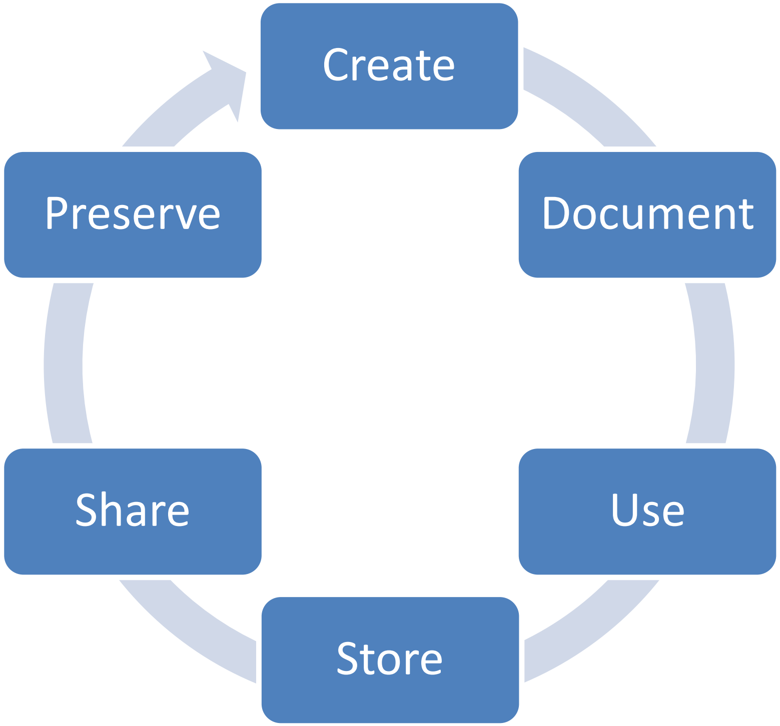
An Introduction to Research Data Management and Open Science

S. Venkataraman
Research Data Specialist
Digital Curation Centre

The Hong Kong University of Science and Technology, 18th March 2019

What is RDM?

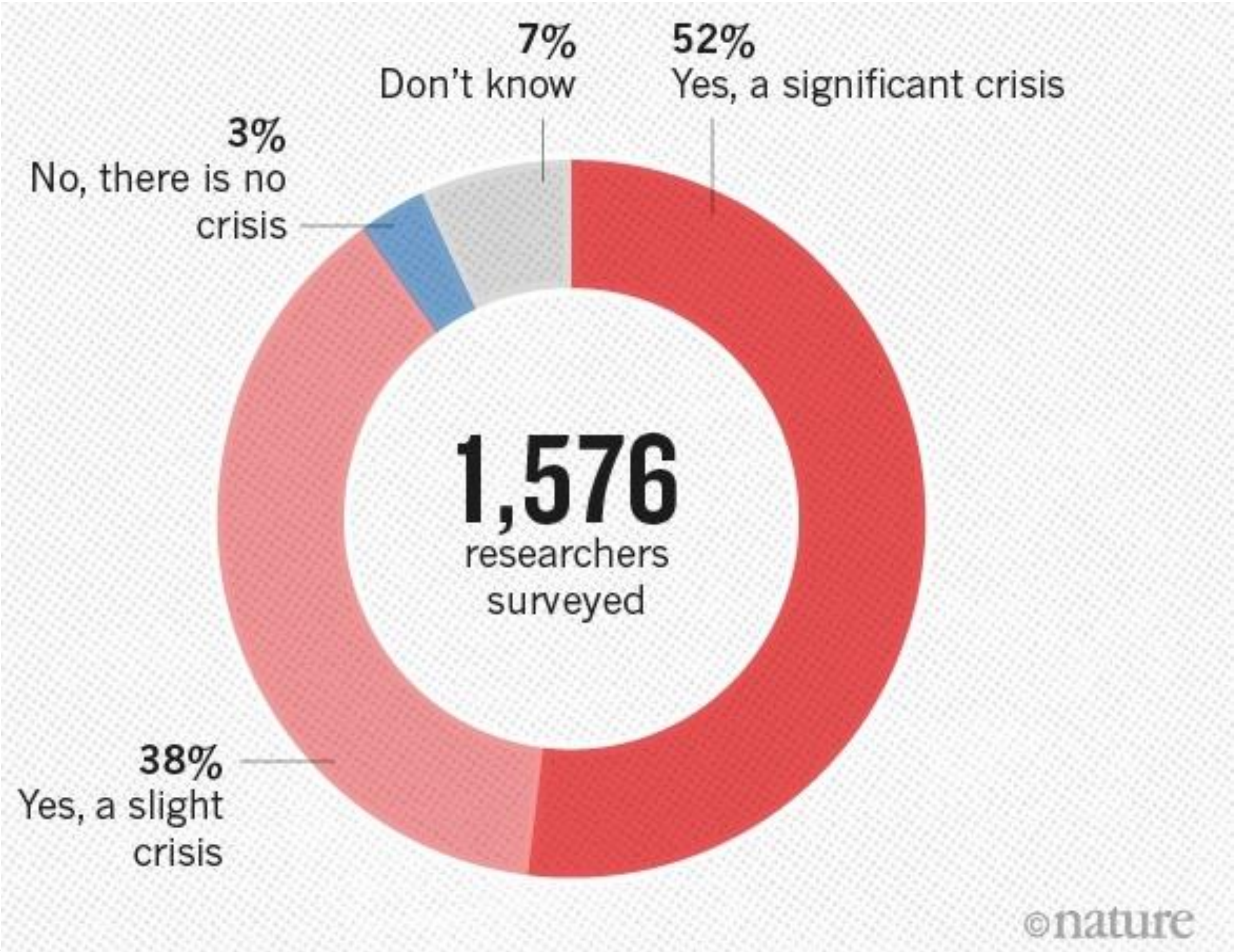
What is Research Data Management?



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

Data management is part of good research practice

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

Creating data

Image CC-SA-ND by Bill Dickinson www.flickr.com/photos/skynoir/8270436894

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

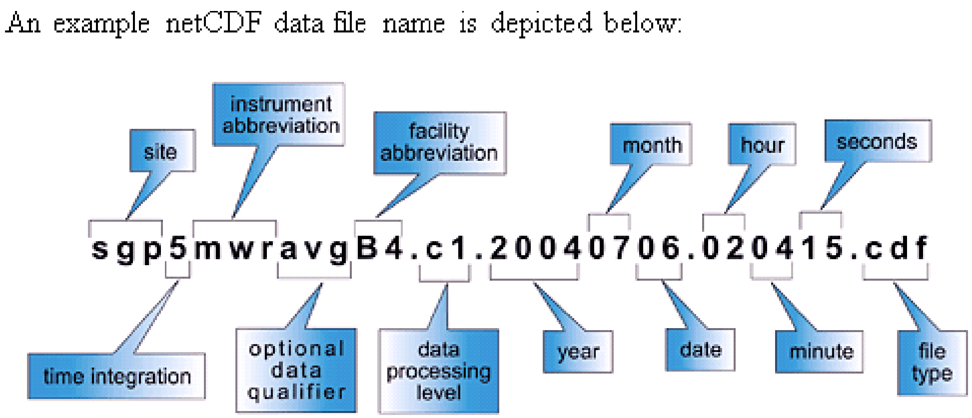
Data centres may suggest preferred formats for deposit

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

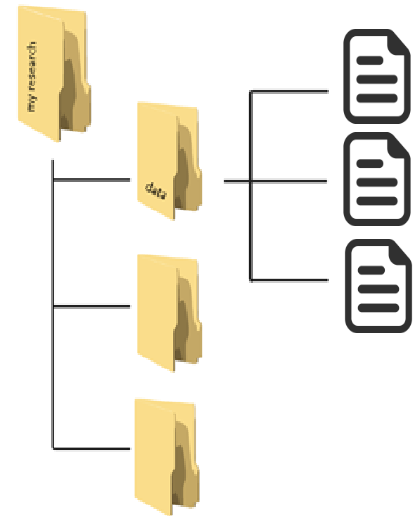
Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g.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

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

How will you organise your data?



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**

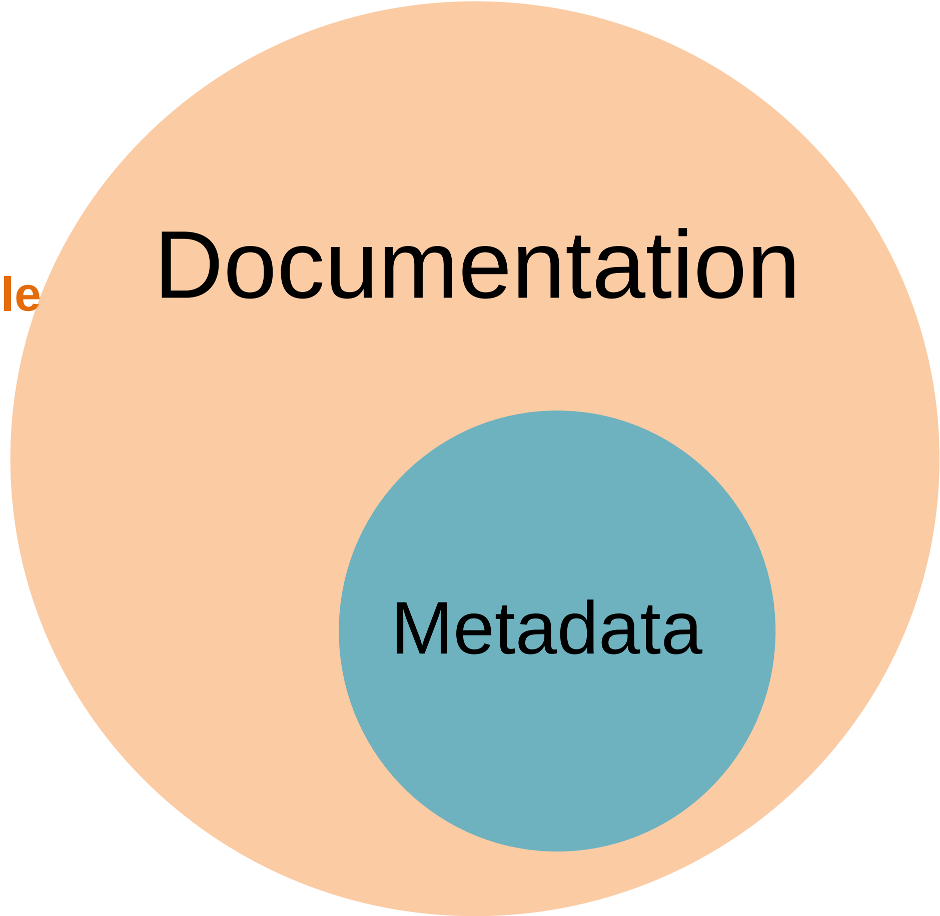
What is metadata?

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

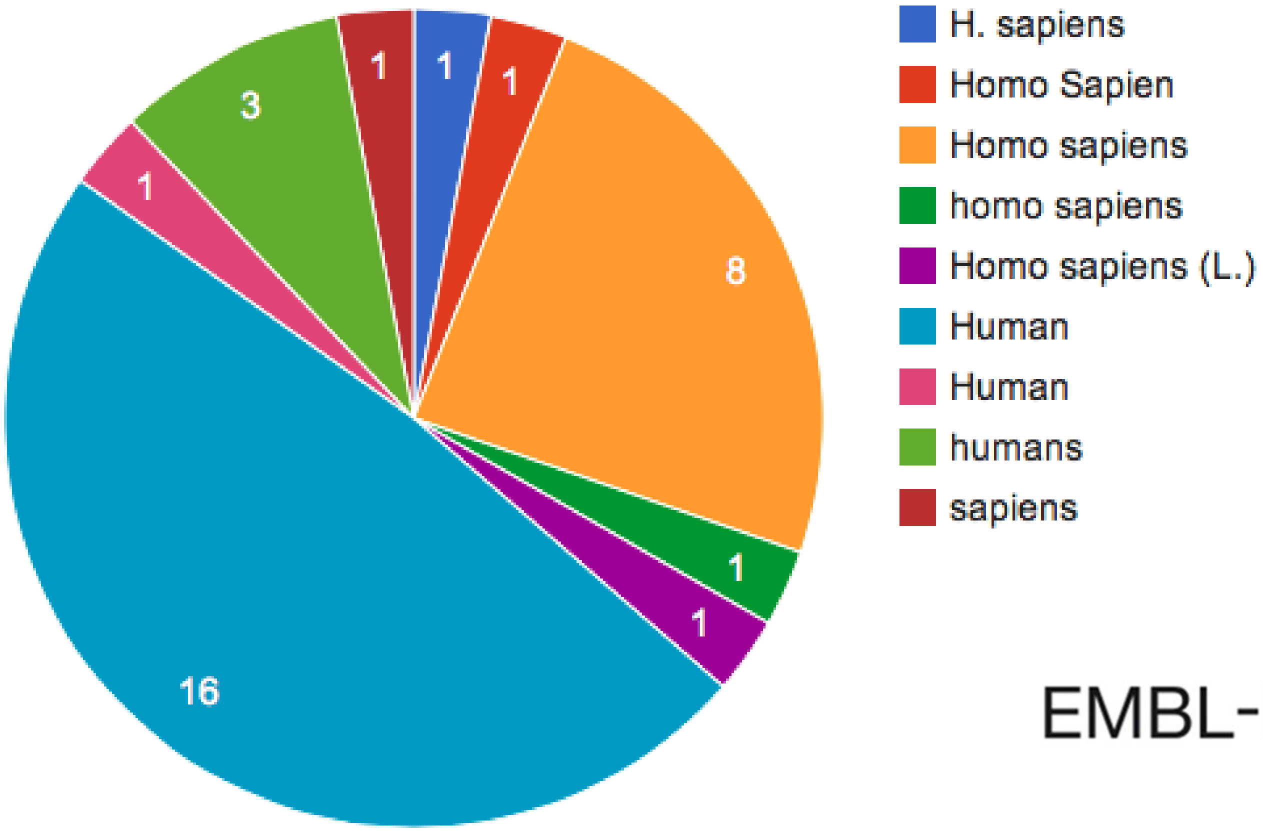
- **Data Documentation Initiative (DDI) – social science**
- **Ecological Metadata Language (EML) - ecology**
- **Flexible Image Transport System (FITS) – astronomy**

Search for standards in catalogues like:

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

Controlled vocabularies

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



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

Why are ontologies important?

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

Include ontologies as well

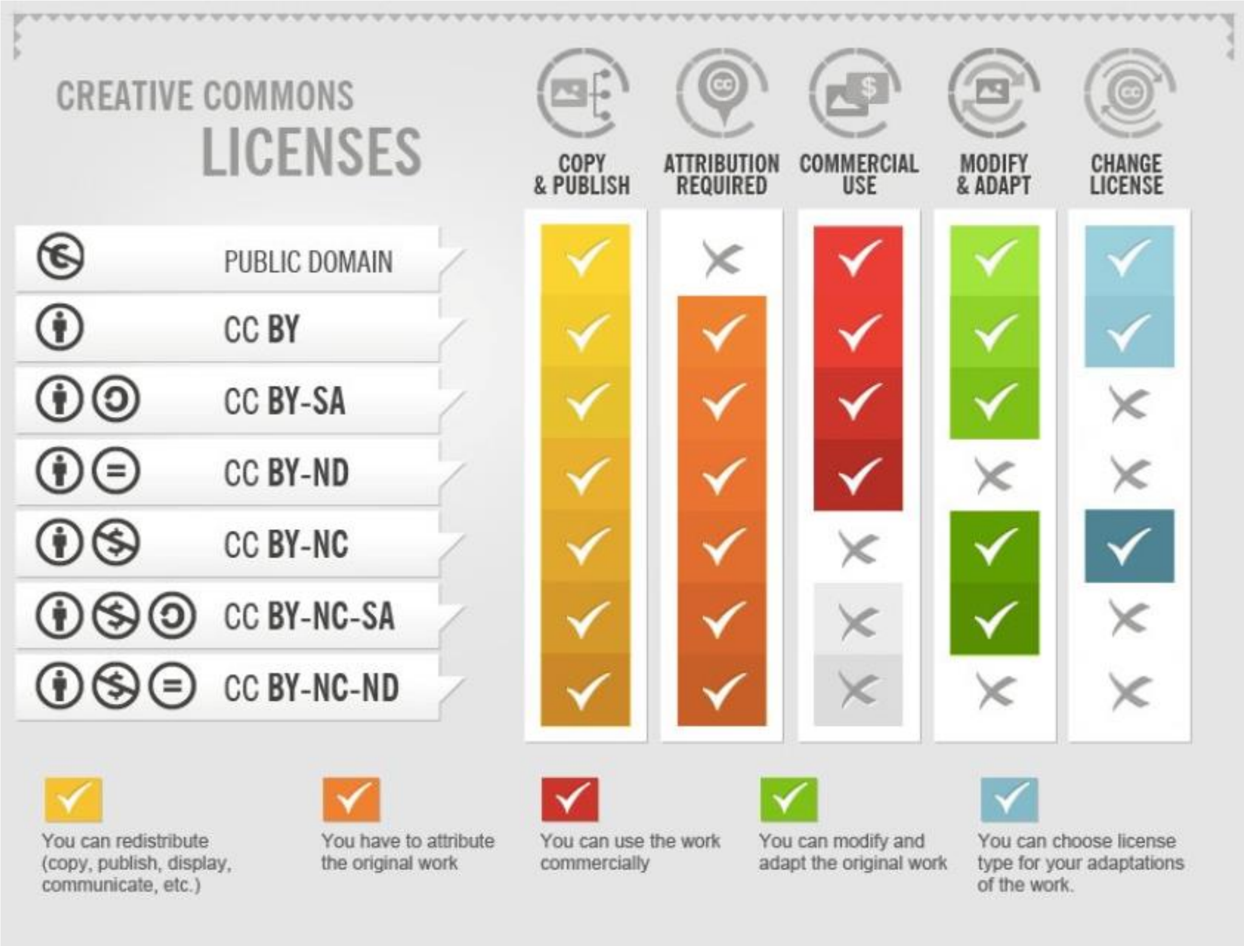
Defined terms + taxonomy

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*

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

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.

https://www.eudat.eu/services/userdoc/license-selector

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

Useful tools for documentation

E-lab notebooks, wikis, etc

- Record experiment procedures and results
- Share protocols

jupyter

RSpace

LabTrove

SCINOTE

http://openwetware.org

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

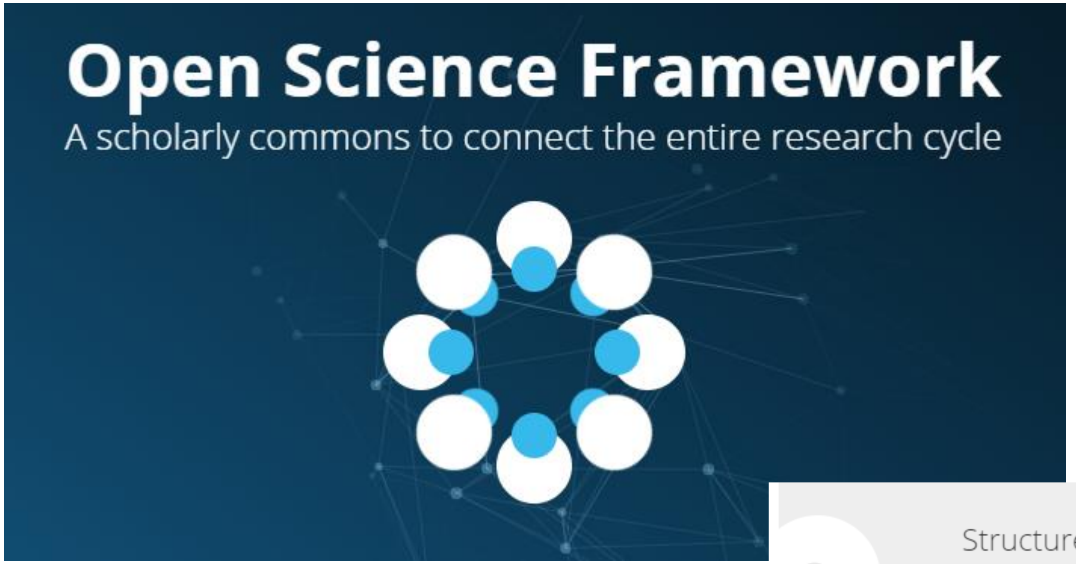
How to keep your data secure?

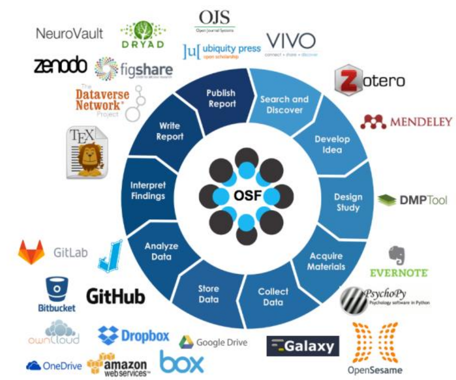
Develop a practical solution that fits your circumstances

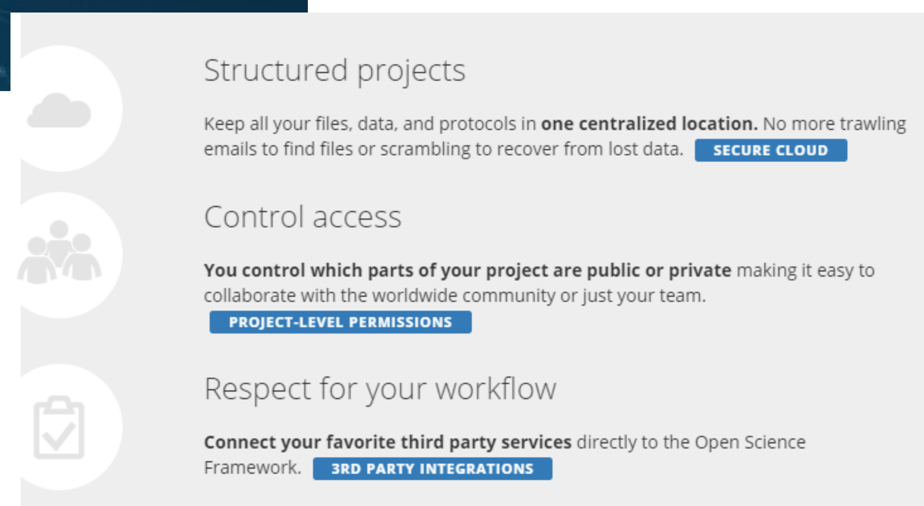
- Store your data on managed servers
- Restrict access to collaborators or smaller subset
- Encrypt mobile devices carrying sensitive information
- Keep anti-virus software up-to-date
- Use secure data services for long-term sharing

www.wsj.com/articles/SB10001424052748703843804575534122591921594

Collaborative platforms e.g. OSF







https://osf.io

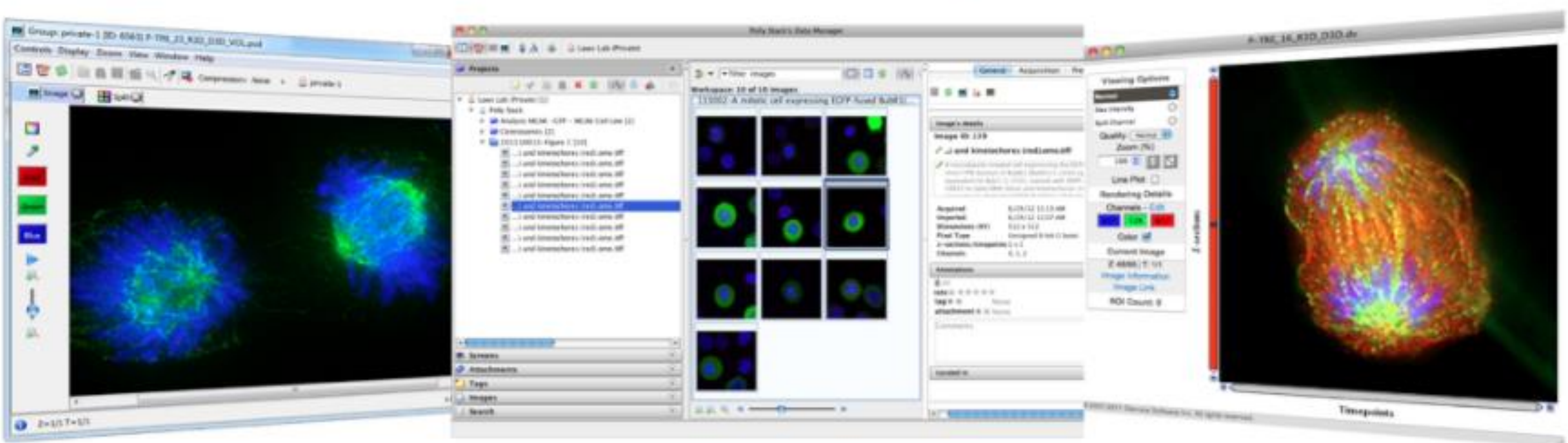
Data-specific platforms e.g. OMERO

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

ownCloud

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

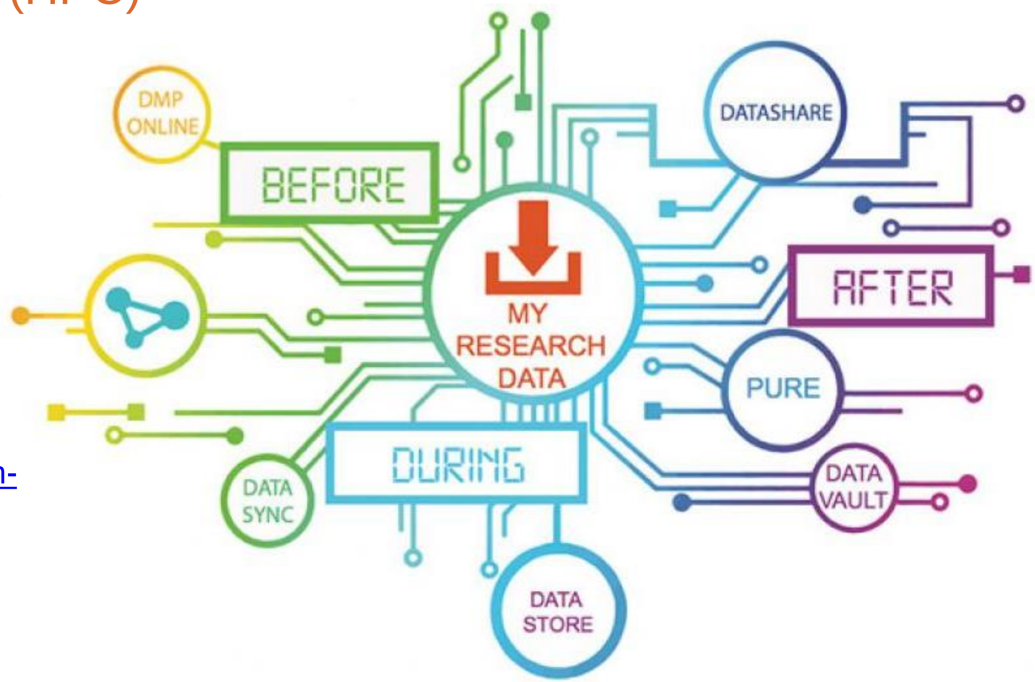
https://owncloud.org

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

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 and being Open

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 ▶

Nature **555**, 403-405 (2018).
https://www.nature.com/articles/d41586-018-03071-1
doi: 10.1038/d41586-018-03071-1

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

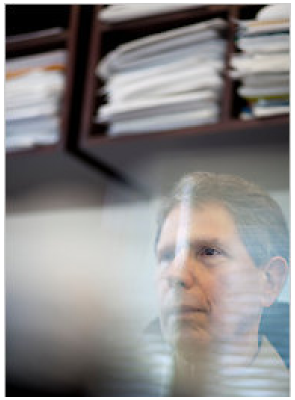
Sharing leads to breakthroughs!

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.

"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

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

Benefits for you: sharing data increases citations!

Want evidence?

Piwowar, Vision – 9% (microarray data)

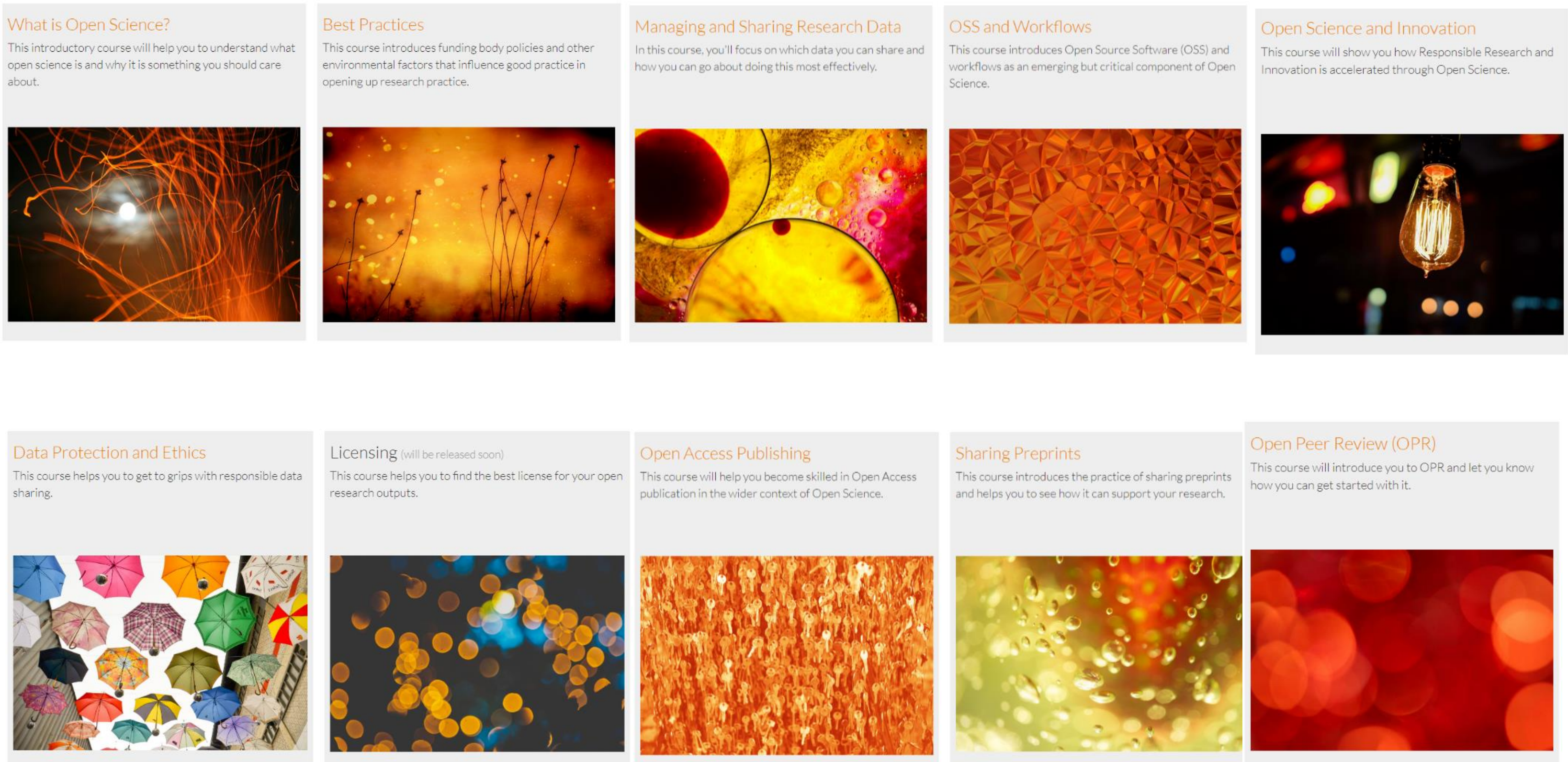
Drachen, Dorch, et al – 25-40%, astronomy

Gleditch, et al – doubling to trebling (international relations)

Open Data Citation Advantage

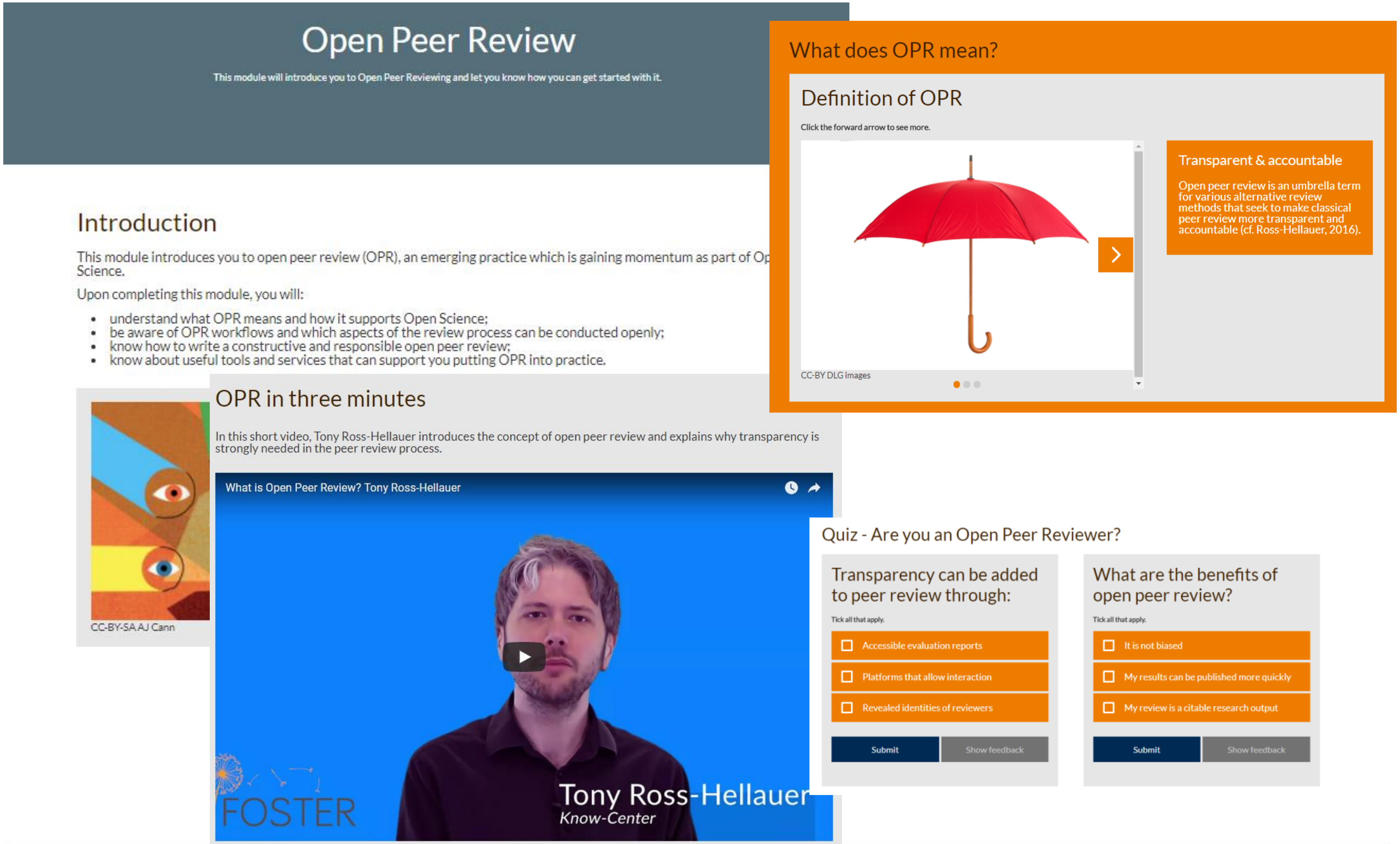
http://sparceurope.org/open-data-citation-advantage

FOSTER Open Science toolkit



https://www.fosteropenscience.eu/toolkit

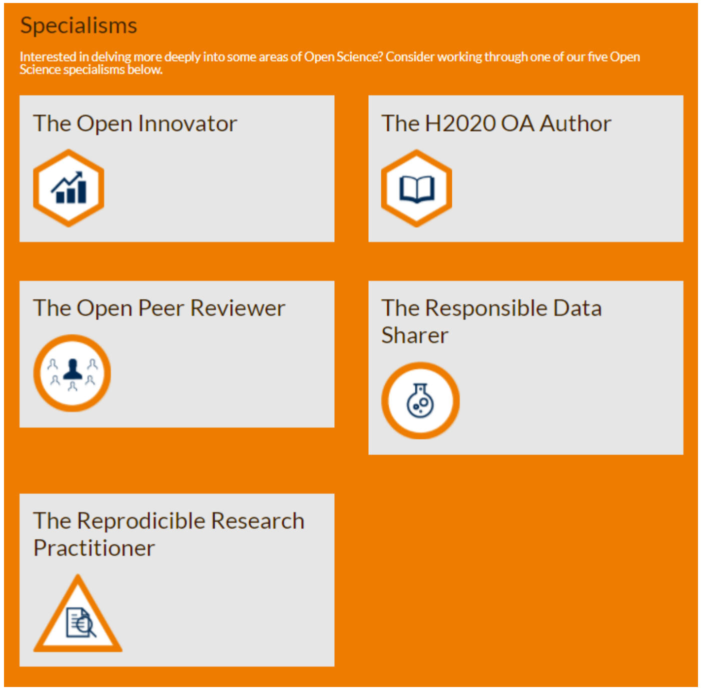
Open Peer Review module example



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

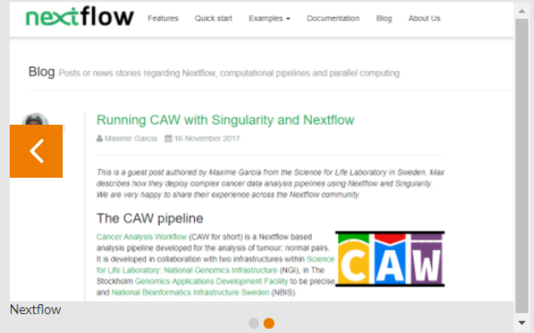


For more information, see www.fosteropenscience.eu/learning-paths

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

docker

Open Peer Review

bioRxiv THE PREPRINT SERVER FOR BIOLOGY

Ethics

EUROPEAN GENOME-PHENOME ARCHIVE 10TH ANNIVERSARY

Open Innovation

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

(Another Set of Guidelines…TOP)

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

http://science.sciencemag.org/content/348/6242/1422

Digital Curation Centre

Thank you!

In collaboration with: