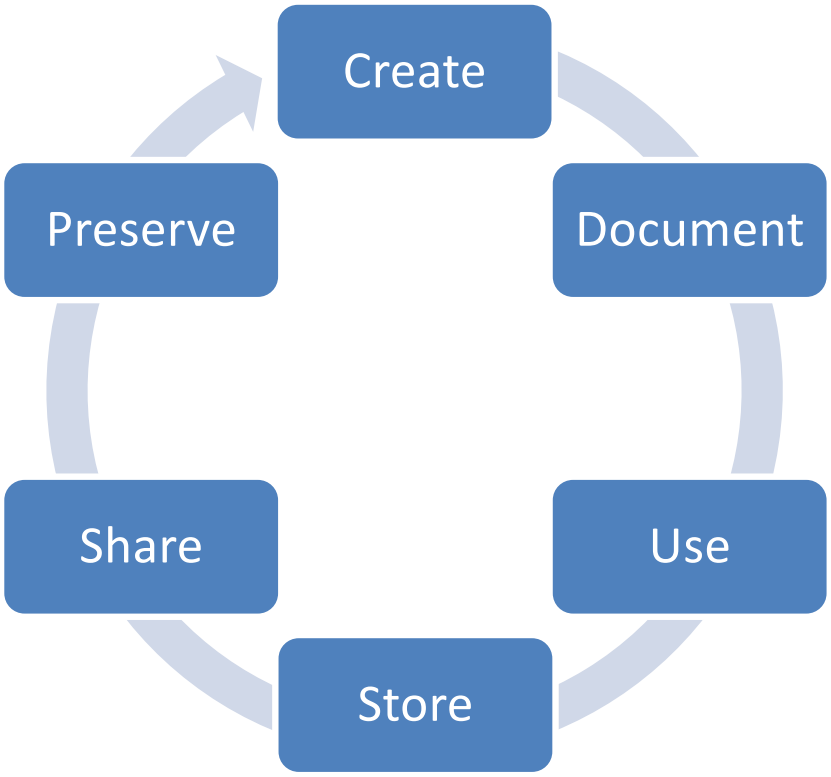
Managing and Sharing Research Data

Material from: Sarah Jones
Digital Curation Centre, Glasgow
Twitter: @sjDCC
Presented by: Marcela Alfaro Córdoba

CODATA Research Data Science School, 6 December 2018, Sao Paulo #dataSaoPaulo18

What is RDM & FAIR?

What is Research Data Management?

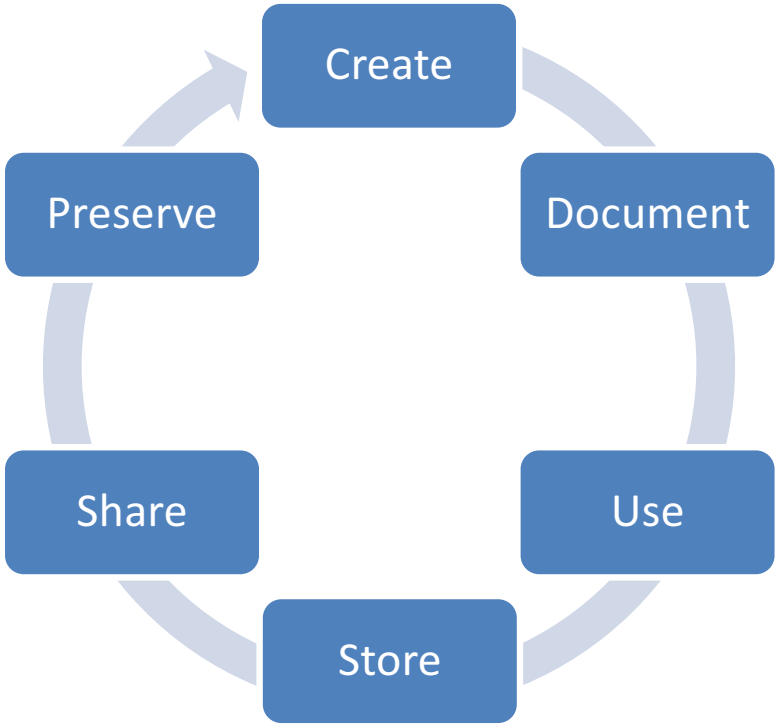


"the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

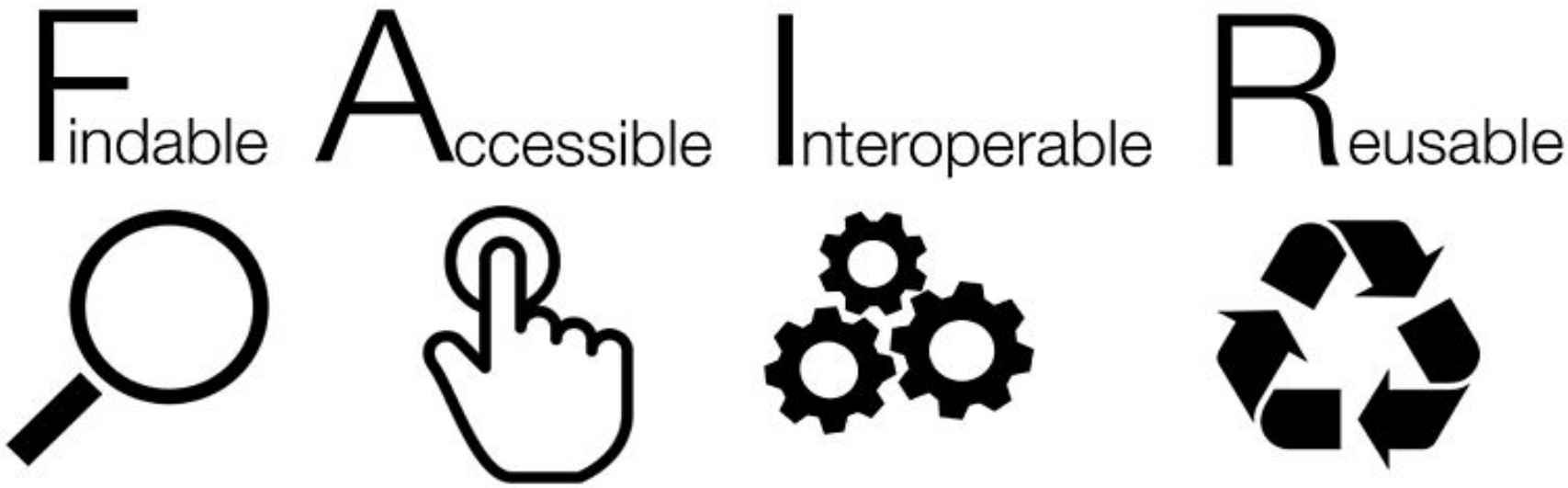
Data management is part of good research practice

What is involved in RDM?

- Data Management Planning
- Data creation
- Annotating / documenting data
- Analysis, use, versioning
- Storage and backup
- Publishing papers and data
- Preparing for deposit
- Archiving and sharing
- Licensing
- Citing…



Who has heard of FAIR?



What FAIR means: 15 principles

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes;

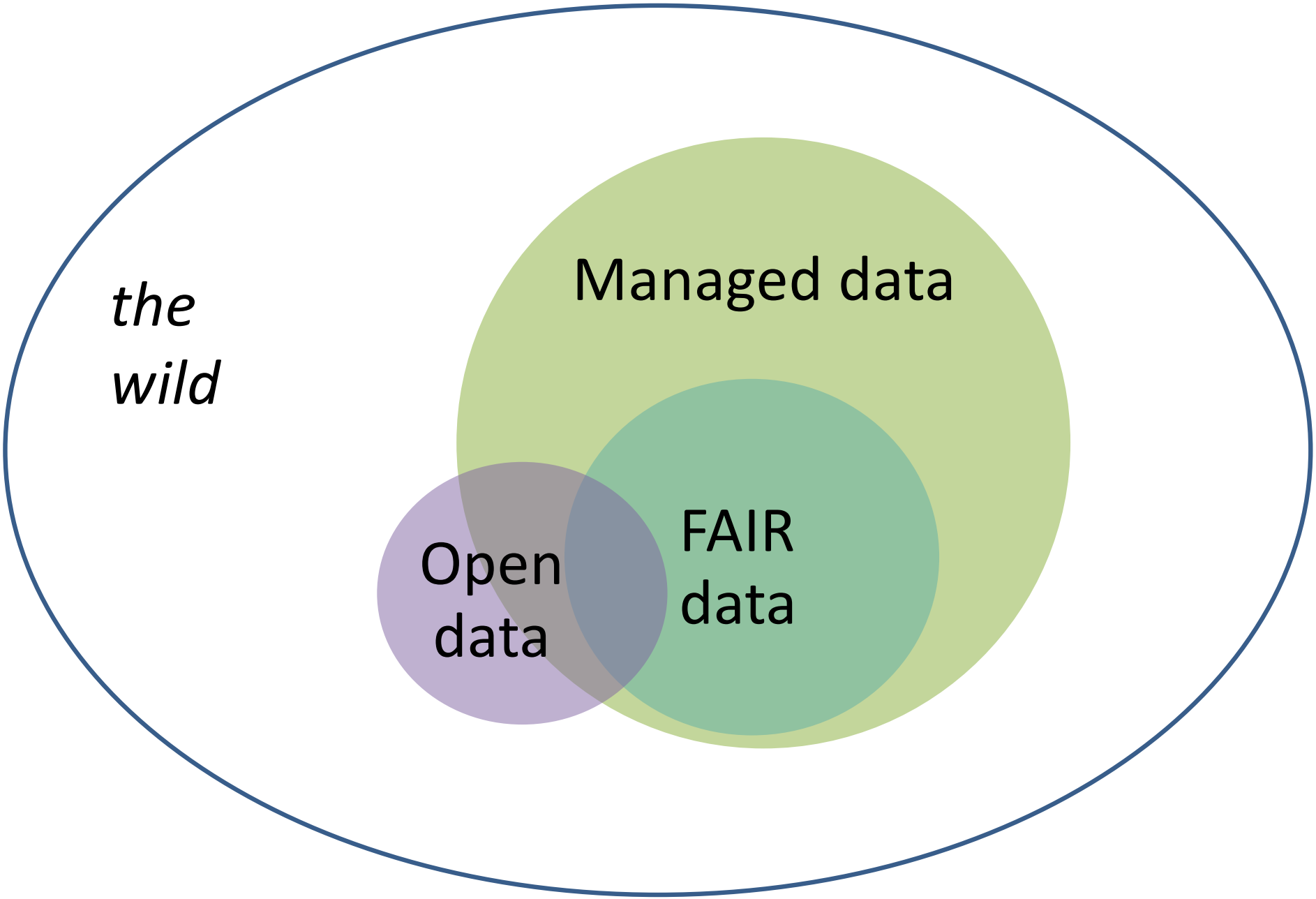
R1.1. (meta)data are released with a clear and accessible data usage license;

R1.2. (meta)data are associated with detailed provenance;

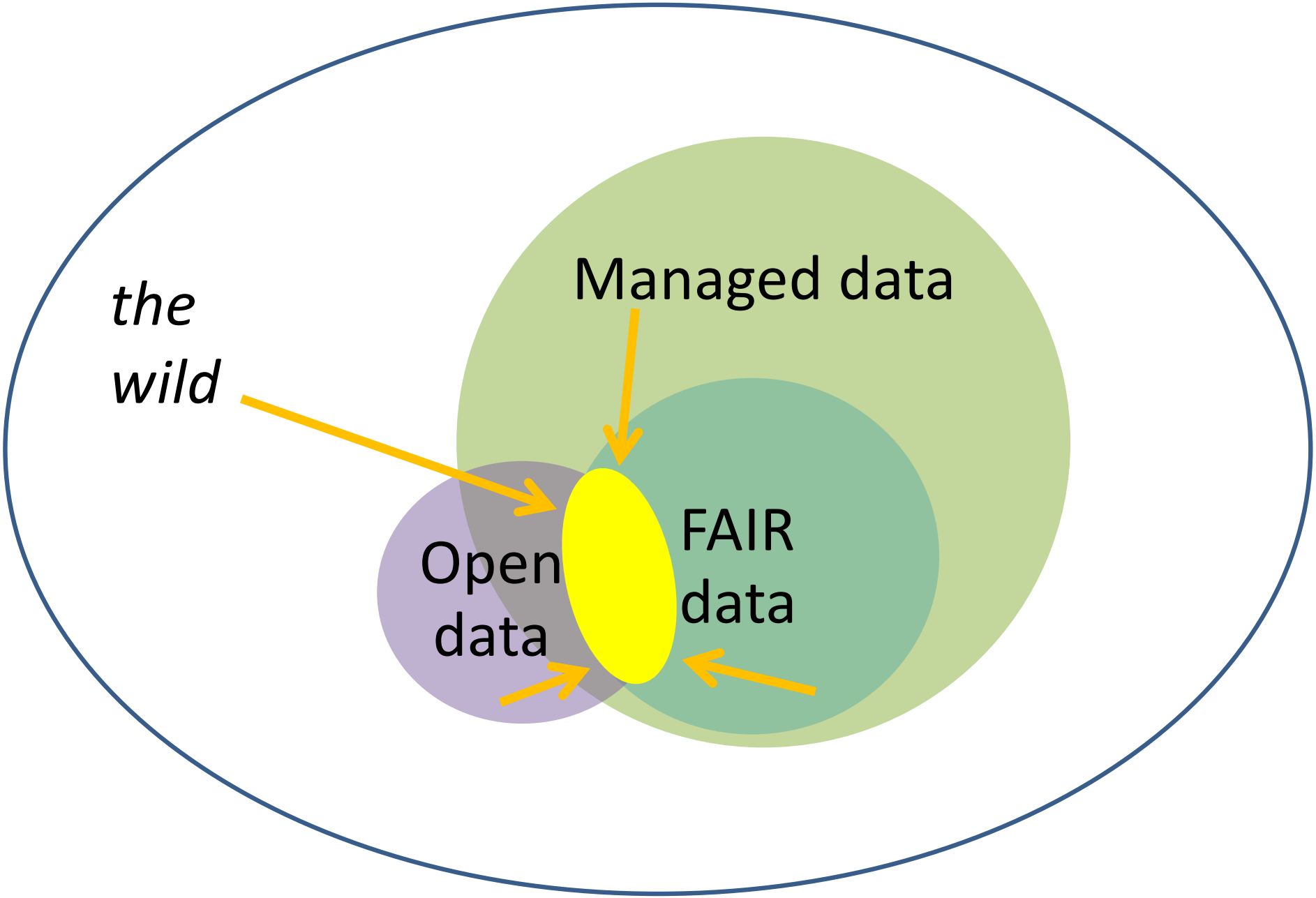
R1.3. (meta)data meet domain-relevant community standards;

doi: 10.1038/sdata.2016.18

All research data



Increasing that which is FAIR & open



Why make data available?

"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

#OpenData
#OpenScience

Original image via doi:10.1038/461145a. "Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science." - *Nature* 461, 145

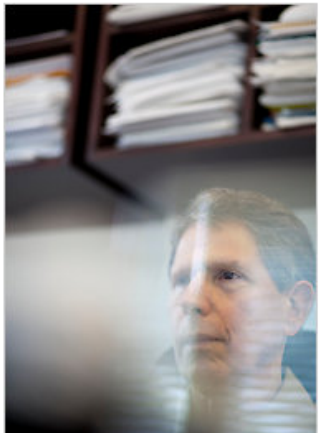
Sharing leads to breakthroughs

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

Enlarge This Image



Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

...and increases the speed of discovery

Benefits for you: 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

Creating data

Data creation tips

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

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.

www.data-archive.ac.uk/create-manage/consent-ethics/consent?index=3

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

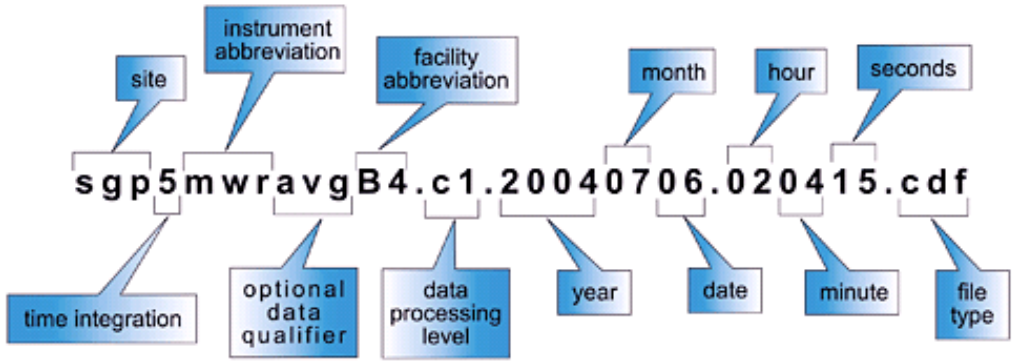
BioformatsConverter batch converts a variety of proprietary microscopy image formats to the Open Microscopy Environment format - OME-TIFF

Data centres may suggest preferred formats for deposit

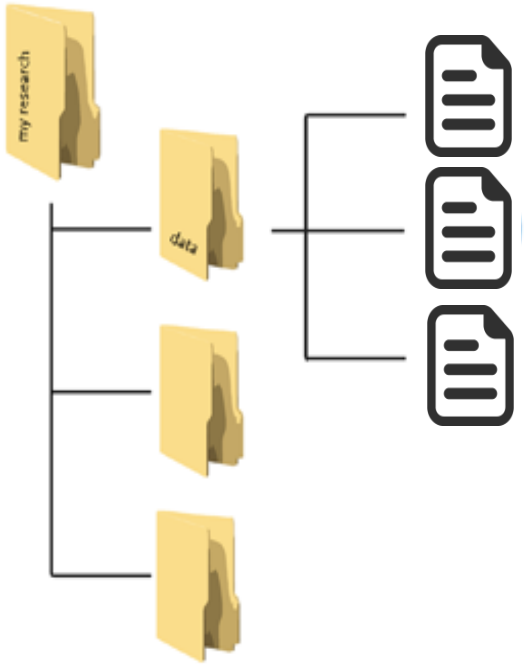
www.data-archive.ac.uk/create-manage/format/formats-table

How will you organise your data?

An example netCDF data file name is depicted below:



Example from ARM Climate Research Facility www.arm.gov/data/docs/plan



- 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

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

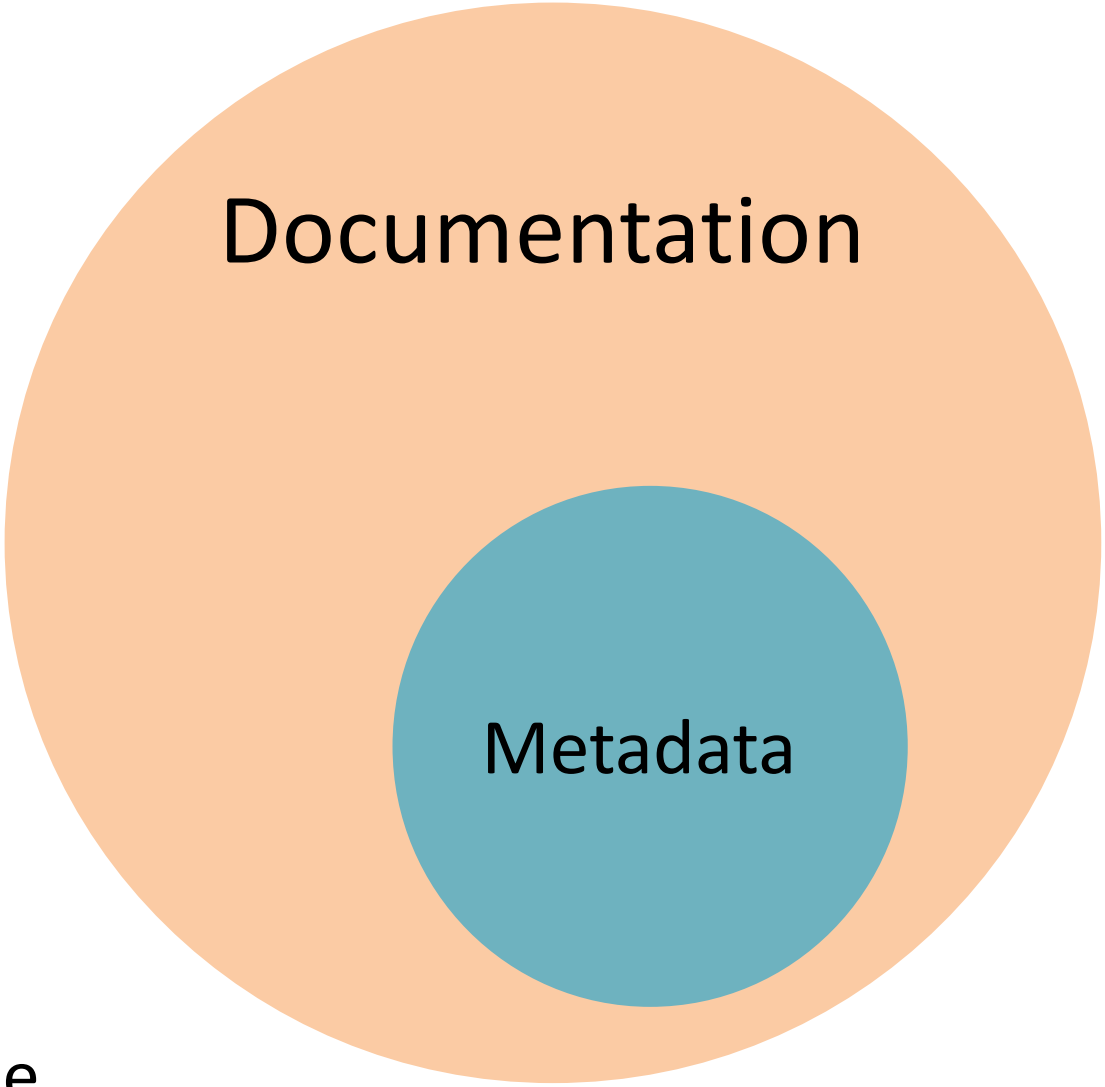
What is metadata?

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

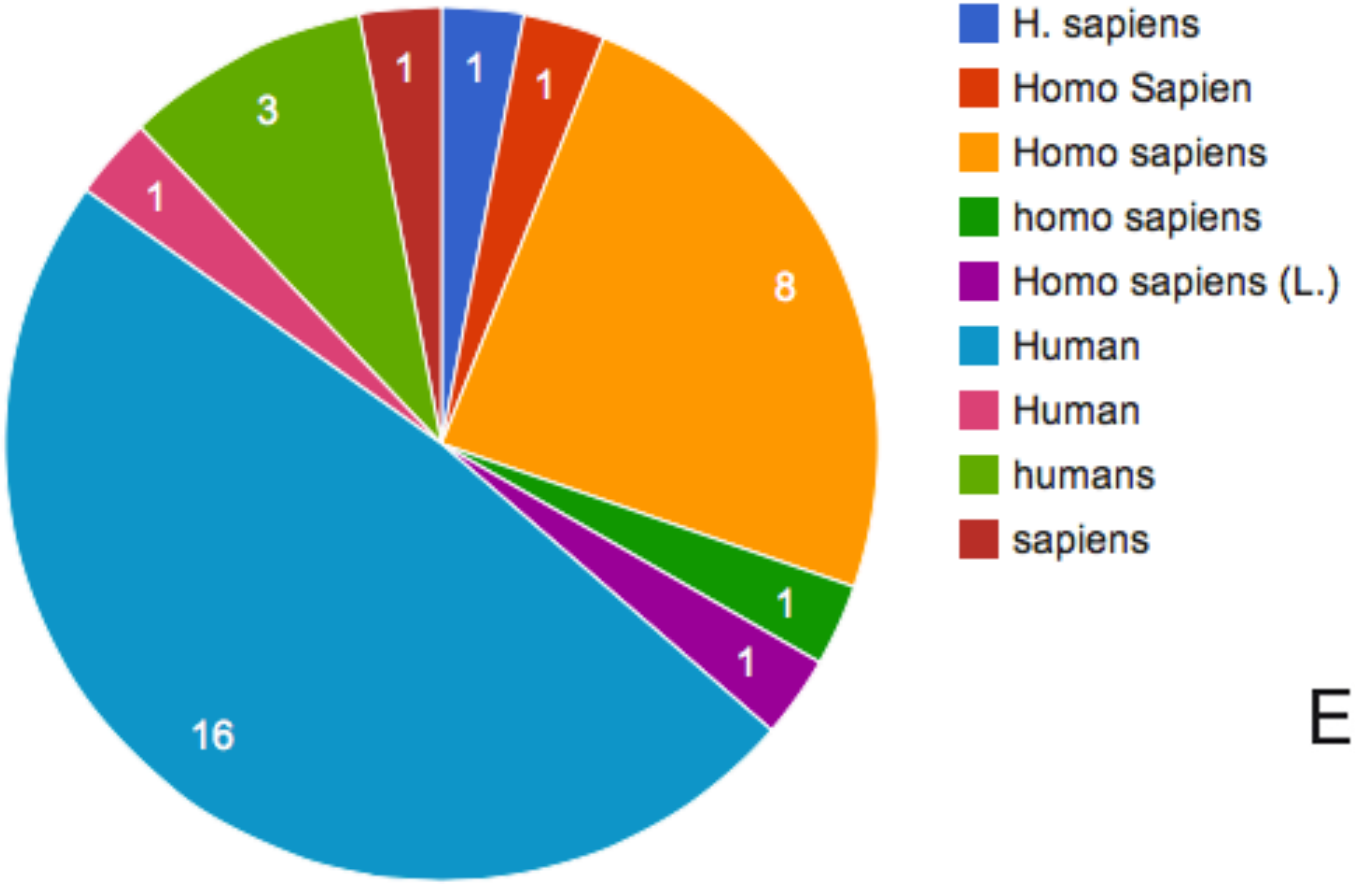
- Data Documentation Initiative (DDI) – social science
- Ecological Metadata Language (EML) - ecology
- Flexible Image Transport System (FITS) – astronomy

Search for standards in catalogues like:

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

Why are ontologies important?

"MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in……"



EMBL-EBI

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

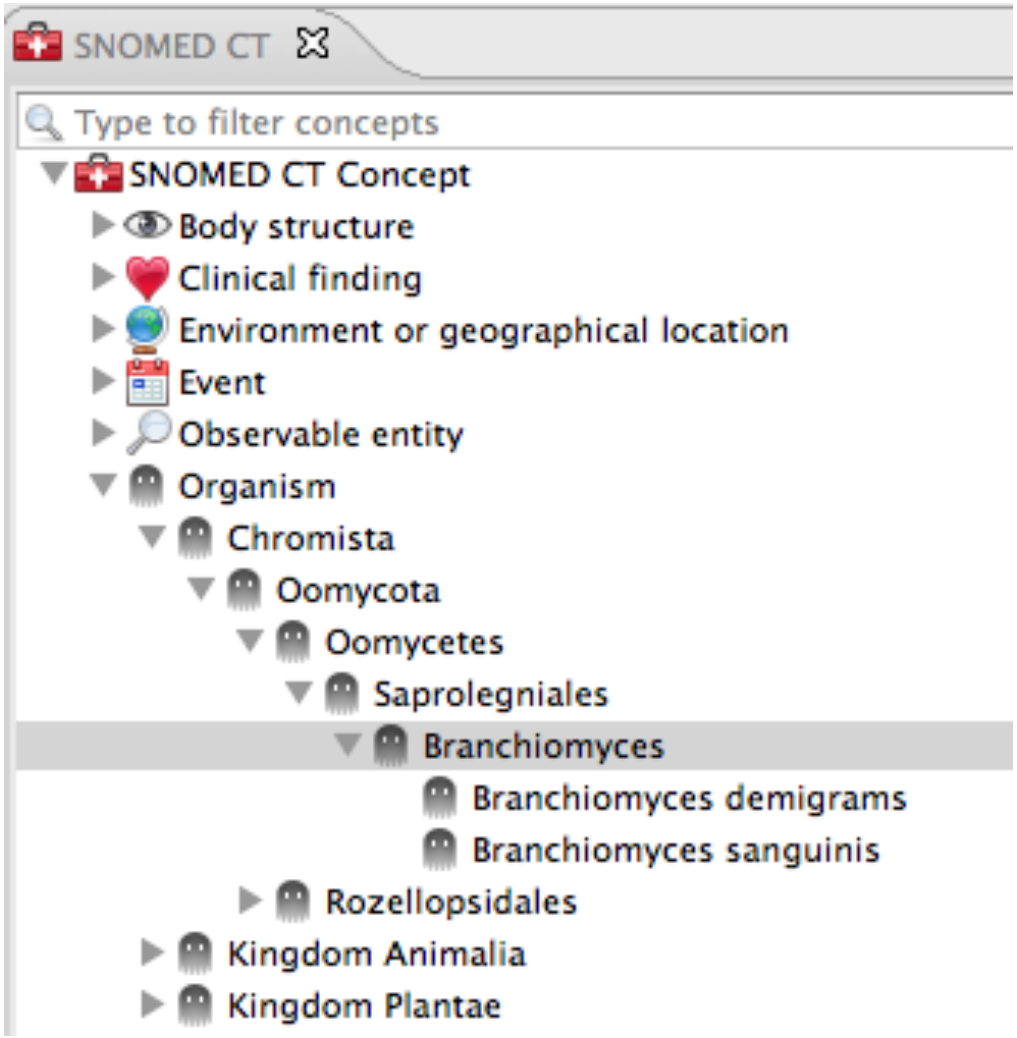
Controlled vocabularies

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

Include ontologies as well

- Defined terms + taxonomy

Useful for selecting keywords to tag datasets



Documentation

Think about what is needed in order to evaluate, understand, and reuse the data.

- Why was the data created?
- Have you documented what you did and how?
- Did you develop code to run analyses? If so, this should be kept and shared too.
- Important to provide wider context for trust

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

Example template: https://www.lib.umn.edu/datamanagement/metadata

Useful tools for documentation

E-lab notebooks, wikis etc

- Record experiment procedures and results
- Share protocols

RSpace

jupyter

LabTrove

SCINOTE

http://openwetware.org

Workflow tools e.g. MyExperiment

Version 7 (latest) (of 7) — View version: 7 (latest)

Version created on: 02/09/11 @ 11:43:00 by: Paul Fisher | Revision comment

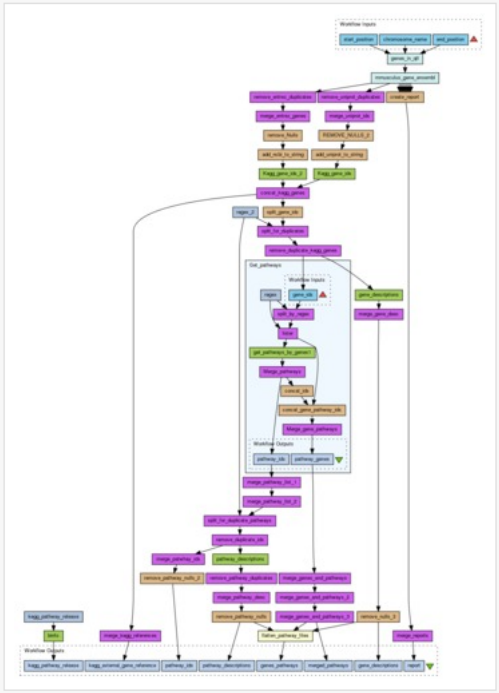
Last edited on: 02/09/11 @ 11:44:57 by: Paul Fisher

Title: Pathways and Gene annotations for QTL region

Type: Taverna 2

Preview

(Click on the image to get the full size)



Download Scalable Diagram (SVG)

Workflow Type
Taverna 2

Original Uploader

Paul Fisher

License
All versions of this Workflow are licensed under:

Credits (1)
(People/Groups)
Paul Fisher

Attributions (0)
(Workflows/Files)
None

Tags (21)

Original Uploader tags

adasd | annotation | chromosome | data-driven | disease | ensembl | entrez | gene | genes | genotype | kegg | mouse | nbiconworkflows | pathway | pathway-driven | pathways | phenotype | qtl | shim | subworkflow | uniprot

Log in to add Tags

Shared with Groups (0)
None

Ratings (10)
Current:
4.6 / 5
(10 ratings)
Log in to rate and see breakdown of ratings

Attributed By (7)
(Workflows/Files)
- The impact of workflow tools on data-centric research
- *Item doesn't exist anymore*
- Pathways and Gene annotations for QTL region
- microRNA to KEGG Pathways and Abstracts
- Pathways and Gene annotations for QTL region
- KEGG Gene IDs to KEGG Pathways
- Pathways and Gene annotations for Arabidopsis affy data

Favourited By (11)
- Katy Wolstencroft
- David Withers
- Taverna
- Xiaoliang
- Kawther
- AbuJarour
- Ali Rezaee
- Delistyle777
- Gamble
- Wotan
- Stian Soiland-Reyes

myexperiment

www.myexperiment.org/workflows/16.html

Managing data

Where will you store the data?

- Your own device (laptop, flash drive, server etc.)
  - And if you lose it? Or it breaks?
- Departmental drives or university servers
- “Cloud” storage
  - Do they care as much about your data as you do?

The decision will be based on how sensitive your data are, how robust you need the storage to be, and who needs access to the data and when

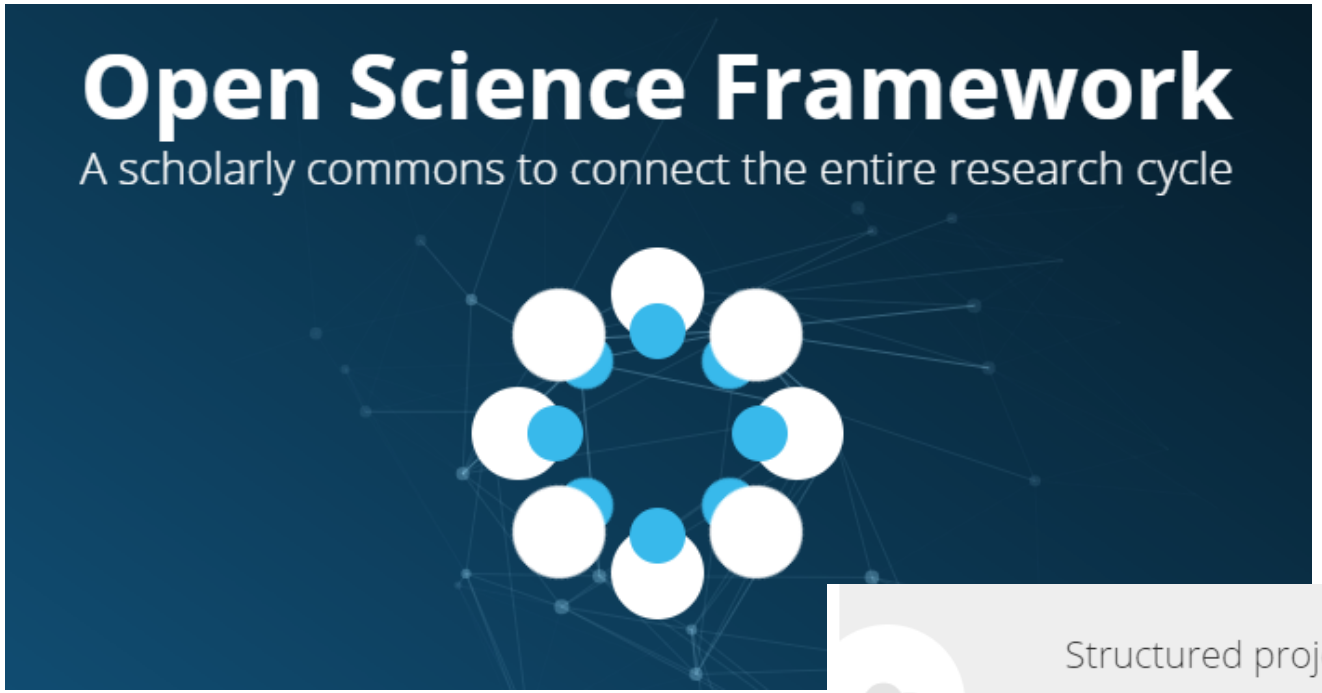
How to keep you data secure?

Develop a practical solution that fits your circumstances

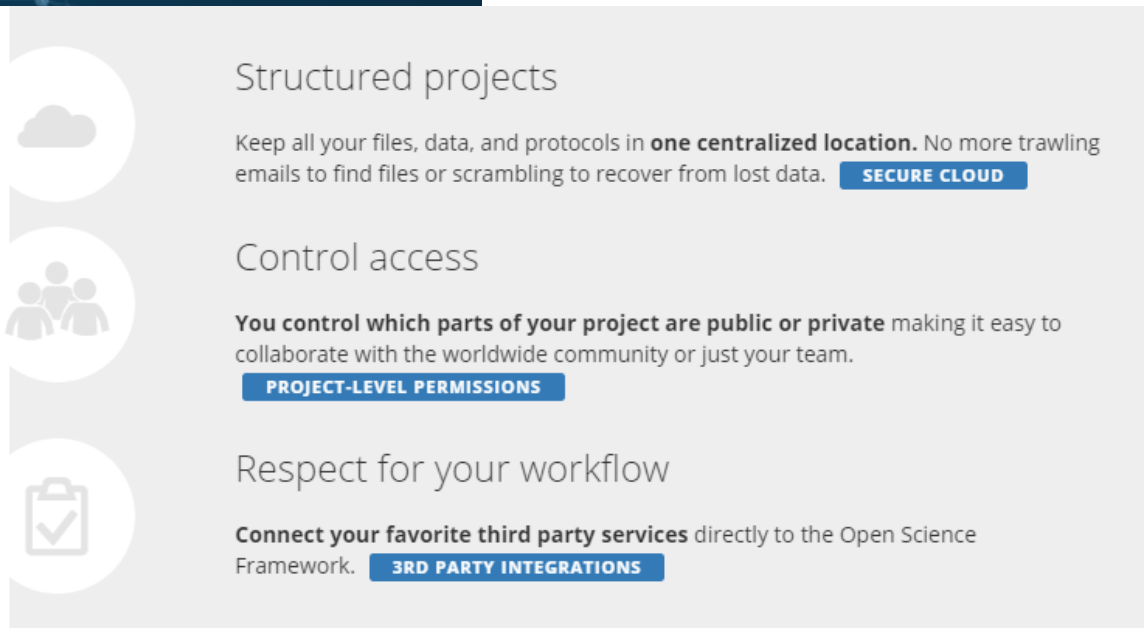
- Store your data on managed servers
- Restrict access to collaborators or smaller subset
- Encrypt mobile devices carrying sensitive information
- Keep anti-virus software up-to-date
- Use secure data services for long-term sharing

www.wsj.com/articles/SB10001424052748703843804575534122591921594

Collaborative platforms e.g. OSF



https://osf.io

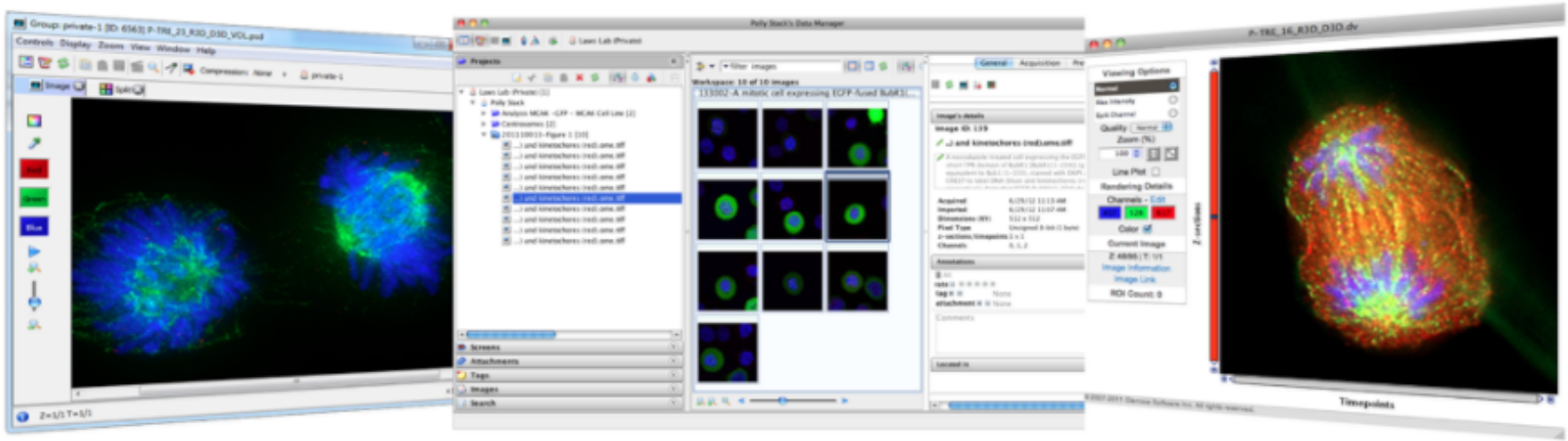


Data-specific platforms e.g. OMERO

OMERO Open Microscopy Environment

What is OMERO?

From the microscope to publication, OMERO handles all your images in a secure central repository. You can view, organize, analyze and share your data from anywhere you have internet access. Work with your images from a desktop app (Windows, Mac or Linux), from the web or from 3rd party software. Over 140 image file formats supported, including all major microscope formats.



Import **Organize** **View** **Analyze** **Publish** **Export**

http://www.openmicroscopy.org/site/products/omero

Third-party tools for collaboration

Dropbox

Using Dropbox and other cloud services – LSE guidelines

http://www.lse.ac.uk/intranet/LSEServices/IMT/guides/softwareGuides/other/usingDropboxCloudStorageServices.aspx

ownCloud

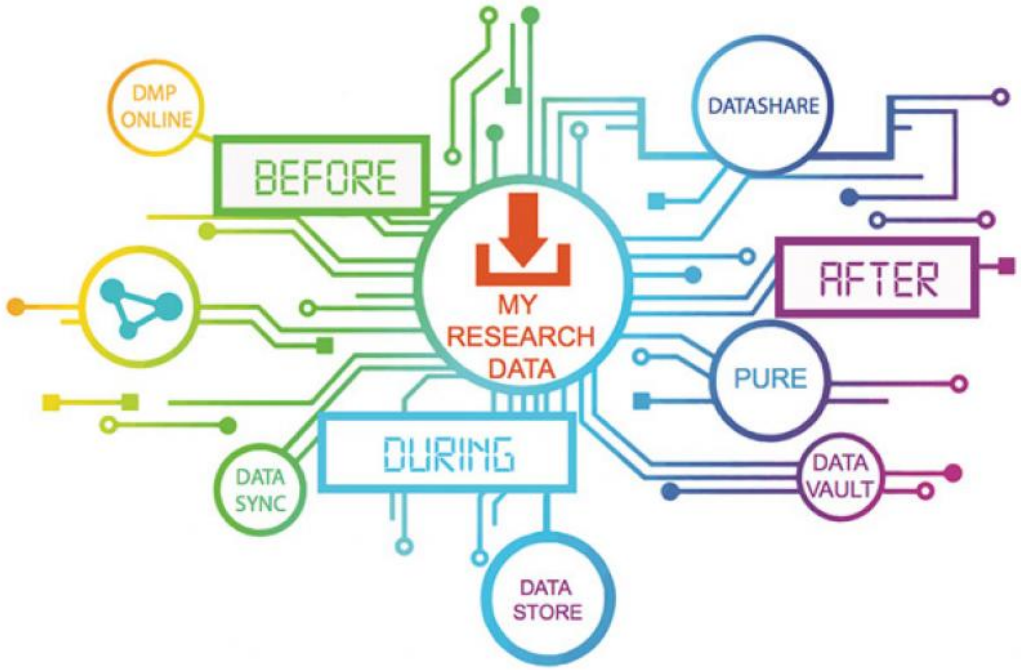
- Open source product with Dropbox-like functionality
- Used by many unis and service providers to offer 'approved' solution

https://owncloud.org

ownCloud

University RDM services e.g. Edinburgh

- DataStore
- Compute & Data Facility (HPC)
- DataSync
- Wiki service
- Subversion
- Electronic Lab Notebook
- DataShare repository
- DataVault
- Pure (research info)
- Secure data service



www.ed.ac.uk/information-services/research-support/research-data-service

One copy = risk of data loss

CC image by Sharyn Morrow on Flickr

CC image by momboleum on Flickr

Who will do the backup?

Use managed services where possible (e.g. University filestores rather than local or external hard drives), so backup is done automatically

3… 2… 1… backup!

at least 3 copies of a file
on at least 2 different media
with at least 1 offsite

Ask central IT team for advice

Backup and preservation – not the same thing!

Backups

- Used to take periodic snapshots of data in case the current version is destroyed or lost
- Backups are copies of files stored for short or near-long-term
- Often performed on a somewhat frequent schedule

Archiving

- Used to preserve data for historical reference or potentially during disasters
- Archives are usually the final version, stored for long-term, and generally not copied over
- Often performed at the end of a project or during major milestones

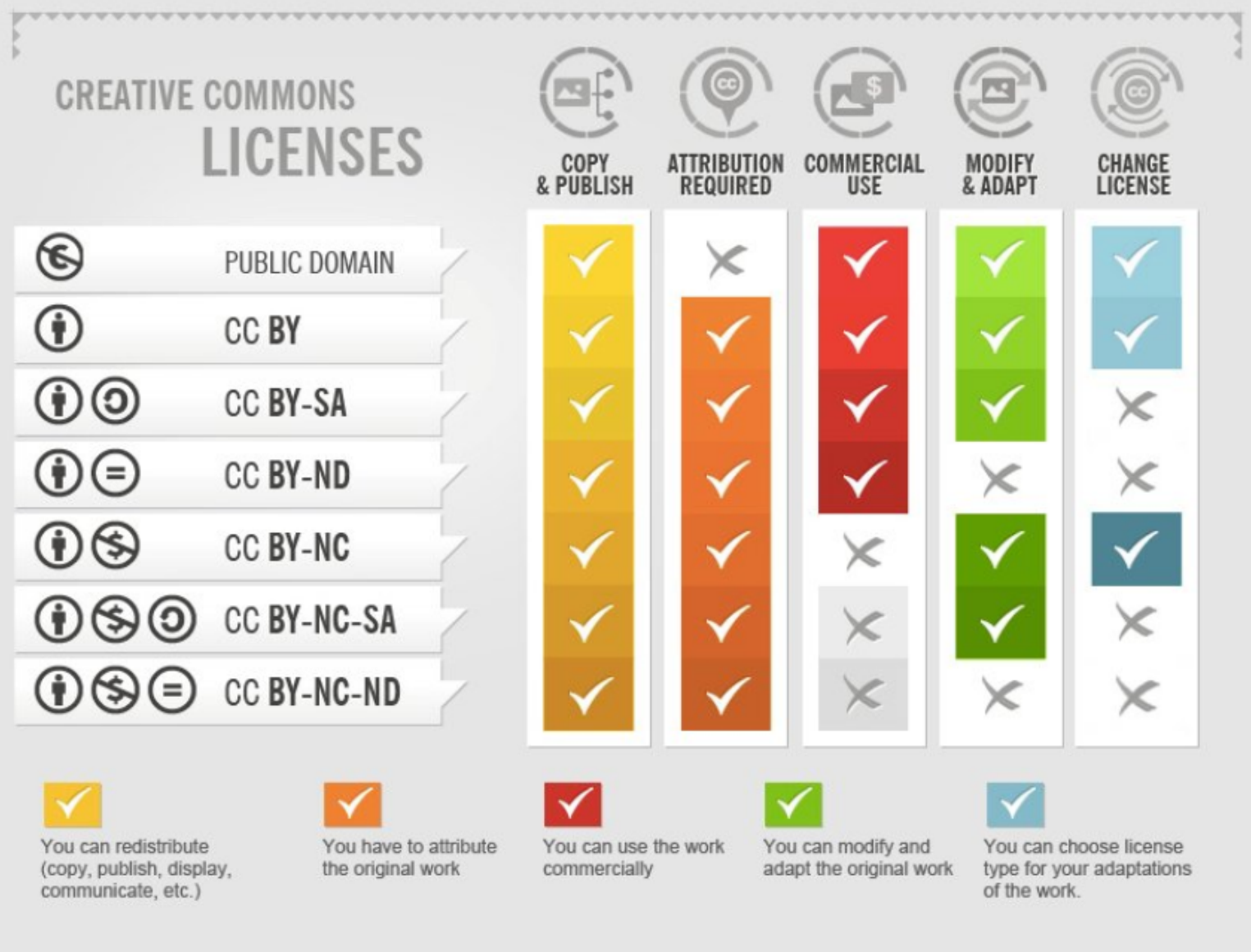
Data sharing

How 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…

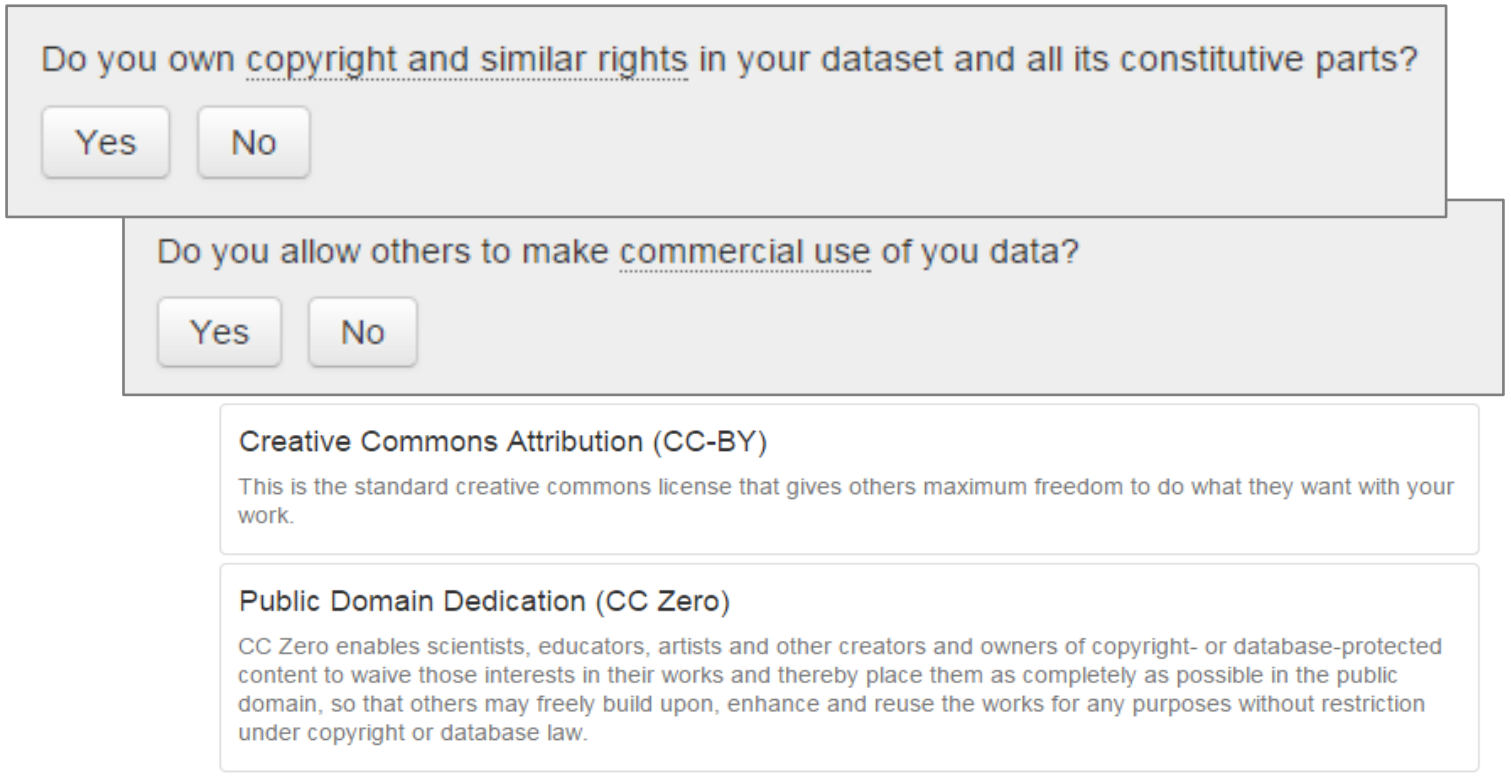
License research data openly



DCC how-to guide: www.dcc.ac.uk/resources/how-guides/license-research-data

EUDAT licensing tool

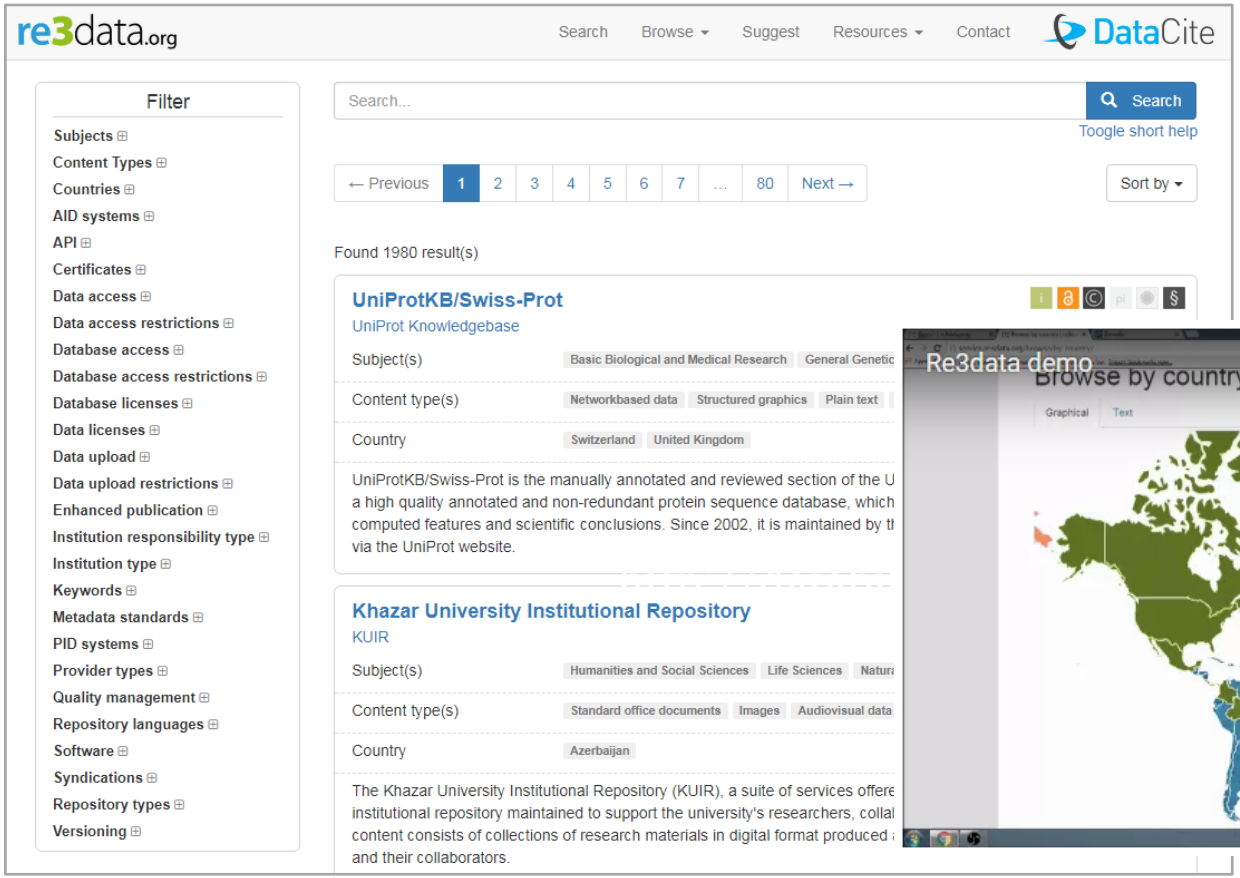
Answer questions to determine which licence(s) are appropriate to use

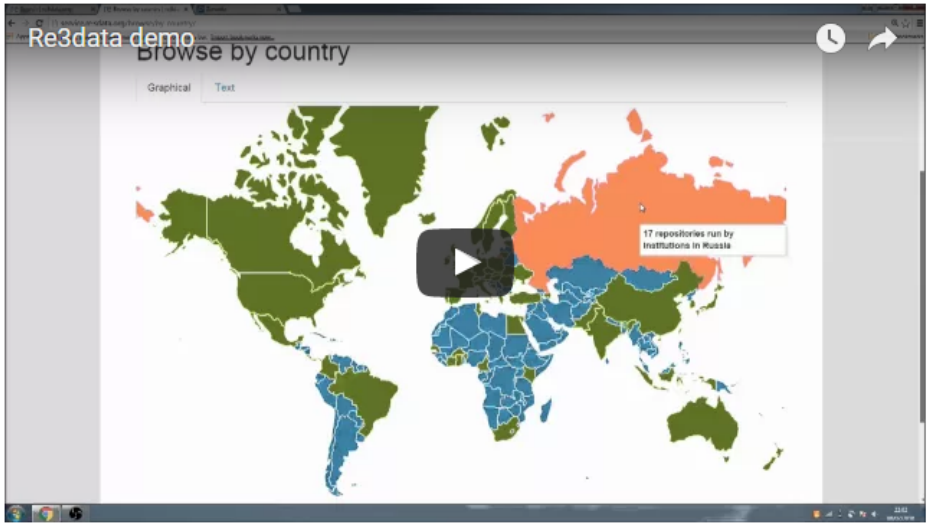


http://ufal.github.io/lindat-license-selector

Deposit in a data repository

The Re3data catalogue can be searched to find a home for data





www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

National / domain repositories

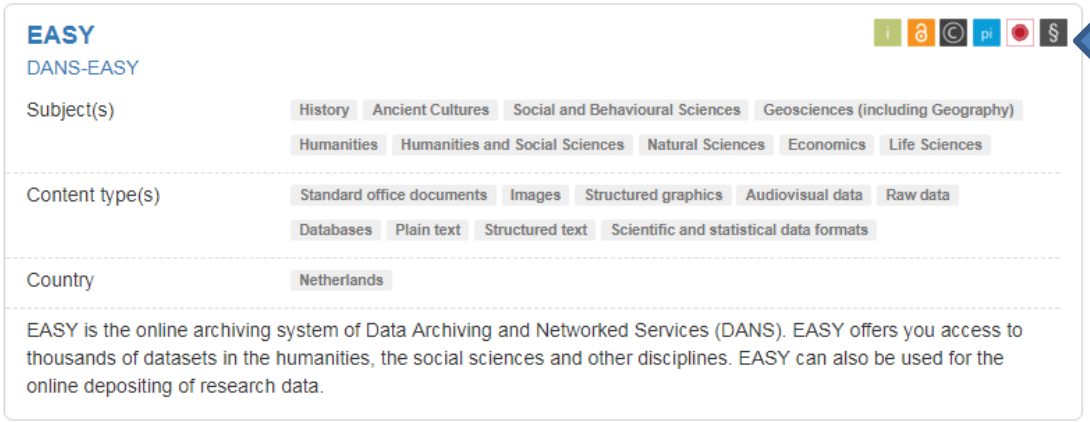
www.re3data.org

BioSharing portal of databases in life sciences

https://biosharing.org

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…

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

Archiving code in Zenodo

Making Your Code Citable

10 minute read

Digital Object Identifiers (DOI) are the backbone of the academic reference and metrics system. If you're a researcher writing software, this guide will show you how to make the work you share on GitHub citable by archiving one of your GitHub repositories and assigning a DOI with the data archiving tool Zenodo.

> **ProTip:** This tutorial is aimed at researchers who want to cite GitHub repositories in academic literature. Provided you've already set up a GitHub repository, this tutorial can be completed without installing any special software. If you haven't yet created a project on GitHub, start first by uploading your work to a repository.

Intro

Choosing Your Repo

Login to Zenodo

Check Repo Settings

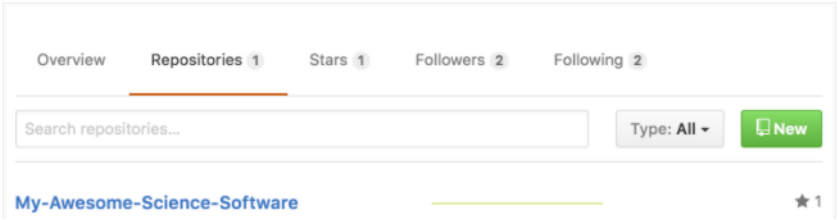
Create a New Release

Minting a DOI

Finishing up

Choose your repository

Repositories are the most basic element of GitHub. They're easiest to imagine as your project's folder. The first step in creating a DOI is to select the repository you want to archive in Zenodo. To do so, head over to your profile and click the **Repositories** tab.



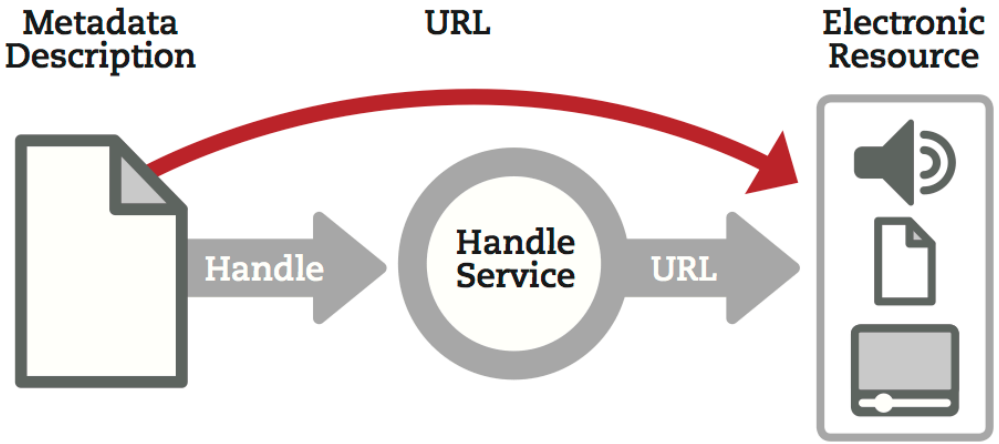
Get a DOI for each release

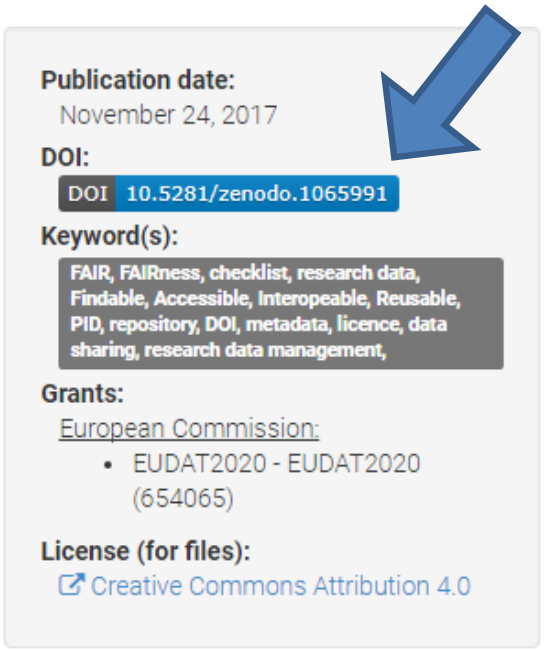
https://guides.github.com/activities/citable-code

What is a Persistent Identifier?

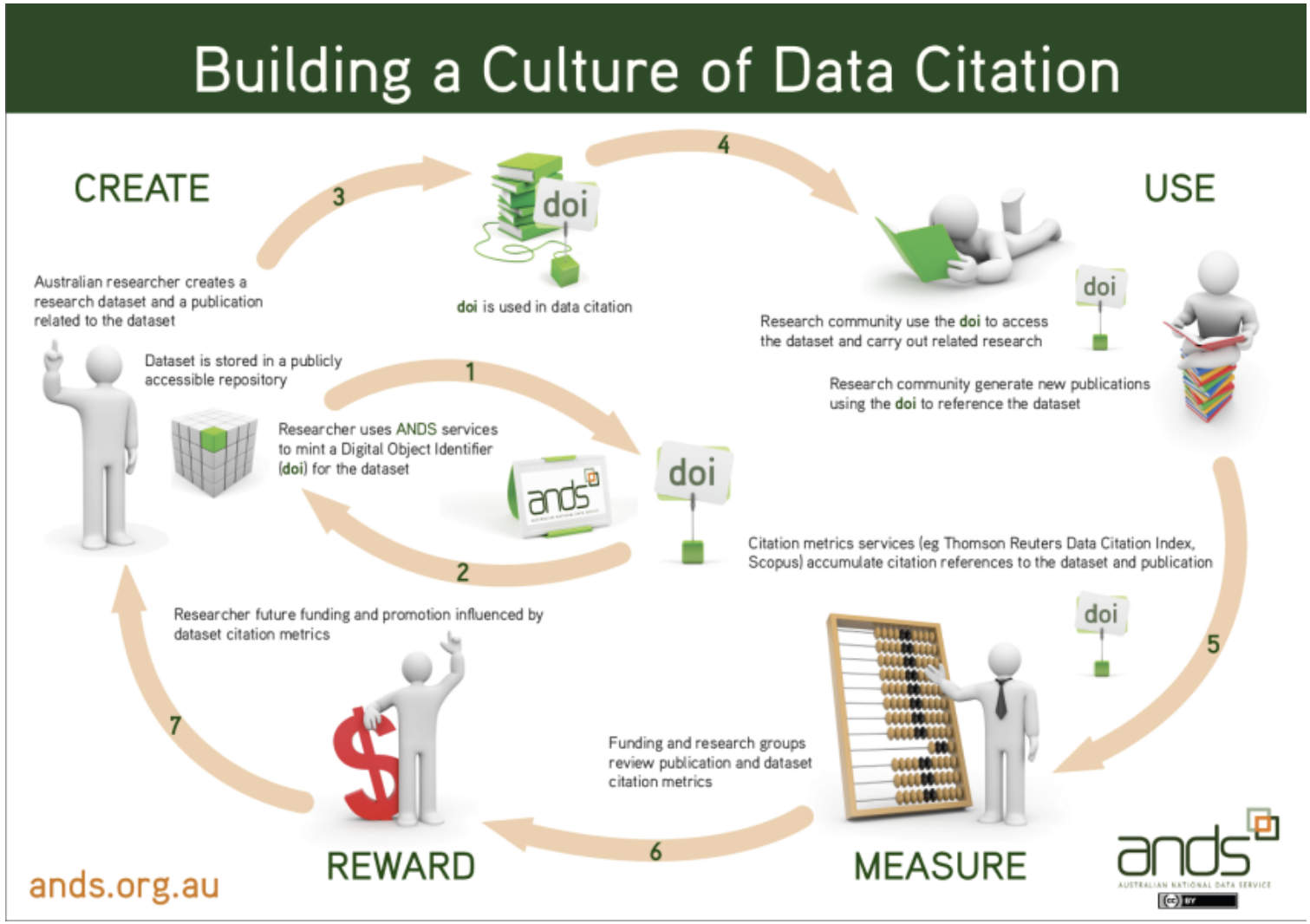
a long-lasting reference to a document, file or other object

- 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





Citing research data: why?

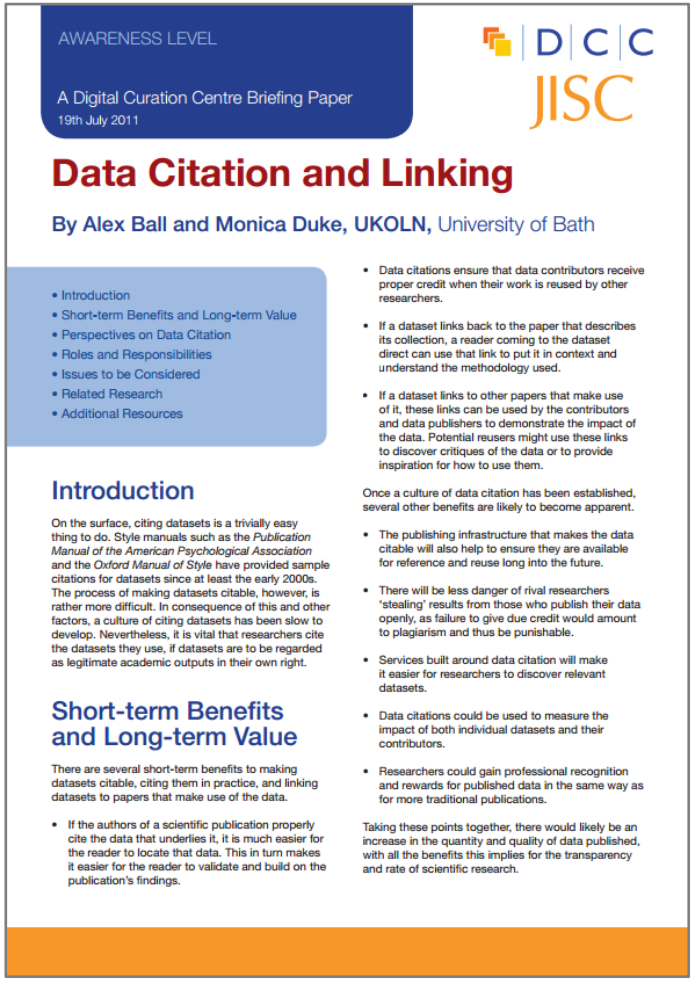


http://ands.org.au/cite-data

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

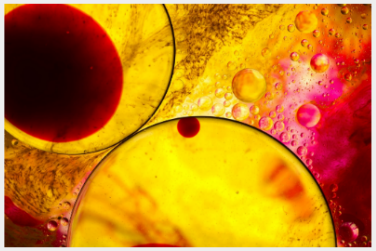
  www.dcc.ac.uk/resources/how-guides/license-research-data

FOSTER Open Science toolkit

What is Open Science?
This introductory course will help you to understand what open science is and why it is something you should care about.

Best Practices
This course introduces funding body policies and other environmental factors that influence good practice in opening up research practice.

Managing and Sharing Research Data
In this course, you'll focus on which data you can share and how you can go about doing this most effectively.



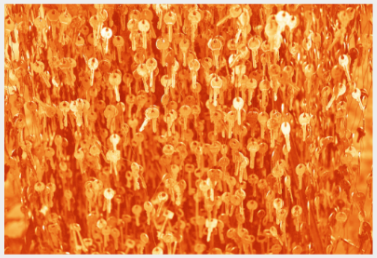
OSS and Workflows
This course introduces Open Source Software (OSS) and workflows as an emerging but critical component of Open Science.

Open Science and Innovation
This course will show you how Responsible Research and Innovation is accelerated through Open Science.

Data Protection and Ethics
This course helps you to get to grips with responsible data sharing.

Licensing (will be released soon)
This course helps you to find the best license for your open research outputs.

Open Access Publishing
This course will help you become skilled in Open Access publication in the wider context of Open Science.



Sharing Preprints
This course introduces the practice of sharing preprints and helps you to see how it can support your research.

Open Peer Review (OPR)
This course will introduce you to OPR and let you know how you can get started with it.

https://www.fosteropenscience.eu/toolkit

Thanks for listening

For DCC resources see:

www.dcc.ac.uk/resources

Follow DCC us on twitter:

@digitalcuration and #ukdcc