Manging and Sharing Research Data

S Venkataraman
Digital Curation Centre, Edinburgh
s.venkataraman@ed.ac.uk

CODATA Research Data Science School, 25-26 October 2018, Kigali, Rwanda

Go to **www.menti.com** and use the code **21 04 1**

What would you like to learn from this session?

Mentimeter

Slide is not active **Activate**

0

What is your background?

Mentimeter

Slide is not active Activate

0

What kinds of data are there?

Mentimeter

Slide is not active Activate

0

Research Data Management

Describe in three words at most what RDM means to you?

Mentimeter

Slide is not active Activate

0

Describe WHY manage your data and the benefits?

Mentimeter

Pause scroll

Slide is not active Activate

0

What benefits are there to freely sharing data?

Mentimeter

Pause scroll

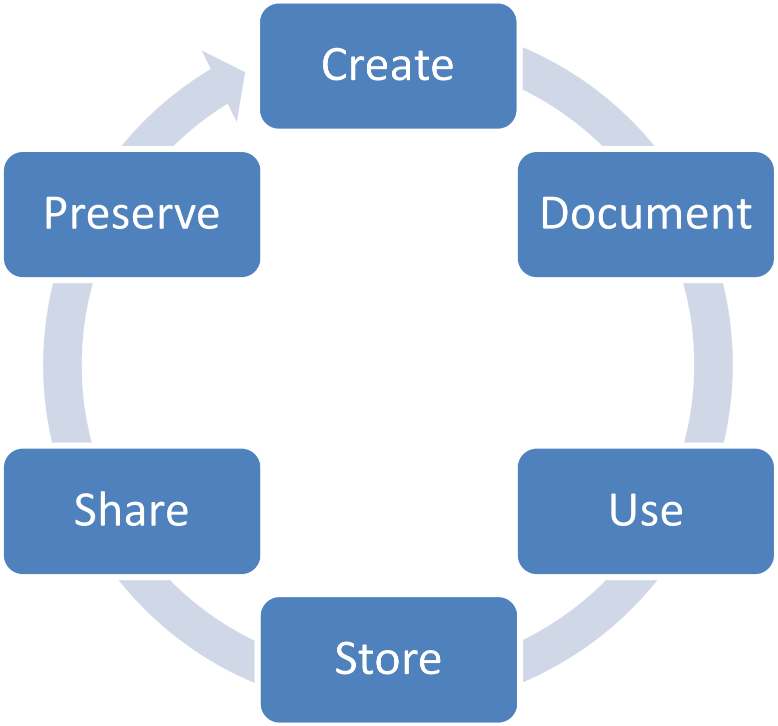
Slide is not active Activate

0

Why manage and share data?

Direct benefits for you	Research integrity	Potential to share data
• To make your research easier! • Stop yourself drowning in irrelevant stuff • Make sure you can understand and reuse your data again later • Advance your career – data is growing in significance	• To avoid accusations of fraud or bad science • Evidence findings and enable validation of research methods • Meet codes of practice on research conduct • Many research funders worldwide now require Data Management and Sharing Plans	• So others can reuse and build on your data • To gain credit – several studies have shown higher citation rates when data are shared • For greater visibility, impact and new research collaborations • Promote innovation and allow research in your field to advance faster

What is Research Data Management?

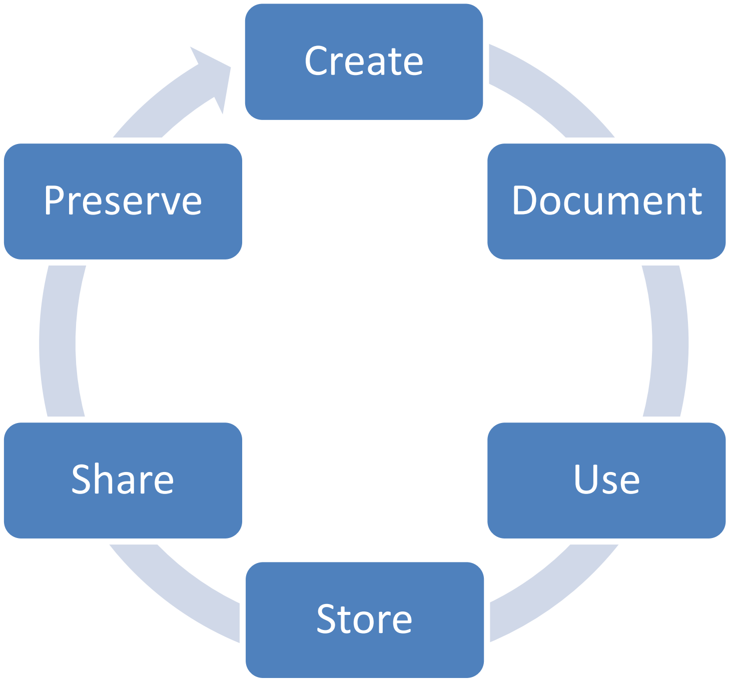


"the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

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…



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

Why make data available?

Making plans

They sound dull, but data-management plans are essential, and funders must explain why.

Data are the alpha and omega of scientific and social research. A versatile good, they exist both as raw material for producing knowledge and, when processed and interpreted with an expert eye, the end product of the exercise.

So it might sound like a truism that researchers should conscientiously handle, preserve and — where appropriate — share the data they generate and use. The problem is that this can be hard to do.

As science produces day by day a huge volume of data, it's a growing challenge to manage and store this information. To encourage this, many funders now ask applicants to submit a concise data-management plan with their grant proposals: effectively, a to-do list that details how they plan to collect, clean, store and share the products of their research.

Such plans are important, and are something that *Nature* supports (we discuss them in detail in a Careers article on page 403). But to accelerate acceptance of what some might deem just another administrative burden, science funders and research institutions must work to streamline the process and to explain the need and benefits.

First, rigorously collected, well-preserved data sets — including meaningful descriptors or metadata — will help the data owners to reach solid, meaningful results. Second, they will help future investigators to make sense of and reuse data, thereby enhancing utility and reproducibility. Preserving comprehensive data, ideally for many years, also reduces the risk of duplicating science done by others.

Still, there is no single recipe for proper data management. The task varies according to the field of science, project size and the specific types of data in question. That makes cross-disciplinary common standards unlikely, so research agencies need to engage with different scientific communities to create formats that best serve specific disciplines. To avoid a hotchpotch of standards, formats and data protocols — undesirable in our increasingly global scientific enterprise — research agencies in all parts of the world must engage.

An initiative for voluntary international alignment of research data-management policies, launched in January by Science Europe and the Netherlands Organisation for Scientific Research, is an important step in that direction. And existing data stewardship in particle physics and genomics shows that internationally aligned data governance not only is perfectly doable, but also has a positive impact on collaborative research. NASA pioneered this approach, setting up a centre in the 1980s to specifically curate the data from the Infrared Astronomical Satellite.

The message must now be passed on to scientists who work in fields less familiar with big data. Many of these, at all career stages, are worryingly unprepared. A survey of European researchers last year revealed that many have never been asked to provide a data-management plan, and that most are unaware of policies and guidelines already in place to help them. Only one-quarter of respondents to the survey, carried out by the European Commission and the European Council of Doctoral Candidates and Junior Researchers, had actually written a data-management plan, with another quarter saying they didn't even know what such a plan might be. There is nothing to suggest Europe is unusual in this.

Funders and universities, then, must ensure that the rationale of data management, and the basic skills of exercising it properly, become part of postgraduate education everywhere. Training and support must go further and be offered at every career level.

The laudable move towards open science — under which data are shared — makes the need for good data management more pressing than ever: there's no point in sharing data if they aren't clean and annotated enough to be reused. If you haven't got a plan for your data, you need one now. ■

CAREERS

PERSONAL ETHICS How a vegetarian biologist balances his beliefs with his work **p.405** | **BLOG** Personal stories and careers counsel **http://blogs.nature.com/naturejobs** | **NATUREJOBS** For the latest career listings and advice **www.naturejobs.com**

ADAPTED FROM GETTY

DATA MANAGEMENT

For the record

Making project data freely available is vital for open science.

BY QUIRIN SCHIERMEIER

When Marjorie Etique learnt that she had to create a data-management plan for her next research project, she was not sure exactly what to do.

The soil chemist, a postdoc at the Swiss Federal Institute of Technology (ETH) in Zurich, studies the interaction of trace elements in sediments and water. While preparing a grant proposal for the Swiss National Science Foundation last October, she learnt of the funder's new data rules. These require applicants to provide a written plan for the organization and long-term storage of their research data, to help minimize the risk of data loss and provide guidance for other scientists on how to use the data in the future.

Etique found the task daunting. "Data management is really not my primary skill," she says. "I had absolutely no idea how to go about it." She was able to get advice from her supervisor and from ETH's digital library service. Other researchers might not be so lucky, and might not even know what a data-management plan is — let alone why they would need one and how to produce it. Here, we answer these questions.

WHAT ARE DATA-MANAGEMENT PLANS?

A data-management plan explains how researchers will handle their data during and after a project, and encompasses creating, sharing and preserving research data of any type, including text, spreadsheets, images, recordings, models, algorithms and software. It does not matter whether the data are generated by large pieces of research equipment, such as imaging tools or particle accelerators, or from straightforward field observation.

Many funders are asking grant applicants to provide data plans. Requirements vary from one discipline to another. But in general, scientists will need to describe — before they begin any research — what data they will generate; how the data will be documented, described, secured and curated; and who will have access to those data after the research is completed. They must also explain any data sharing and reuse restrictions, such as legal and confidentiality issues. Researchers can consult their funder and their host institute's digital library services for assistance. Colleagues who have previously produced data plans may also be able to help (see 'Keeping stock').

WHO NEEDS THEM?

Data management is one example of the way in which public research sponsors and research institutions are implementing 'open science', the push to make scientific research and data freely accessible. Many funding agencies have made data-management plans mandatory for grant applicants in the past decade or so. All US federal agencies, including the National Science Foundation and the National Institutes of Health, have such policies. Data-management plans must also now be included in grant proposals to the European Research Council and other European Union-funded research programmes. And many national funding agencies in Europe — including the UK research councils and the London-based Wellcome Trust, world's largest biomedical research charity — also ask for data plans.

Many scientists already practise data management by default. Astronomers, for example, have been doing so for decades when calibrating their observations and archiving huge amounts of telescope-survey data in standardized, machine-readable catalogues for reuse.

Geneticists, too, use special data repositories to archive the vast amounts of DNA and genome-sequencing data (see go.nature.com/2omlrbe). But less data-intensive fields of science and social research also benefit from data management. For example, geochemists analysing soil bacteria and mineral products in different environments can use it to ▶

https://www.nature.com/articles/d41586-018-03071-1

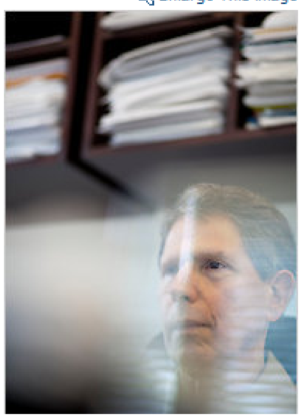
Sharing leads to breakthroughs

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?

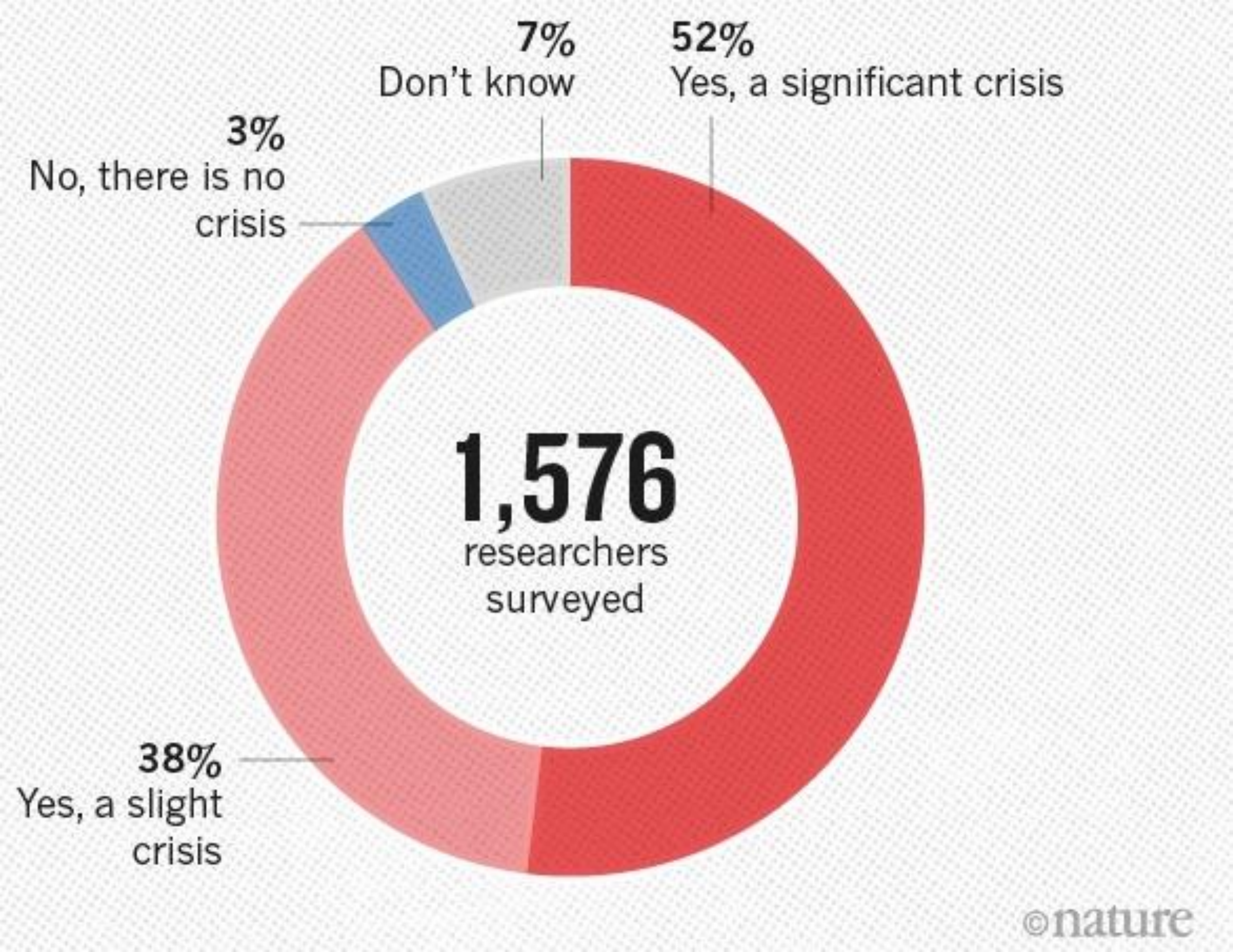
- Piwowar, Vision – 9% (microarray data)
- Drachen, Dorch, et al – 25-40%, astronomy
- Gleditch, et al – doubling to trebling (international relations)

Open Data Citation Advantage

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

FAIR Principles

Is there a reproducibility crisis?



Baker, M. (2016) "1,500 scientists lift the lid on reproducibility", *Nature*, *533:7604*, http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

RESEARCH DATA - OPEN BY DEFAULT

Who has heard of FAIR?

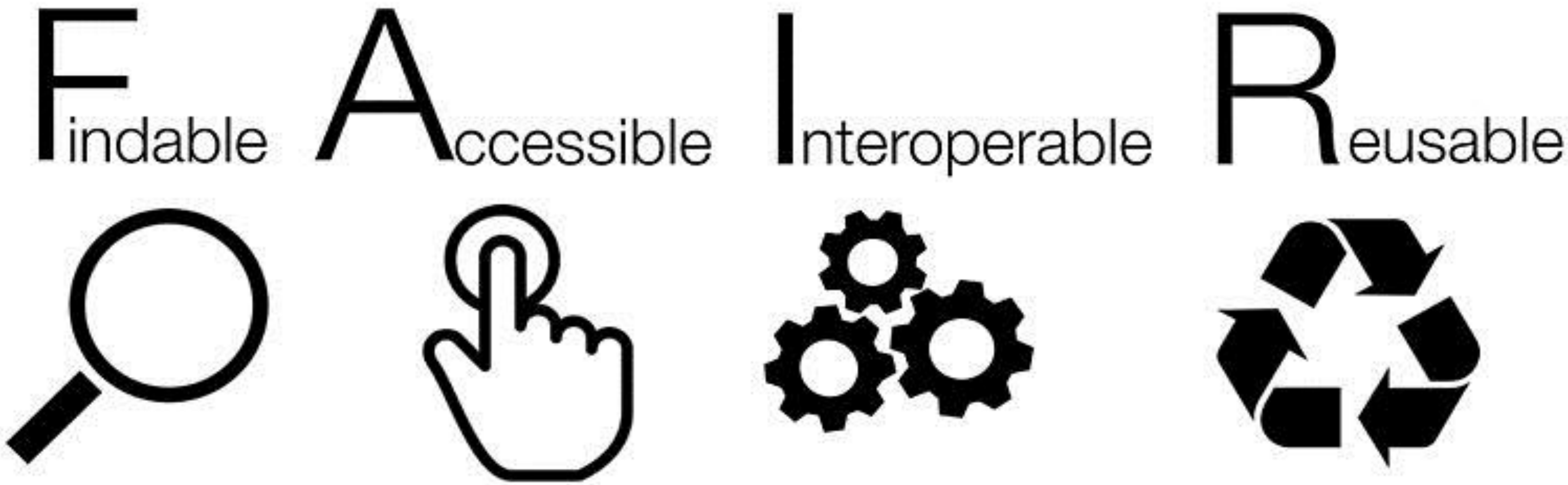


Image CC-BY-SA by [SangyaPundir]

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;

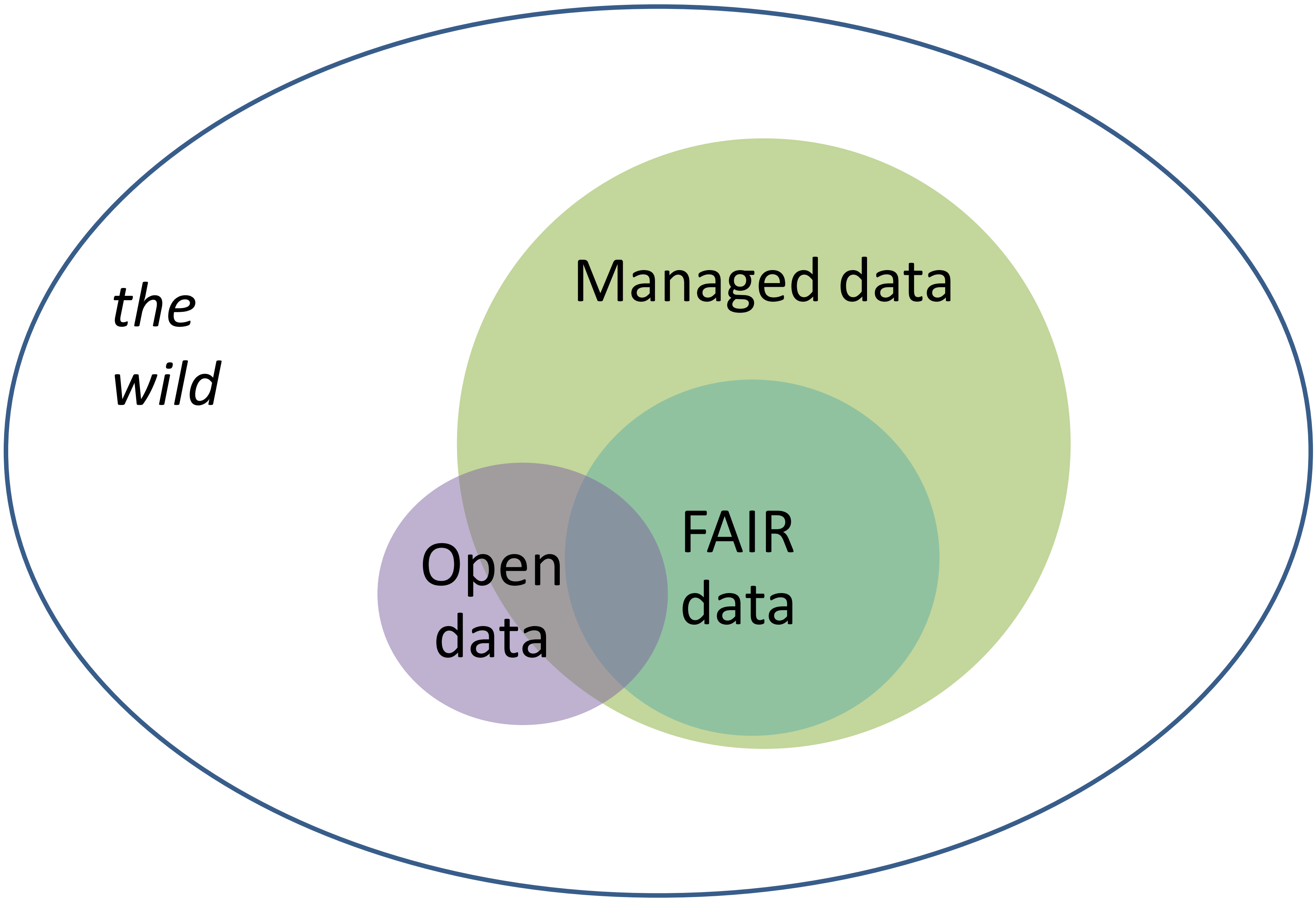
R1.1. (meta)data are released with a clear and accessible data usage license;

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

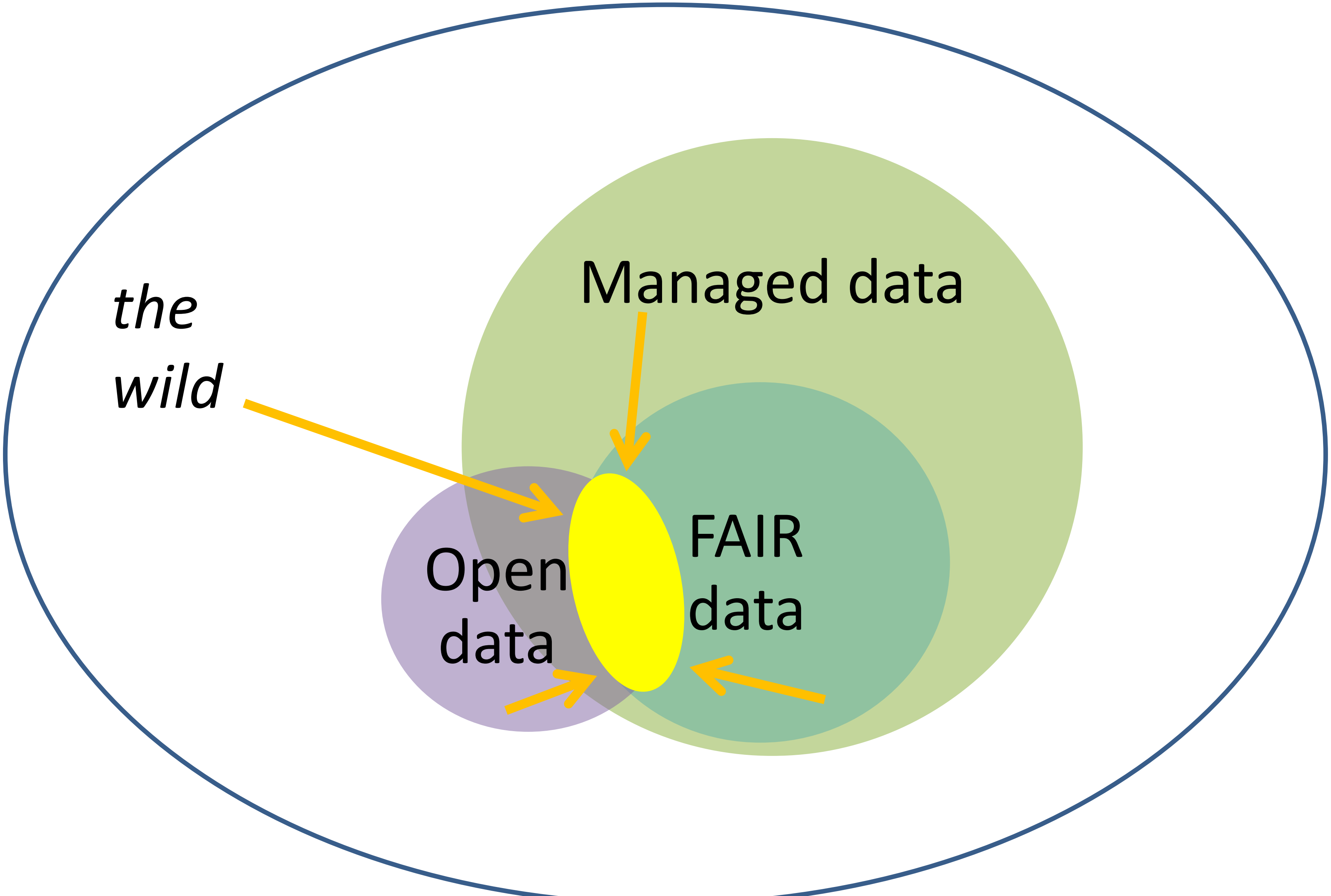
R1.3. (meta)data meet domain-relevant community standards;

doi: 10.1038/sdata.2016.18

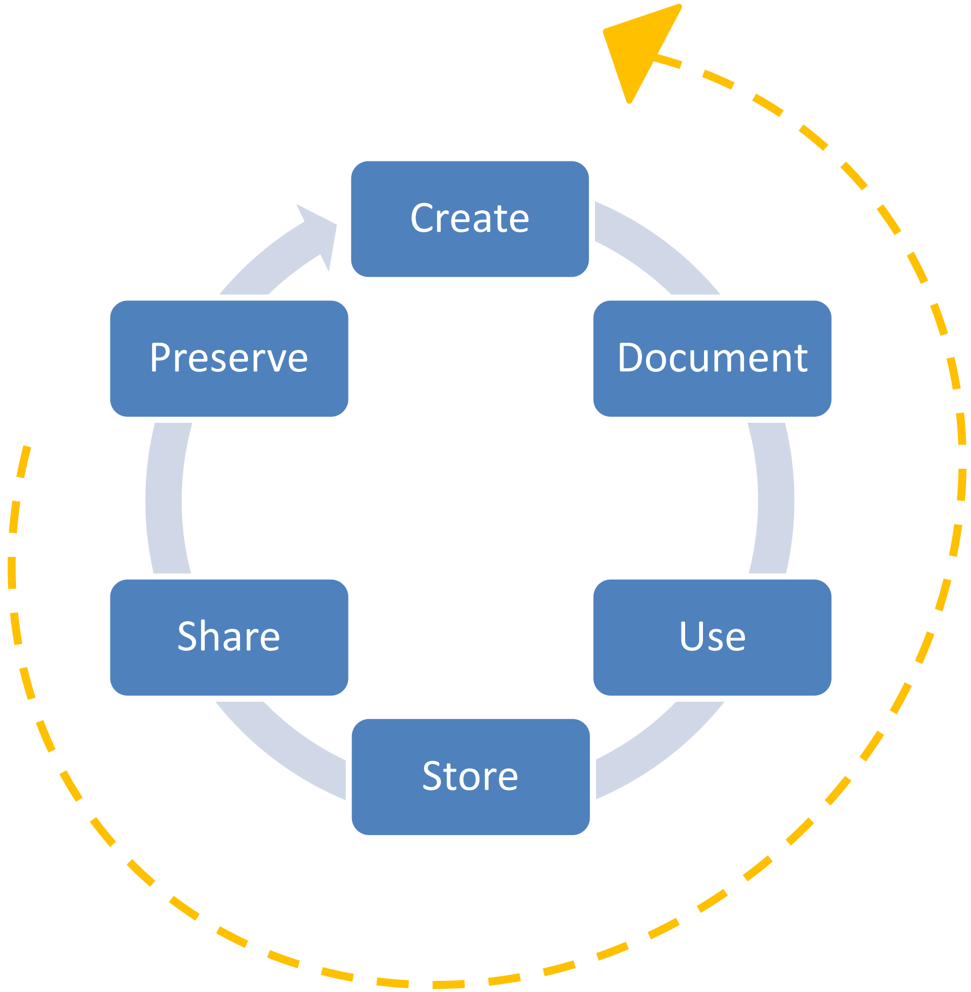
All research data



Increasing that which is FAIR & open



Openness at every stage

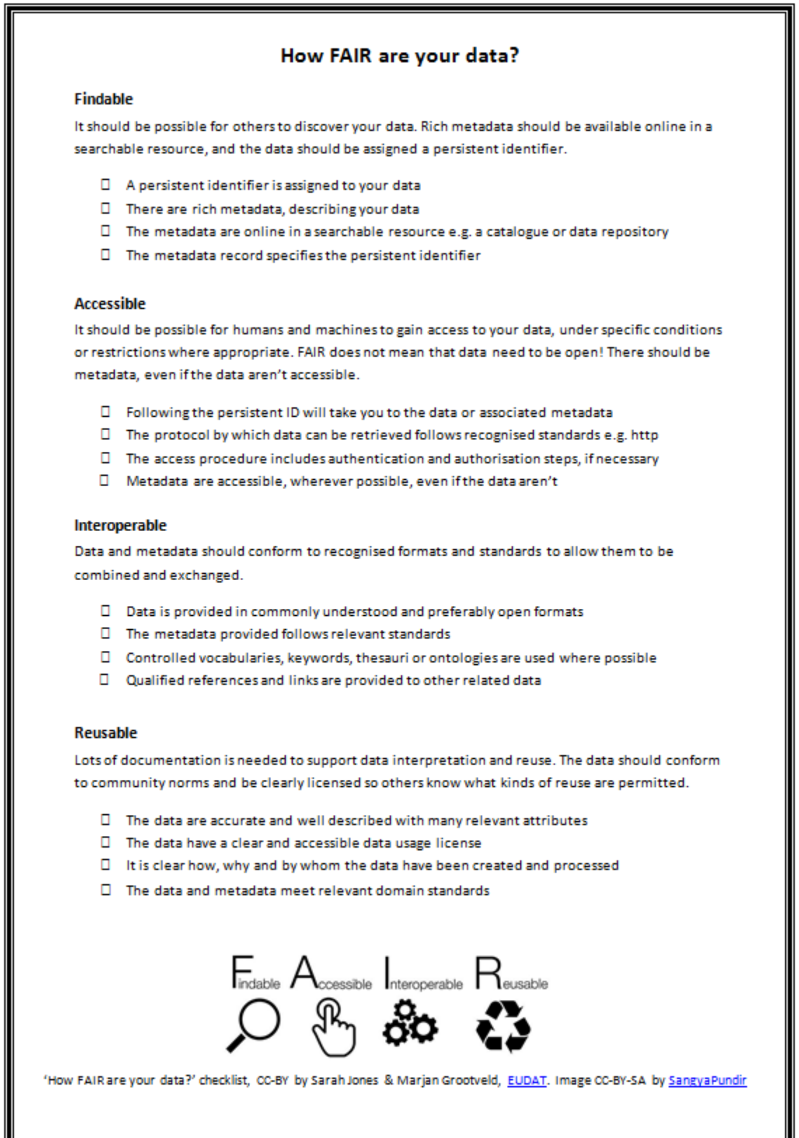


- Change the typical lifecycle
- Publish earlier and release more
- Papers + Data + Methods + Code…
- Support reproducibility

as open as possible, as closed as necessary

Image: 'Balancing rocks' by Viewminder CC-BY-SA-ND
www.flickr.com/photos/light_seeker/7780857224

How FAIR are your data?



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

- Complete the FAIR data checklist
- Base decisions on how you currently manage and share your data
- Which are the most challenging aspects of FAIR to meet?

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

What file formats do you use?

Mentimeter

Slide is not active Activate

0

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

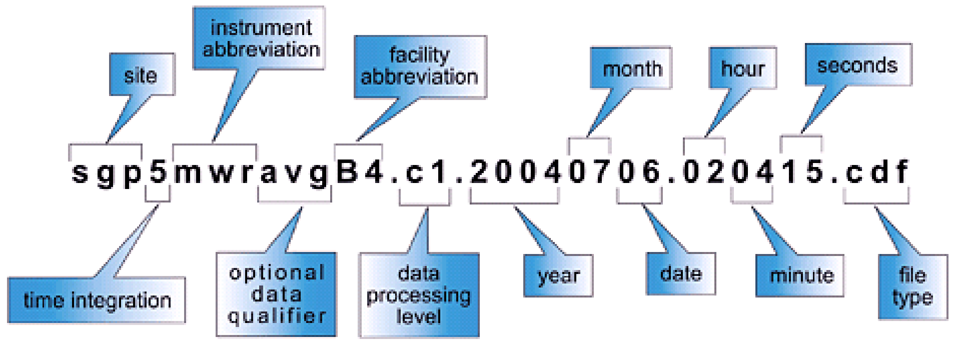
Data centres may suggest preferred formats for deposit

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

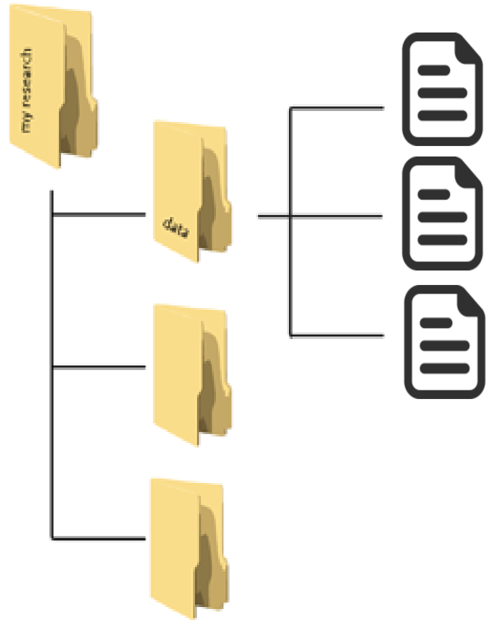
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

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

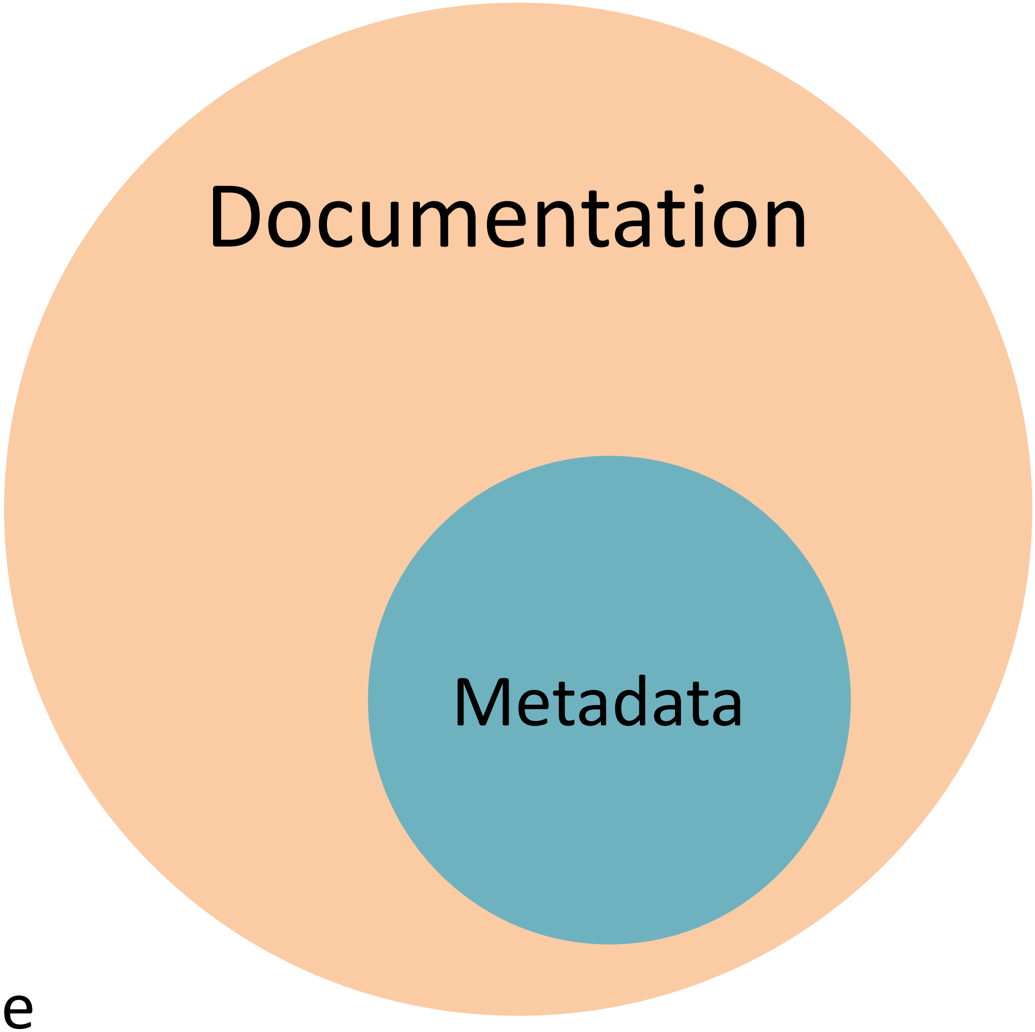
What is metadata?

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

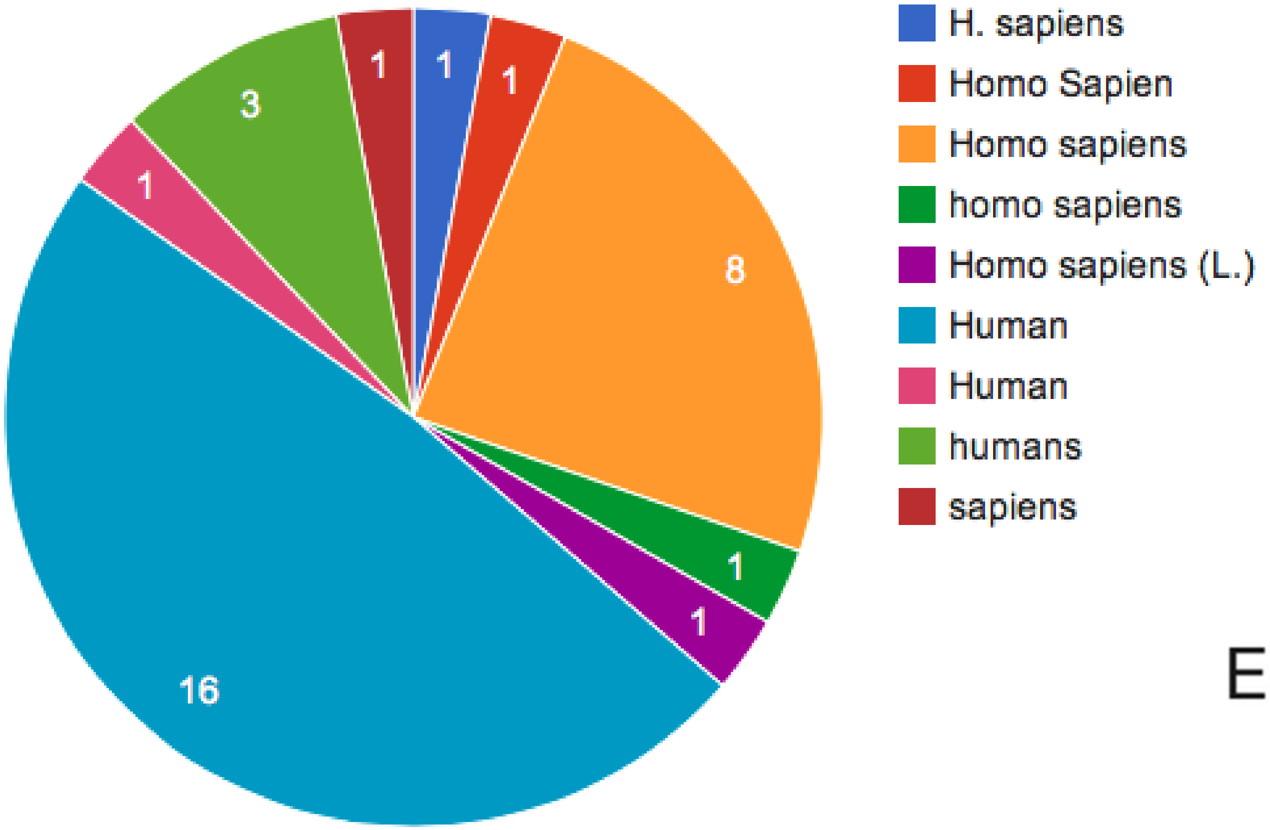
- Data Documentation Initiative (DDI) – social science
- Ecological Metadata Language (EML) - ecology
- Flexible Image Transport System (FITS) – astronomy

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……"



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

- **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*

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

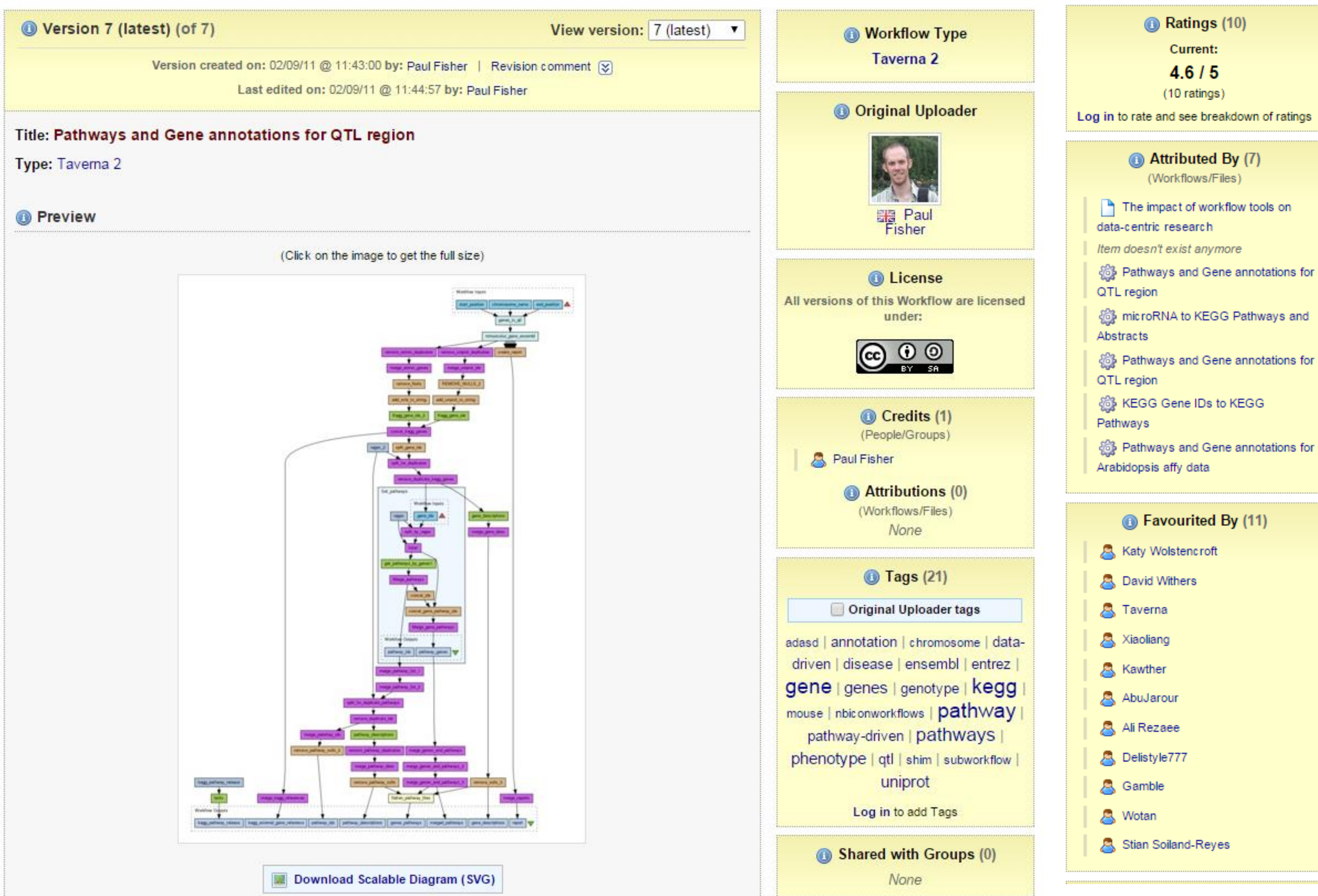
jupyter

LabTrove

SCINOTE

http://openwetware.org

Workflow tools e.g. MyExperiment



my experiment

www.myexperiment.org/workflows/16.html

Managing and sharing data

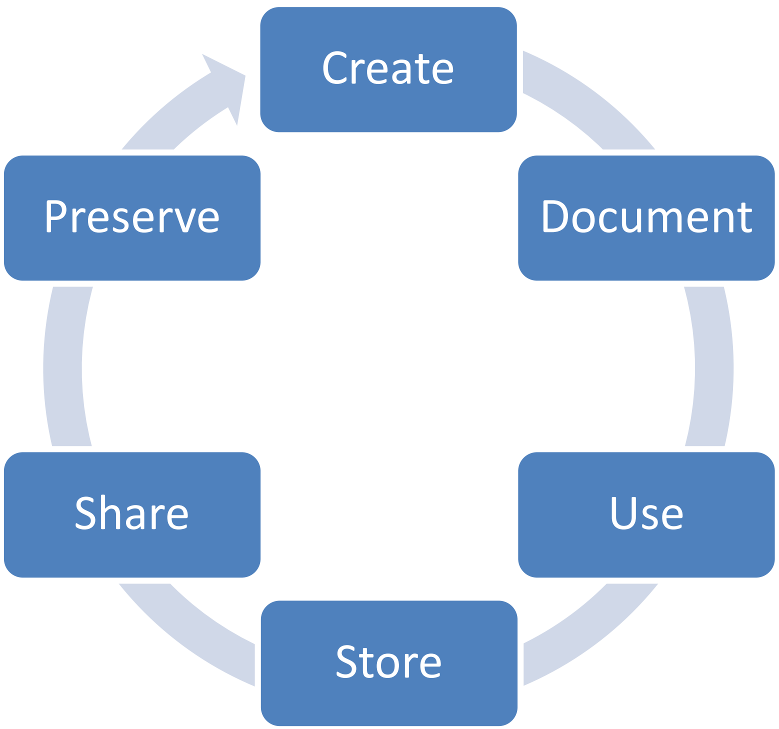
What tools are you using to analyse, store and share data?

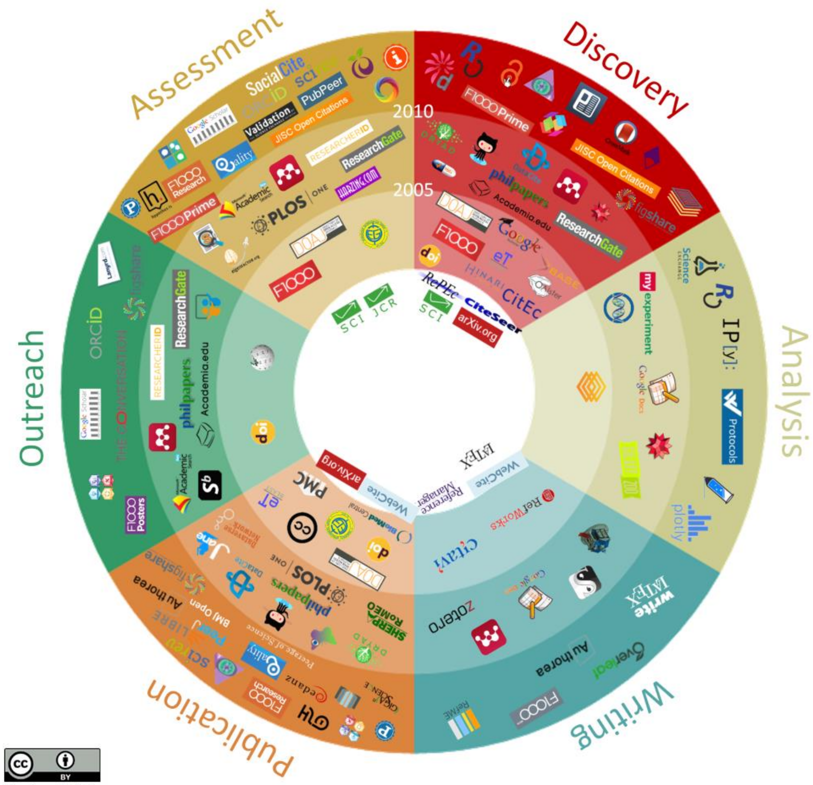
Mentimeter

Slide is not active Activate

0

Different tools in different stages of the lifecycle…





101 Innovative tools and sites in 6 research workflow phases (< 2000 - 2015)

Most important developments in 6 research workflow phases

	Discovery	Analysis	Writing	Publication	Outreach	Assessment
Trends	social discovery tools	datadriven & crowdsourced science	collaborative online writing	Open Access & data publication	scholarly social media	article level (alt)metrics
Expectations	growing importance of data discovery	more online analysis tools	more integration with publication & assessment tools	more use of "publish first, judge later"	use of altmetrics for monitoring outreach	more open and post-publication peer review
Uncertainties	support for full-text search and text mining	willingness to share in analysis phase	acceptance of collaborative online writing	effect of journal/publisher status	requirements of funders & institutions	who pays for costly qualitative assessment?
Opportunities	discovery based on aggregated OA full text	open labnotes	semantic tagging while writing/citing	reader-side paper formatting	using repositories for institutional visibility	using author-, publication- and affiliation-IDs
Challenges	real semantic search (concepts & relations)	reproducibility	safety/privacy of online writing	globalization of publishing/access standards	making outreach a two-way discussion	quality of measuring tools

Most important long-term development	multidisciplinary + citation-enhanced databases	collaboration + data-driven	online writing platforms	Open Access	more & better connected researcher profiles	importance of societal relevance + non-publication contributions
Potentially most disruptive development	semantic/concept search + contextual/social recommendations	open science	collaborative writing + integration with publishing	circumventing traditional publishers	public access to research findings, also for agenda setting	moving away from simple quantitative indicators

Typical workflow examples

https://figshare.com/articles/101_Innovations_in_Scholarly_Communication_the_Changing_Research_Workflow/1286826

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

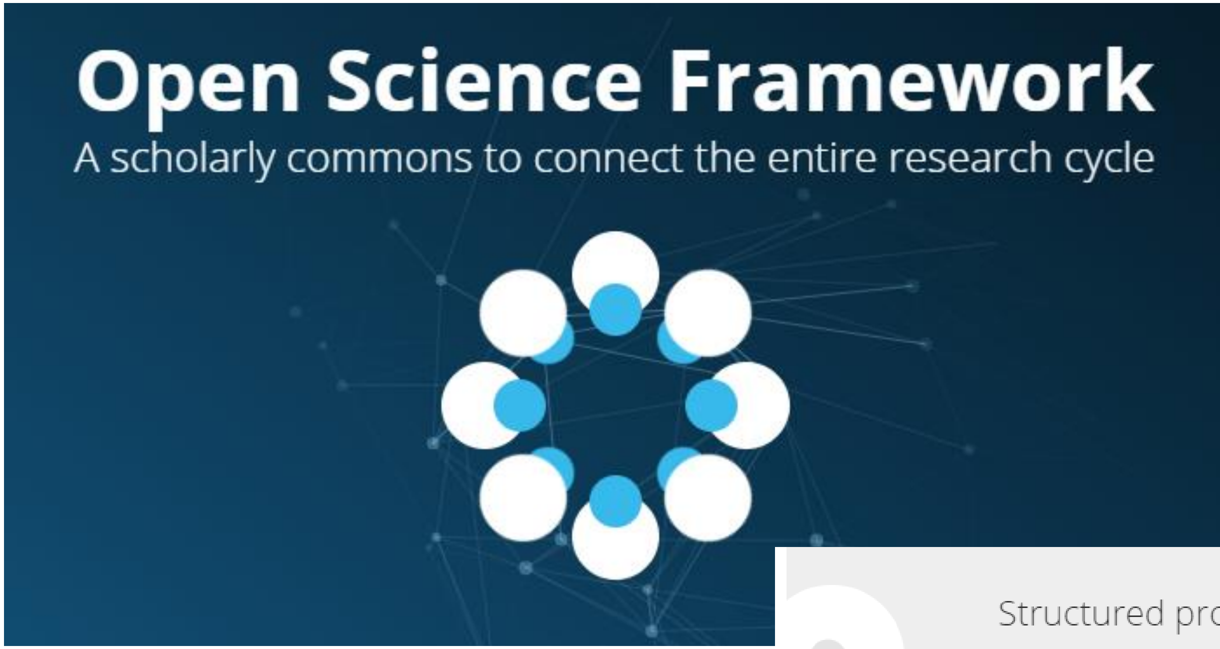
How to keep your data secure?

Develop a practical solution that fits your circumstances

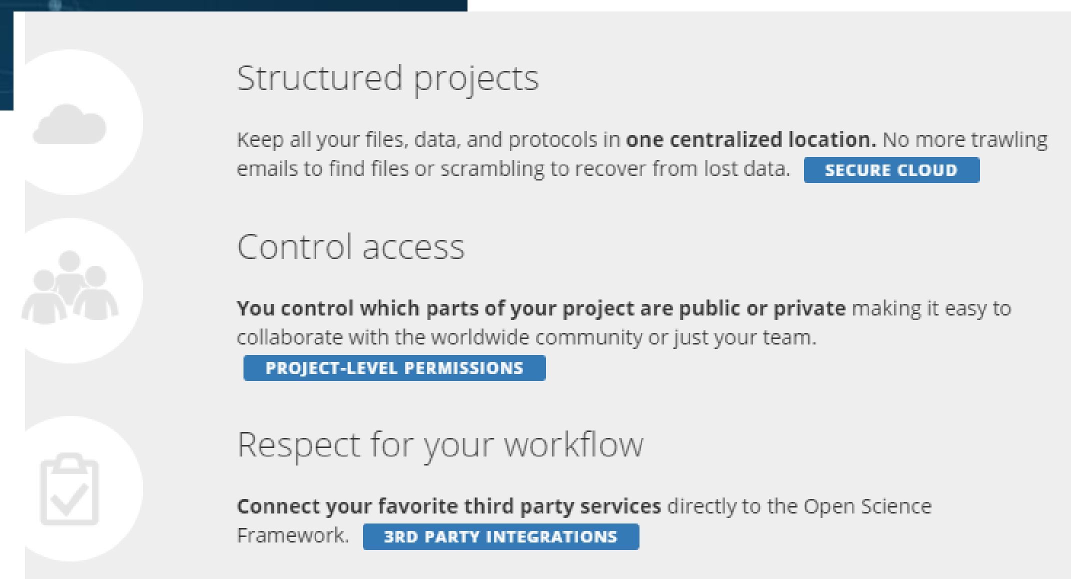
– Store your data on managed servers
– Restrict access to collaborators or smaller subset
– Encrypt mobile devices carrying sensitive information
– Keep anti-virus software up-to-date
– Use secure data services for long-term sharing

www.wsj.com/articles/SB10001424052748703843804575534122591921594

Collaborative platforms e.g. OSF



https://osf.io

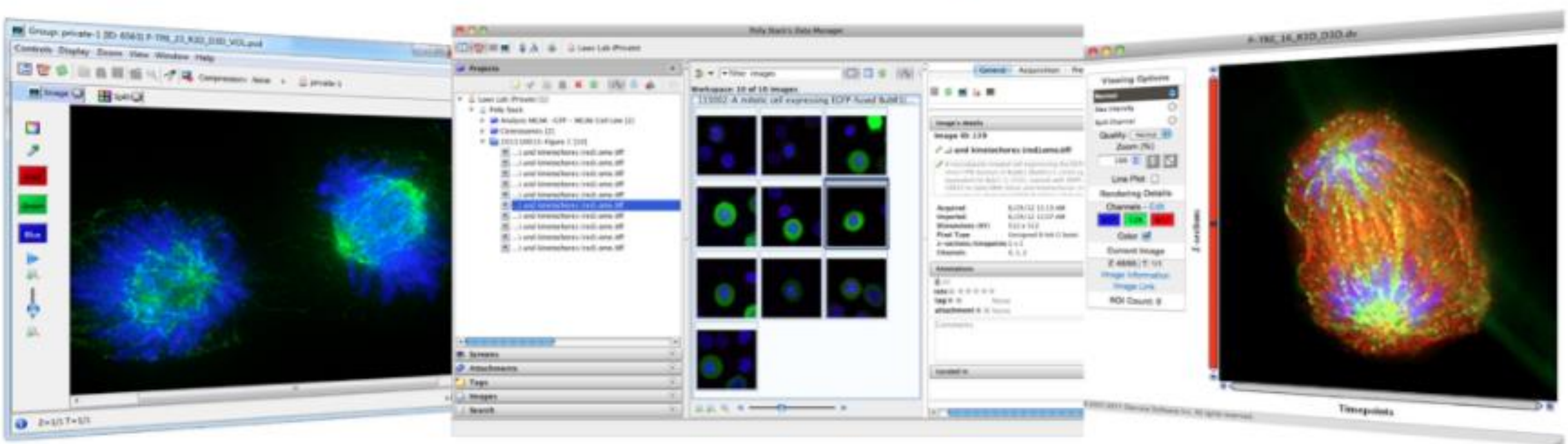


Data-specific platforms e.g. OMERO

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

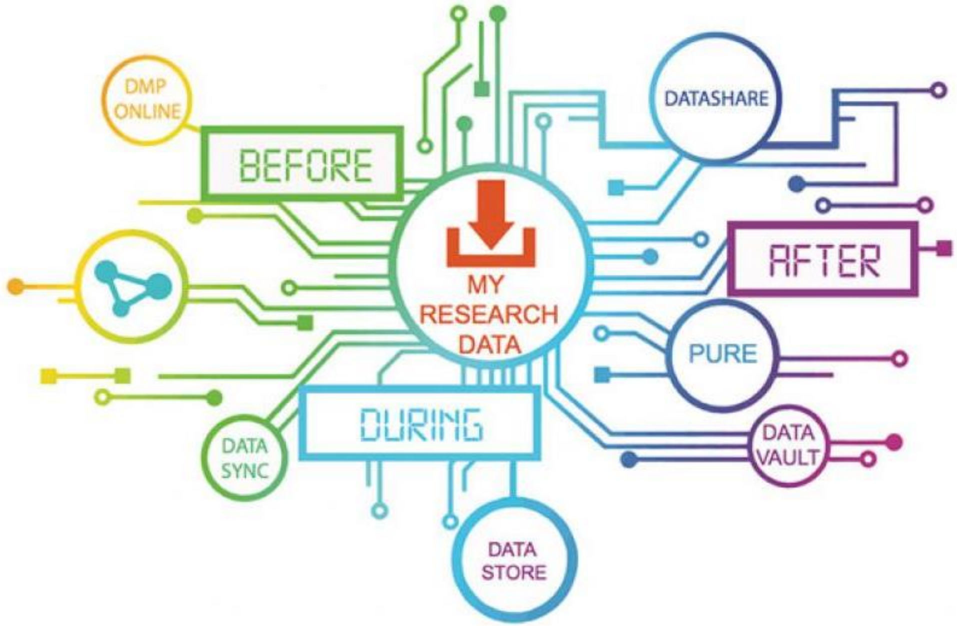
- Open source product with Dropbox-like functionality
- Used by many unis and service providers to offer ‘approved’ solution

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!

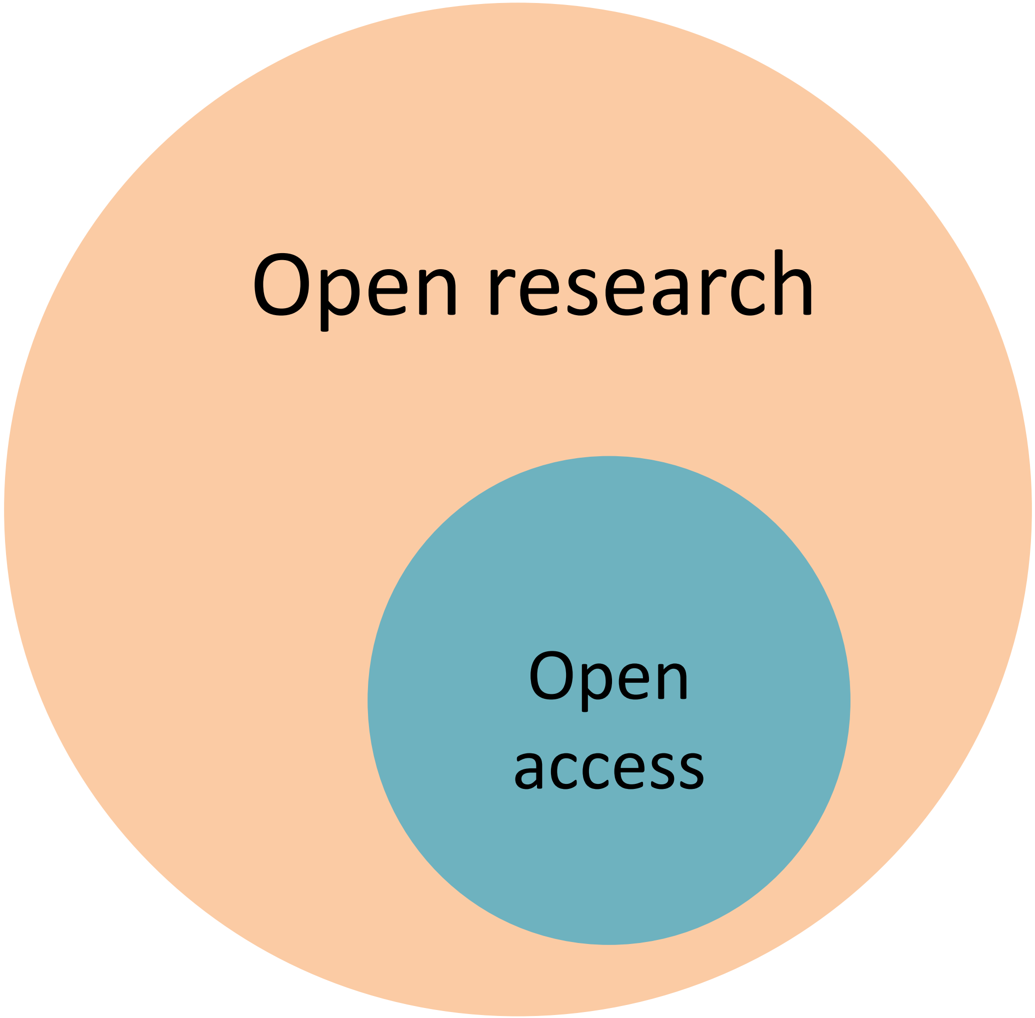
Backups

- 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

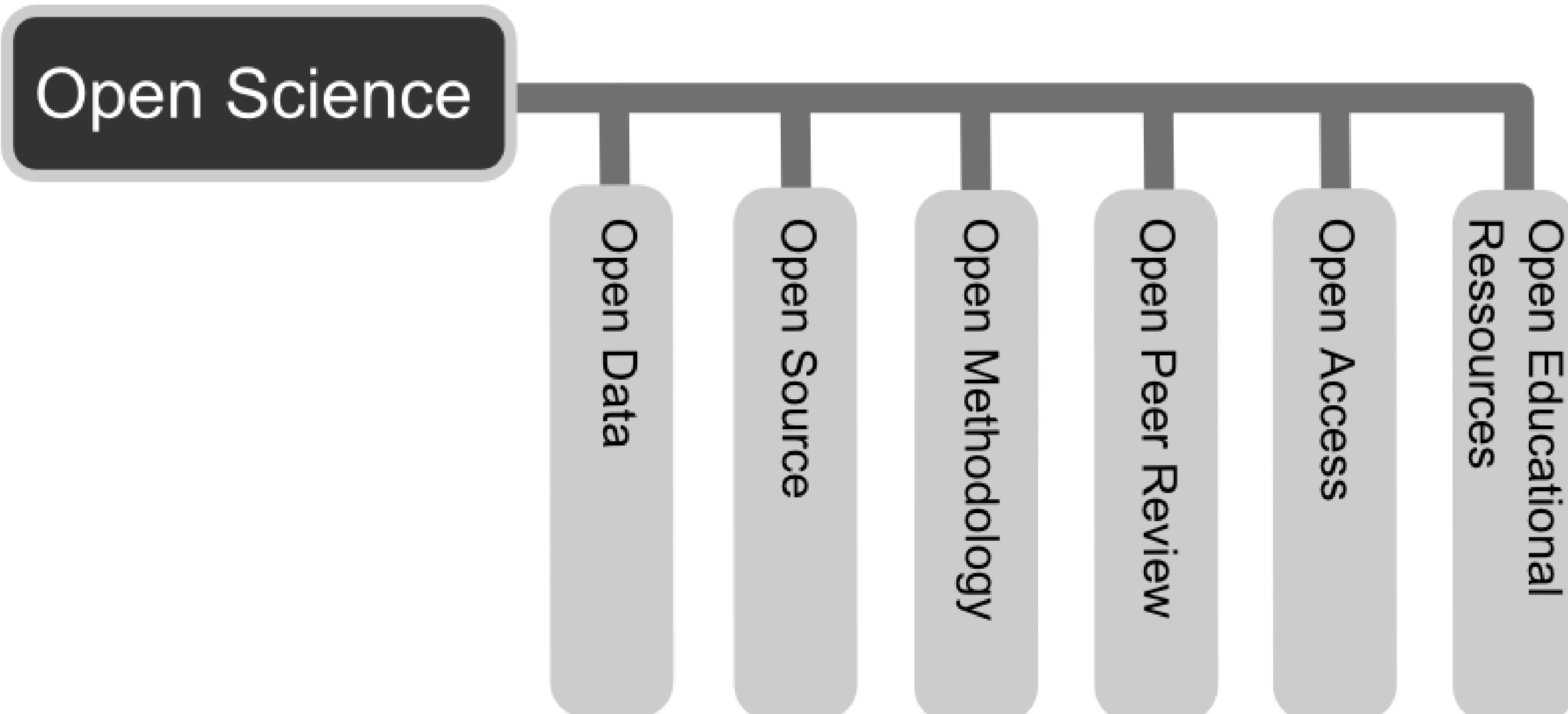
Archiving

- 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

Open access ≠ open research



Open access ≠ open research



Degrees of openness

Five star open data

★★★★★

Unable to share

Under embargo

Open — Restricted — Closed

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

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

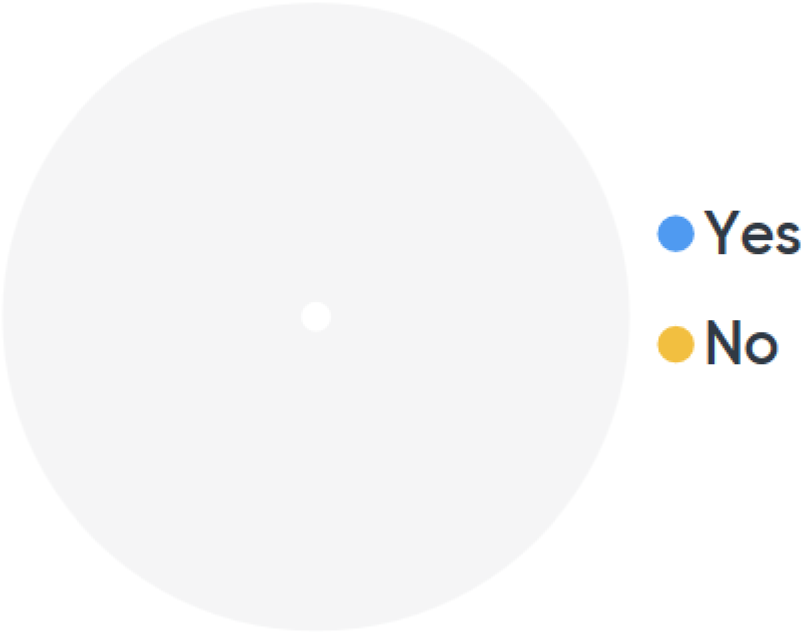
- Charges for use
- Data sharing agreements
- Restrictive licences
- Peer-to-peer exchange
- …

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…

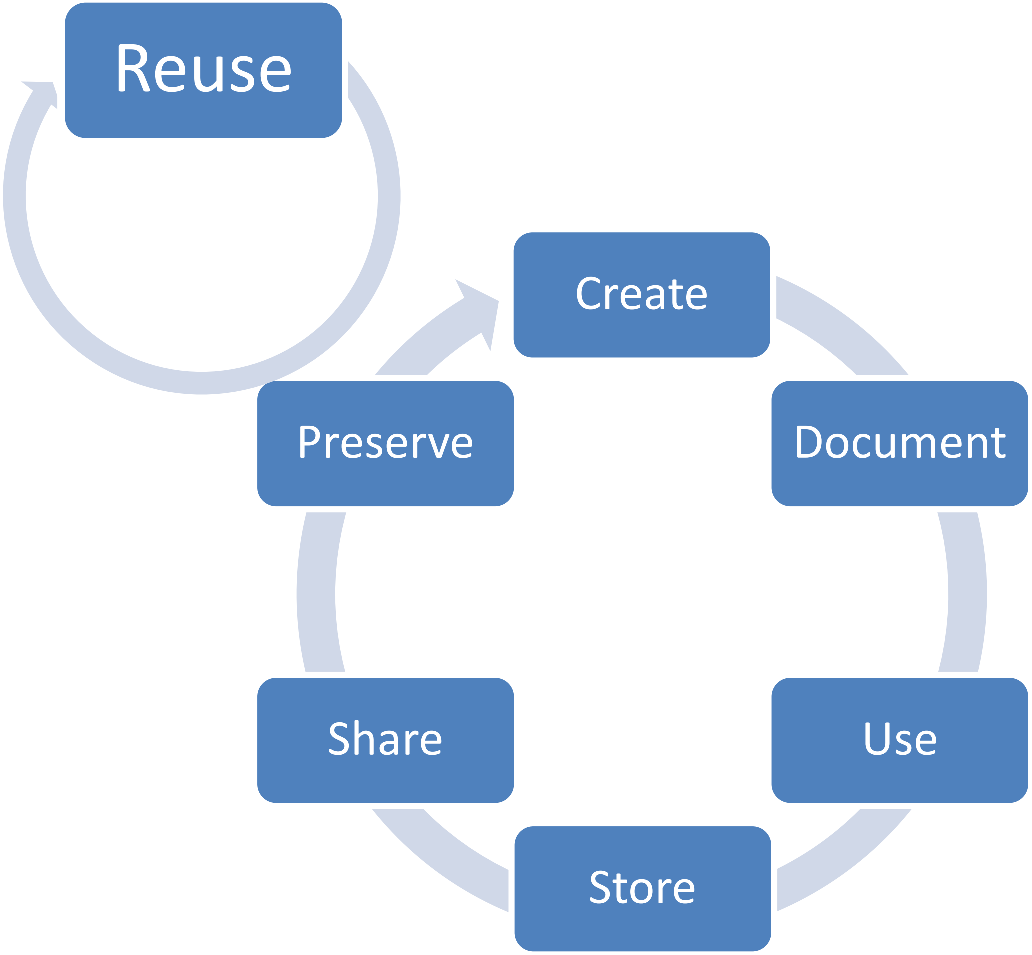
Have you reused other people's data?



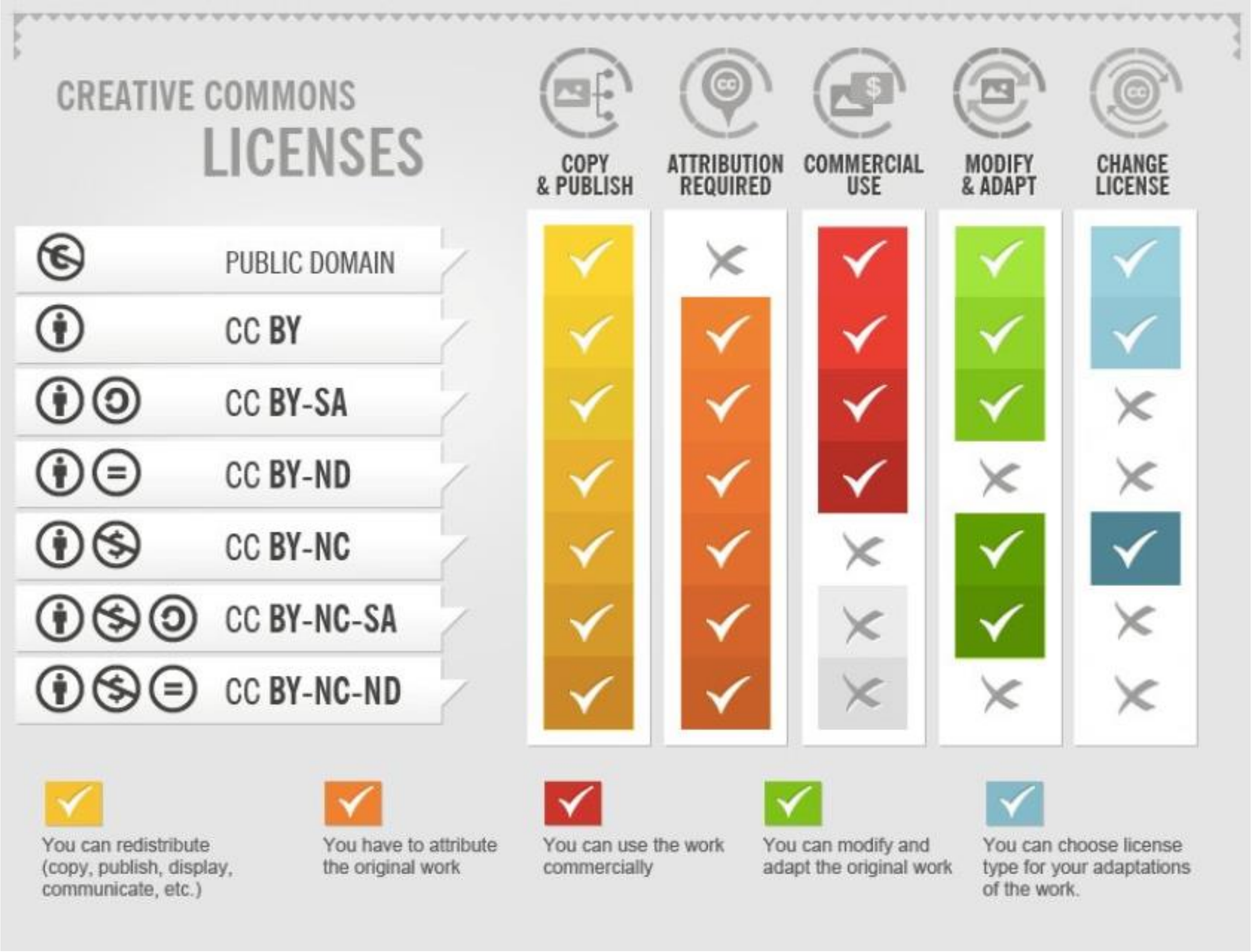
Slide is not active Activate

0

Primary and secondary data



License research data openly



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

License research data openly

Key Points

- Some research outputs that we want to share are protected legally as intellectual property (e.g., Copyright, EU Database Protection laws). o This means they are distributed with "All Rights Reserved", creating barriers to access, reuse, remixing, and redistribution.
- Creative Commons (CC) licenses have been developed by legal experts as a convenient and legally sound mechanism to openly share protected works with "Some Rights Reserved." All CC licenses require attribution. Some types of CC license offer other terms and conditions under which sharing may take place.
- Creative Commons waivers (CC0) allow the rights holder to dedicate the work to the public domain, removing all copyright and database protection restrictions. Waiving rights using CC0 is the mechanism recommended by the CODATA-RDA group on Legal Interoperability of Research Data.

Adapted from Gail Clement

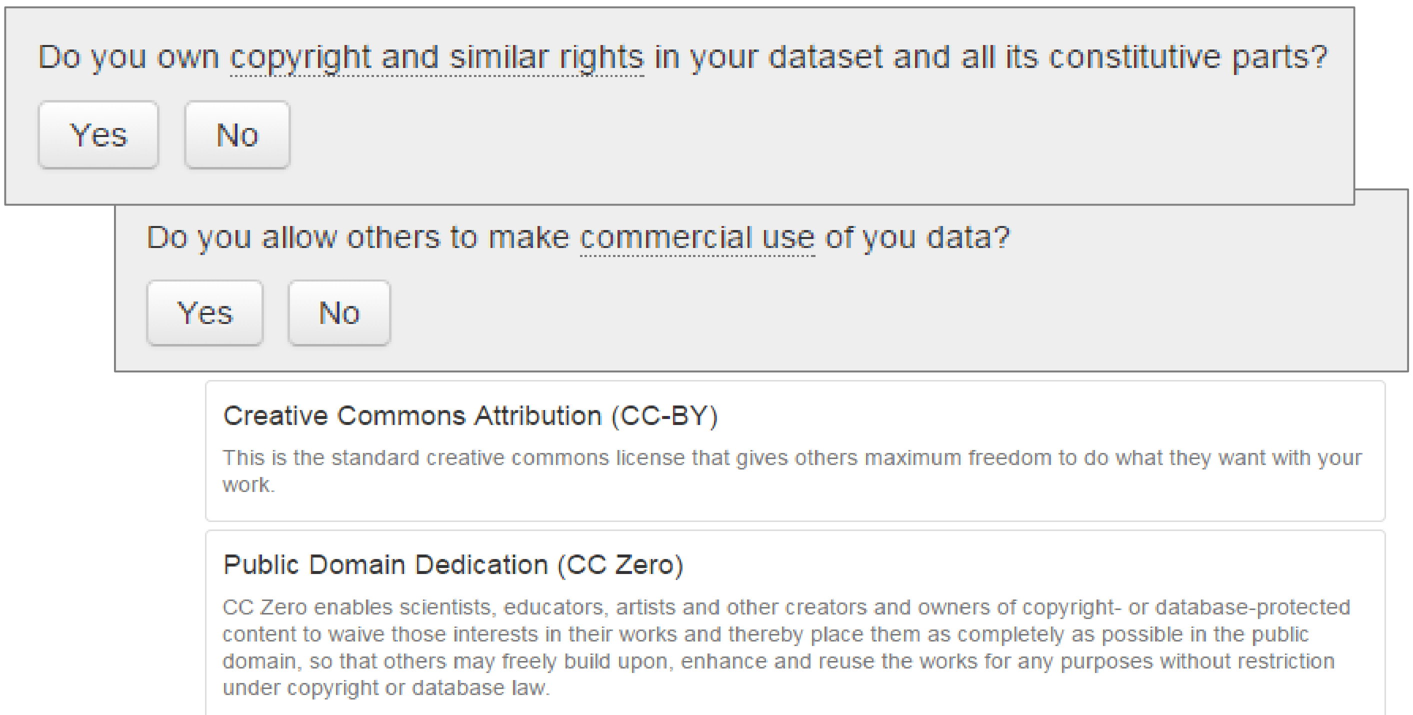
License research data openly

- Creative Commons licenses and waivers are both human and machine readable. Attaching licenses to your works means that you retain copyright in the work but allow sharing under the terms and conditions specified. You also must be attributed in appropriate fashion as a condition of using your work. If a user fails to adhere to the terms of the Creative Commons license, s/he violates copyright law.
- There may be additional legal issues governing the sharing of research data beyond copyright and database protection. Laws regarding Patents, Trademarks, Trade secrets, Privacy, National Security, may place restrictions on how research data may be distributed and reused. These legal issues are not addressed through Creative Commons licenses and waivers.

Adapted from Gail Clement

EUDAT licensing tool

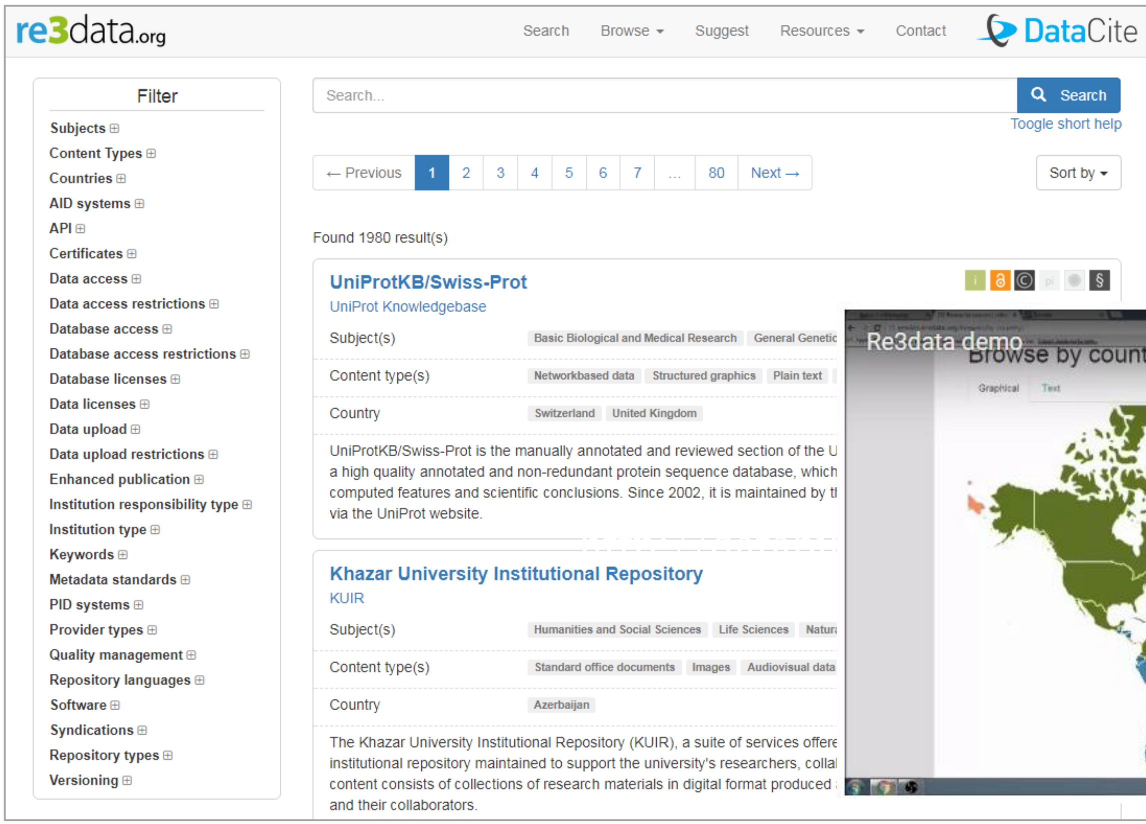
Answer questions to determine which licence(s) are appropriate to use



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

Deposit in a data repository

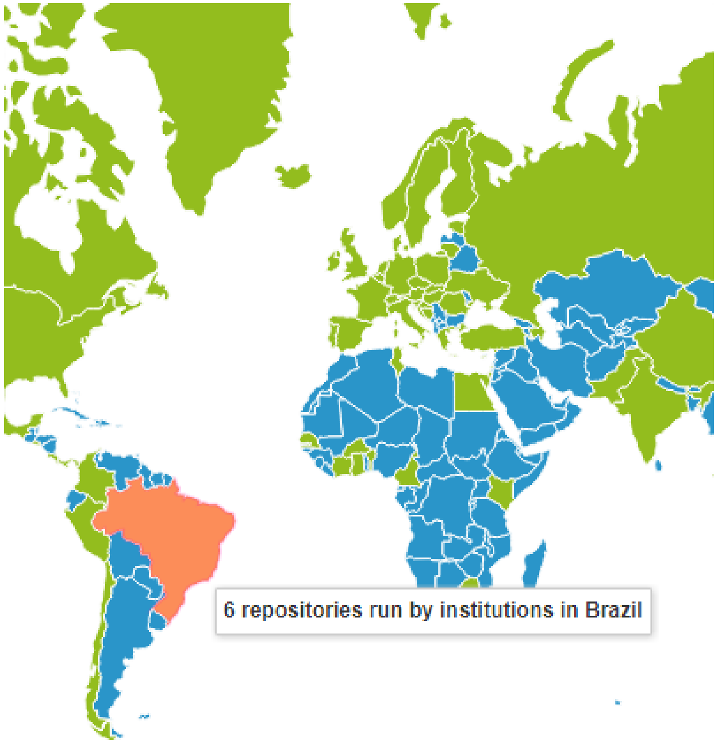
The Re3data catalogue can be searched to find a home for data



www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

National / domain repositories



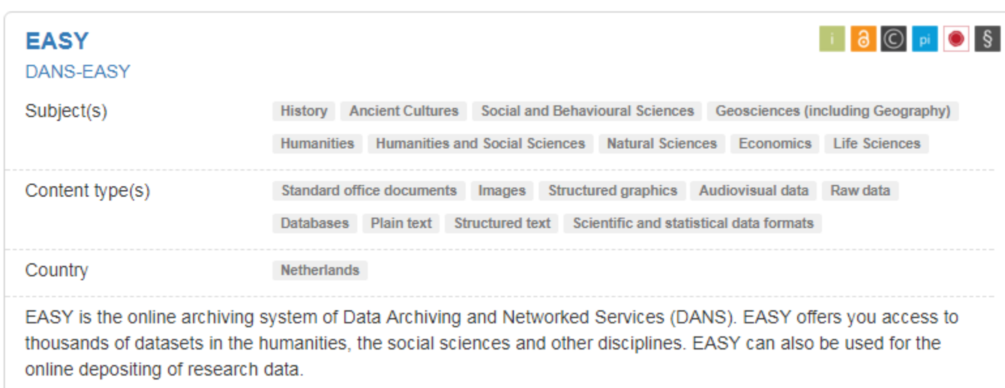
BioSharing portal of databases in life sciences

www.re3data.org

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.

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.

Overview Repositories 1 Stars 1 Followers 2 Following 2

Search repositories... Type: All New

My-Awesome-Science-Software ★1

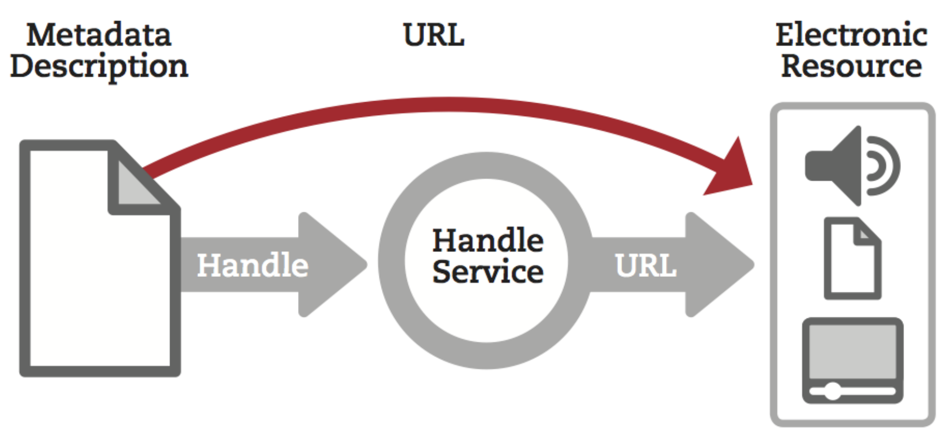
Get a DOI for each release

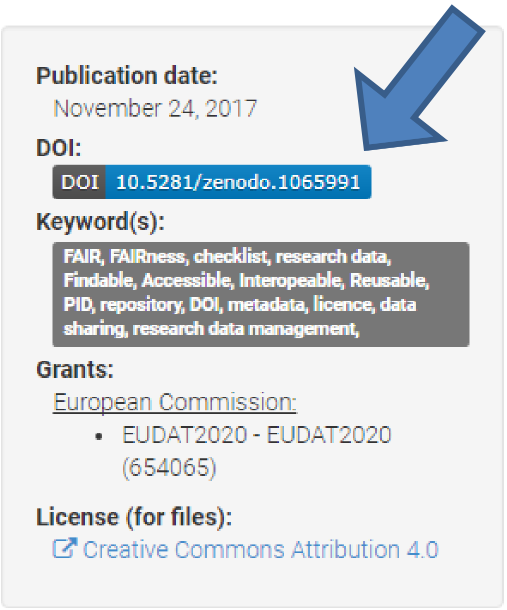
https://guides.github.com/activities/citable-code

What is a Persistent Identifier?

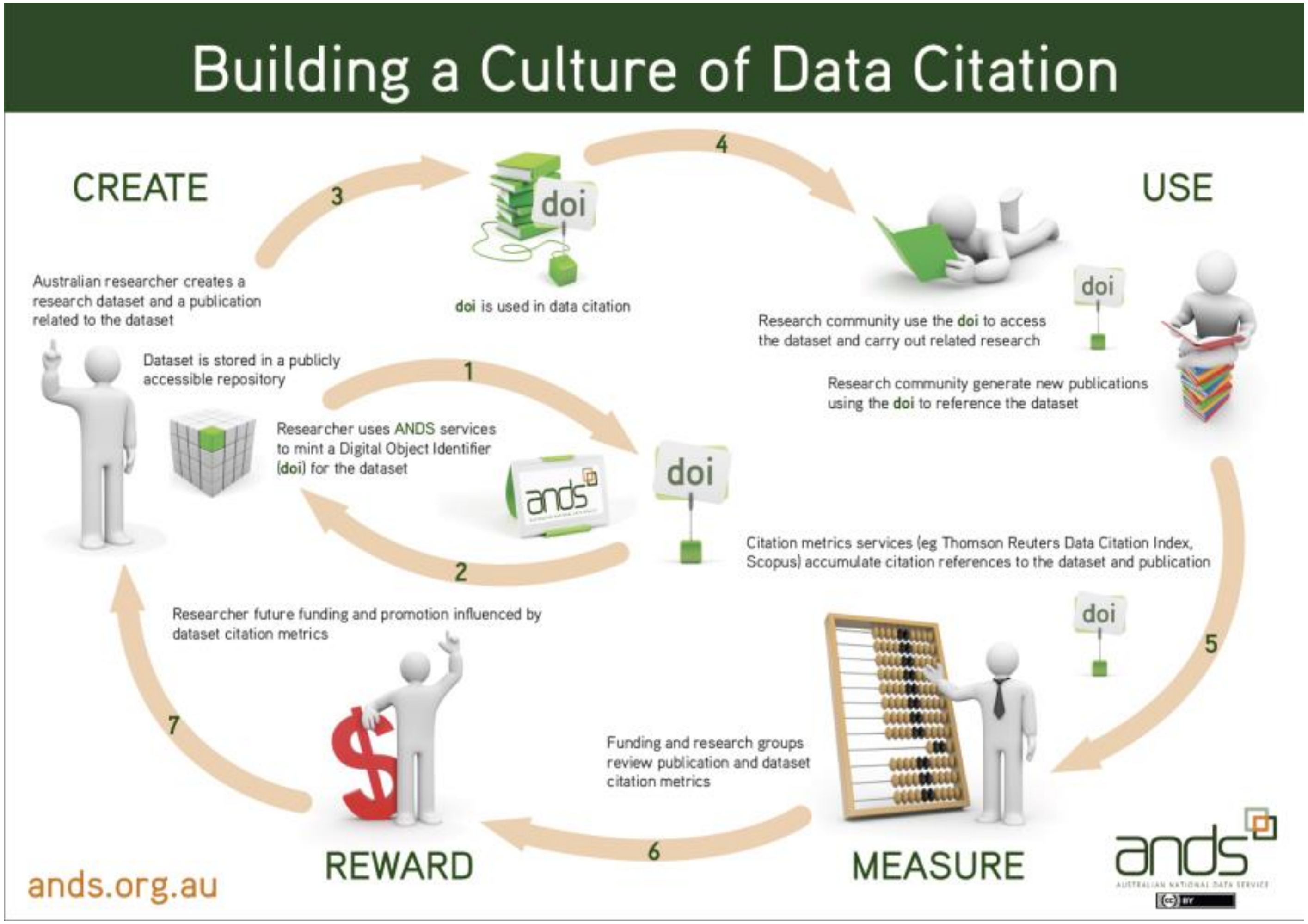
a long-lasting reference to a document, file or other object

- 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?

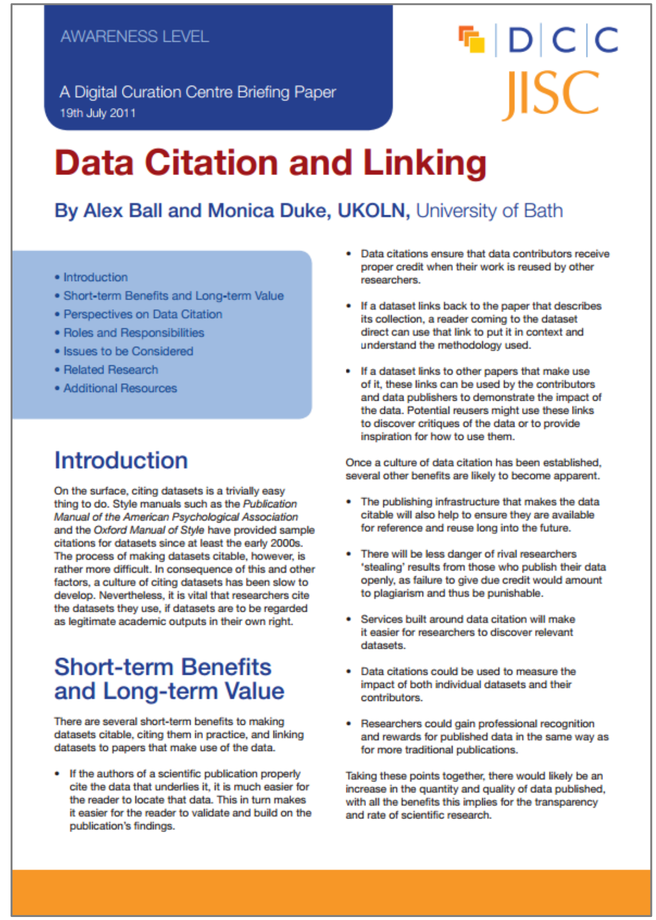


http://ands.org.au/cite-data

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

Open Research = sharing our knowledge freely and globally

To make our shared global knowledge reliable, trustworthy, and fair, we need 3 ingredients

1. Certainty that the information is accurate and reliable
2. Confidence that creators and contributors get credit for their hard work
3. Legal clarity that the sharing activity is ‘ok’

Adapted from Gail Clement

Certainty that the information is accurate and reliable

- DOIs guarantee that the content is available ..always
- ORCiDs guarantee the author/creator/contributor is right
- Open licenses (Creative Commons) guarantee the ©rights are right

Adapted from Gail Clement

Confidence that creators and contributors get credit for their hard work

- DOI's reference the specific work being used and cited
- ORCiD's remove ambiguity of researchers' names
- Open licenses build attribution into the conditions of use

Adapted from Gail Clement

Legal clarity that the sharing activity is ‘ok’

- DOIs are assigned to works that are trustworthy: they are being curated and taken care of ..for the long term
- ORCiDs are assigned to creators and contributors who manage their own profiles so they are authoritative
- Open licenses like Creative Commons are 100% compliant with copyright law

Adapted from Gail Clement

Some European Initiatives

EOSCpilot

The European Open Science Cloud for Research Pilot Project

OpenAIRE

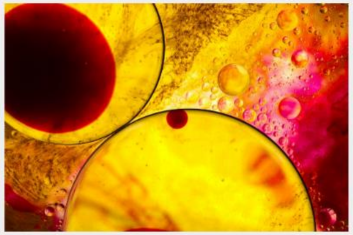
FOSTER

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.

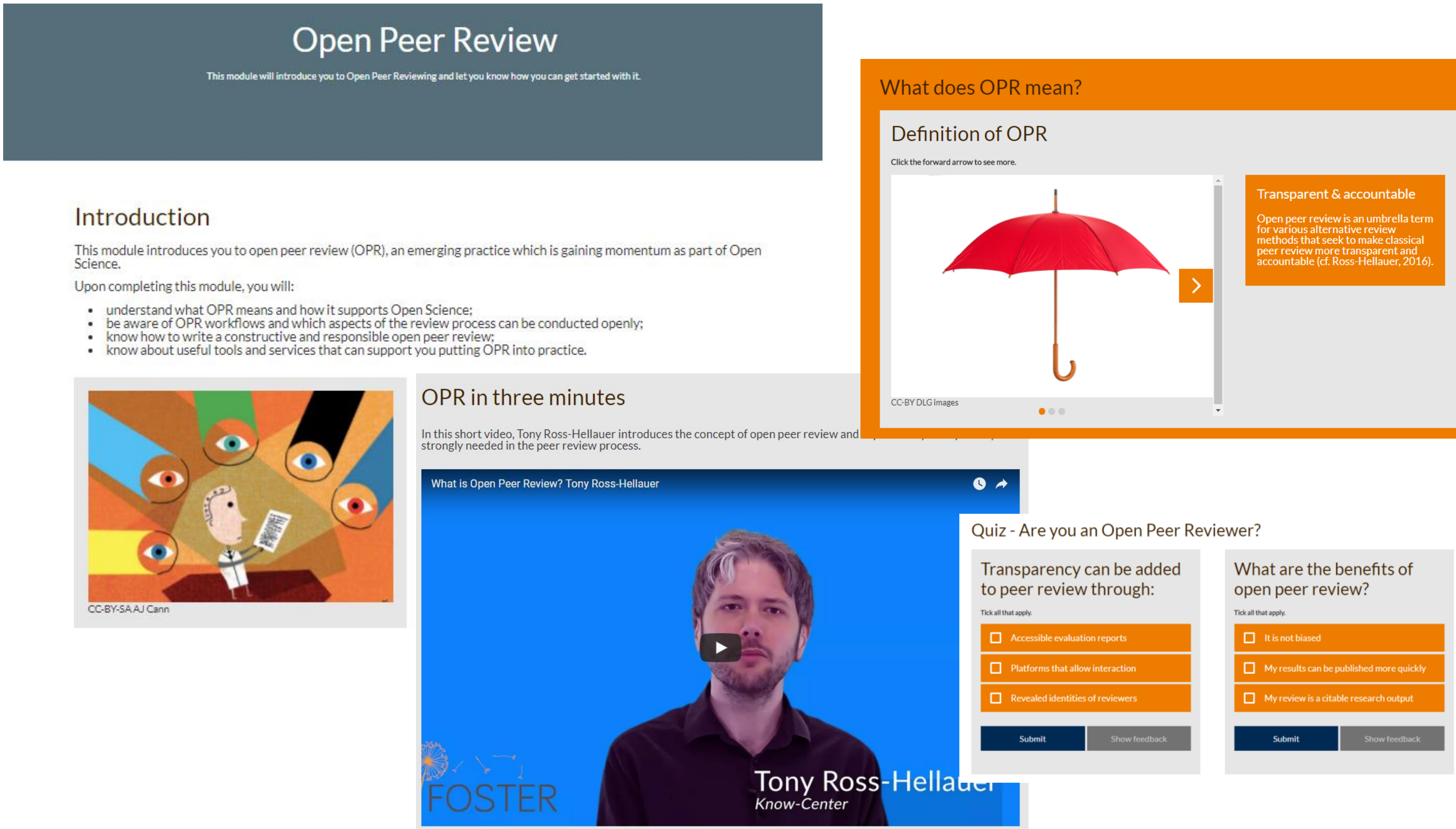
Open Access Publishing
This course will help you become skilled in Open Access publication in the wider context of Open Science.

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

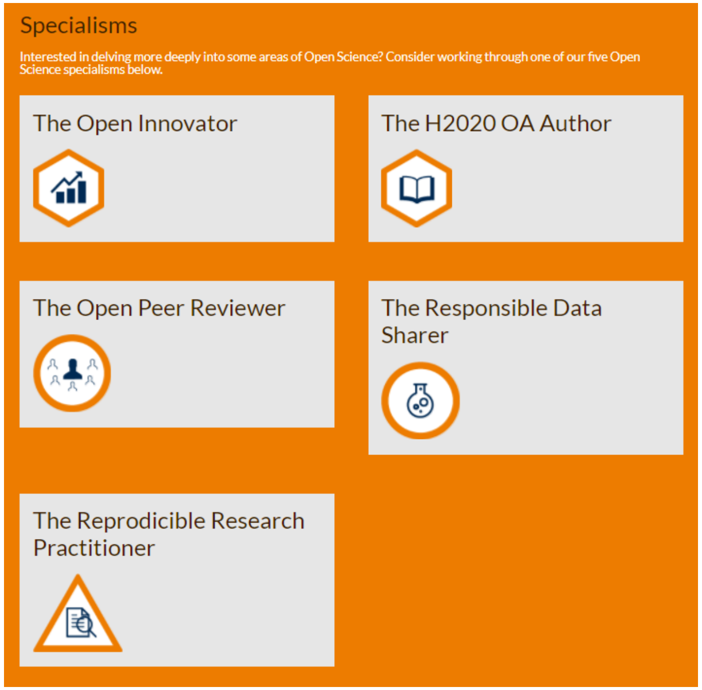
Open Peer Review module example



Specialisation pathways

2-4 hours of content

- The reproducible research practitioner
- The responsible data sharer
- The Open Access Author
- The open peer reviewer
- The open innovator

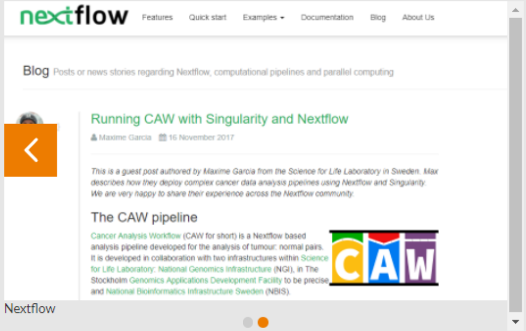


For more information, see www.fosteropenscience.eu/learning-paths

Case study approach

Using the EC Open Science Monitor approach to share practical examples of activity from the Life Sciences, Social Sciences and Humanities.

Life Sciences: Nextflow for reproducible in silico genomics



Nextflow

Why?

The analysis of big data in a performant and reproducible manner is an increasing pressing issue in many scientific fields including and mostly in life science disciplines. This problem has been fuelled by the combined reliance on increasingly complex data analysis methods and the exponential growth of biological datasets. When considering the installation, deployment and maintenance of bioinformatic pipelines, an even more challenging picture emerges due to the lack of community standards. Moreover, the effect of limited standards on reproducibility is amplified by the very diverse range of computational platforms and configurations on which these applications are expected to be applied (workstations, clusters, HPC, clouds, etc.). The Nextflow open source technology provides a simple but yet effective solutions to many of these problems.

Open Access

eLIFE

Open Source Licensing

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Open Research Data

Example use of EBI metagenomics

Open Peer Review

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Ethics

EUROPEAN GENOME-PHENOME ARCHIVE
10TH ANNIVERSARY

Open Innovation

docker

Digital Curation Centre

Thank you!

For DCC resources see:
www.dcc.ac.uk/resources

Follow us on twitter:
@digitalcuration and #ukdcc

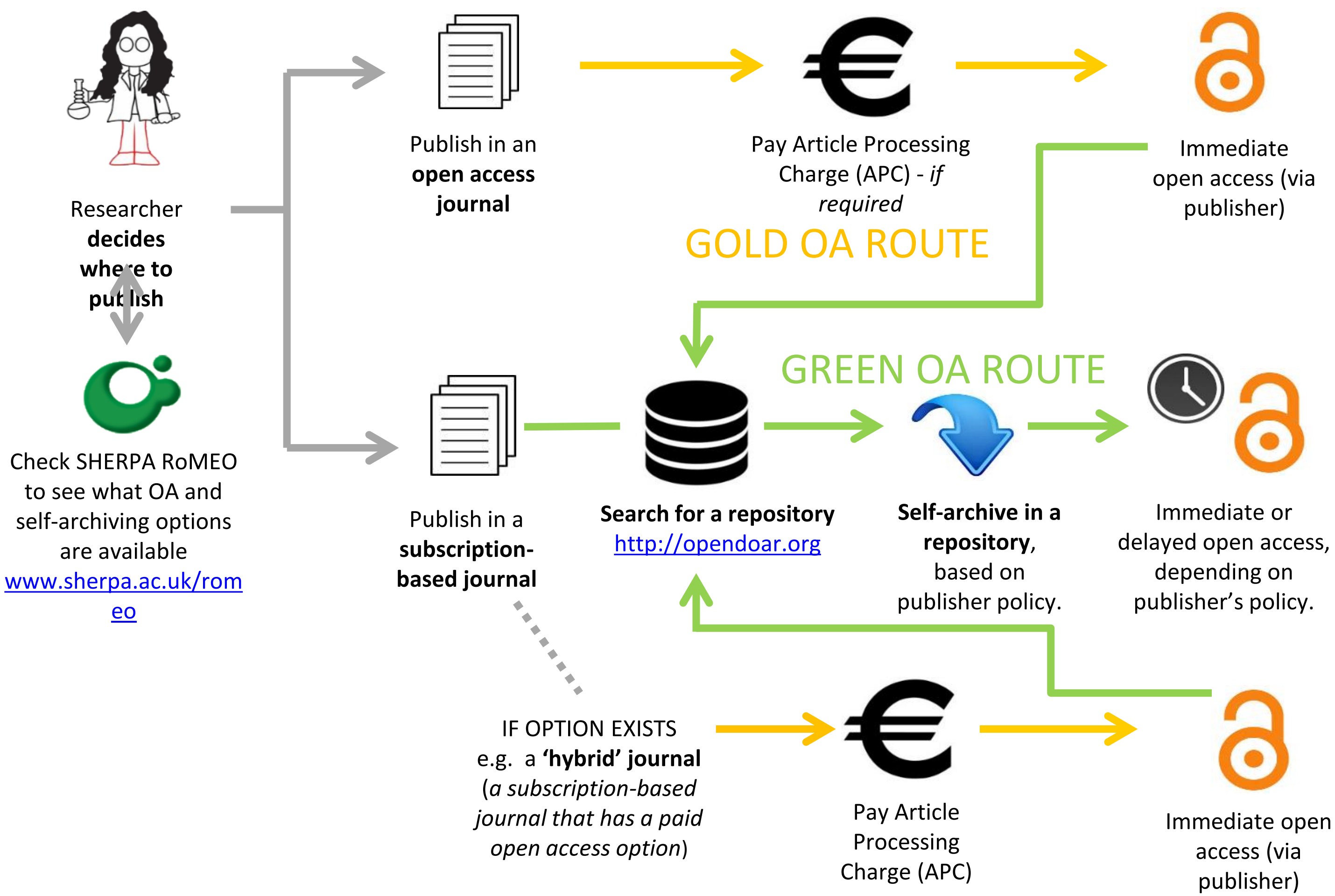
In collaboration with:

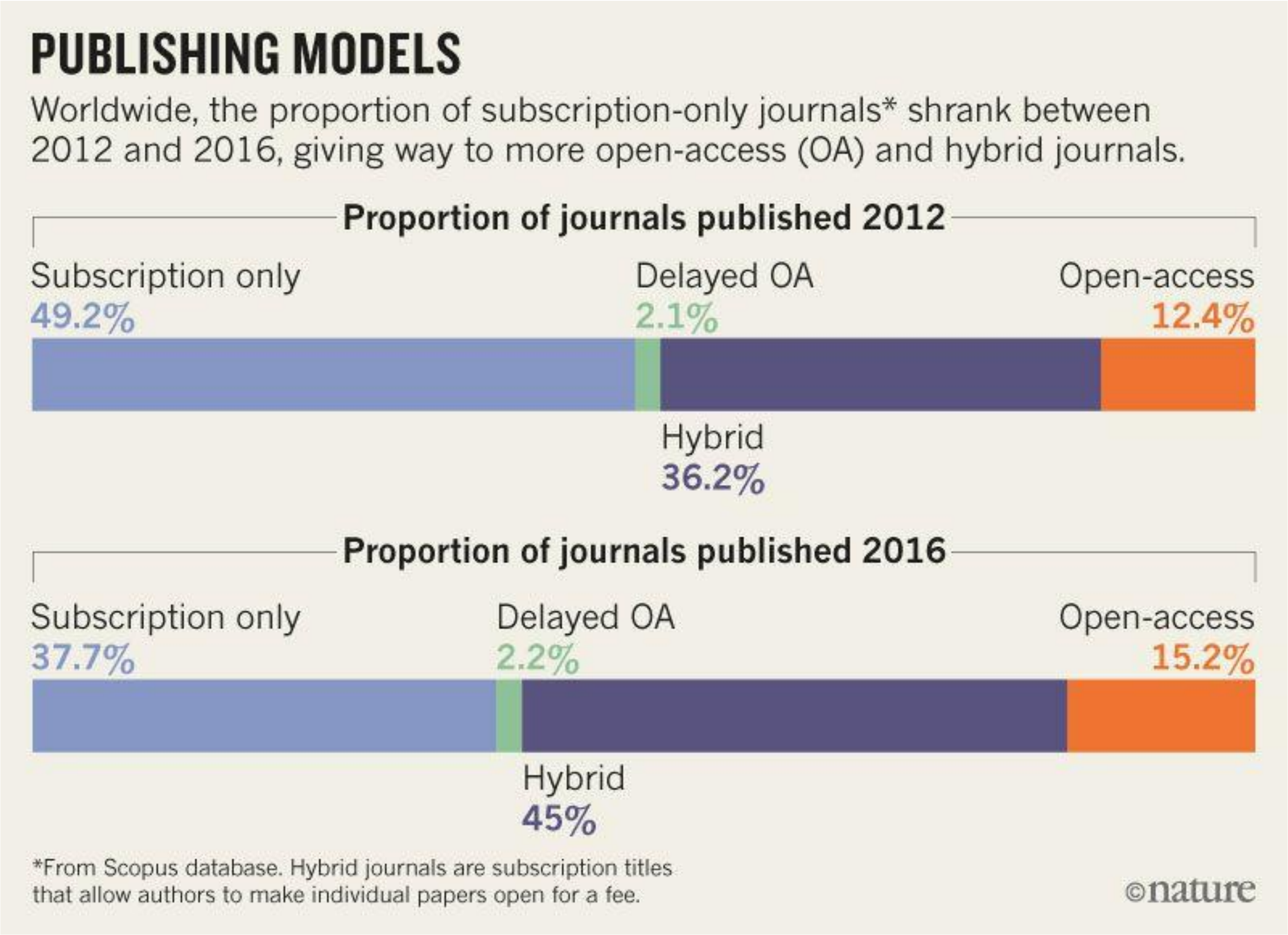
THE UNIVERSITY OF EDINBURGH

Open Access Week

Paywall: The Business of Scholarship

Open access publication





https://www.nature.com/articles/d41586-018-06178-7