Introduction to Research Data Management and Open Science (aka Research)

S. Venkataraman, PhD
s.venkataraman@dans.knaw.nl

28th and 29th July 2025, Trieste, Italy

Agenda

	Day 1 (28th July)
14:00	Introduction to research data management (RDM)
15:00	Exercise: Practical session on RDM
15:30	Introduction to Open Science (Research)
16:00	**Break**
16:30	Introduction to Open Science (Research) (cont’d)
17:00	Exercise: Open Science
18:00	**End Day 1**
	Day 2 (29th July)
08:30	Introduction to DMPs
09:30	**End**

Learning outcomes

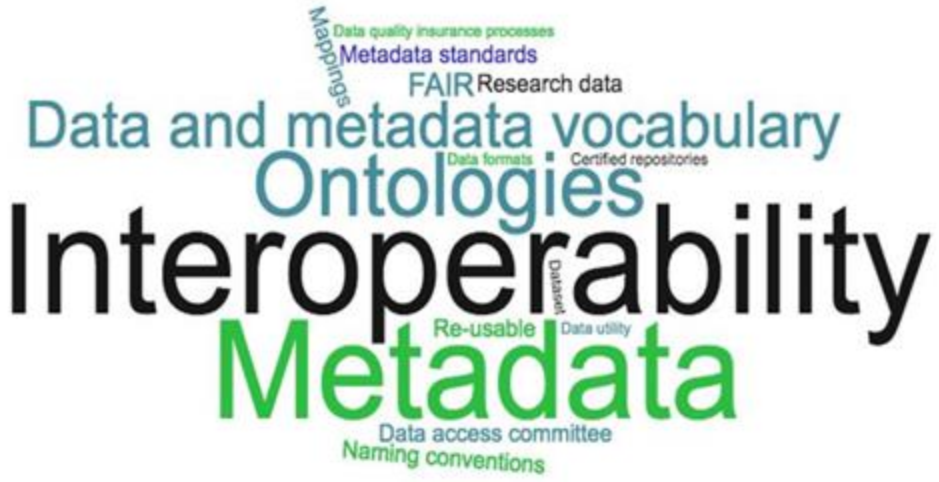
- Be familiar with the curation lifecycle.
- Understand the standardisation methods and principles available to add value to your data.
- Learn about resources to aid your workflows.
- Increase/encourage your level of openness.
- Learn about data management plans and the value in implementing them.

Language is a barrier...

Respondents mentioned 40 terms which were unclear to them in European Commission DMP:

“Researchers are not familiar with the following terms/phrases : Metadata, standards for metadata/data, ontologies, mapping with ontologies, interoperability, ... All the ICT jargon”

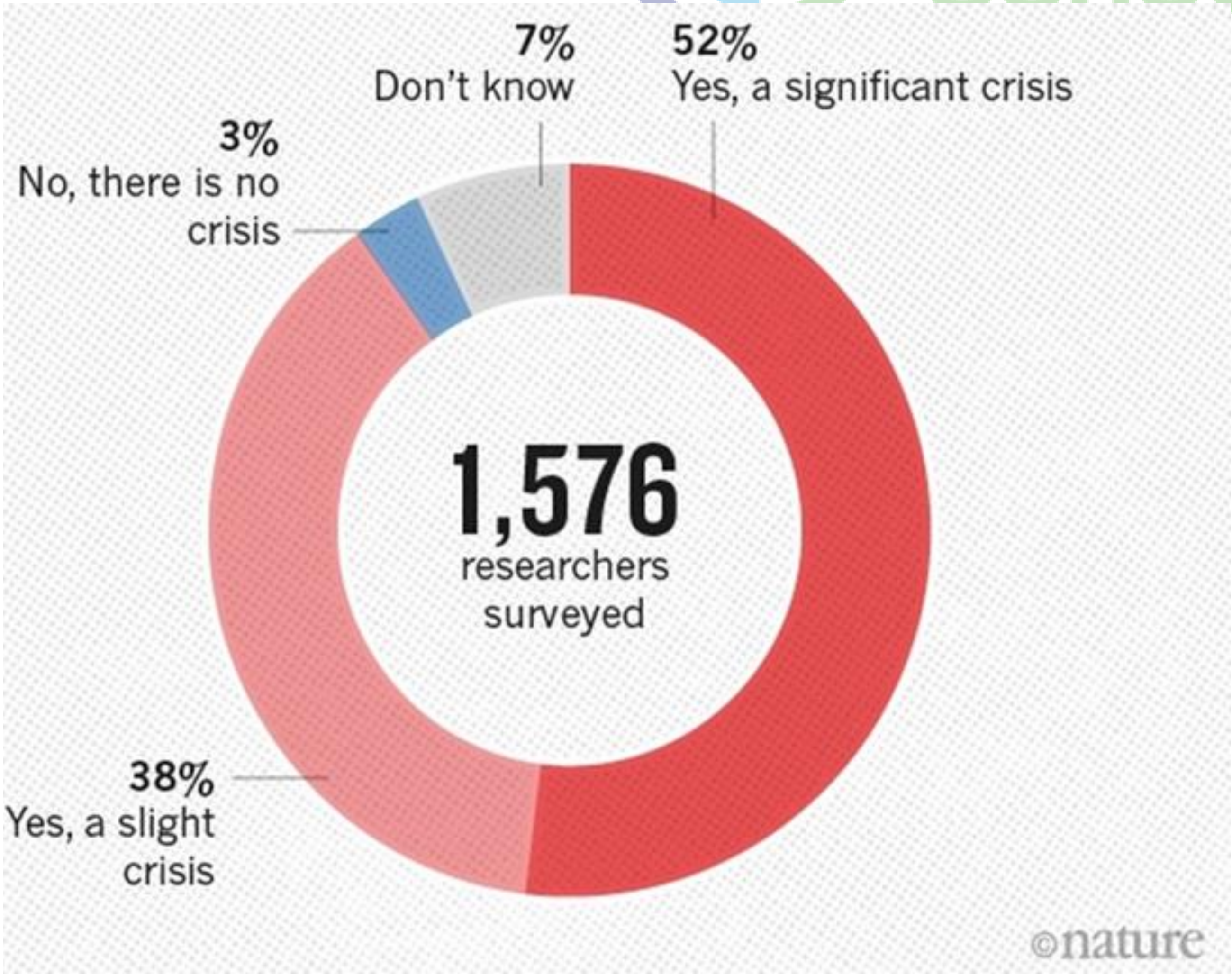
“With the help from Swedish National Data Service we could clarify many questions. Without this help we would not be able to finish the DMP.”



Grootveld et al. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans http://doi.org/10.5281/zenodo.1120245

Is there a reproducibility crisis?

Baker, M. “1,500 scientists lift the lid on reproducibility” *Nature* 533: 452-454 (2016).



Is there a reproducibility crisis?

Kupferschmidt, K. *Tide of Lies* Science 361: 636-641 (2018)

- 5 out of the top 10 in the Retraction Watch Leaderboard are Japanese researchers.
- This article tells a story of one of the researchers in this list and how their research misconduct was uncovered.

Is there a reproducibility crisis?

Kupferschmidt, K. *Tide of Lies* Science 361: 636-641 (2018)

- But this points to cultural issues that could affect the scientific process.
- We need to instil a *culture change*.

The “data deluge”

- The volume of data is growing exponentially with >90% of all data in the world having been generated in just the last few years.
- **How to safeguard for the future?**
  - **Good RDM is essential!**
- **And what about the environmental impact??**

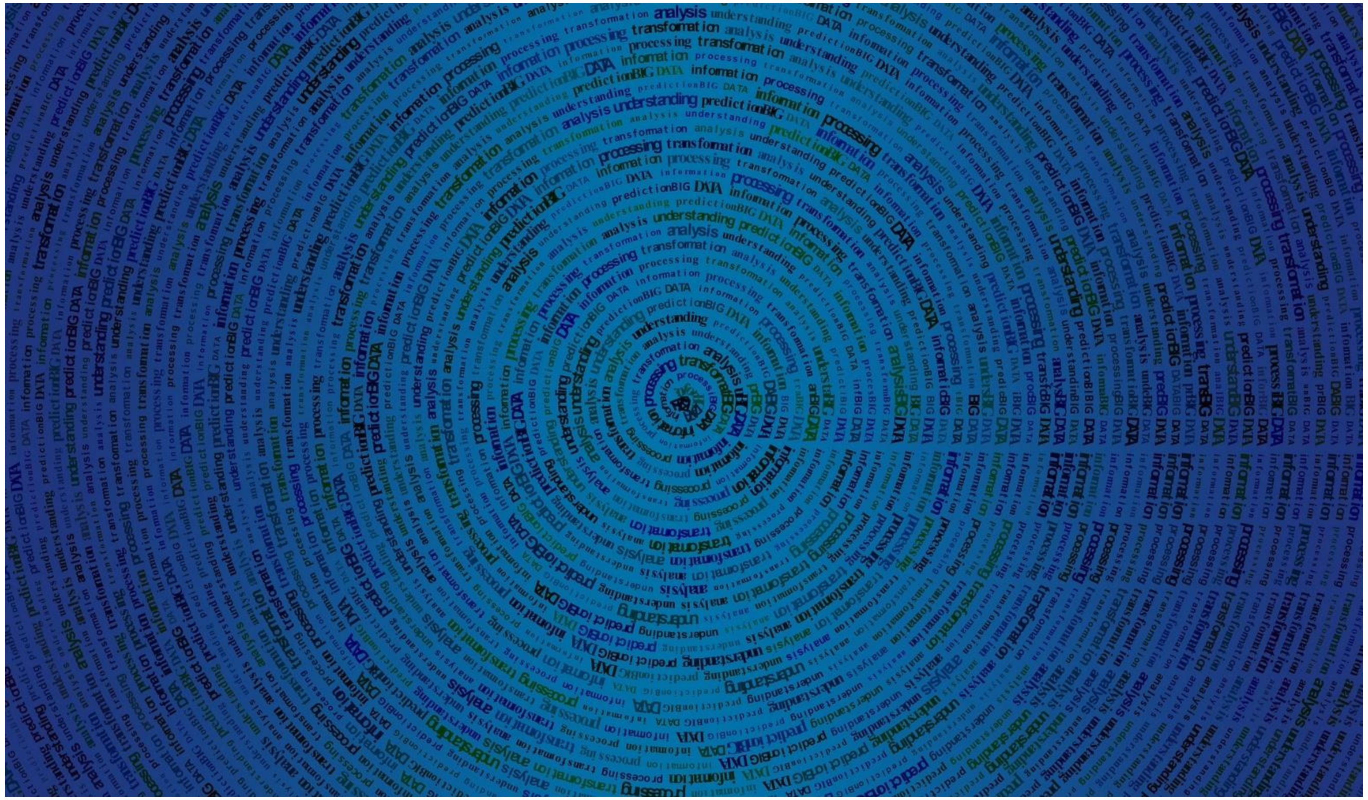
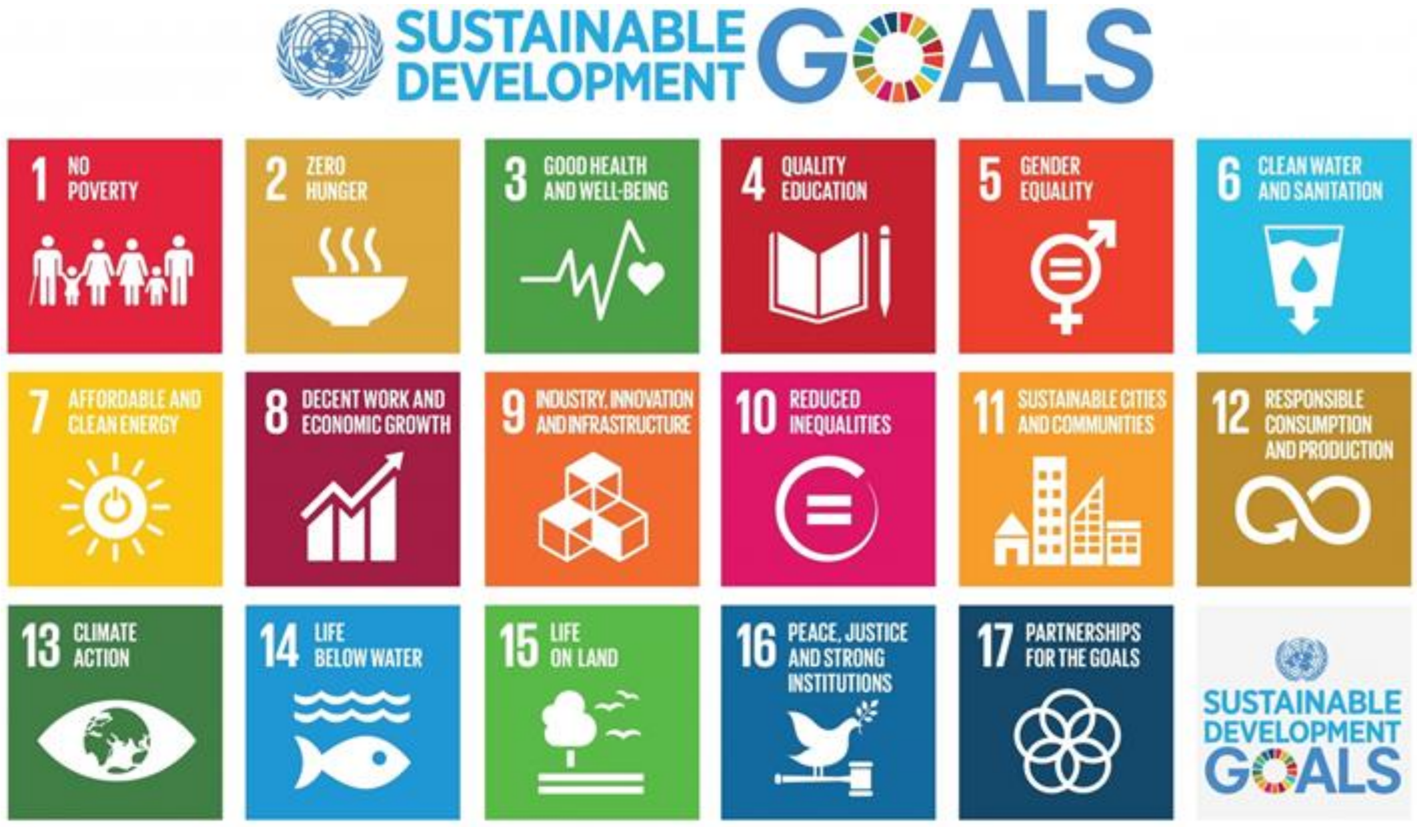


Image by Pete from Pixabay

The wider context

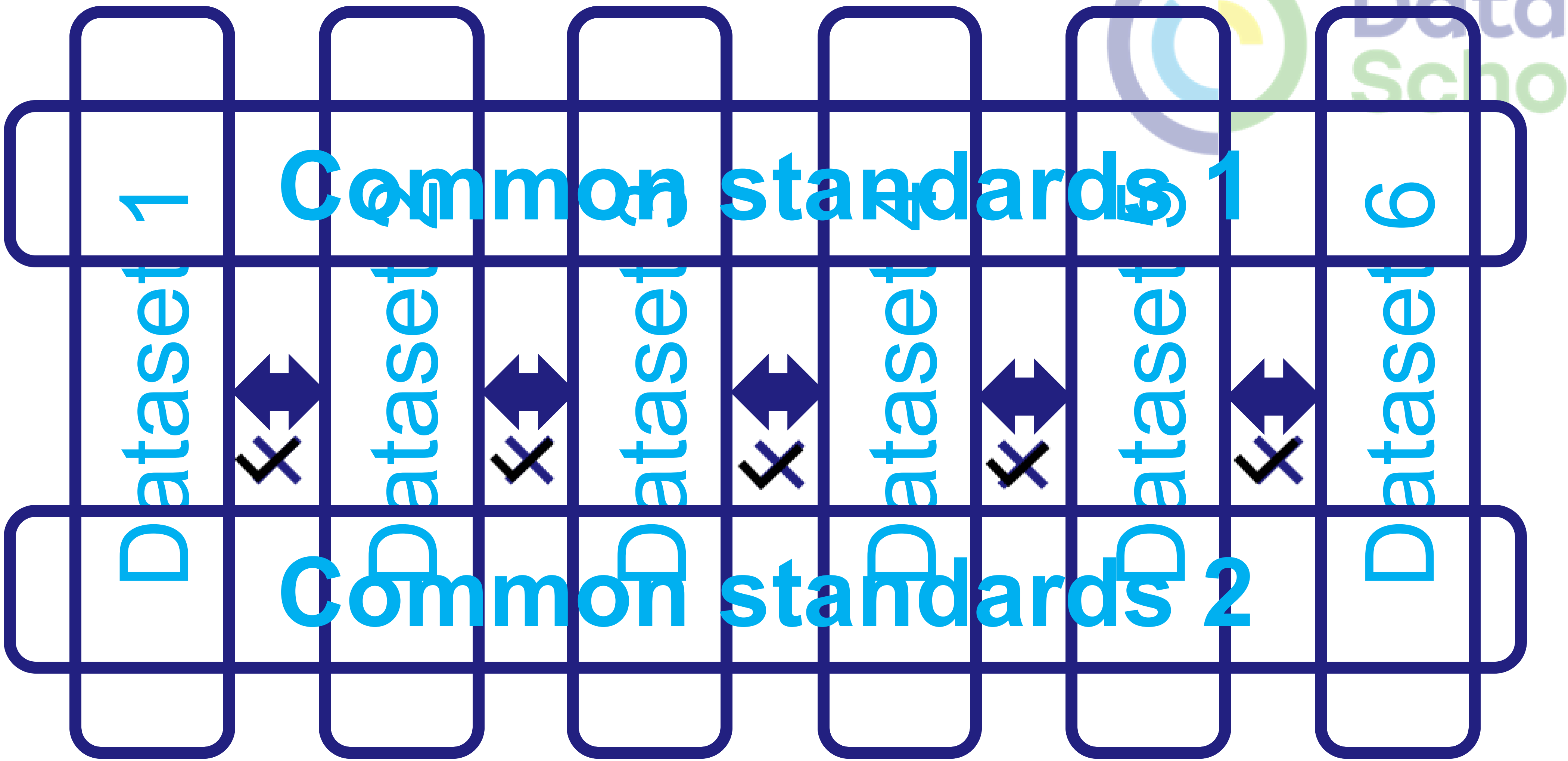
Set of goals outlined by the United Nations


SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Developed in collaboration with **TROLLBÄCK + COMPANY** | TheGlobalGoals@trollback.com | +1.212.529.1010
For queries on usage, contact: dpicampaigns@un.org

Data Schools

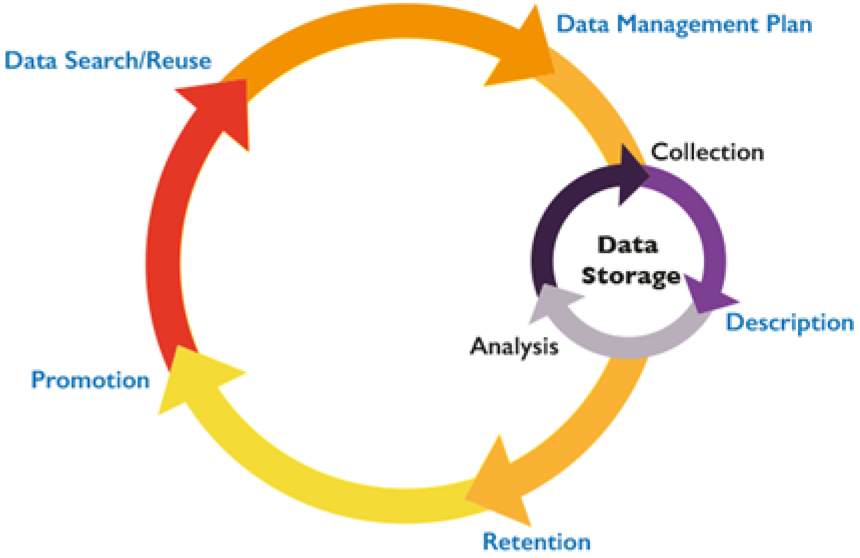
RDM & the Data Lifecycle

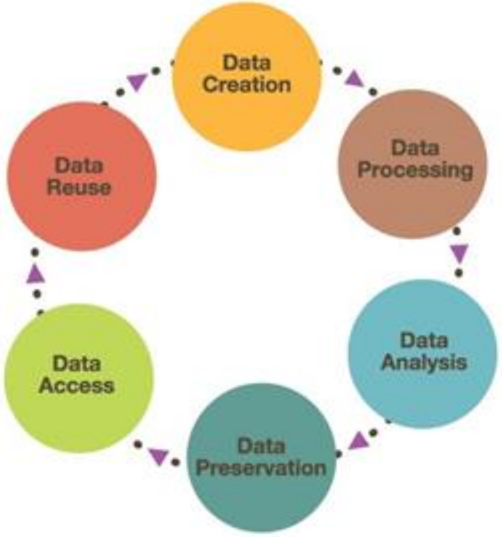


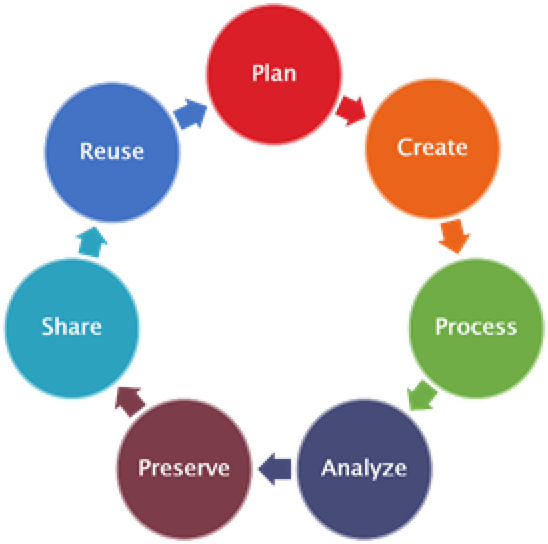
Common standards 1
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Dataset 5
Dataset 6
✓
✓
✓
✓
✓
Common standards 2

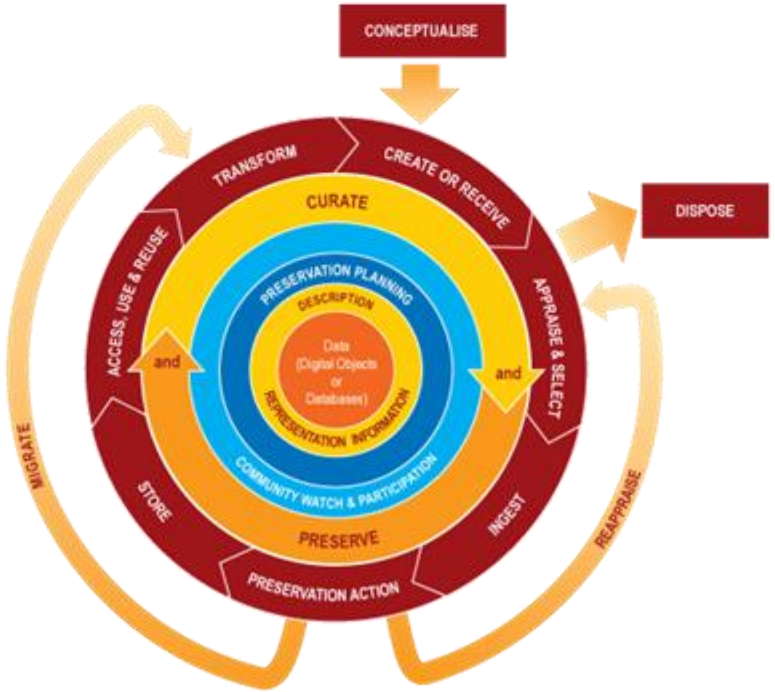
Data silos

RDM lifecycles

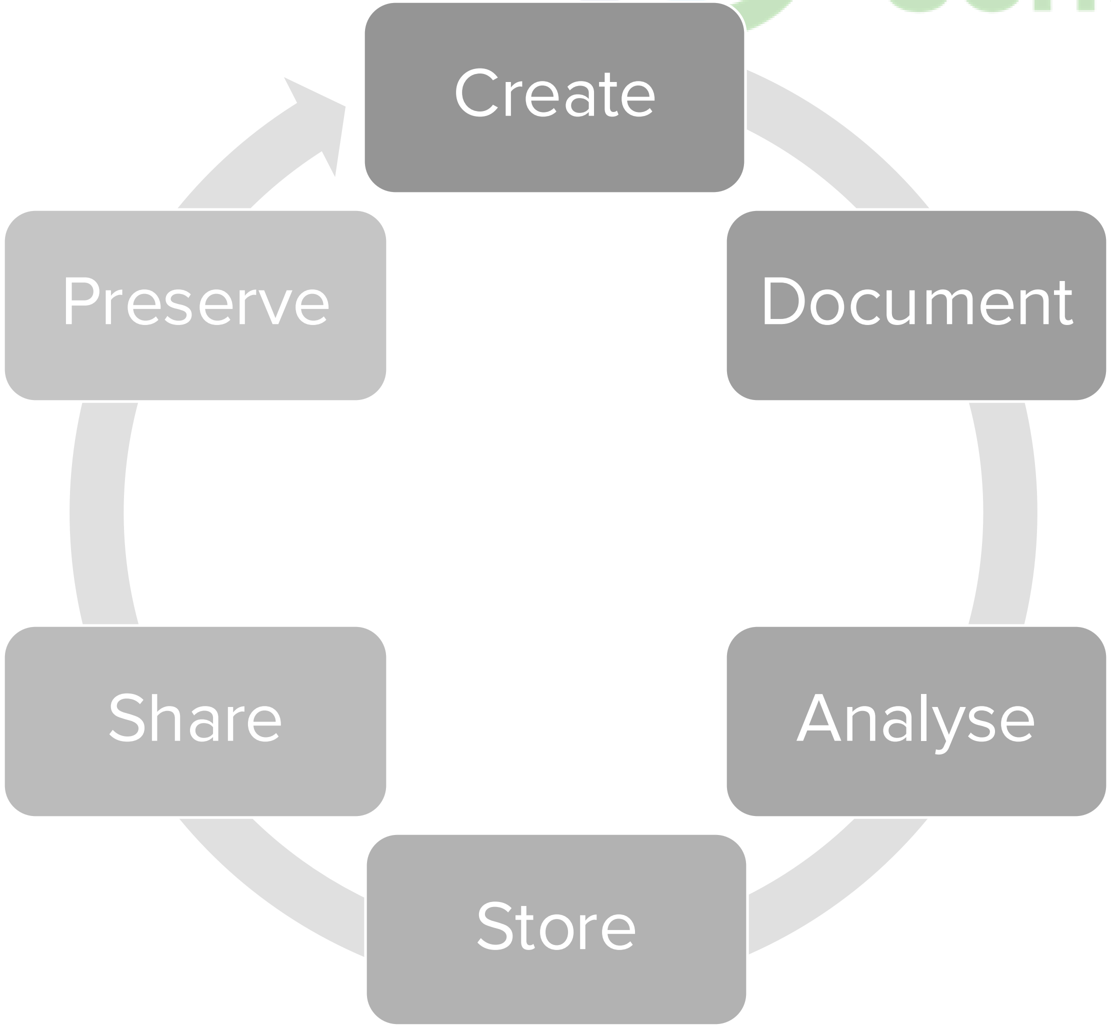








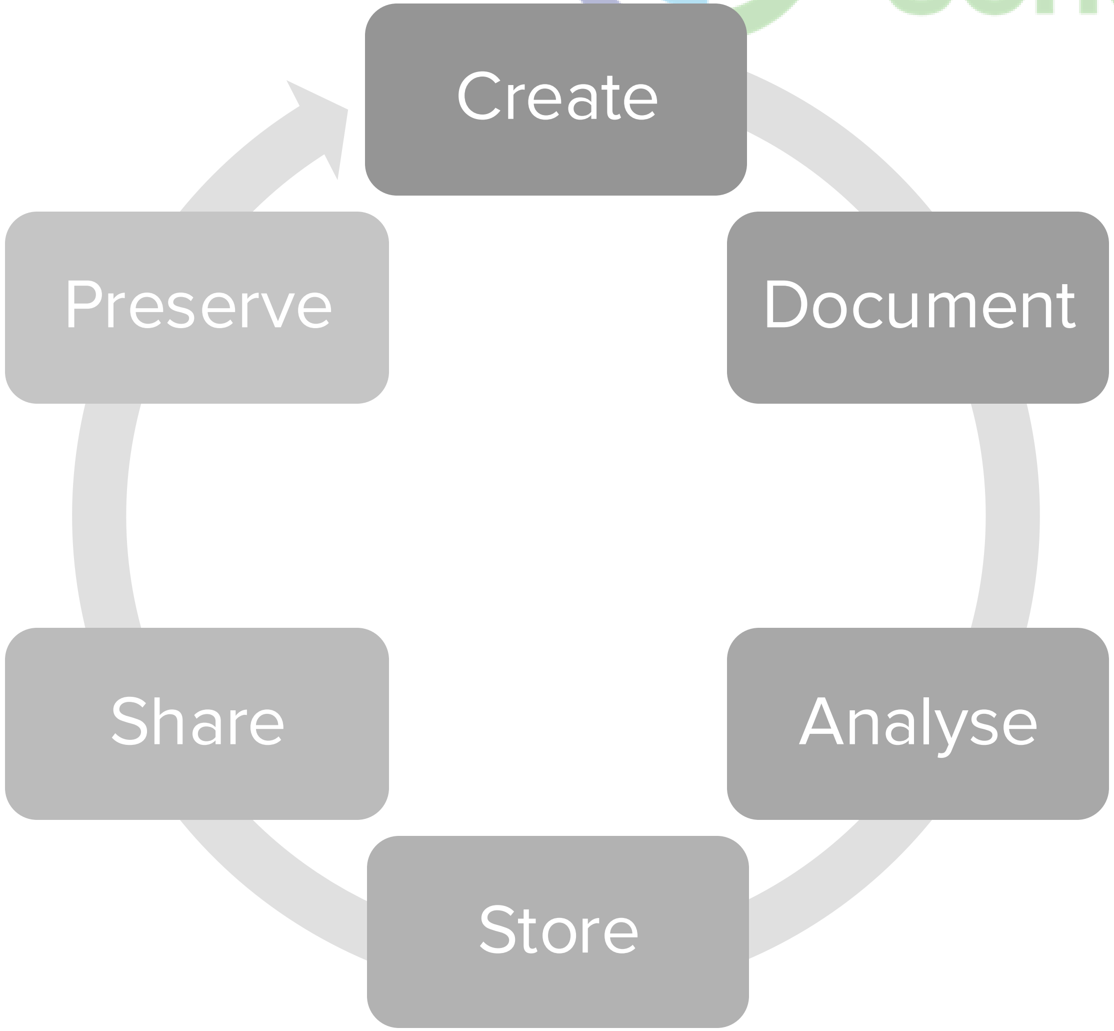
The curation lifecycle



What is Research Data Management?

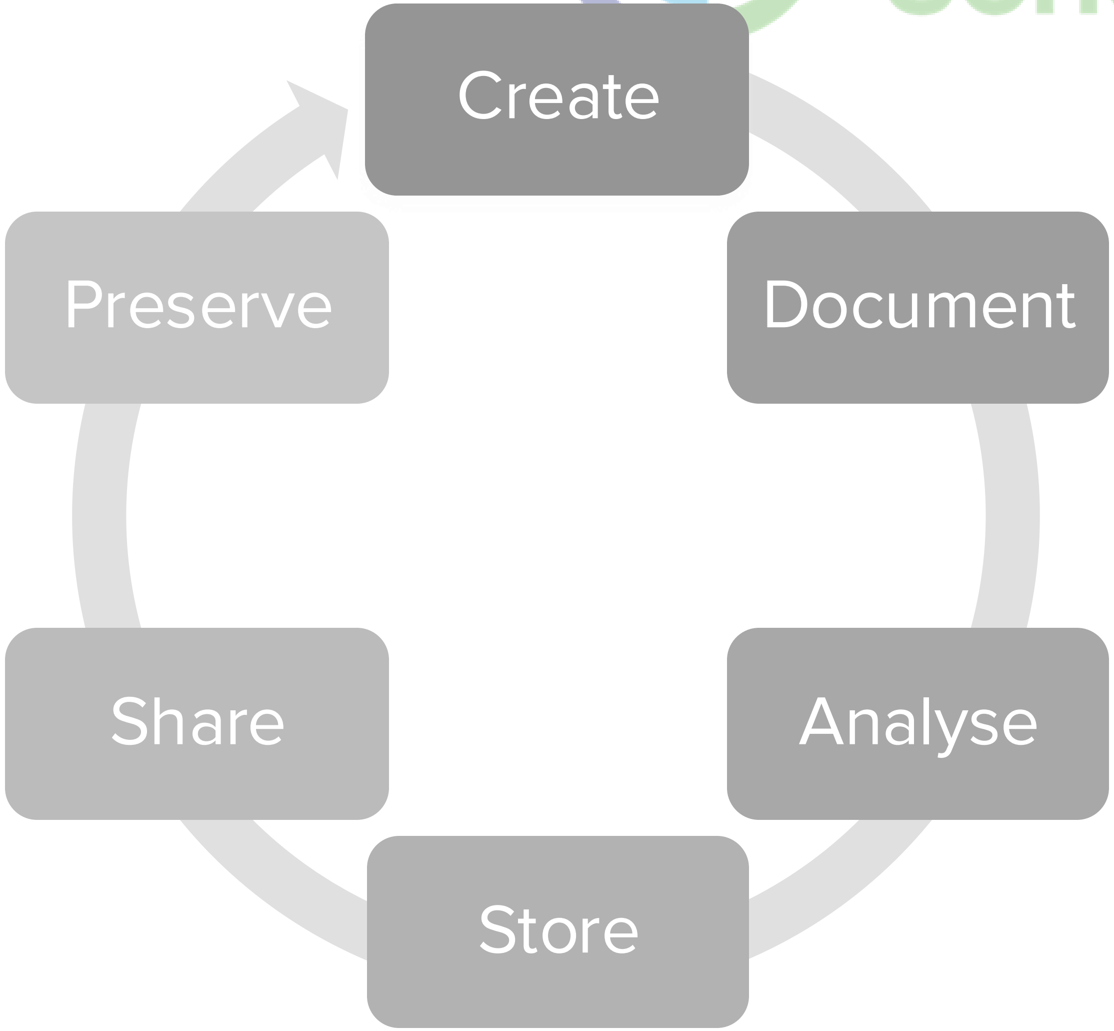
“the active management and appraisal of data over the lifecycle of scholarly and scientific interest”

Data management is part of good research practice.



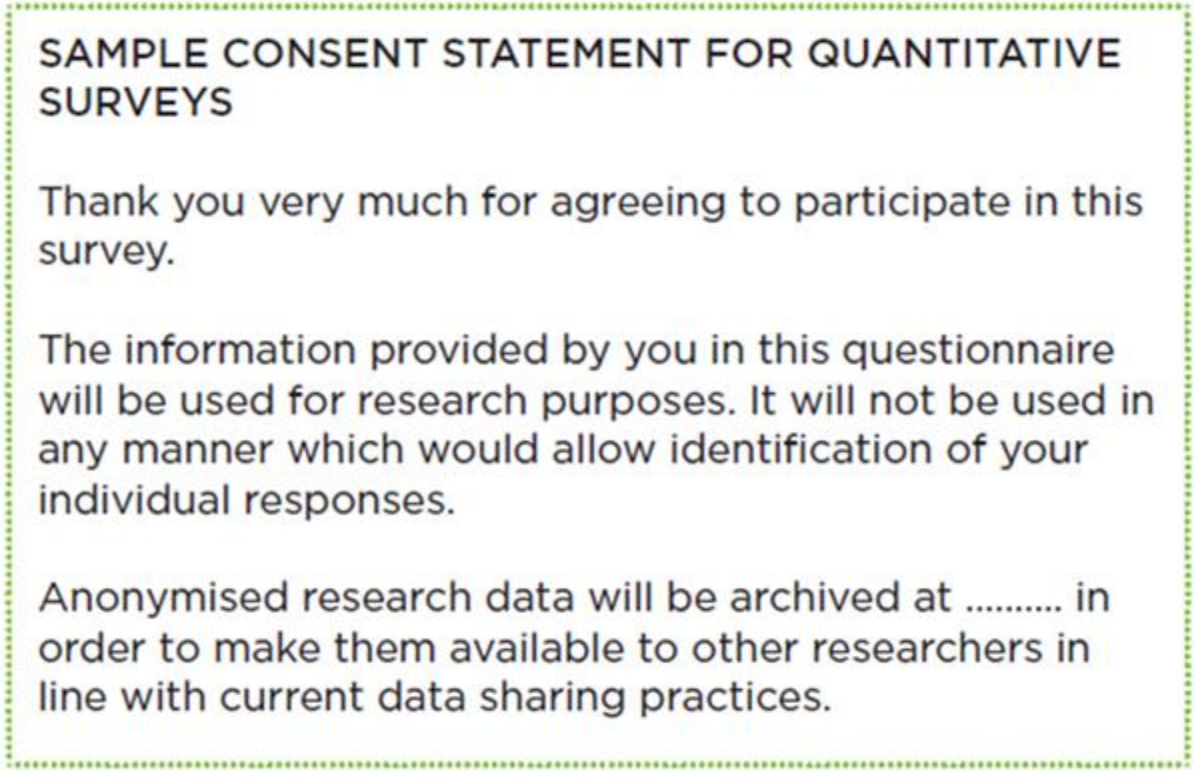
Data creation tips

- 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.

Choose appropriate file formats

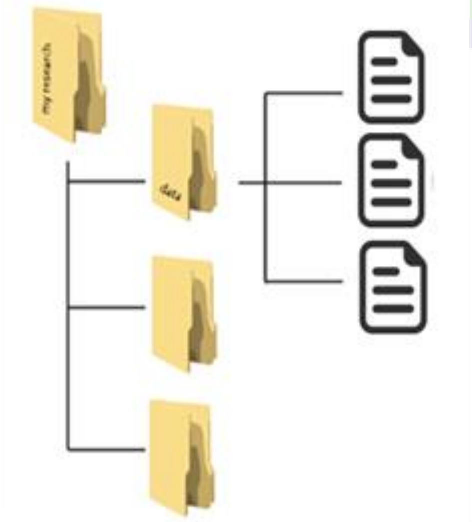
- Different formats are good for different things.
- *open*, *lossless* formats are more sustainable e.g. rtf, xml, tif, wav.
- proprietary and/or compressed formats are less preservable but are often in widespread use e.g. doc, jpg, mp3.
- One format for analysis then convert to a standard format.
- 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

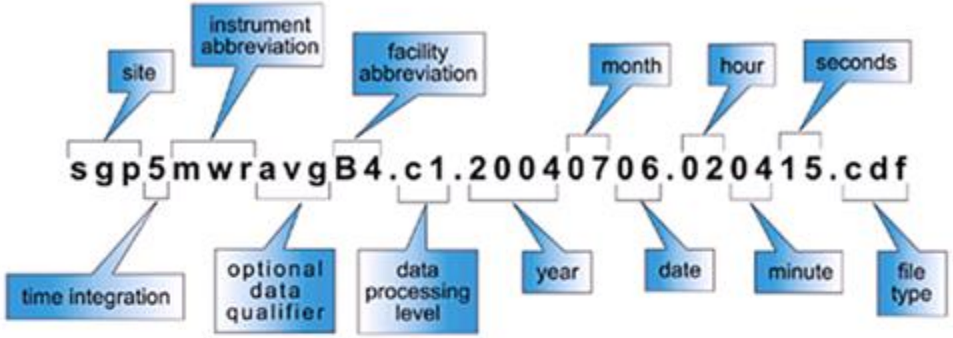
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

How will you organise your data?

- 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.
- Also consider data cleaning!



An example netCDF data file name is depicted below:



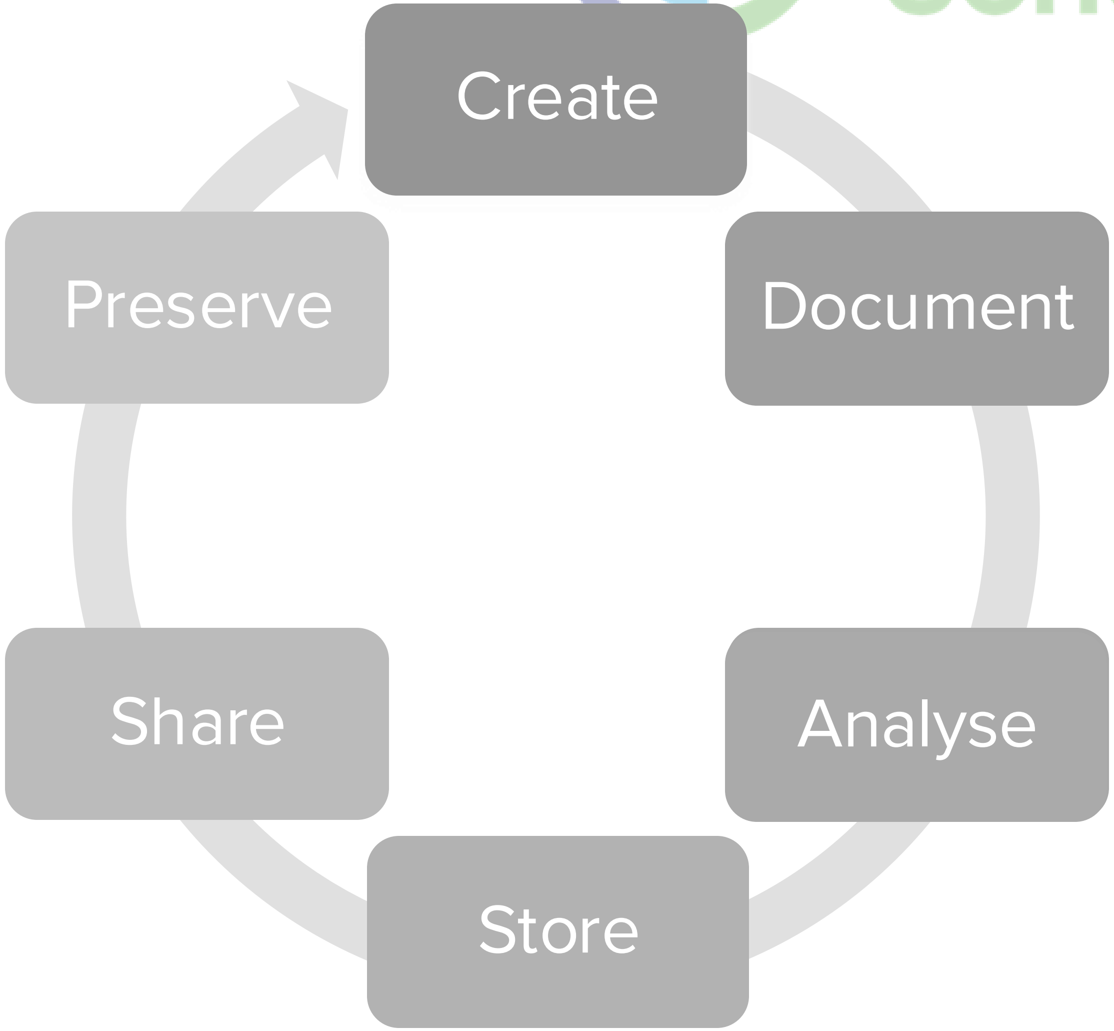
[OpenRefine](#)

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

Documentation

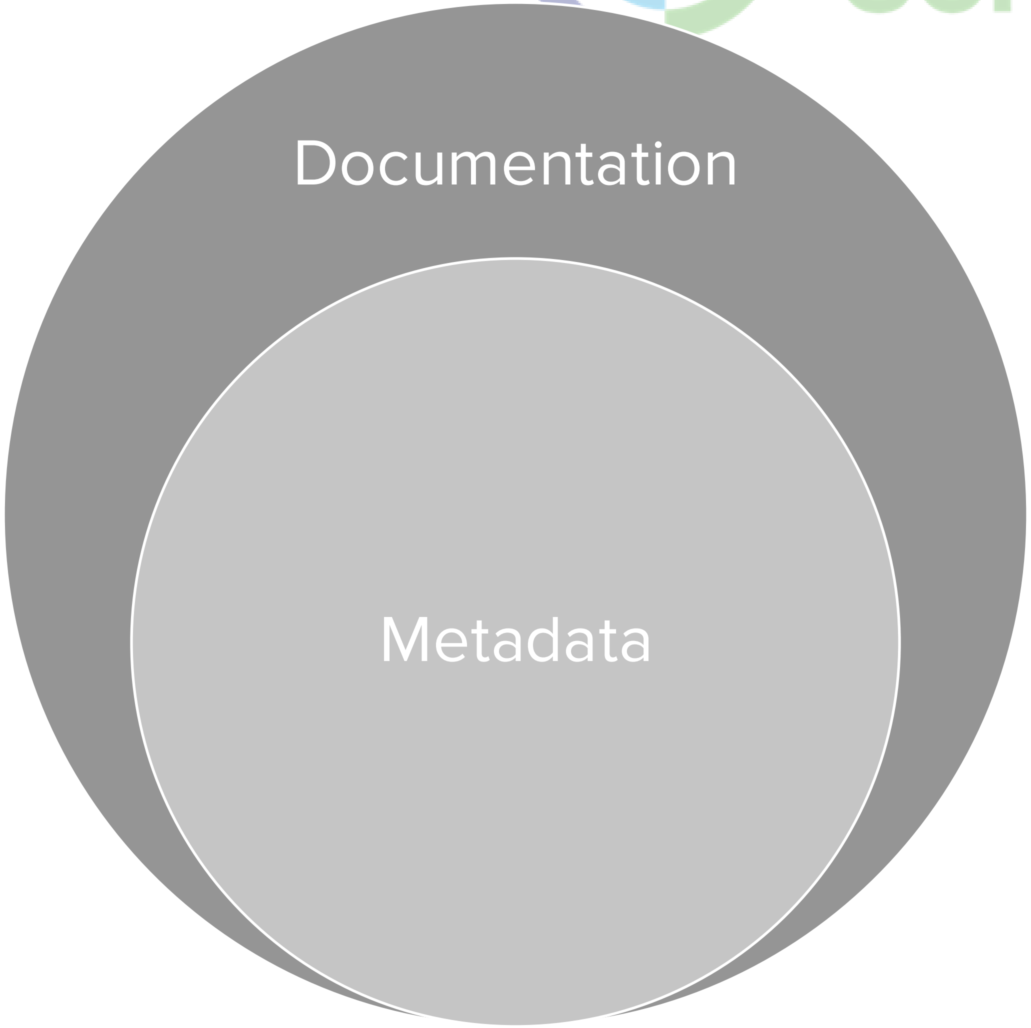
Think about what is needed in order to evaluate, understand, and reuse the data.

- 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.



What are metadata?

- Metadata
  - Standardised
  - Structured
  - Machine and human readable
- Metadata helps to cite and disambiguate data.
- Documentation aids reuse.



Metadata standards

These can be general – such as Dublin Core

Or discipline specific

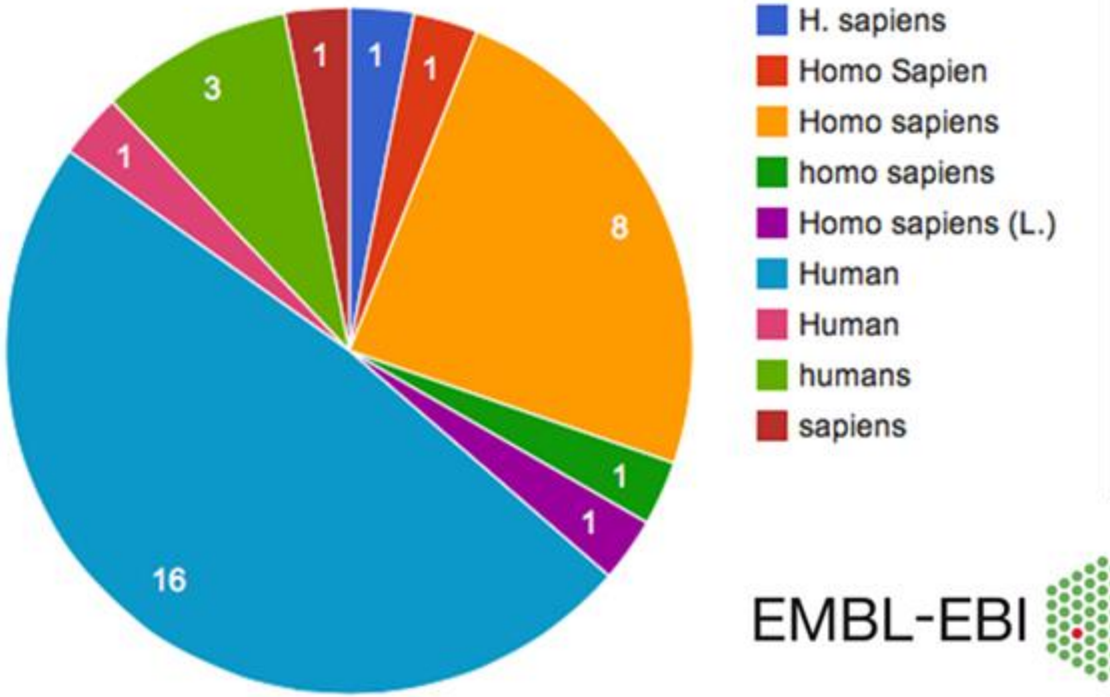
- Data Documentation Initiative (DDI) – social science
- Ecological Metadata Language (EML) - ecology
- Flexible Image Transport System (FITS) – astronomy

Search for standards in catalogues like:

- http://rd-alliance.github.io/metadata-directory/
- https://rdamsc.dcc.ac.uk/
- http://www.fairsharing.org

Controlled vocabularies

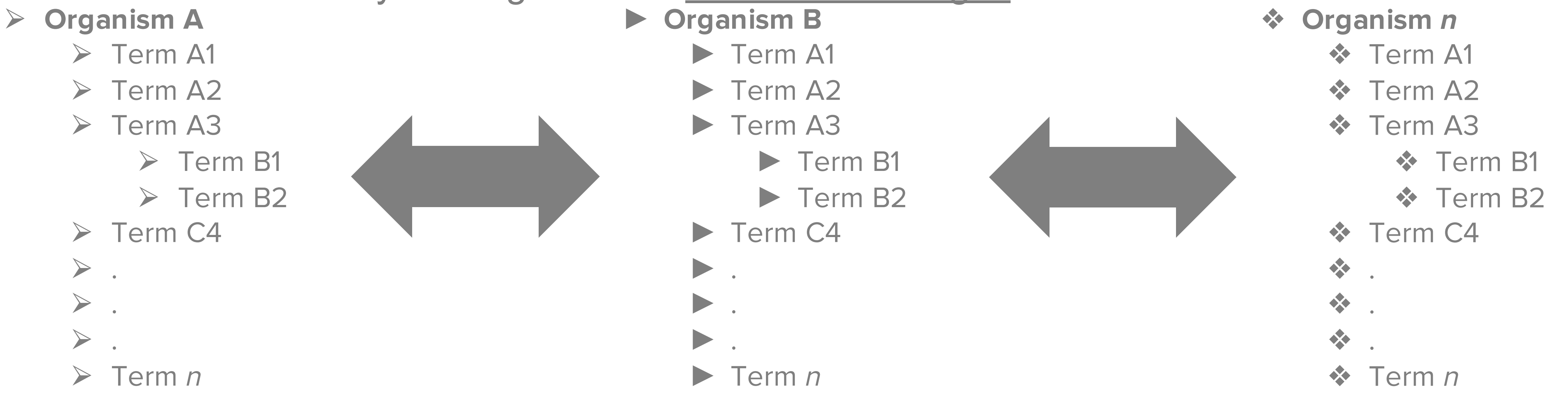
“MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in……”



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

…and ontologies?

- e.g. SNOMED CT (clinical terms) or MeSH
  - Defined terms + taxonomy.
- Useful for selecting keywords to tag datasets.
- You can find many ontologies in the BARTOC catalogue and elsewhere.



➢ **Organism A**
- ➢ Term A1
- ➢ Term A2
- ➢ Term A3
  - ➢ Term B1
  - ➢ Term B2
- ➢ Term C4
- ➢ .
- ➢ .
- ➢ .
- ➢ Term *n*

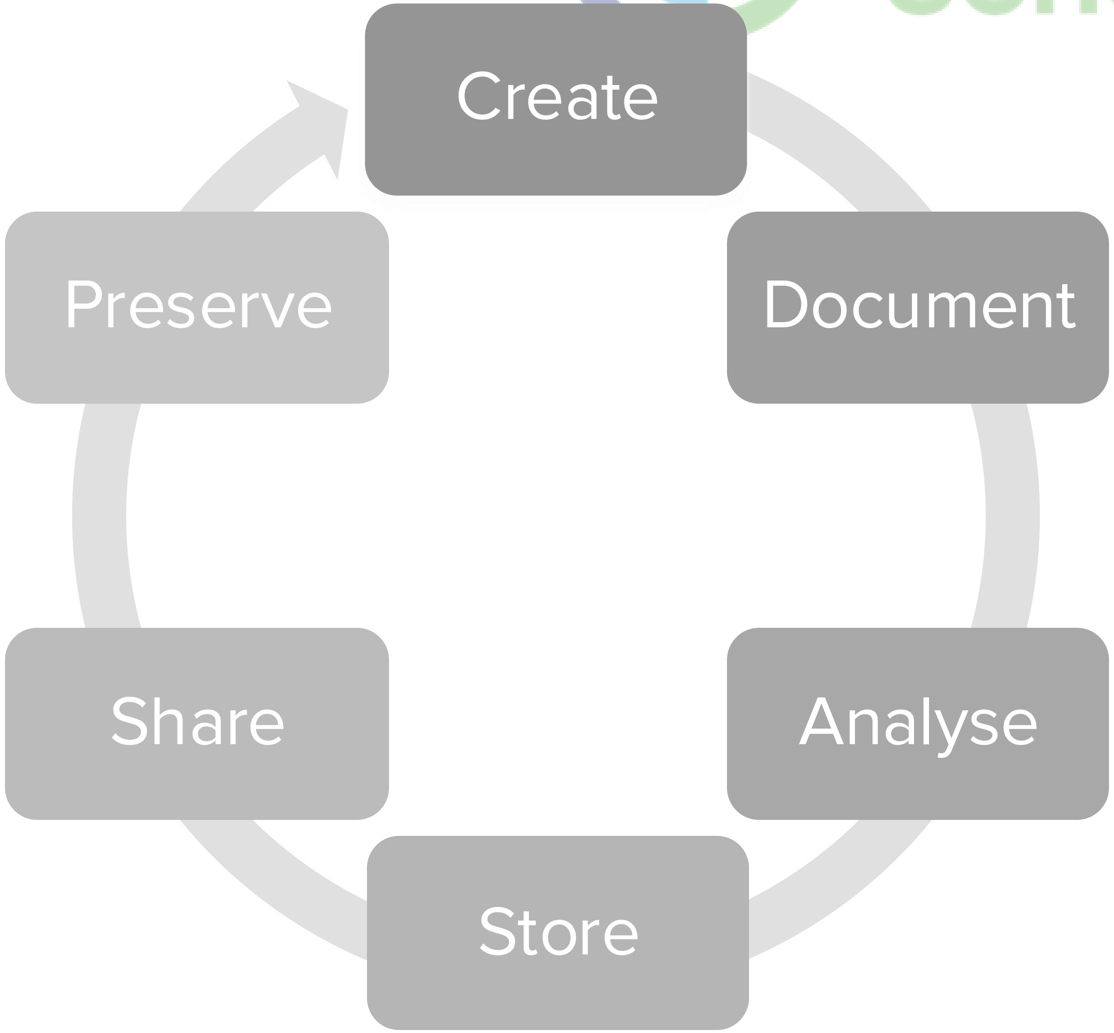
▶ **Organism B**
- ▶ Term A1
- ▶ Term A2
- ▶ Term A3
  - ▶ Term B1
  - ▶ Term B2
- ▶ Term C4
- ▶ .
- ▶ .
- ▶ .
- ▶ Term *n*

❖ **Organism *n***
- ❖ Term A1
- ❖ Term A2
- ❖ Term A3
  - ❖ Term B1
  - ❖ Term B2
- ❖ Term C4
- ❖ .
- ❖ .
- ❖ .
- ❖ Term *n*

Where will you store the data?

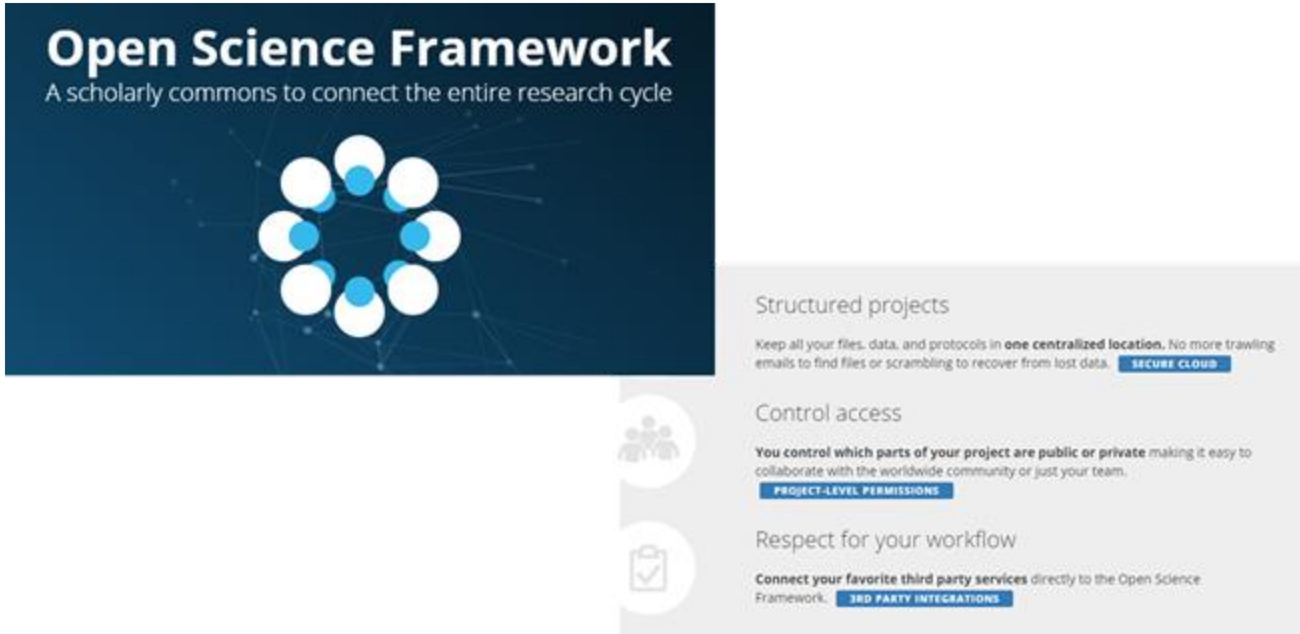
- Your own device (laptop, flash drive, server etc.)
  - And if you lose it? Or it breaks?
- Departmental drives or university servers.
- “Cloud” storage.
- Do they care as much about your data?

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.



Collaborative platforms e.g. OSF

Open platform for sharing data in active phase with fellow researchers and others in secure environment.



Open Science Framework

A scholarly commons to connect the entire research cycle

Structured projects

Keep all your files, data, and protocols in **one centralized location.** No more trawling emails to find files or scrambling to recover from lost data. SECURE CLOUD

Control access

You control which parts of your project are public or private making it easy to collaborate with the worldwide community or just your team.

PROJECT-LEVEL PERMISSIONS

Respect for your workflow

Connect your favorite third party services directly to the Open Science Framework. 3RD PARTY INTEGRATIONS

https://osf.io

Third-party tools for collaboration

Dropbox, Google Drive, OneDrive and other cloud services

- Commercial
- Who owns your data?

ownCloud

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

https://owncloud.org

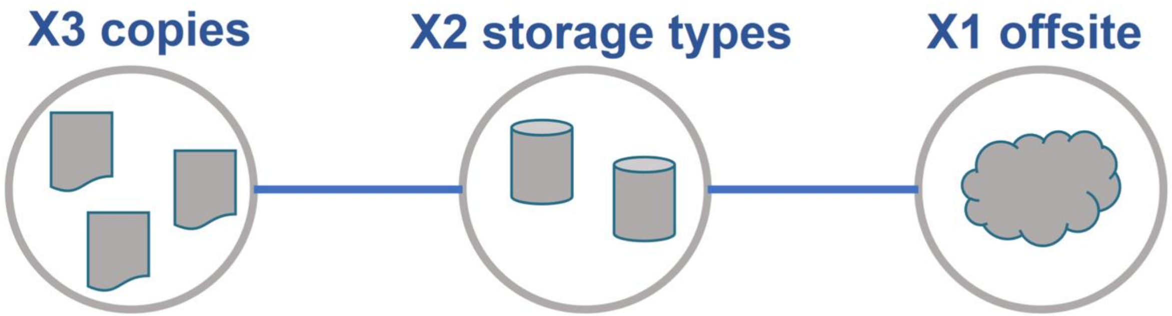
Backup and preservation – not the same thing!

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.

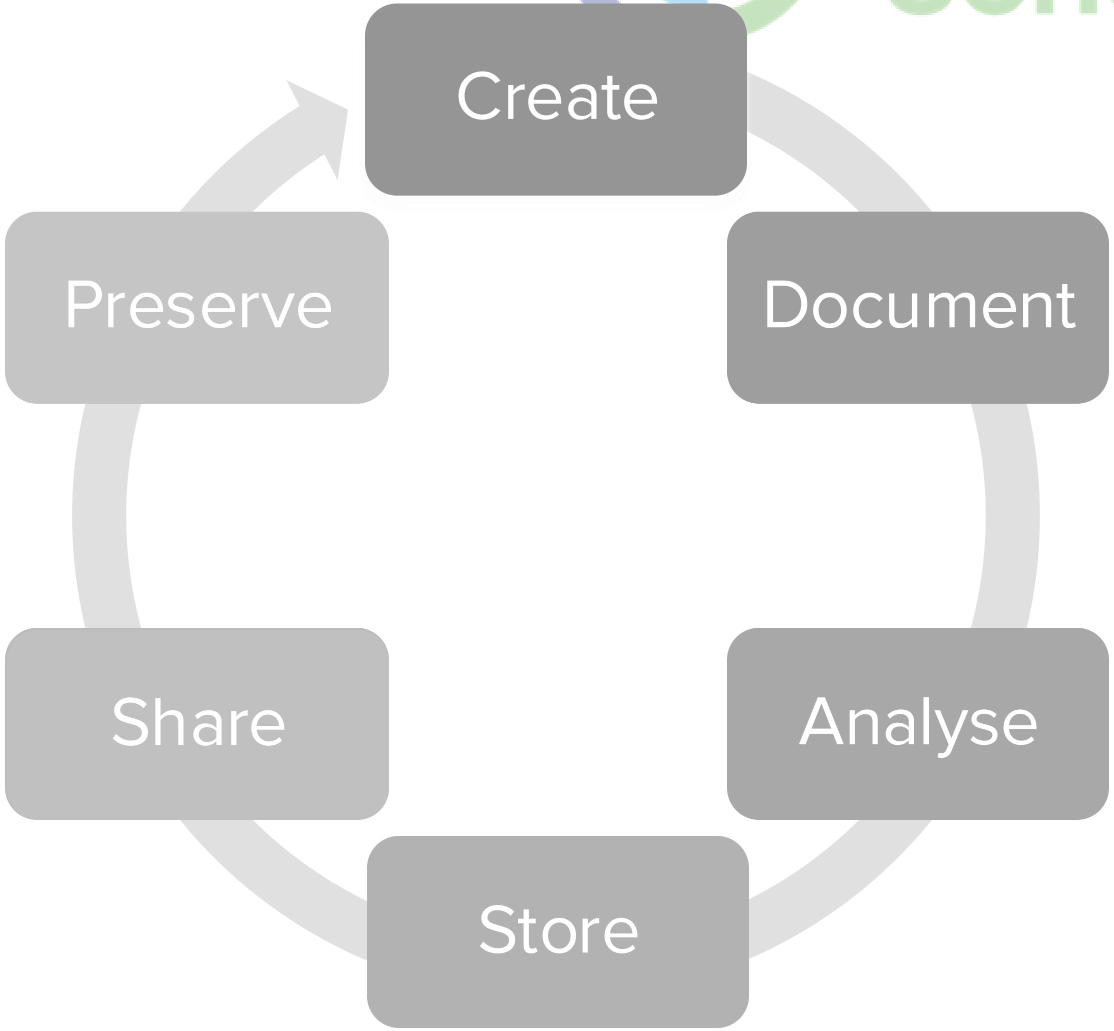
Archiving

- ○ 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.

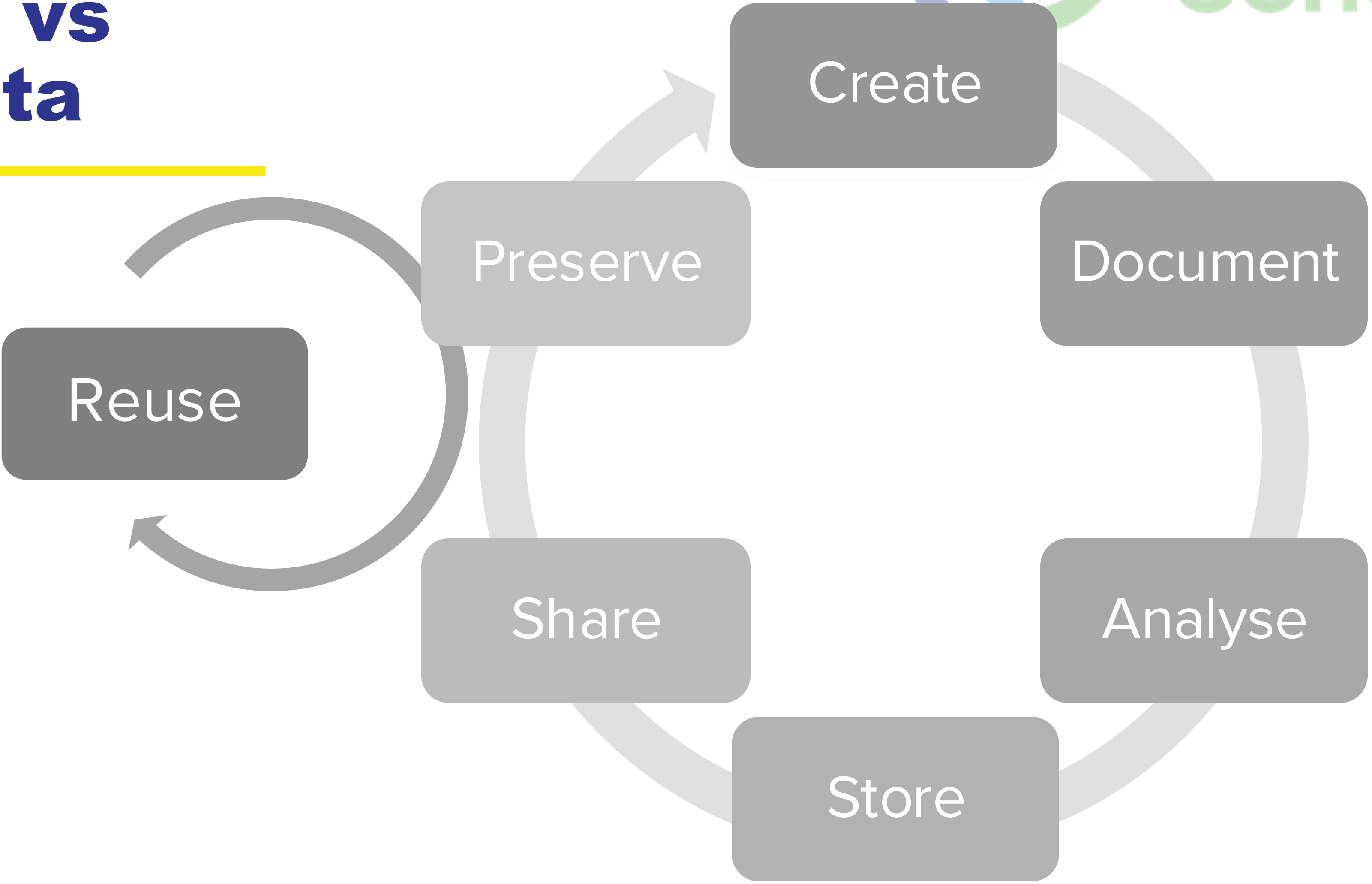


How will you allow others to use your data?

Apply licences to disambiguate reuse restrictions.

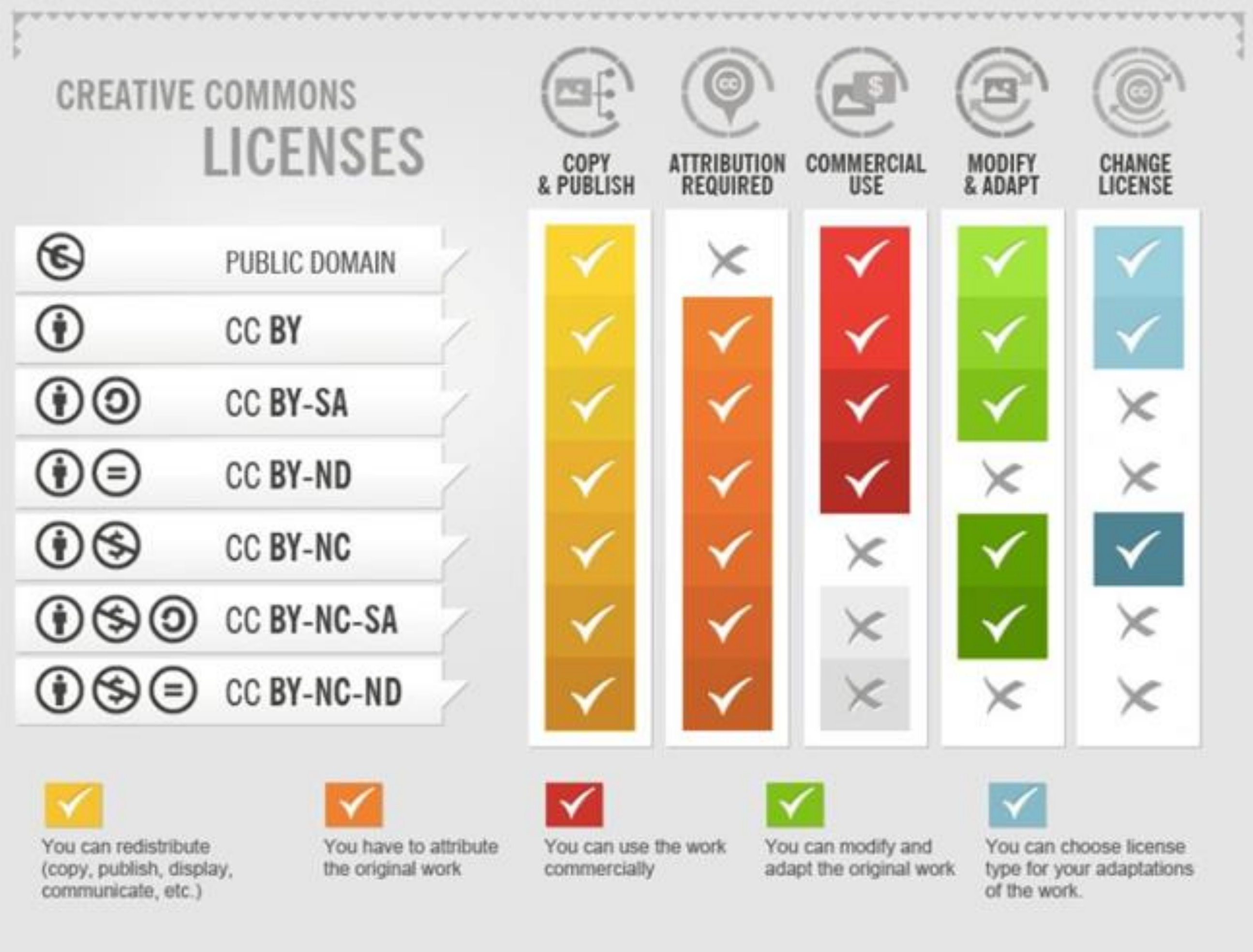


Secondary vs primary data



License research data openly

- For research data, the most common licence that is used is Creative Commons.
- For more detailed explanations, see the descriptions on the Creative Commons website.
- See also GNU licences for software and ODbL for repos.



CREATIVE COMMONS LICENSES

	COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN	✓	✗	✓	✓	✓
CC BY	✓	✓	✓	✓	✓
CC BY-SA	✓	✓	✓	✓	✗
CC BY-ND	✓	✓	✓	✗	✗
CC BY-NC	✓	✓	✗	✓	✓
CC BY-NC-SA	✓	✓	✗	✓	✗
CC BY-NC-ND	✓	✓	✗	✗	✗

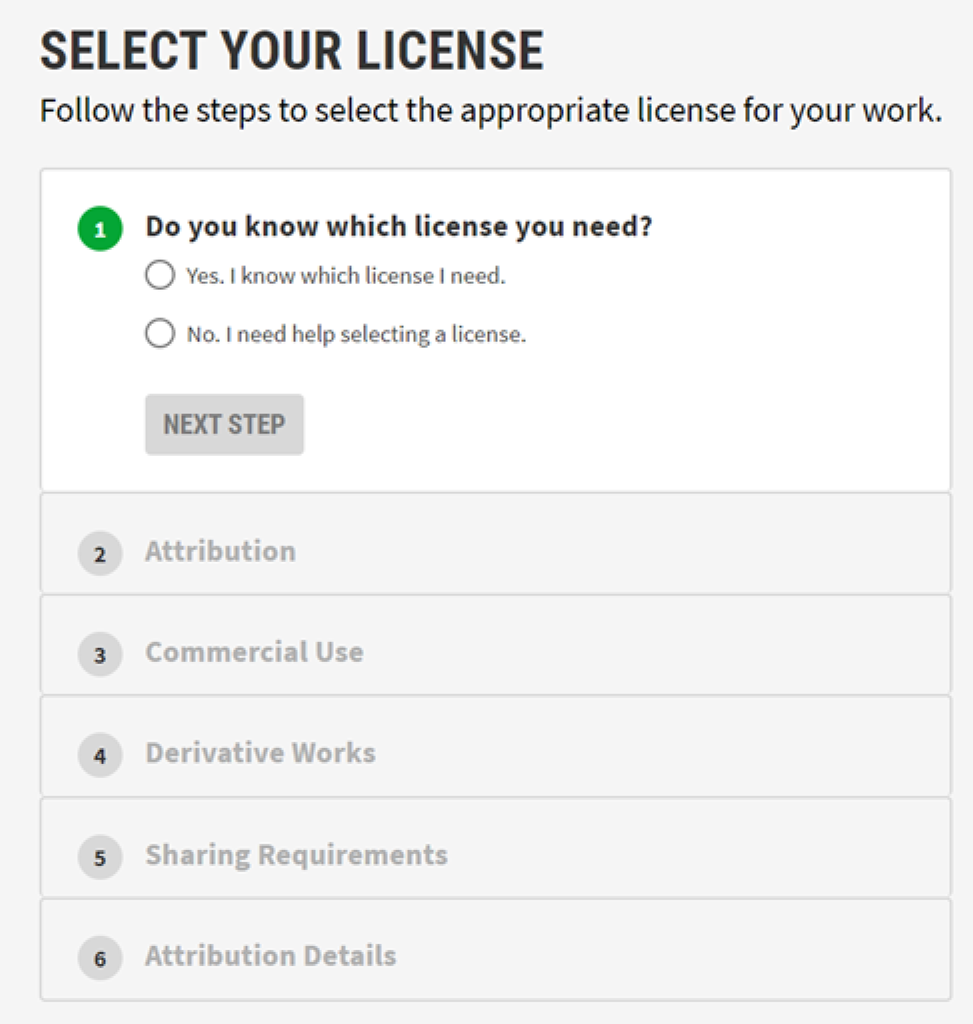
- 🟨 You can redistribute (copy, publish, display, communicate, etc.)
- 🟧 You have to attribute the original work
- 🟥 You can use the work commercially
- 🟩 You can modify and adapt the original work
- 🟦 You can choose license type for your adaptations of the work.

Part of How To Attribute Creative Commons Photos by Foter, licensed CC BY SA 3.0

Tools to decide which license to choose

Choose a license for your data

Check other researchers’ license to know how to re-use their work



SELECT YOUR LICENSE

Follow the steps to select the appropriate license for your work.

1 **Do you know which license you need?**

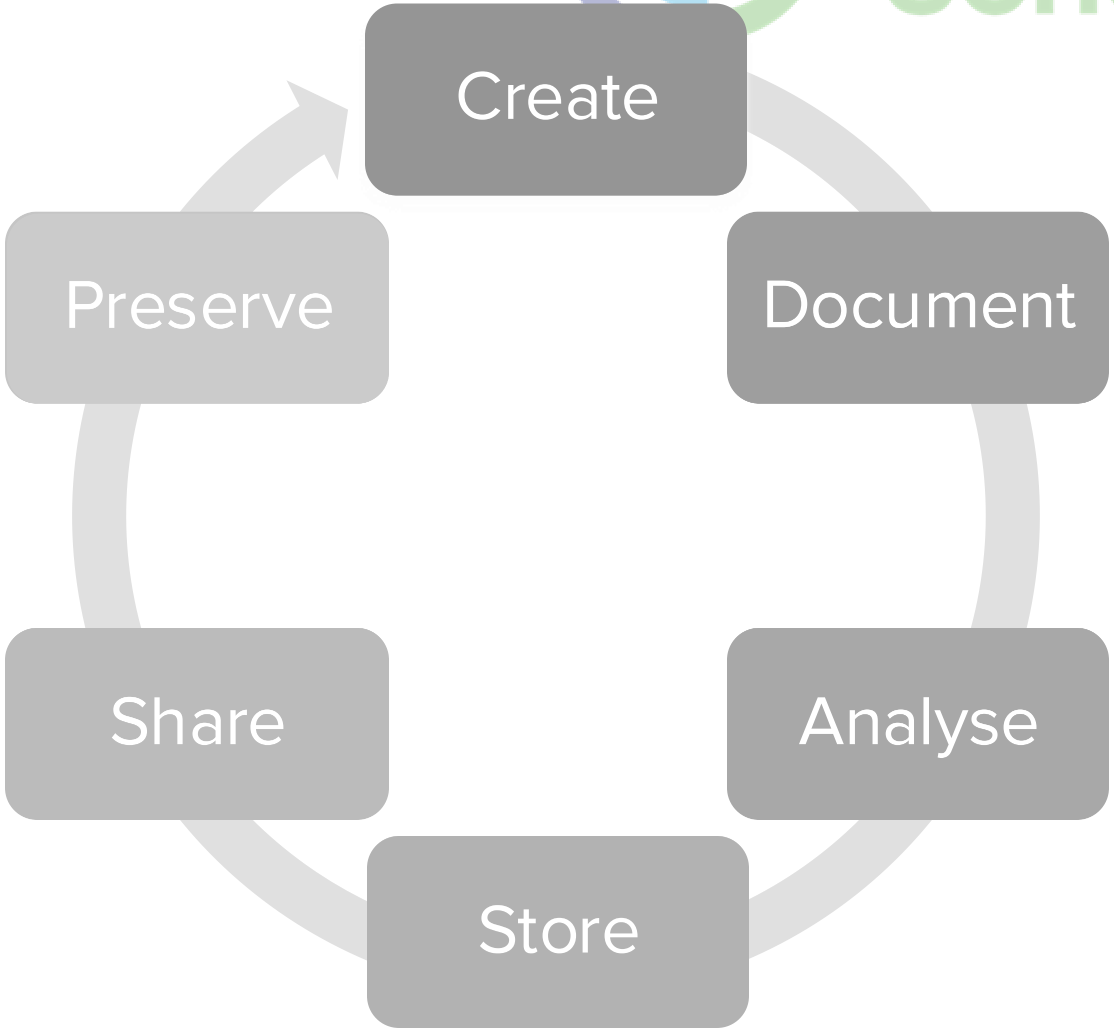
○ Yes. I know which license I need.

○ No. I need help selecting a license.

https://chooser-beta.creativecommons.org/

Deposit in a data repository

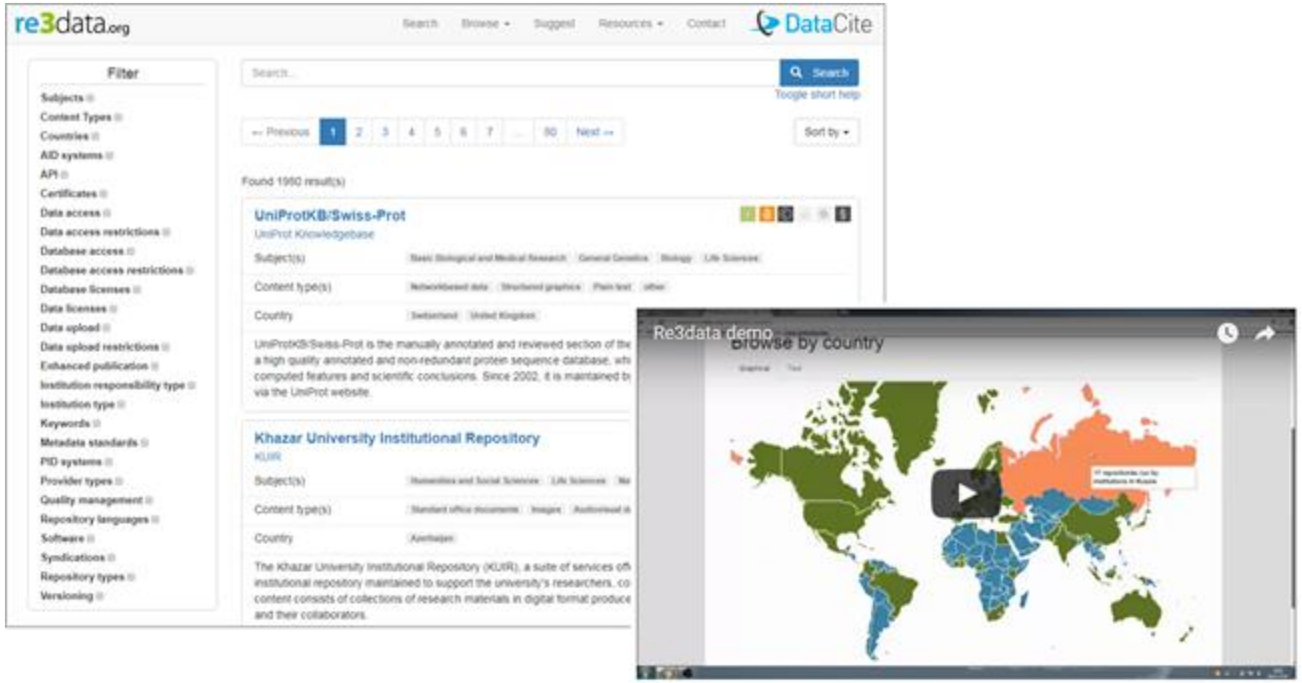
Long-term preservation of data.



Deposit in a data repository

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

www.fosteropenscience.eu/content/re3data-demo



Found 1980 result(s)

UniProtKB/Swiss-Prot

UniProt Knowledgebase

Subject(s): Basic Biological and Medical Research, General Genetics, Biology, Life Sciences

Content type(s): Networkbased data, Structured graphics, Plain text, other

Country: Switzerland, United Kingdom

UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the a high quality annotated and non-redundant protein sequence database, whi computed features and scientific conclusions. Since 2002, it is maintained b via the UniProt website.

Khazar University Institutional Repository

KUIR

Subject(s): Humanities and Social Sciences, Life Sciences, Na

Content type(s): Standard office documents, Images, Audiovisual d

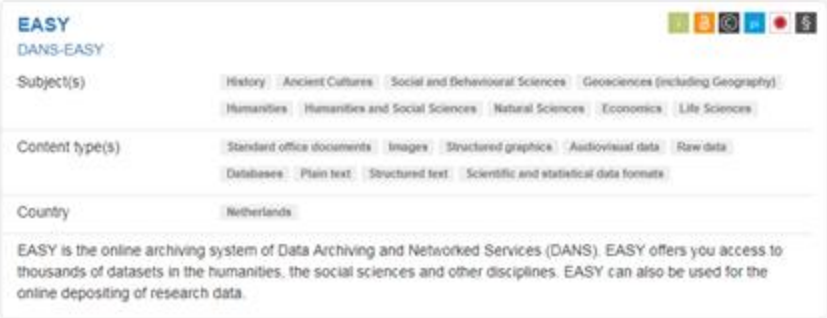
Country: Azerbaijan

The Khazar University Institutional Repository (KUIR), a suite of services off institutional repository maintained to support the university's researchers, co content consists of collections of research materials in digital format produce and their collaborators.

www.re3data.org

Criteria for selecting a repository

- Better to use a domain 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.



EASY

DANS-EASY

Subject(s)	History, Ancient Cultures, Social and Behavioural Sciences, Geosciences (including Geography), Humanities, Humanities and Social Sciences, Natural Sciences, Economics, Life Sciences
Content type(s)	Standard office documents, Images, Structured graphics, Audiovisual data, Raw data, Databases, Plain text, Structured text, Scientific and statistical data formats
Country	Netherlands

EASY is the online archiving system of Data Archiving and Networked Services (DANS). EASY offers you access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data.

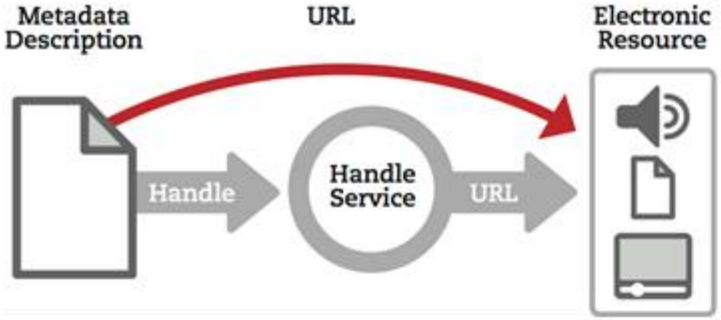
Icons to note open access, licences, PIDs, certificates…

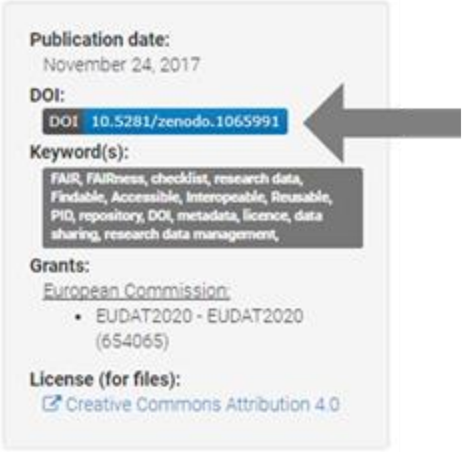
www.re3data.org

What is a Persistent Identifier (PID)?

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

- PIDs come in various forms e.g. ORCID, DOI, ISBN...
- Typically they're actionable i.e. type it into web browser to access.
- Many repositories will assign them on deposit.



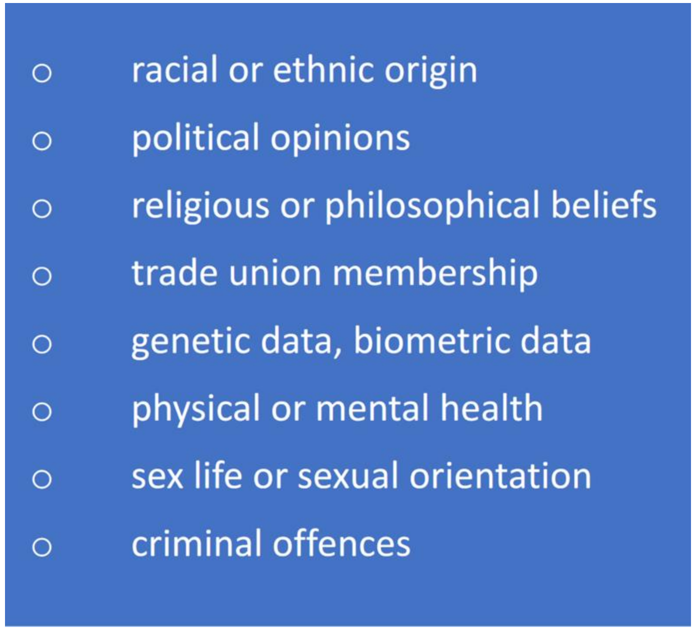

Publication date:
November 24, 2017
DOI:
DOI 10.5281/zenodo.1065991
Keyword(s):
FAIR, FAIRness, checklist, research data, Findable, Accessible, Interopeable, Reusable, PID, repository, DOI, metadata, licence, data sharing, research data management,
Grants:
European Commission:
- EUDAT2020 - EUDAT2020 (654065)

License (for files):
Creative Commons Attribution 4.0

www.re3data.org

Sensitive data

- ○ Personal data (and metadata)
- ○ Confidential data (trade secrets, investigations,...)
- ○ Security data (passwords, financial information, national safety, military,….)
- ○ Data protected by Intellectual Property Rights (IPR)
- ○ Location Data/GPS/mobile phones
- ○ Endangered (plant or animal) species, where their survival is dependent on the protection of their location data (biodiversity community)
- ○ Combination of different datasets could lead to sensitive data?



- o racial or ethnic origin
- o political opinions
- o religious or philosophical beliefs
- o trade union membership
- o genetic data, biometric data
- o physical or mental health
- o sex life or sexual orientation
- o criminal offences

Sensitive data best practices

- Access controls
  passwords, firewall (viruses, hacking)
- Anonymisation
  removing or aggregating variables or reducing the precision or detailed textual meaning of a variable
- Encryption
  encoded digital information
- Share in a secure place
  no cloud drives
- Store in an isolated machine
  server not connected to Internet
- Secure disposal
  no data recovery is possible (uninstall)

Exercise - 25 min (+ 20 min discussion)

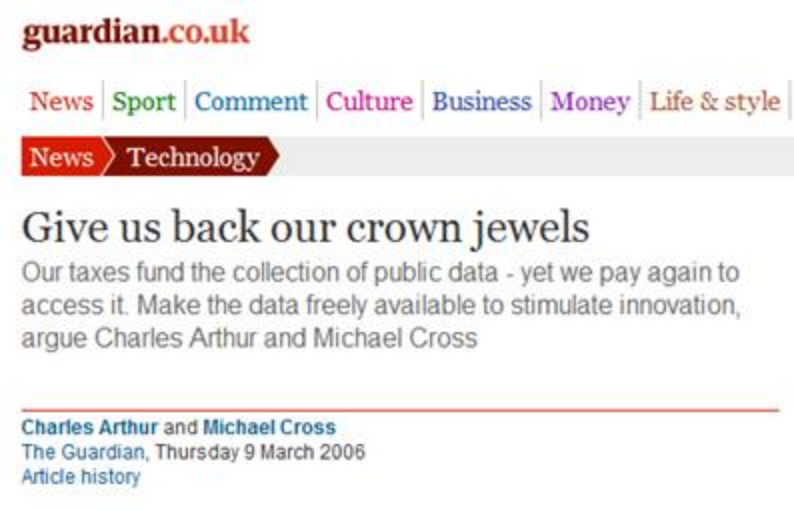
Imagine you are a biologist who is doing microscopy experiments imaging tissue specimens. The data captured by the imaging is 100s of GB in size and is then cleaned and analysed to produce derivatives of the original captured data. Some of these derivatives may eventually be published. In preparation for publication, the data will also be segmented and annotated using standard ontologies. Documentation will also include metadata standards that will sufficiently describe the experimental procedure to allow reproducibility. Publication of the data is mandatory due to funder policy and must be deposited in a repository within 3 years of data production and must use an open licence without restrictions on reuse.

Now…please split into groups and see if you can answer the following questions using the tools and guidelines that have been described:

- What **file format(s)** should data be captured/preserved in?
- Which **metadata standard(s)** should be used?
- What **ontology(ies)** should be used?
- Which **licence(s)** should be used?
- Which **repository** would be the best fit for these data?
- Do you foresee any problems with the data?
- (Hint: not all the questions can be answered definitively! – but why not?)

Data Schools

Data sharing and openness



guardian.co.uk

News | Sport | Comment | Culture | Business | Money | Life & style

News › Technology

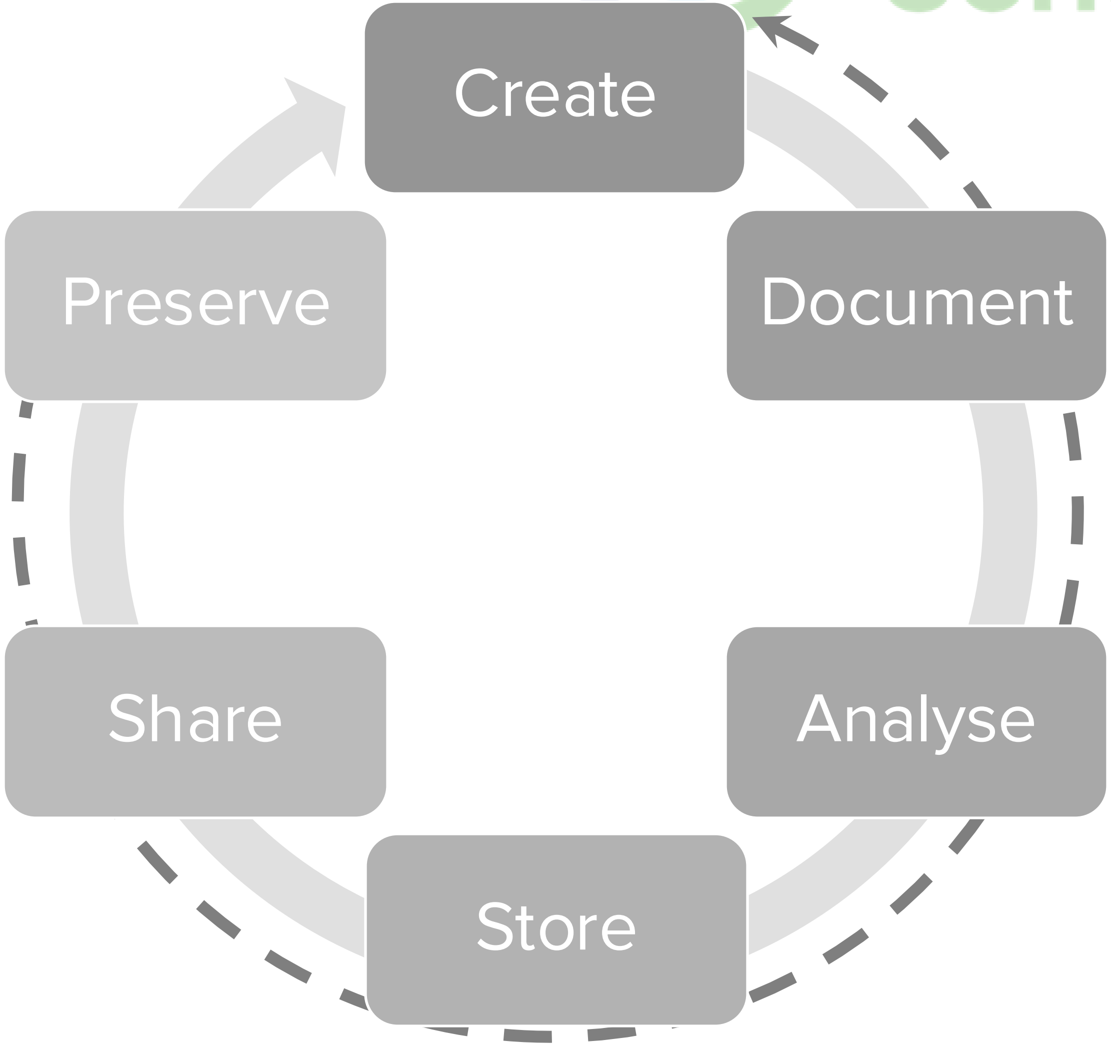
Give us back our crown jewels

Our taxes fund the collection of public data - yet we pay again to access it. Make the data freely available to stimulate innovation, argue Charles Arthur and Michael Cross

Charles Arthur and Michael Cross
The Guardian, Thursday 9 March 2006
Article history

And open research...

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



Why make data available?

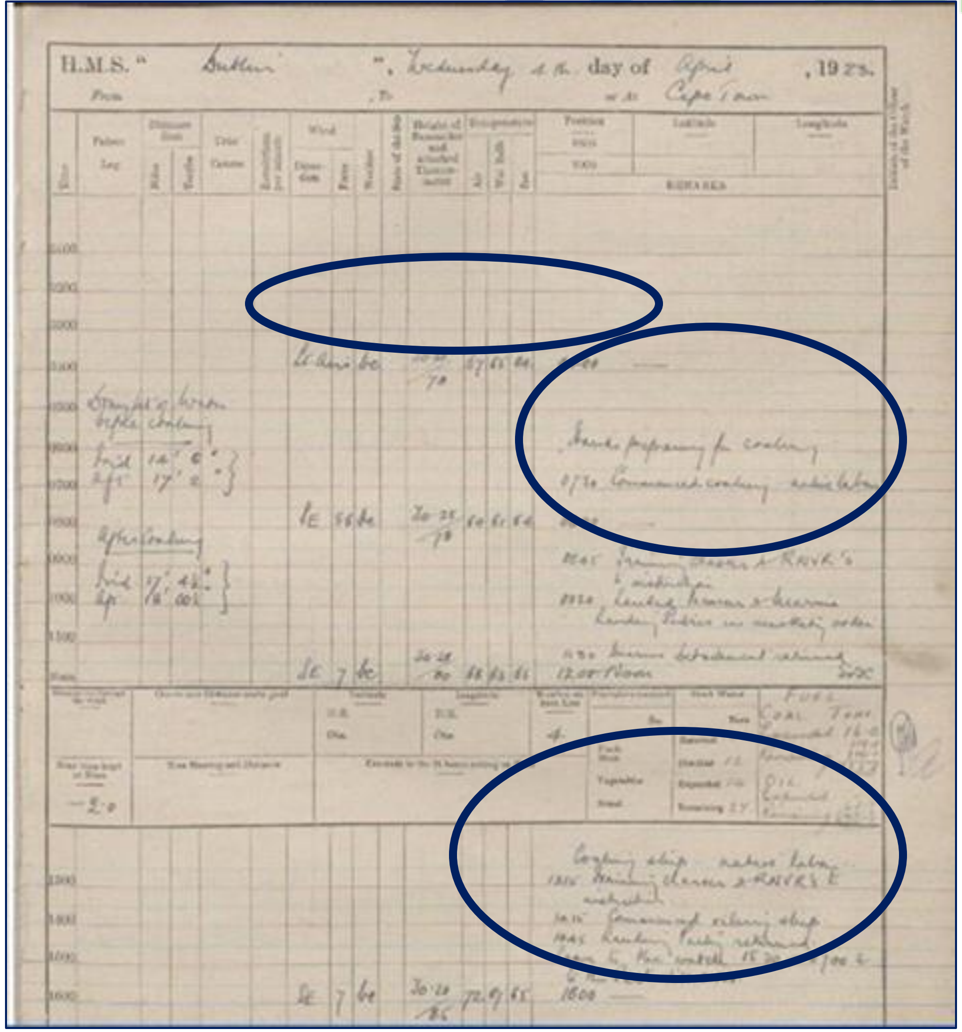
"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

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

The Old Weather Project

Data for research, not from research



H.M.S. "Dublin", Wednesday 4th day of April, 1923.

From , To or At Cape Town

Wind | Direction | Force | Weather | Temperature | Position | Latitude | Longitude | REMARKS

Draught of water before coaling

Fwd 14' 6"

Aft 17' 2"

After coaling

Fwd 17' 4½"

Aft 18' 00"

Hands preparing for coaling

0730 Commenced coaling with native labour

0845 Training classes & RNVR's to instruction

0930 Landing Marines & Seamen Landing Parties in marching order

11.30 Marines detachment returned

12.00 Noon

FUEL Coal Tons

Coaling ship native labour

1315 Training classes & RNVR's to instruction

1415 Commenced cleaning ship

1445 Landing Party returned

Leave to the watch 1530 to 0700

1600

Increased use and economic benefit

The case of NASA Landsat satellite imagery of the Earth’s surface

Up to 2008

- Sold through the US Geological Survey for US$600 per scene
- Sales of 19,000 scenes per year
- Annual revenue of $11.4 million

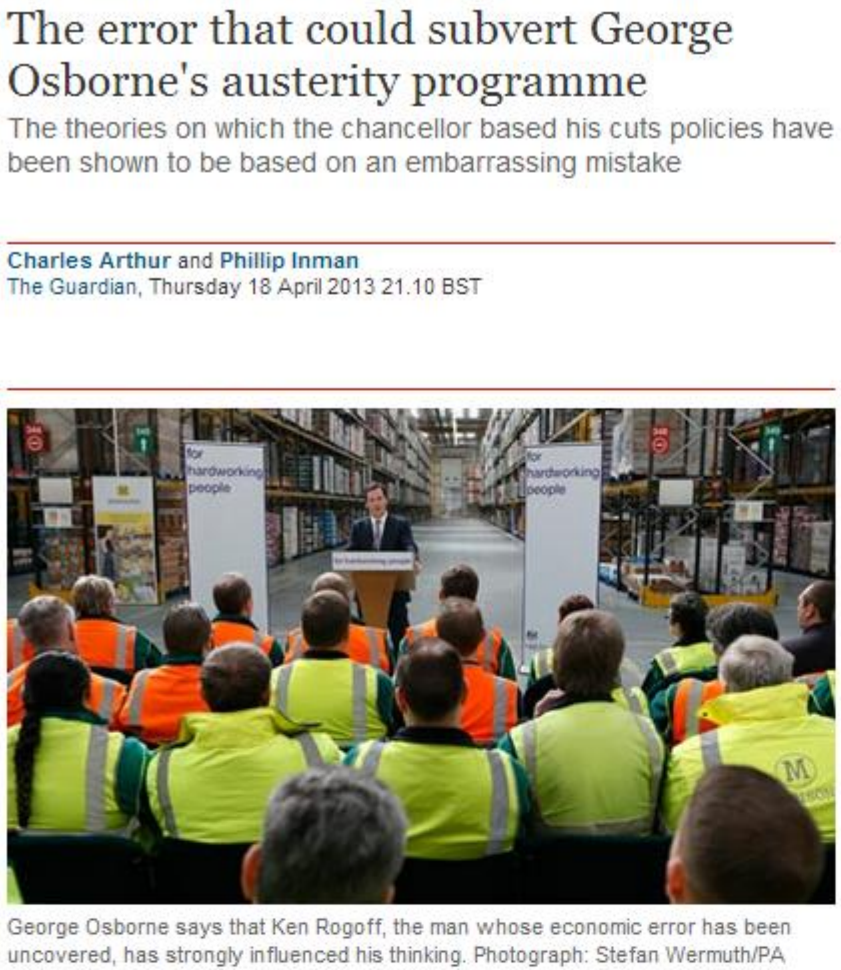
Since 2009

- Freely available over the internet.
- Google Earth now uses the images.
- Transmission of 2,100,000 scenes per year.
- Estimated to have created value for the environmental management industry of $935 million, with direct benefit of more than $100 million per year to the US economy.
- Has stimulated the development of applications from a large number of companies worldwide.
- http://earthobservatory.nasa.gov/IOTD/view.php?id=83394&src=ve

Validation of results

“It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that have arguably led to riots, poverty and lost jobs.”



The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

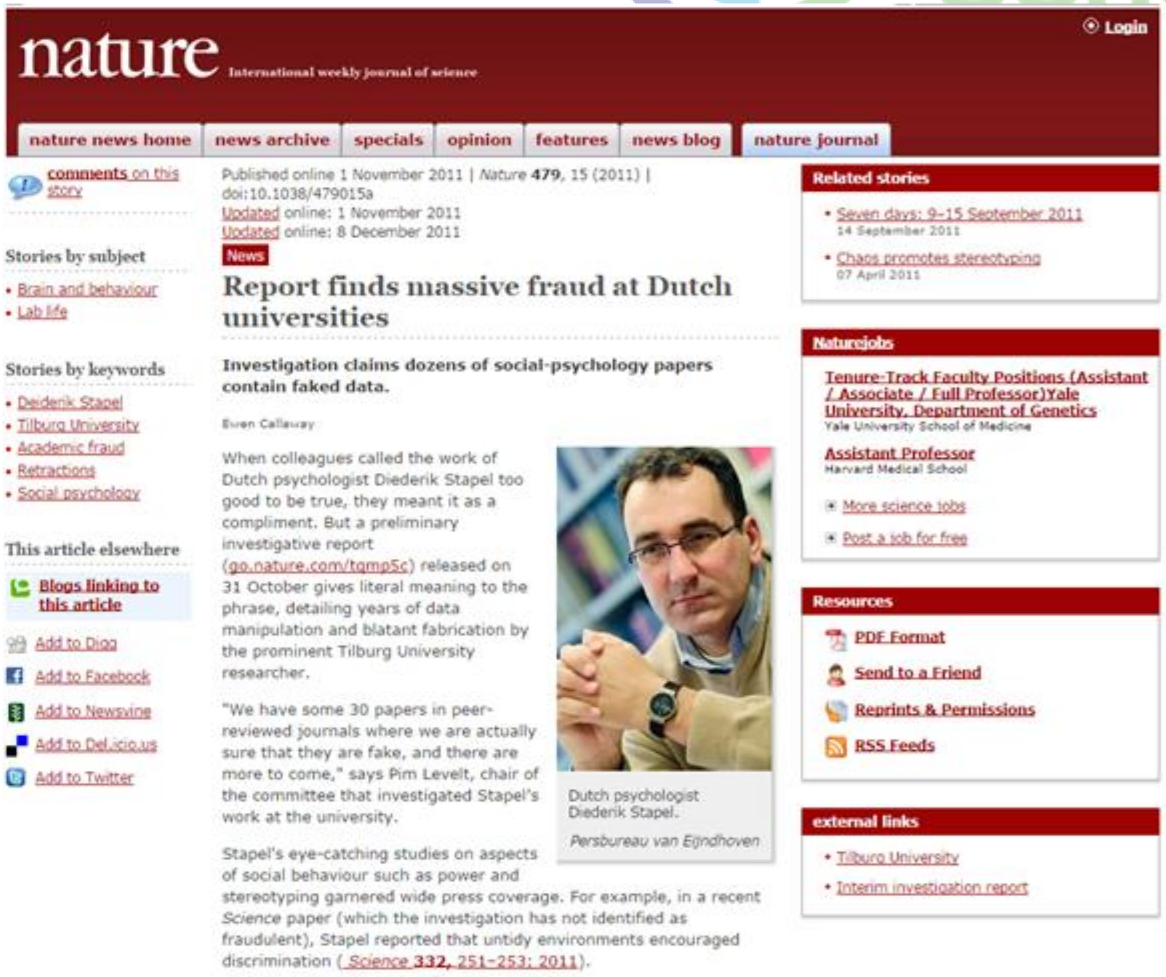
Charles Arthur and Phillip Inman
The Guardian, Thursday 18 April 2013 21.10 BST

George Osborne says that Ken Rogoff, the man whose economic error has been uncovered, has strongly influenced his thinking. Photograph: Stefan Wermuth/PA

www.guardian.co.uk/politics/2013/apr/18/uncovered-error-george-osborne-austerity

Cut down on academic fraud

Stapel – 55 publications – “fictitious data”



nature International weekly journal of science

Login

nature news home | news archive | specials | opinion | features | news blog | nature journal

comments on this story

Stories by subject

- Brain and behaviour
- Lab life

Stories by keywords

- Deiderik Stapel
- Tilburg University
- Academic fraud
- Retractions
- Social psychology

This article elsewhere

Blogs linking to this article

Add to Digg

Add to Facebook

Add to Newsvine

Add to Del.icio.us

Add to Twitter

Published online 1 November 2011 | *Nature* **479**, 15 (2011) | doi:10.1038/479015a

Updated online: 1 November 2011

Updated online: 8 December 2011

News

Report finds massive fraud at Dutch universities

Investigation claims dozens of social-psychology papers contain faked data.

Ewen Callaway

When colleagues called the work of Dutch psychologist Diederik Stapel too good to be true, they meant it as a compliment. But a preliminary investigative report (go.nature.com/tqmp5c) released on 31 October gives literal meaning to the phrase, detailing years of data manipulation and blatant fabrication by the prominent Tilburg University researcher.

"We have some 30 papers in peer-reviewed journals where we are actually sure that they are fake, and there are more to come," says Pim Levelt, chair of the committee that investigated Stapel's work at the university.

Dutch psychologist Diederik Stapel.

Persbureau van Eijndhoven

Stapel's eye-catching studies on aspects of social behaviour such as power and stereotyping garnered wide press coverage. For example, in a recent *Science* paper (which the investigation has not identified as fraudulent), Stapel reported that untidy environments encouraged discrimination (*Science* **332,** 251–253; 2011).

Related stories

- Seven days: 9–15 September 2011
  14 September 2011
- Chaos promotes stereotyping
  07 April 2011

Naturejobs

Tenure-Track Faculty Positions (Assistant / Associate / Full Professor)Yale University, Department of Genetics
Yale University School of Medicine

Assistant Professor
Harvard Medical School

More science jobs

Post a job for free

Resources

PDF Format

Send to a Friend

Reprints & Permissions

RSS Feeds

external links

- Tilburg University
- Interim investigation report

www.nature.com/news/2011/111101/full/479015a.html

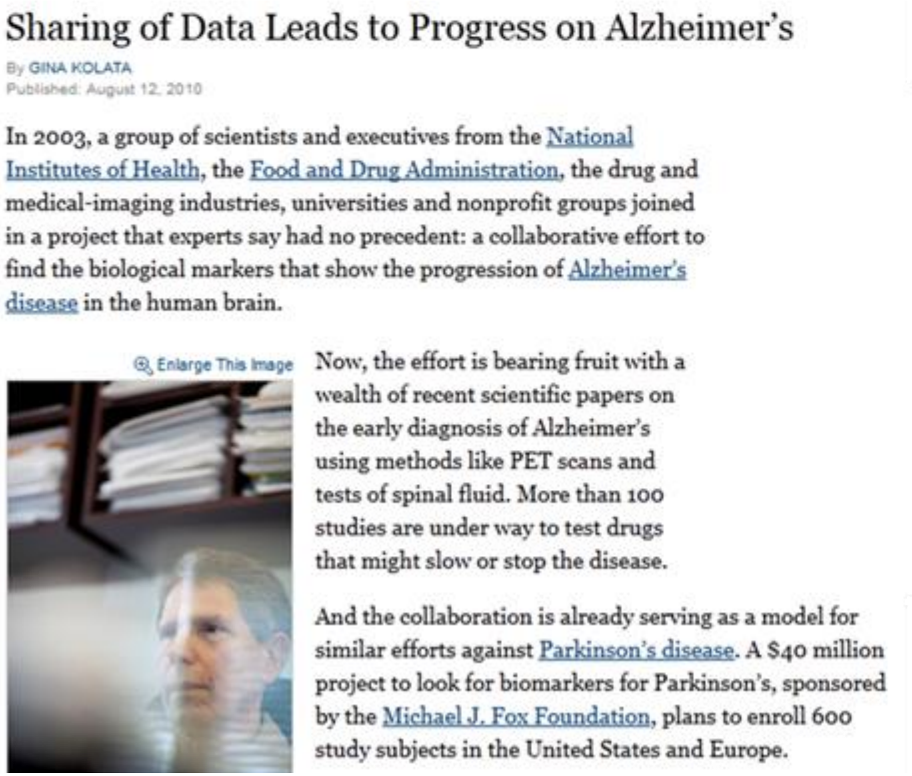
Sharing leads to breakthroughs!

...and increases the speed of discovery

“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

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



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.

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.

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

- License it so others can reuse:

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

Who has heard of this before…?

Findable

- Metadata
- PIDs
- Repositories

Accessible

- Metadata
- Open file formats and software

Interoperable

- Metadata
- Ontologies
- Repositories

Reusable

- Metadata
- Licences

European perspective...

European Commission

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

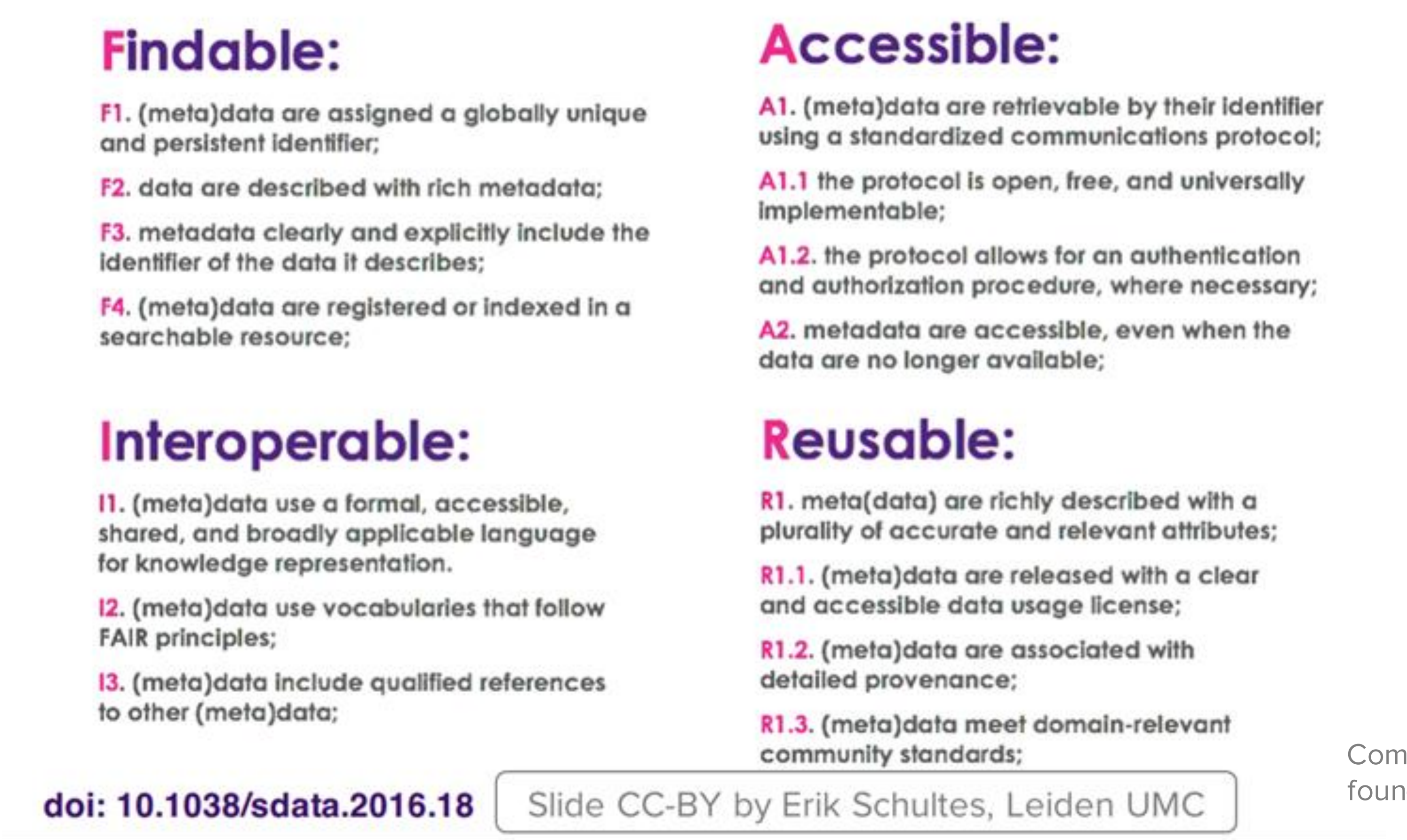
TURNING FAIR INTO REALITY

2018

Research and Innovation

eosc

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

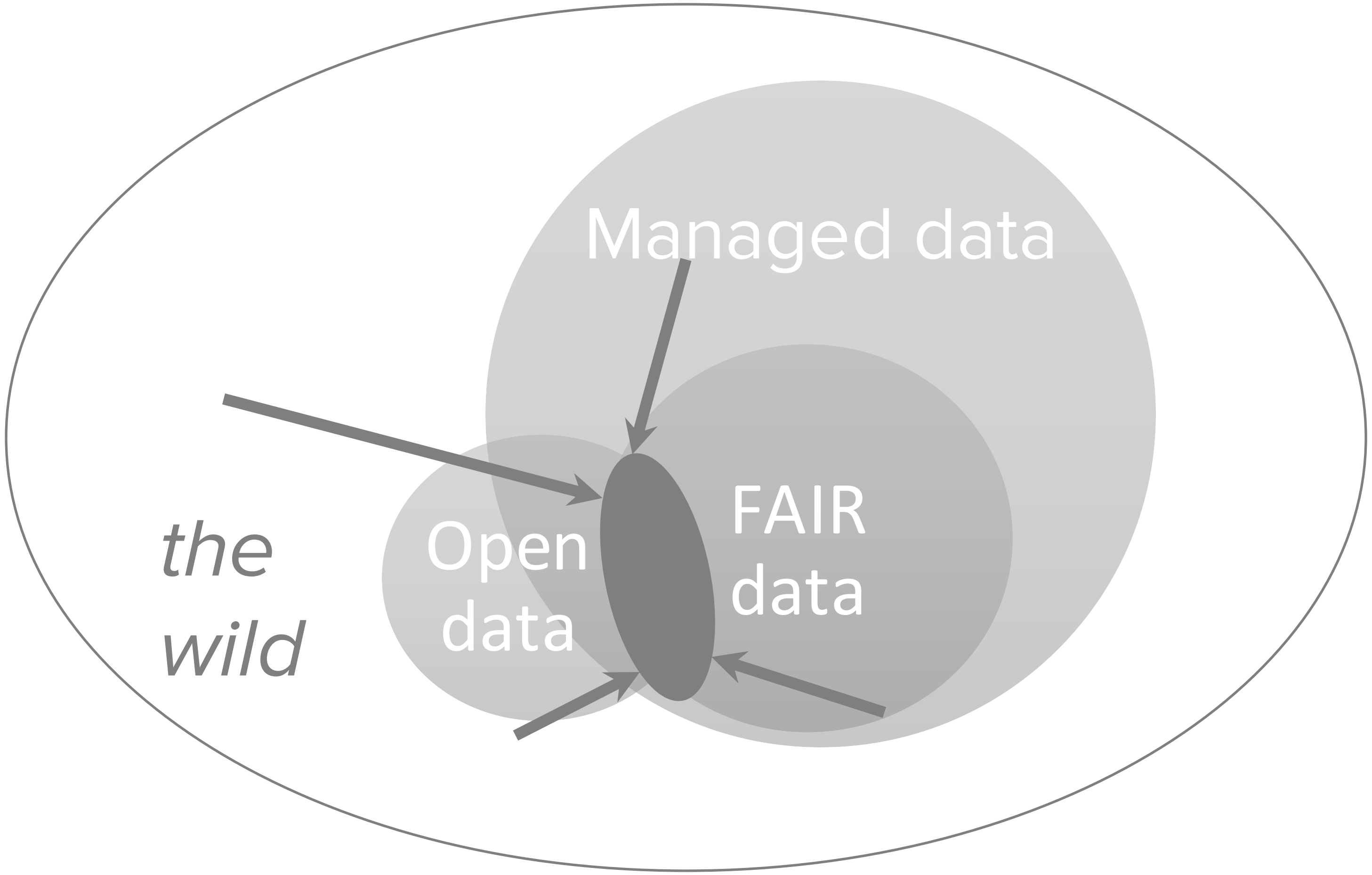
Slide CC-BY by Erik Schultes, Leiden UMC

Comprehensive descriptions can be found at https://www.go-fair.org/fair-principles/

Common misconceptions

- FAIR data does not have to be open.
- The principles do not specify particular technologies or implementations e.g. semantic web.
- FAIR is not a standard to be followed or strict criteria – it's a spectrum/continuum.
- It doesn't only apply to the life sciences.

Increasing that which is FAIR & open



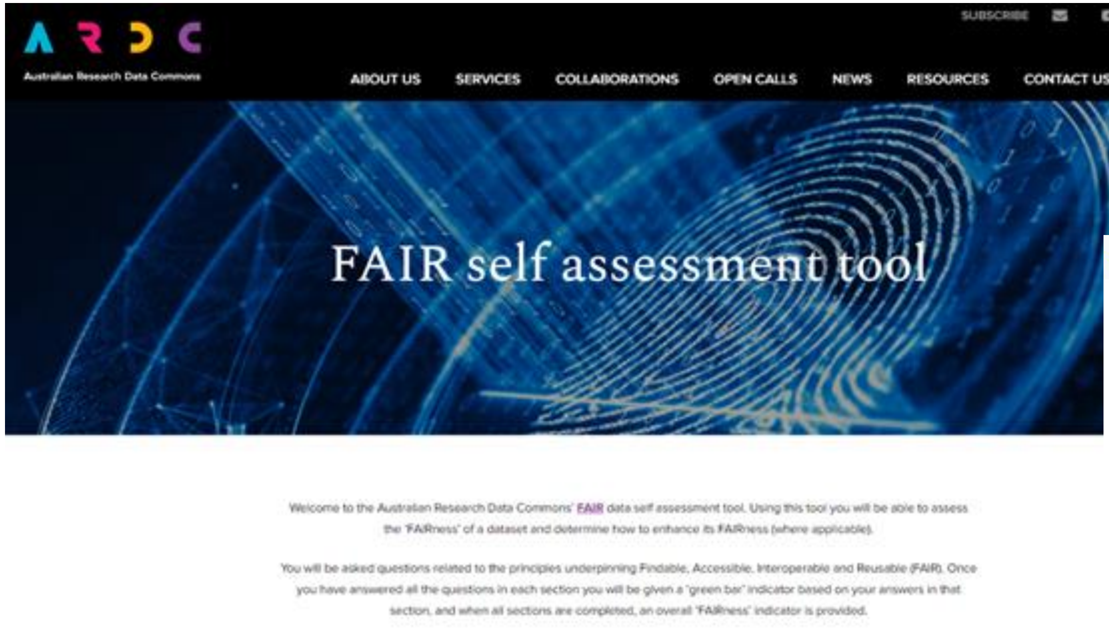
Adapted from DCC.

FAIR ≠ Open

as open as possible, as closed as necessary

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

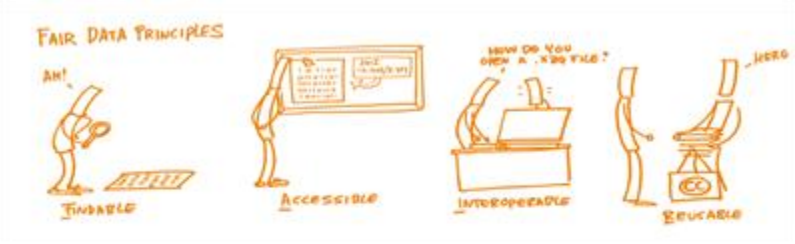
Check how FAIR is your data

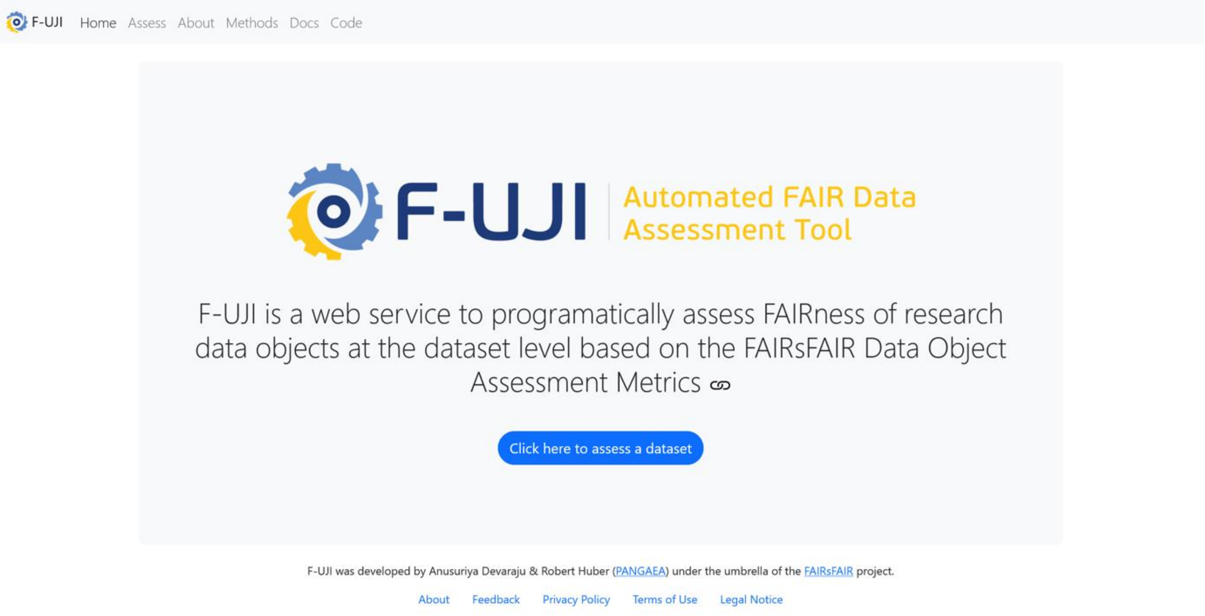


FAIR self assessment tool

Welcome to the Australian Research Data Commons' FAIR data self assessment tool. Using this tool you will be able to assess the 'FAIRness' of a dataset and determine how to enhance its FAIRness (where applicable).

You will be asked questions related to the principles underpinning Findable, Accessible, Interoperable and Reusable (FAIR). Once you have answered all the questions in each section you will be given a "green bar" indicator based on your answers in that section, and when all sections are completed, an overall 'FAIRness' indicator is provided.



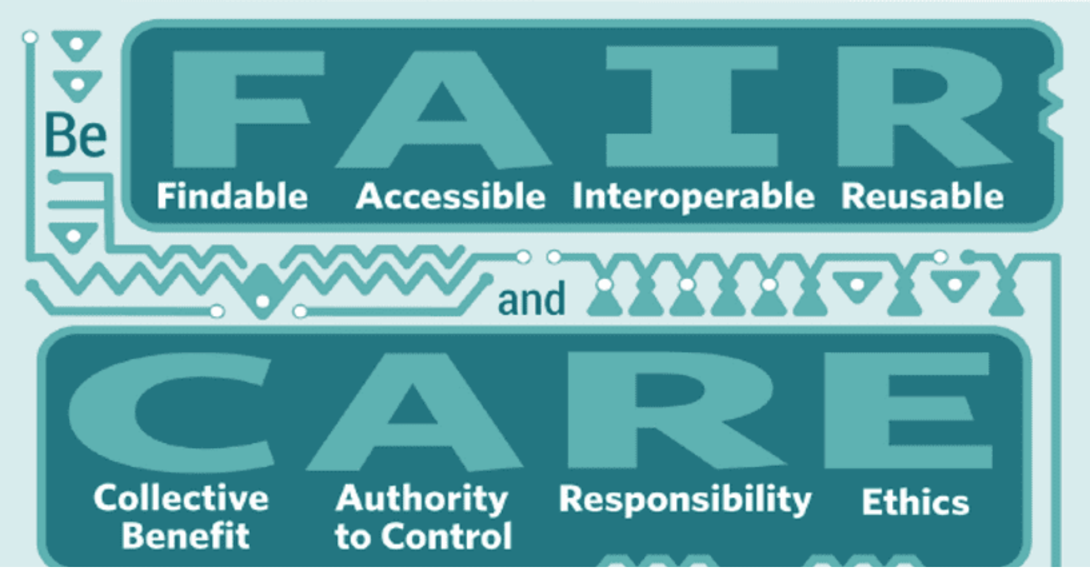

F-UJI
Home
Assess
About
Methods
Docs
Code
F-UJI
Automated FAIR Data Assessment Tool
Click here to assess a dataset

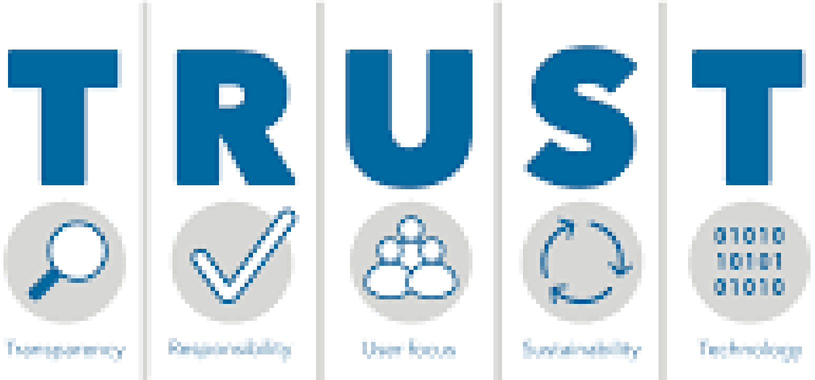
F-UJI is a web service to programatically assess FAIRness of research data objects at the dataset level based on the FAIRsFAIR Data Object Assessment Metrics

F-UJI was developed by Anusuriya Devaraju & Robert Huber (PANGAEA) under the umbrella of the FAIRsFAIR project.

https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/
https://www.f-uji.net/

FAIR isn’t the only consideration…





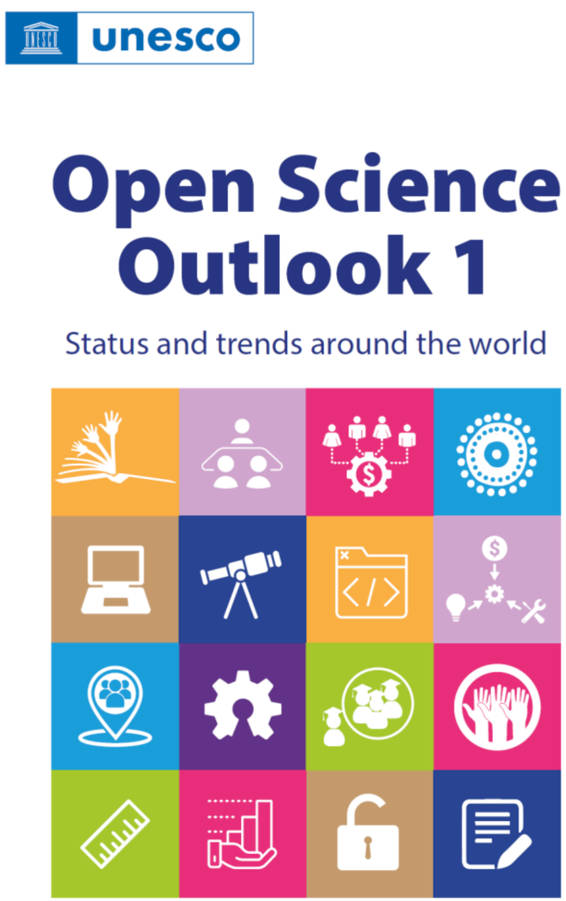
New(ish) frontiers…

- Collaboration
- Reproducibility
- Transparency
- Trust

“Open science practices are on the rise but access to, participation in and sharing of the benefits from open science are uneven across the world.”

unesco

UNESCO Recommendation on Open Science



unesco

Open Science Outlook 1

Status and trends around the world

New(ish) frontiers...

- **Support is not making its way to those who need it**
  Almost three-quarters of respondents had never received support with making their data openly available.
- **One size does not fit all**
  Variations in responses from different subject expertise and geographies highlight a need for a more nuanced approach to research data management support globally.
- **Challenging stereotypes**
  Are later career academics really opposed to progress? The results of the 2023 survey indicate that career stage is not a significant factor in open data awareness or support levels.
- **Credit is an ongoing issue**
  For eight years running, our survey has revealed a recurring concern among researchers: the perception that they don't receive sufficient recognition for openly sharing their data.
- **AI awareness hasn't translated to action**
  For the first time, this year we asked survey respondents to indicate if they were using ChatGPT or similar AI tools for data collection, processing and metadata creation.

A Digital Science Report

The State
Open Dat

The longest-running longitudinal survey

With opening remarks from Springer Nature's CPO, Hars
Authors Mark Hahnel, Graham Smith, Niki Scaplehorn,

DIGITAL science SPRINGE

The State of Open Data 2024: Special Report December 2024

Bridging policy and practice in data sharing

An investigation into what is driving successful data sharing in repositories

Mark Hahnel, Digital Science, Graham Smith, Springer Nature, Ann Campbell, Digital Science

DIGITAL science SPRINGER NATURE figshare

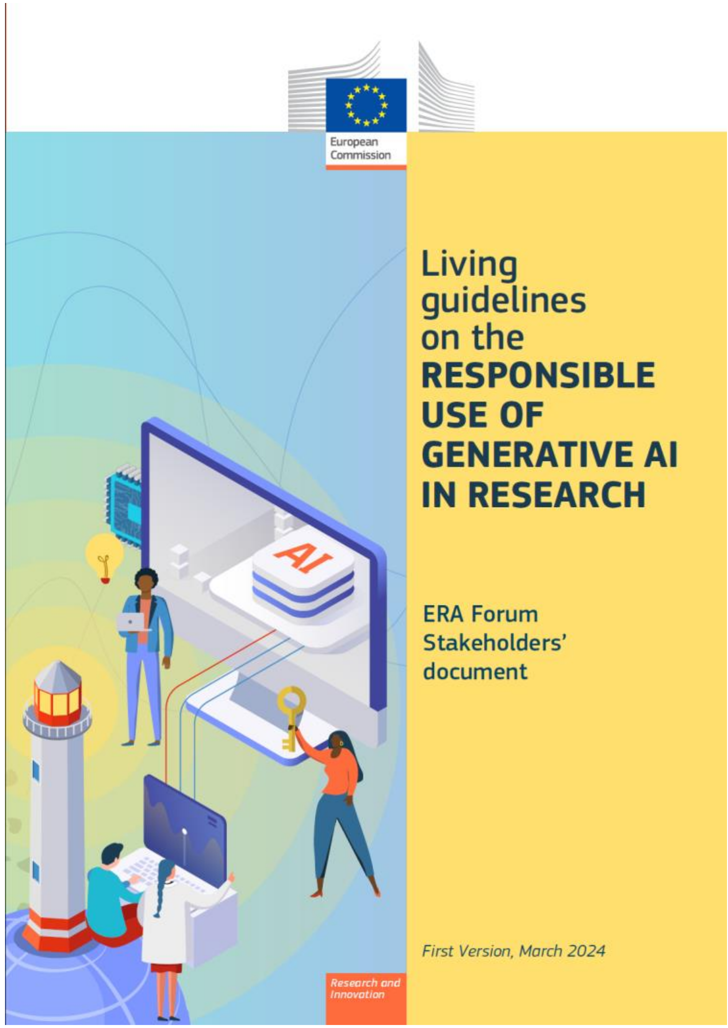
April 2024

ens:
ces in
pen data

R NATURE figshare

Data Schools

New(ish) frontiers...

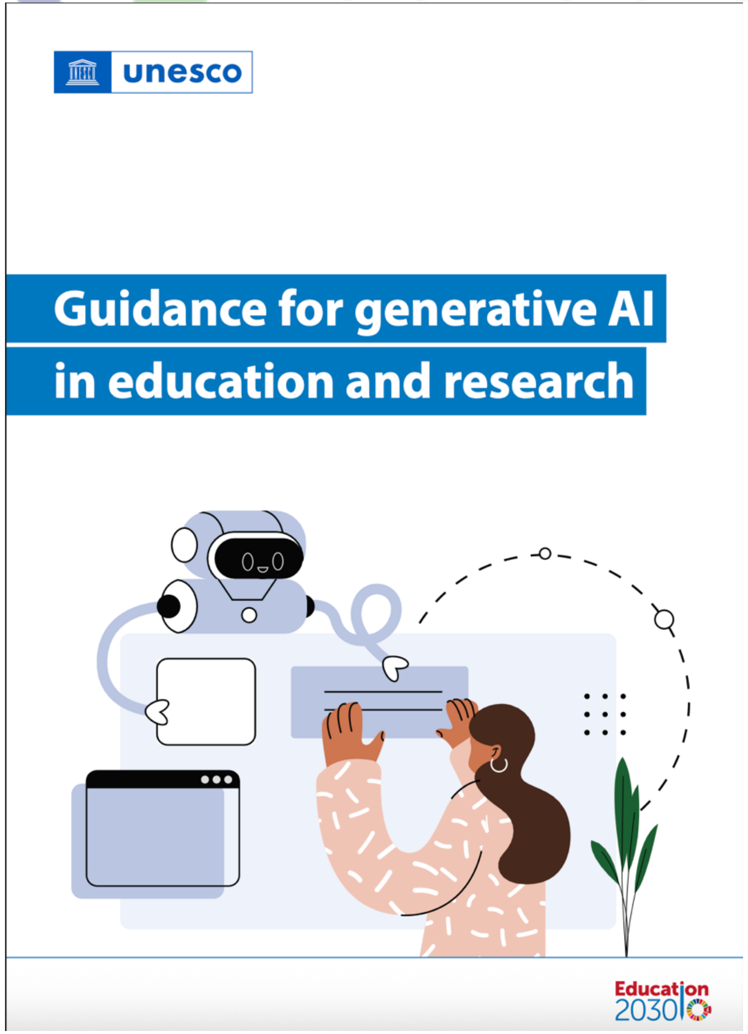

European Commission

Living guidelines on the RESPONSIBLE USE OF GENERATIVE AI IN RESEARCH

ERA Forum Stakeholders' document

First Version, March 2024

Research and Innovation


unesco

Guidance for generative AI in education and research

Education 2030

FOSTER Open Science

What is Open Science?

Best Practice in Open Research

Open Access Publishing

Open Peer Review

Sharing Preprints

Data Protection & Ethics

Open Source Software & Workflows

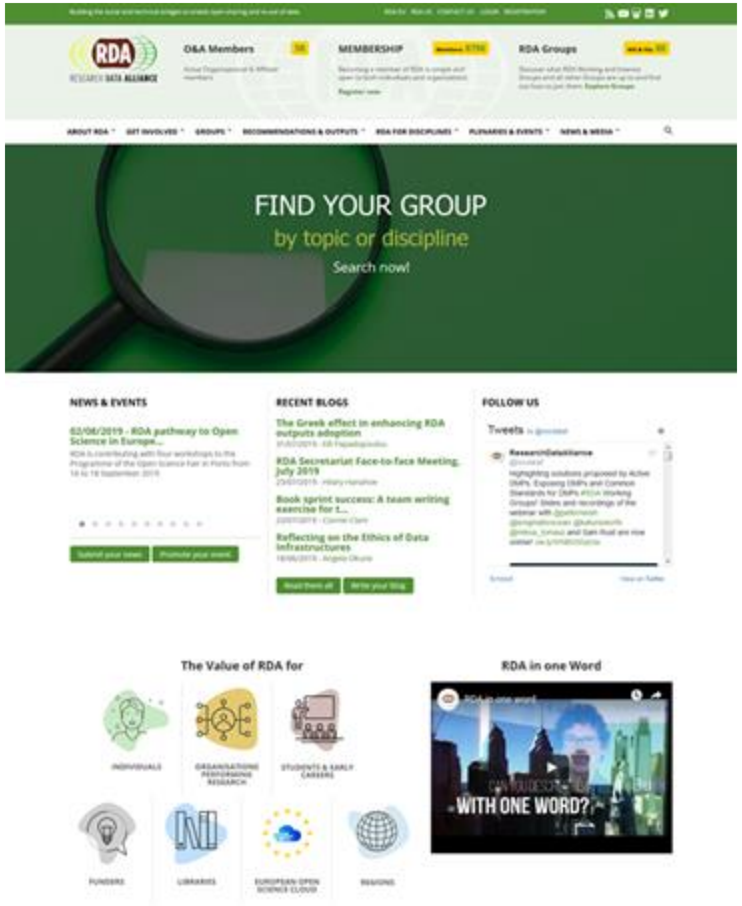
Managing & Sharing Research Data

Open Science & Innovation

Open Licensing

https://www.fosteropenscience.eu/toolkit

Research Data Alliance



O&A Members

MEMBERSHIP

RDA Groups

ABOUT RDA ˅ GET INVOLVED ˅ GROUPS ˅ RECOMMENDATIONS & OUTPUTS ˅ RDA FOR DISCIPLINES ˅ PLENARIES & EVENTS ˅ NEWS & MEDIA ˅

FIND YOUR GROUP

by topic or discipline

Search now!

NEWS & EVENTS

02/08/2019 - RDA pathway to Open Science in Europe...

RECENT BLOGS

The Greek effect in enhancing RDA outputs adoption

RDA Secretariat Face-to-face Meeting, July 2019

Book sprint success: A team writing exercise for t...

Reflecting on the Ethics of Data Infrastructures

FOLLOW US

Tweets

ResearchDataAlliance

The Value of RDA for

INDIVIDUALS

ORGANISATIONS PERFORMING RESEARCH

STUDENTS & EARLY CAREERS

FUNDERS

LIBRARIES

EUROPEAN OPEN SCIENCE CLOUD

REGIONS

RDA in one Word

https://www.rd-alliance.org

Data Schools

Data Management Plans

Bringing together what you’ve learnt

- Make informed decisions to anticipate and avoid problems.
- Avoid duplication, data loss and security breaches.
- Develop procedures early on for consistency.
- Ensure data are accurate, complete, reliable and secure.
- Save time and effort to make your life easier!
- Useful both to researchers and institutions

Making plans

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

[illegible]

CAREERS

DATA MANAGEMENT

For the record

Making project data freely available is vital for open science.

[illegible]

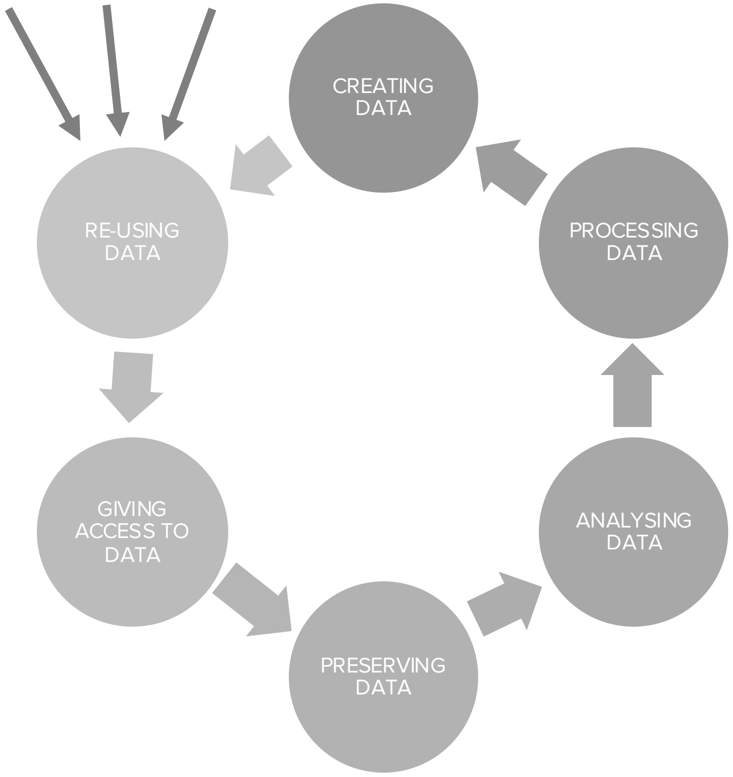
Schiermeier, Q. “Data management made simple” *Nature* **555**, 403-405 (2018).
https://www.nature.com/articles/d41586-018-03071-1
doi: 10.1038/d41586-018-03071-1

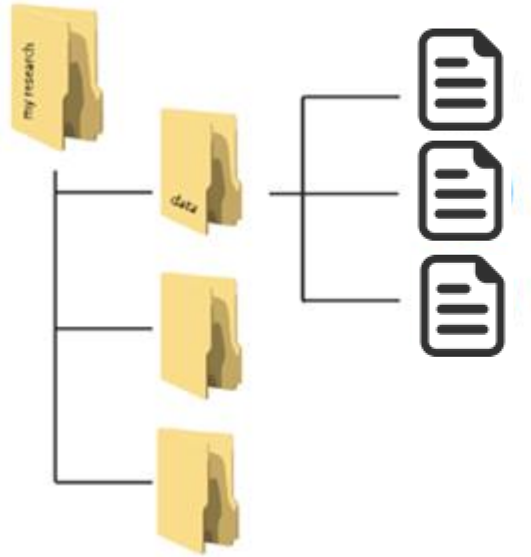
Common themes in DMPs

1. Description of data to be collected / created (i.e. content, type, format, volume...).
2. Standards/methodologies for data collection & management.
3. Ethics and Intellectual Property (highlight any restrictions on data sharing e.g. embargoes, confidentiality).
4. Plans for data sharing and access (i.e. how, when, to whom).
5. Strategy for long-term preservation.

Planning trick 1: think backwards

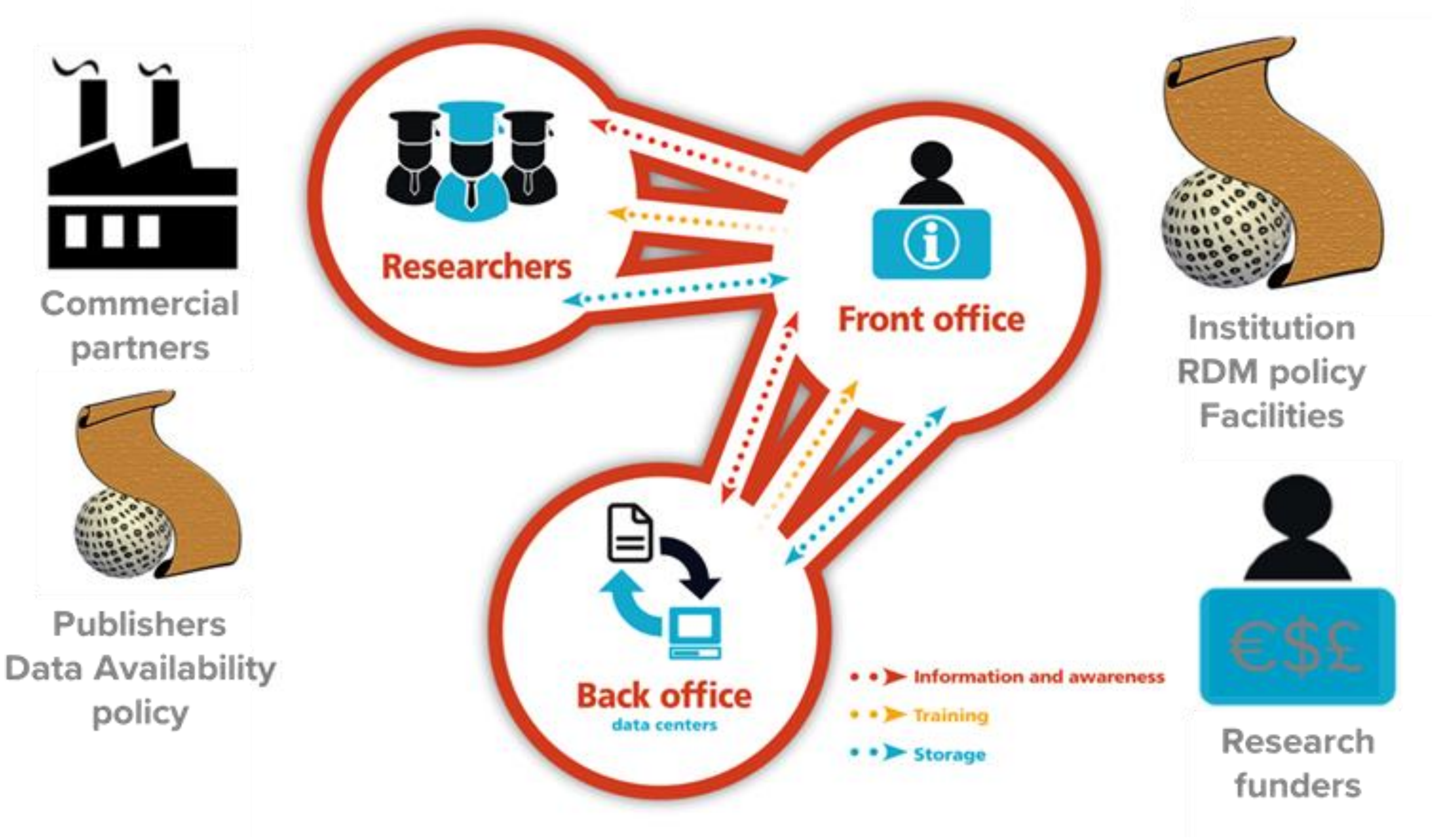
What data organisation would a re-user like?





Design how you will organise data in the project (folder structure, file naming convention, …)

Planning trick 2: include RDM stakeholders



www.openaire.eu/briefpaper-rdm-infonoads

Planning trick 3: ground your plan in reality

Base plans on available skills, support and good practice for the field – show it's feasible to implement.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

cc 0 ZERO

What makes a good DMP?

- Clear, detailed information that is relevant to the science:
  - adopting recognised standards.
  - practices in line with norms for that field.
  - use of support services e.g. university storage, subject repositories…
- Realistic approach that is feasible to implement.
- Evidence of consultation and seeking advice.
- Proper justification of restrictions and costs.
- **Have you taken time to reflect on what to do?**

Is the information specific enough?

“we will use suitable formats to ensure that our data can be preserved and sustained over the long term”

- Which standards? Name them!
- Show that you know which are suitable.
- Does your chosen repository have preferences?

Are decisions justified?

“data will be made available upon request to bona fide medieval historians”

- Why is it restricted?
- Could other communities not reuse the data?
- Will the research team be around to handle access requests in the future?

A better response...

“We will provide MP3 audio files for online dissemination. While this is not an open format, it is well-established and the most widely supported. High-resolution WAV files will be used for the archival master recordings.”

○ Be clear, specific and detailed.

○ Justify decisions.

Example plans

- Plans from several funders and disciplines via DCC www.dcc.ac.uk/resources/data-management-plans/guidance-examples

- Scientific DMPs submitted to the NSF (USA) provided by DataOne https://www.dataone.org/data-management-planning

- DMPs published in RIO journal http://riojournal.com/browse_user_collection_documents.php?collection_id=3&journal_id=17

- Share yours! - www.dcc.ac.uk/share-DMPs

DCC Checklist for a DMP

- The DCC assessed existing funder requirements, DMP templates and other best practice to see what should be included in plans. This was synthesised down into common themes and questions.
- 13 questions on what’s asked across the board.
- Prompts/pointers to help researchers get started.
- Guidance on how to answer.

DCC Checklist for a Data Management Plan

DCC

www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

Thank you!

Questions?

(Please get in touch!)