Introduction to Research Data Management and Open Research

S. Venkataraman, Training Officer, OpenAIRE

s.venkataraman@openaire.eu

31st July 2022, ICTP, Trieste

Agenda

	Day 1 (31st July)
14:00	Introduction to research data management (RDM)
14:45	Exercise: Practical session on RDM
15:30	Introduction to open science (research)
16:00	**Break**
16:30	Exercise and discussion: Open science
18:00	**End Day 1**
	Day 2 (1st August)
08:30	Introduction to DMPs
09:15	**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.

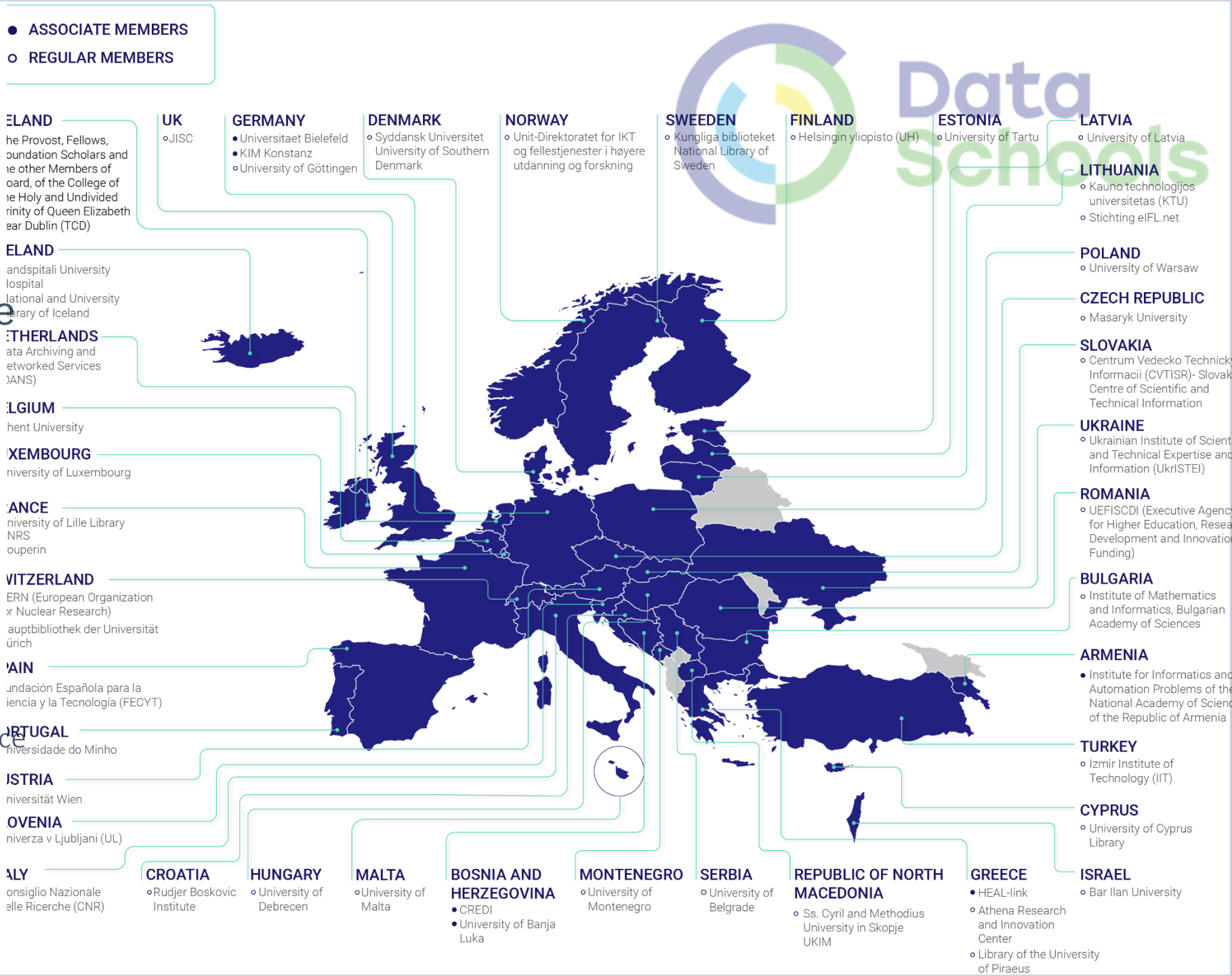
WHAT IS OPENAIRE

A European infrastructure on open scholarly communication

Mission: Shift scholarly communication towards openness and transparency and facilitate innovative ways to communicate and monitor research.

Non-profit organisation

- Established Oct 2018
- Headquarter Greece, virtual office
- 47 members
- From 34 countries

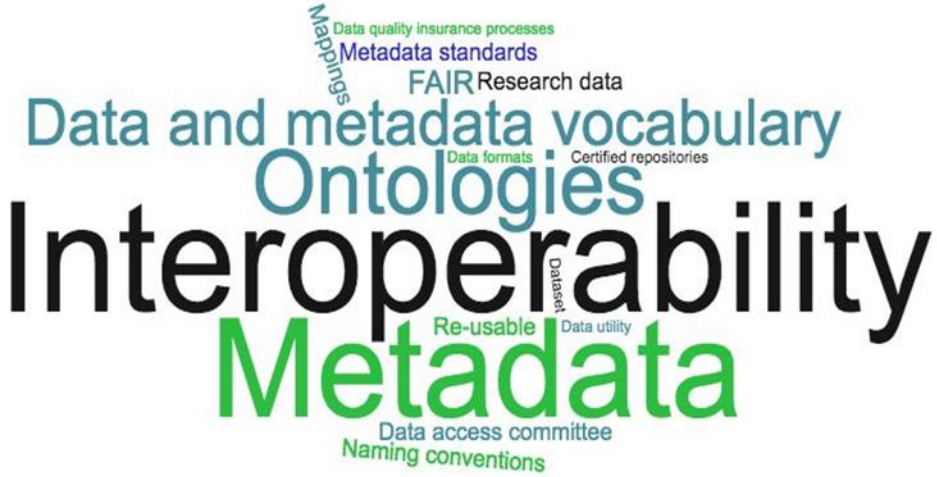


Language is a barrier…

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."



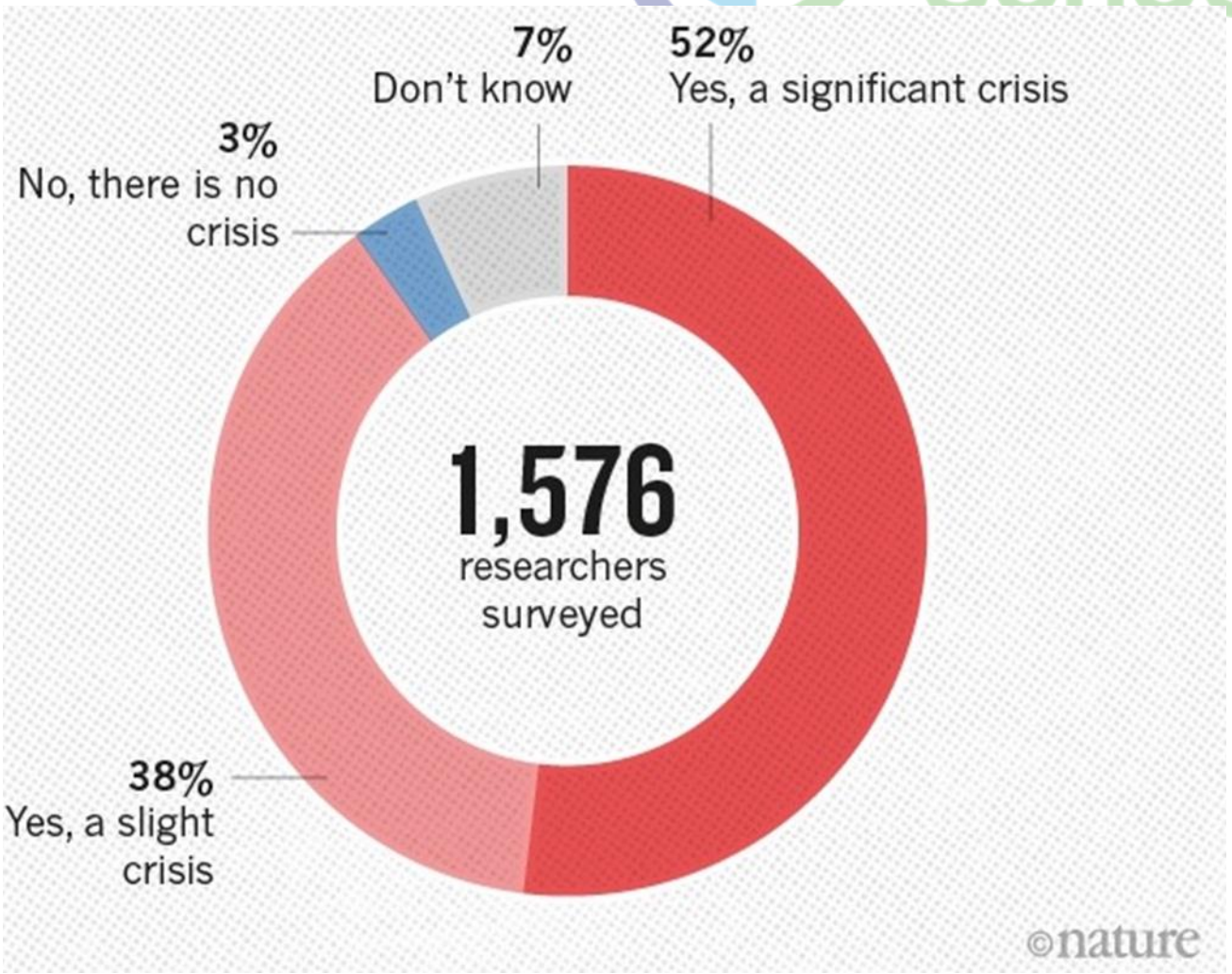
Grootveld et al. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans http://doi.org/10.5281/zenodo.1120245

Is there a reproducibility crisis?

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

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

doi:10.1038/533452a

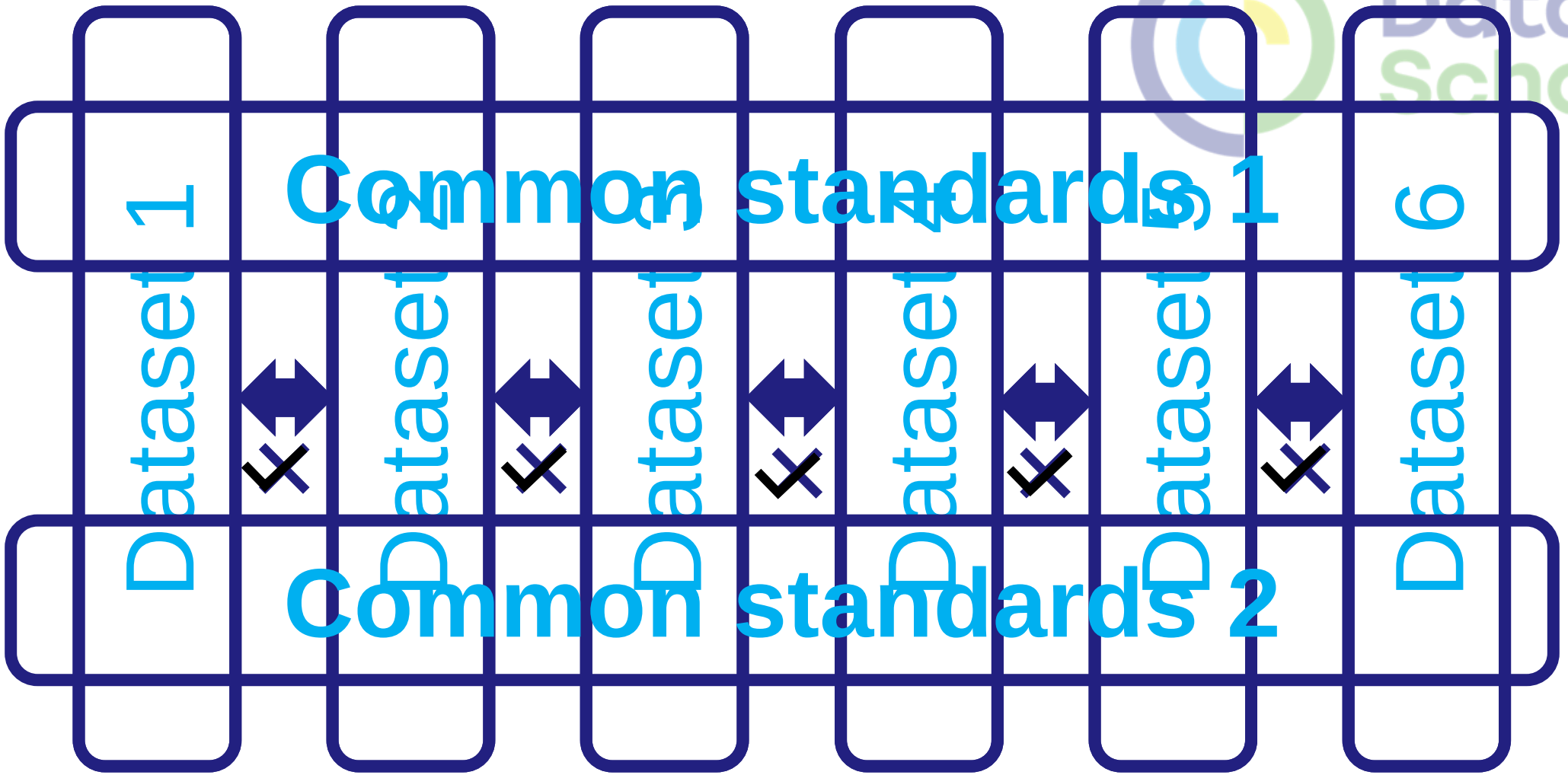


The wider context

Set of goals outlined by the United Nations

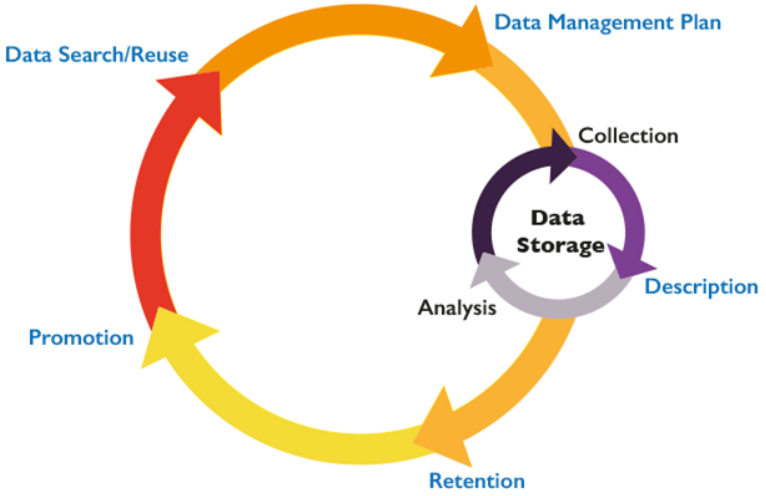
Data Schools

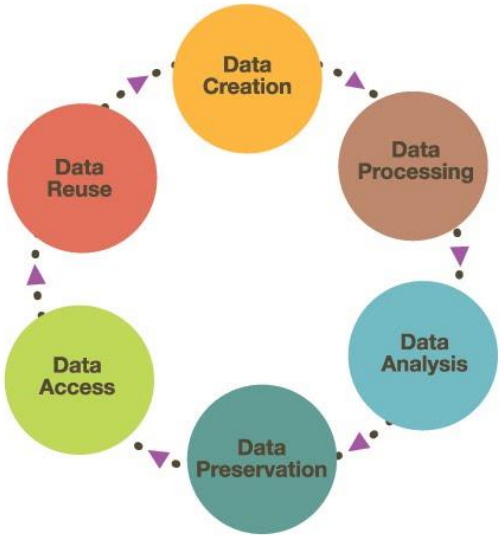
RDM & the Data Lifecycle

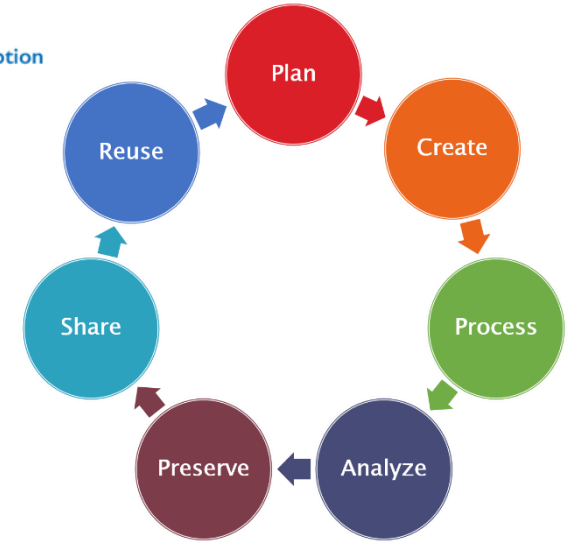

Common standards 1
Common standards 2
Dataset 1
Dataset 2
Dataset 3
Dataset 4
Dataset 5
Dataset 6

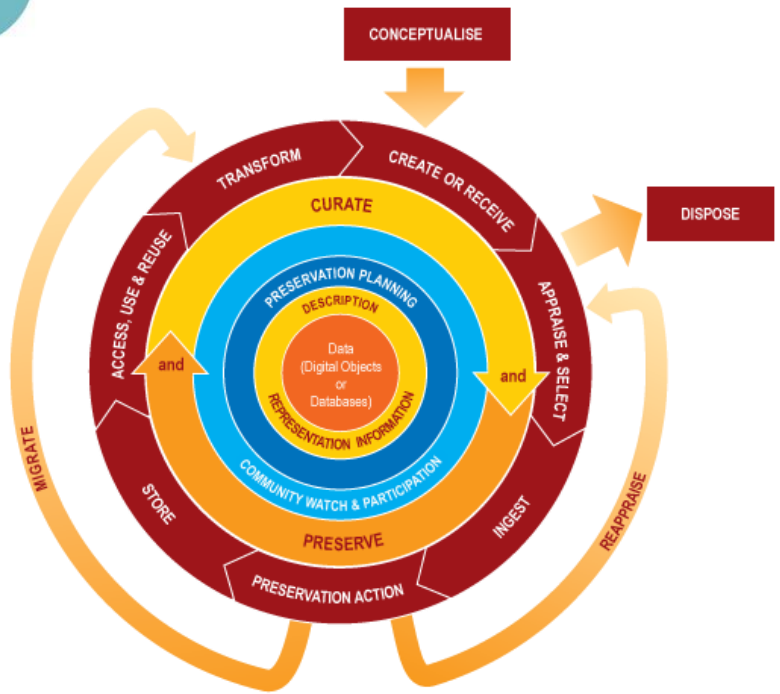
Data silos

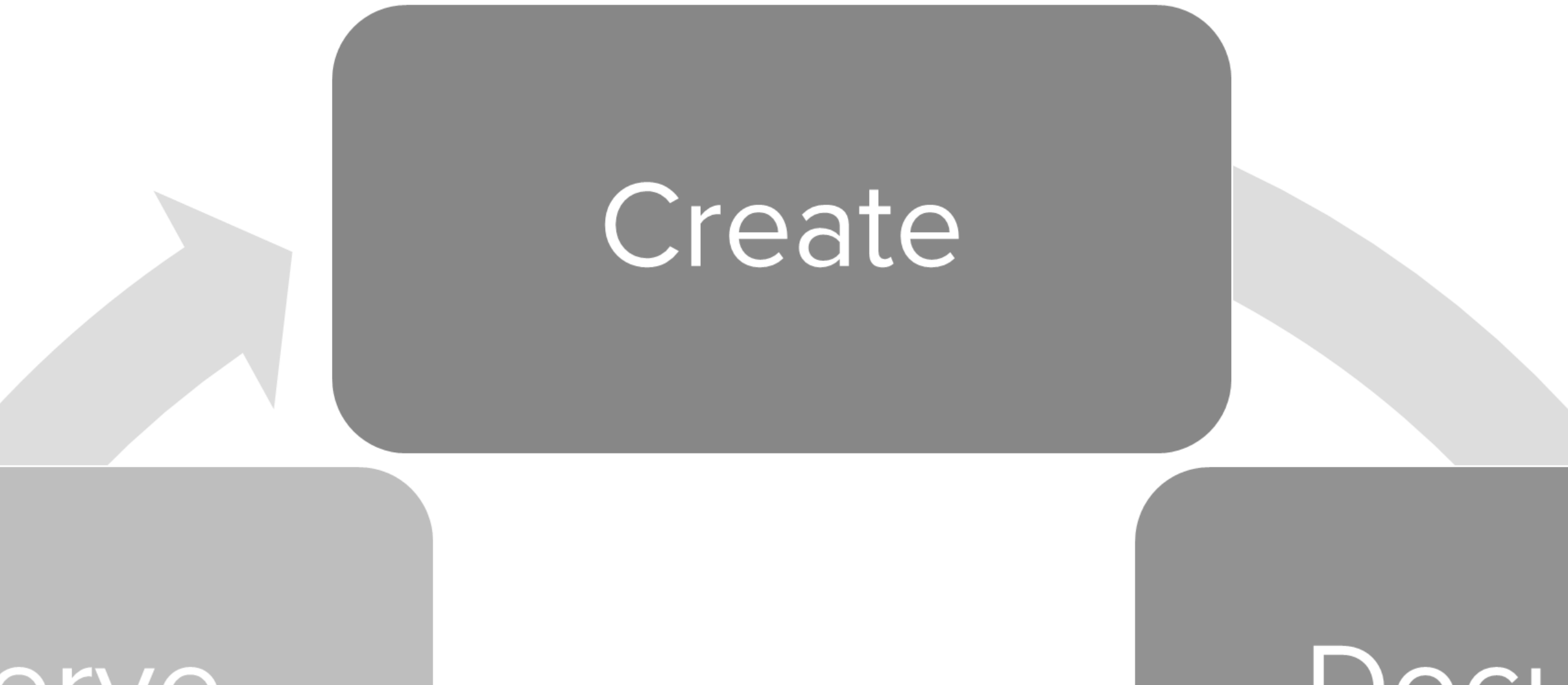
RDM lifecycles










Create

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

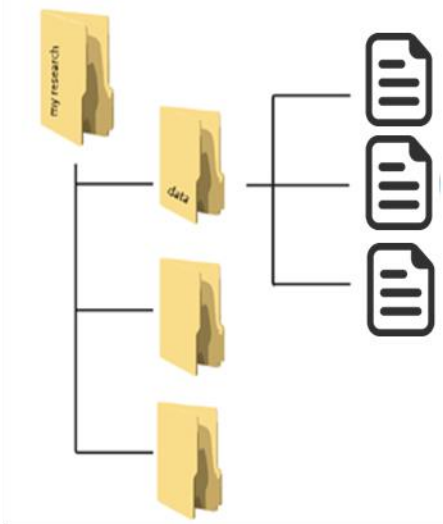
- Different formats are good for different things.
- *open*, *lossless* formats are more sustainable e.g. rtf, xml, tif, wav.
- proprietary and/or compressed formats are less preservable but are often in widespread use e.g. doc, jpg, mp3.
- One format for analysis then convert to a standard format.
-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

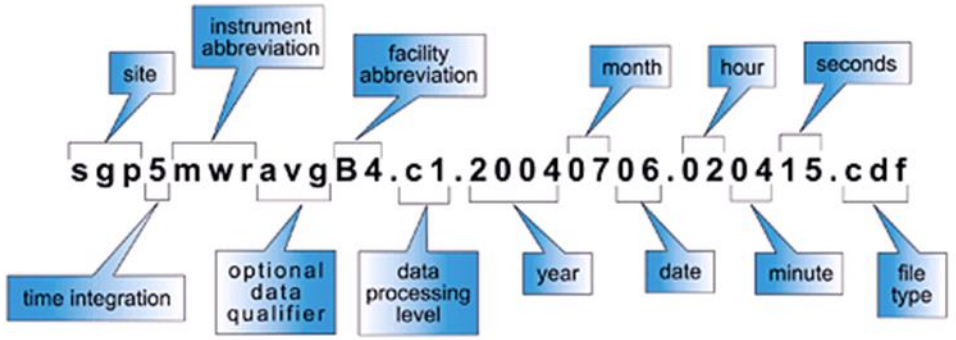
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

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.



An example netCDF data file name is depicted below:



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

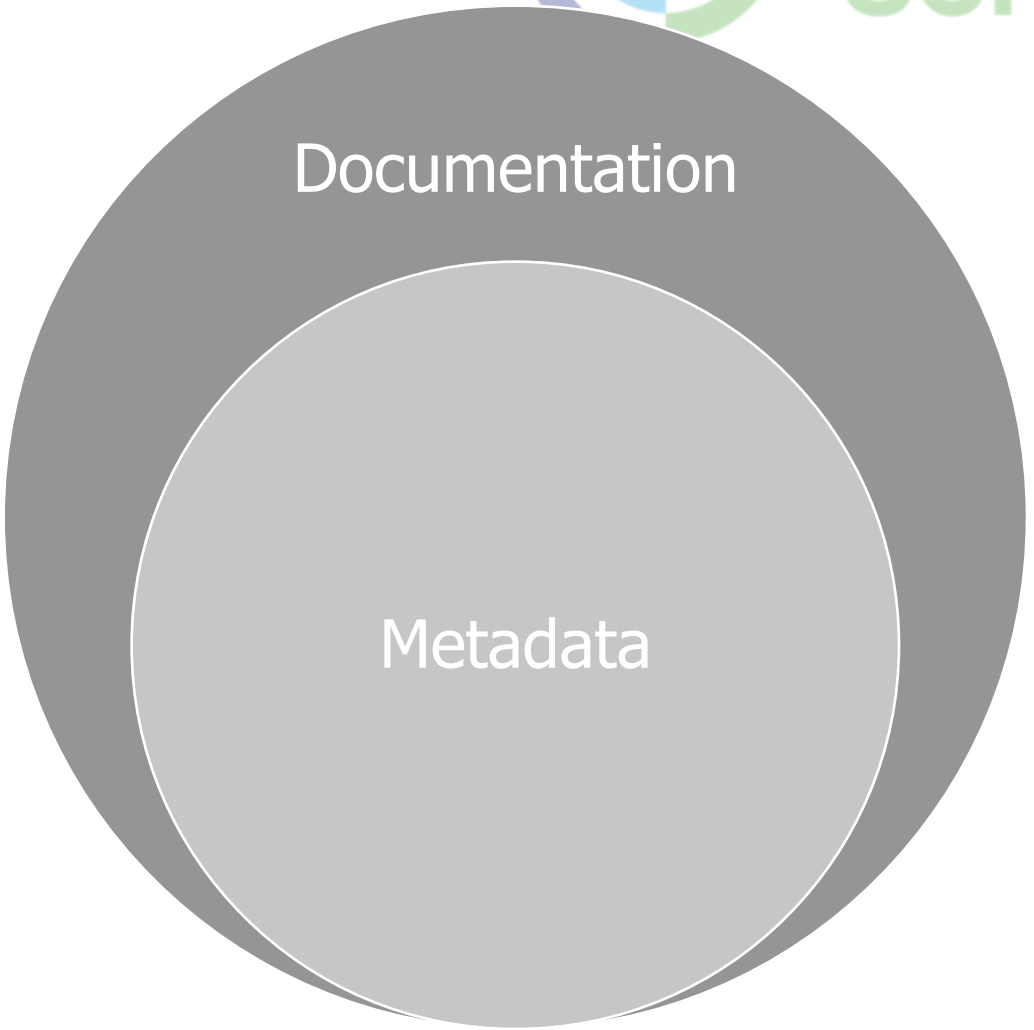
Documentation

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

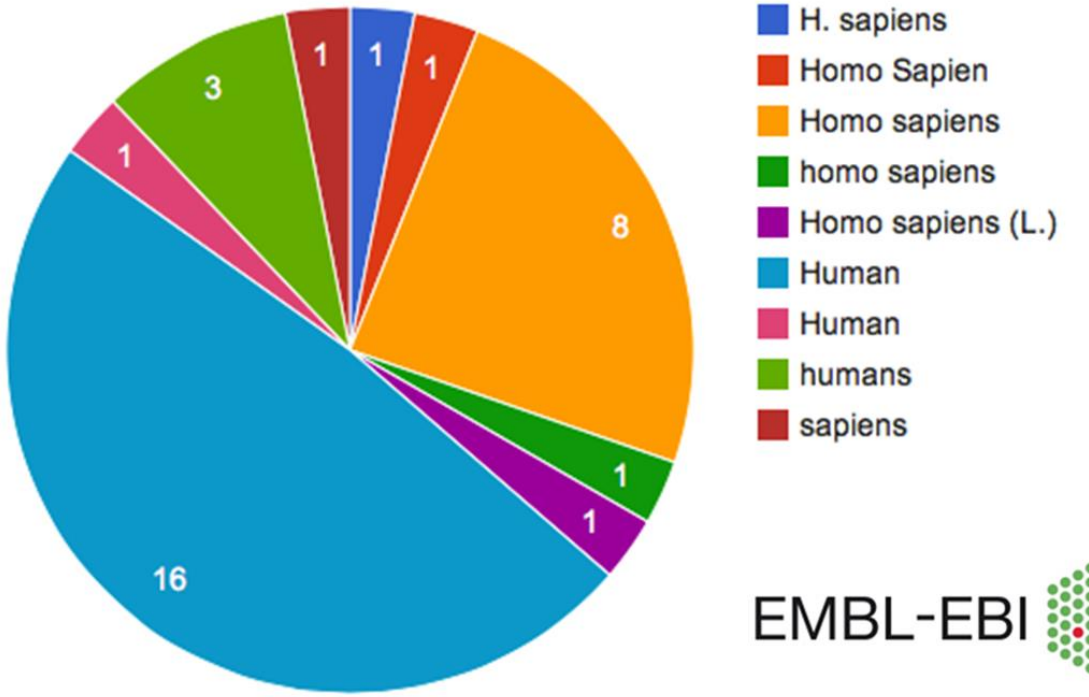
- Data Documentation Initiative (DDI) – social science
- Ecological Metadata Language (EML) - ecology
- Flexible Image Transport System (FITS) – astronomy

Search for standards in catalogues like:

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

Controlled vocabularies

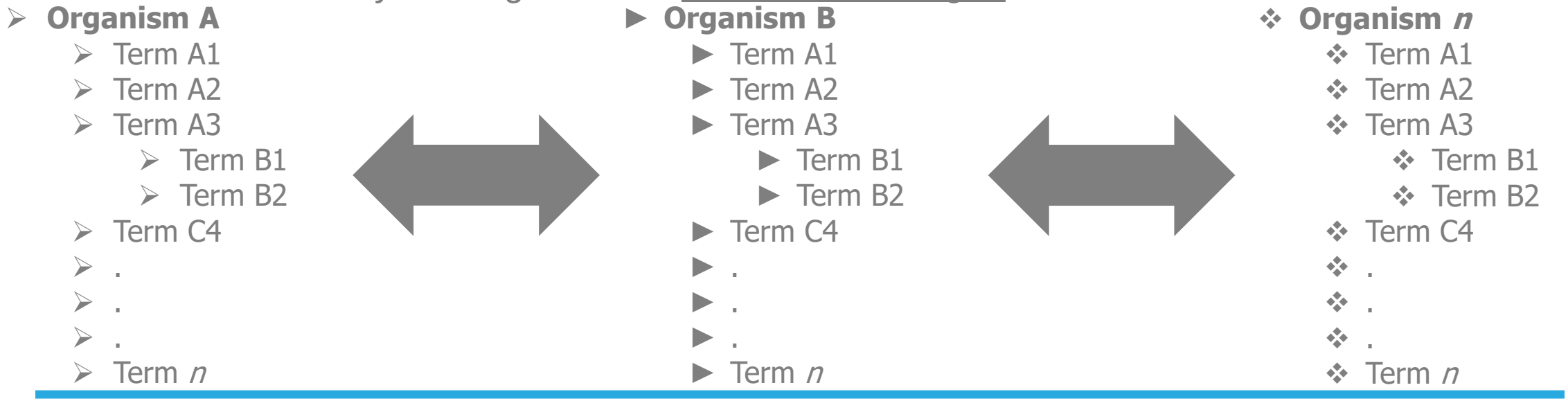
"MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in……"



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.



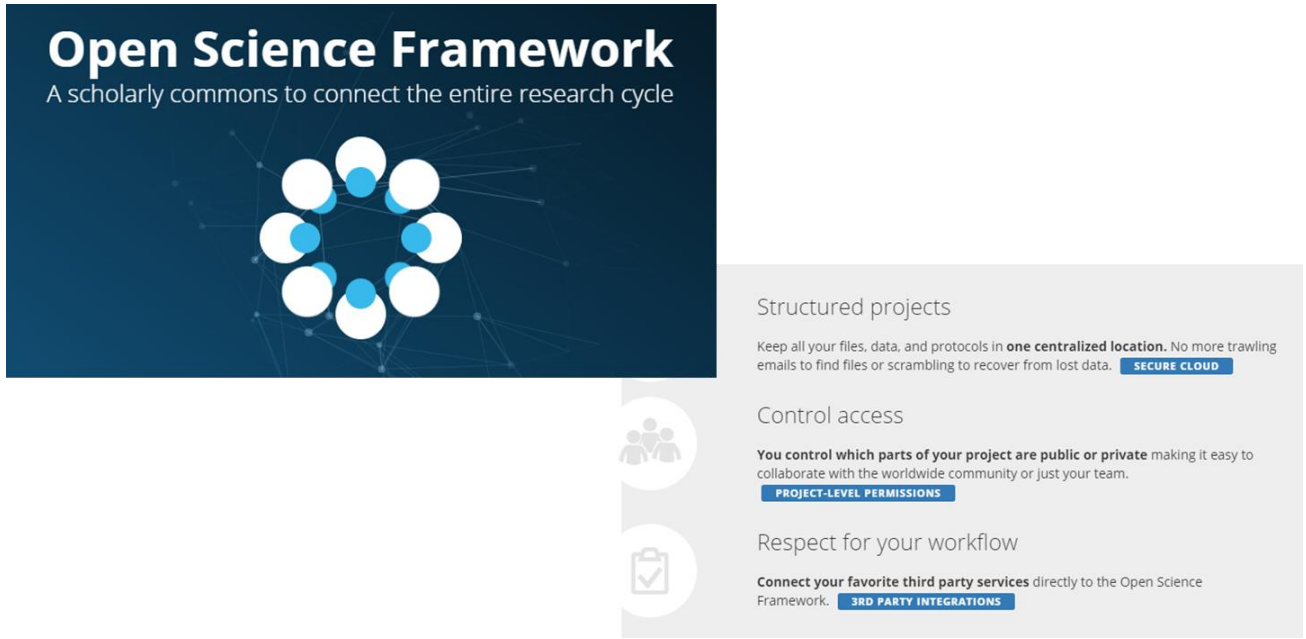
Where will you store the data?

- 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.



https://osf.io

Third-party tools for collaboration

Dropbox, Google Drive, OneDrive and other cloud services

- Commercial
- Who owns your data?

Dropbox

ownCloud

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

ownCloud

https://owncloud.org

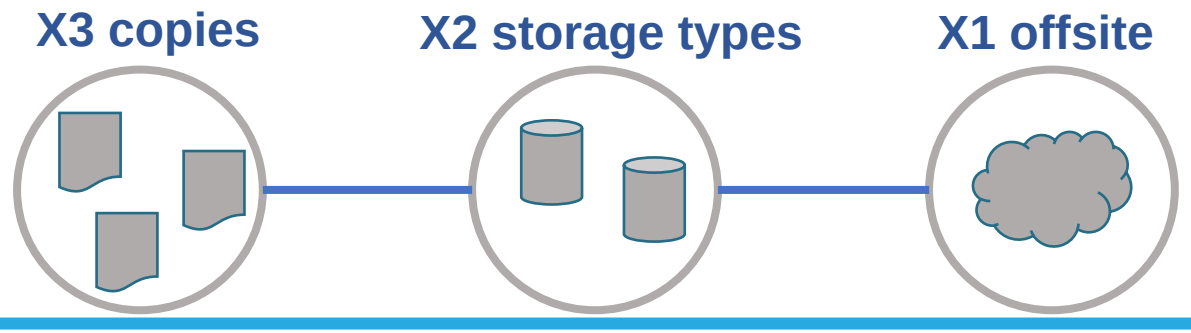
Backup and preservation – not the same thing!

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

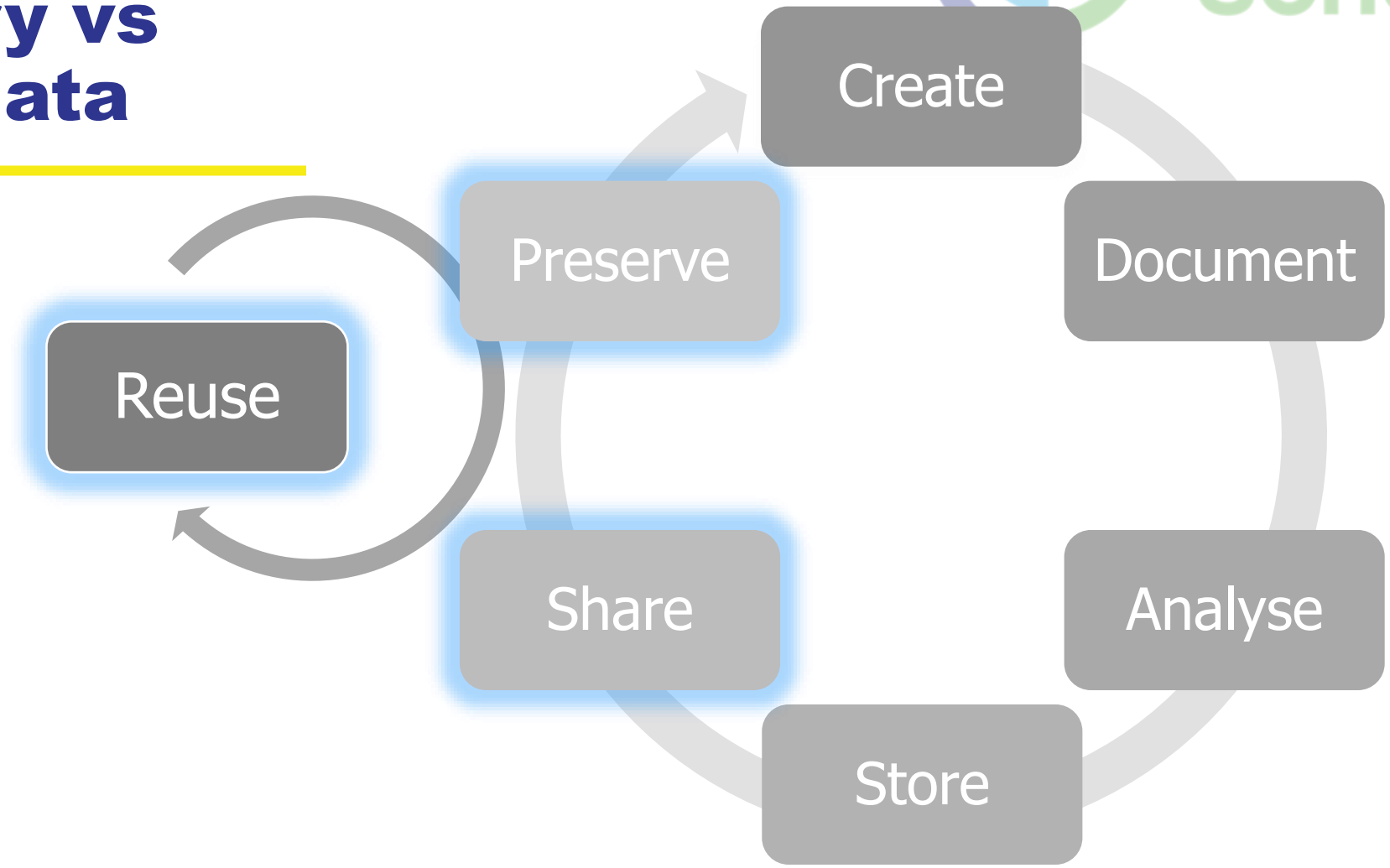
- Used to preserve data for historical reference or potentially during disasters.
- Archives are usually the final version, stored for long-term, and generally not copied over.
- Often performed at the end of a project or during major milestones.



How will you allow others to use your data?

Apply licences to disambiguate reuse restrictions.

Secondary vs primary data



Reused data

EXISTING DATA

Data with PIDs, in repositories

How can I reuse existing data?

☞ Check copyright, licenses etc.

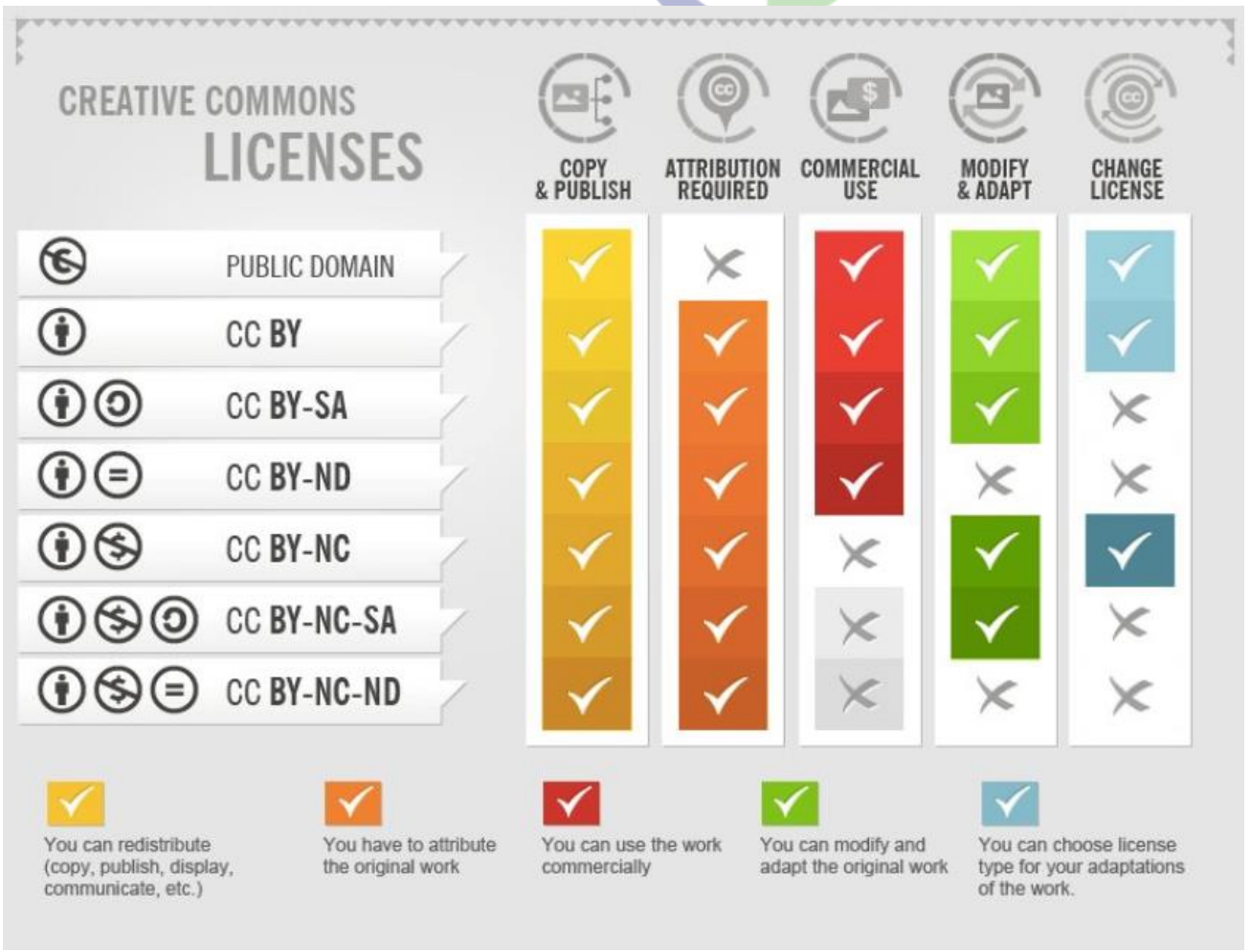
NEW DATA

Derived from current activities

How can others reuse my data?

☞Add licenses, access conditions etc.

License research data openly

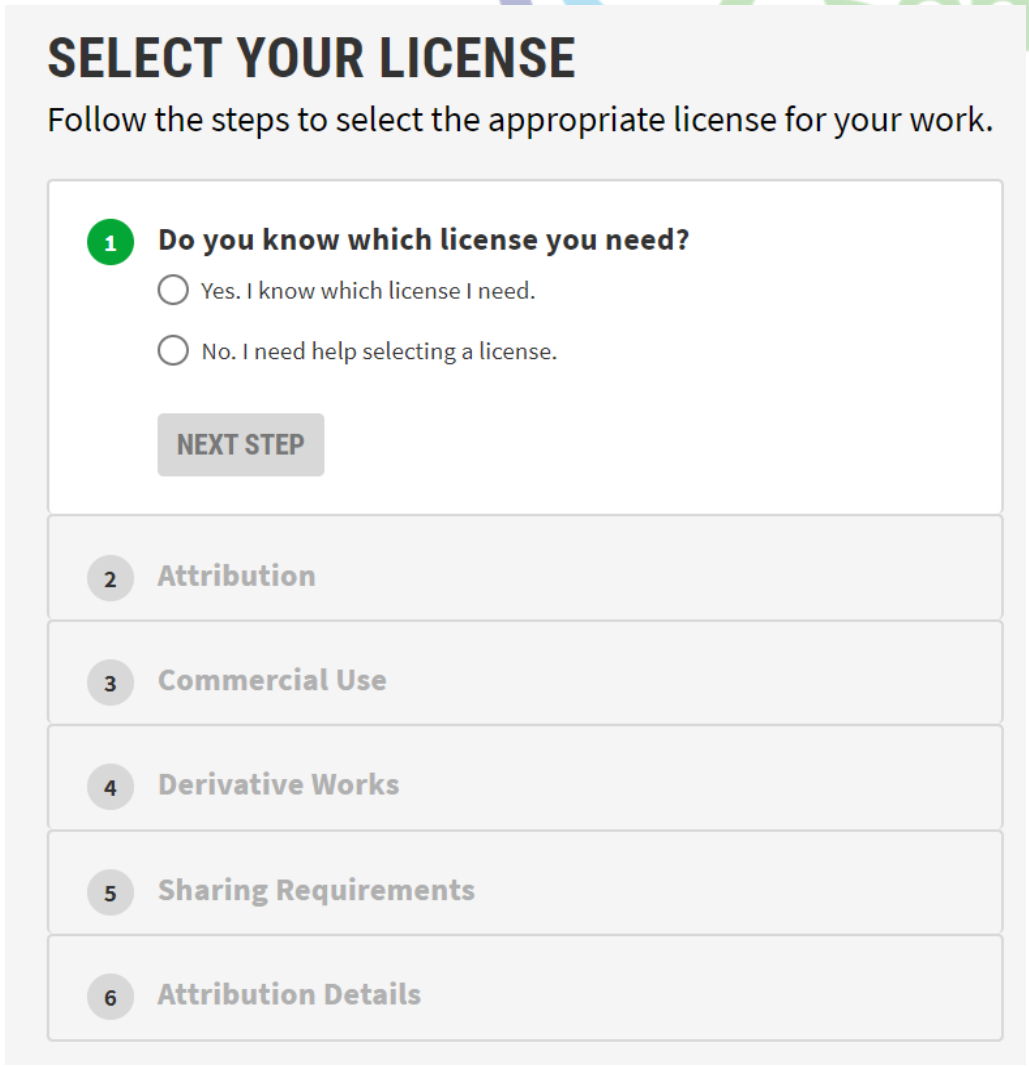


Part of How To Attribute Creative Commons Photos by Foter, licensed CC BY SA 3.0

Tools to decide which licence to choose

Choose a license for your data

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



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

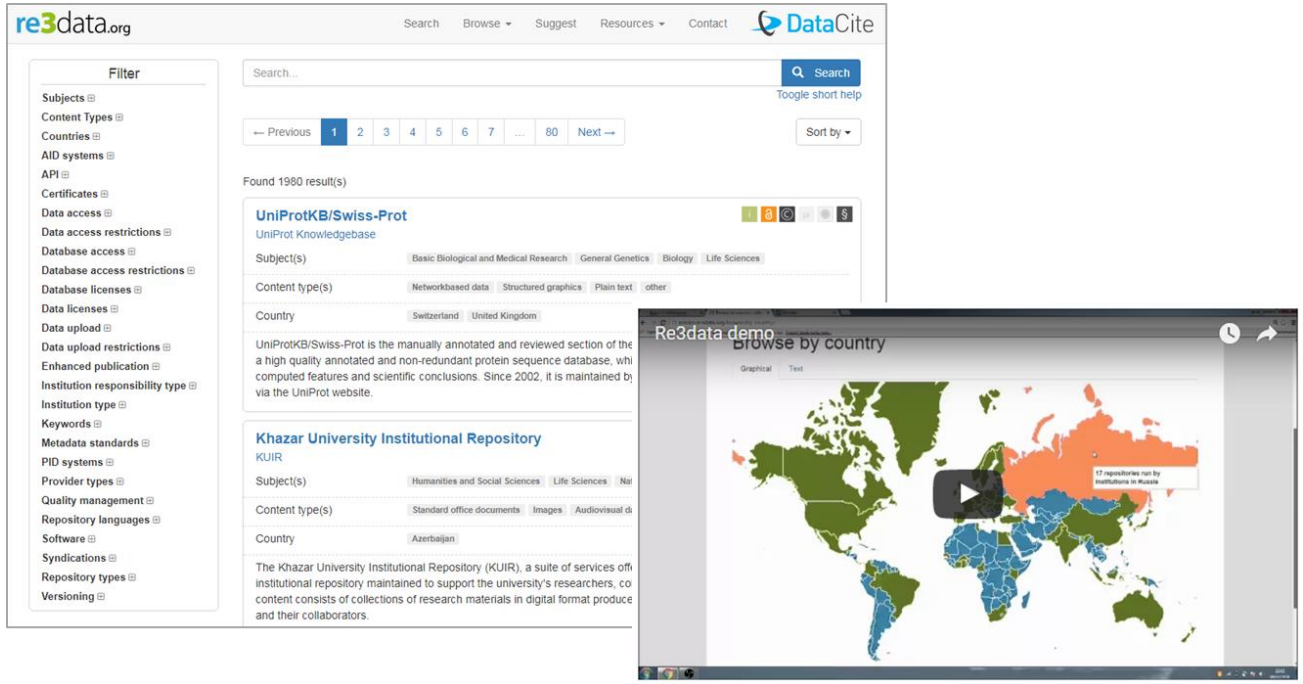
Deposit in a data repository

Long-term preservation of data.

Deposit in a data repository

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

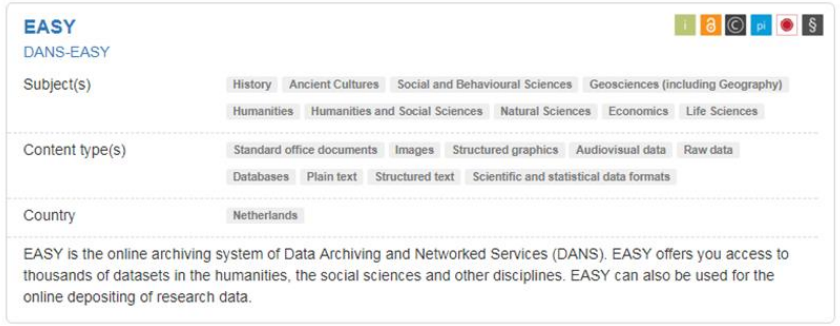
www.fosteropenscience.eu/content/re3data-demo



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.



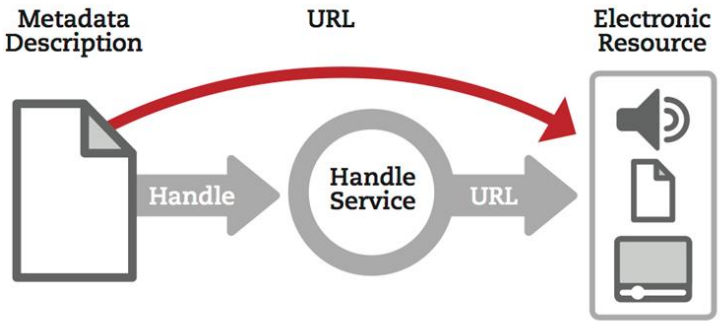
Icons to note open access, licences, PIDs, certificates...

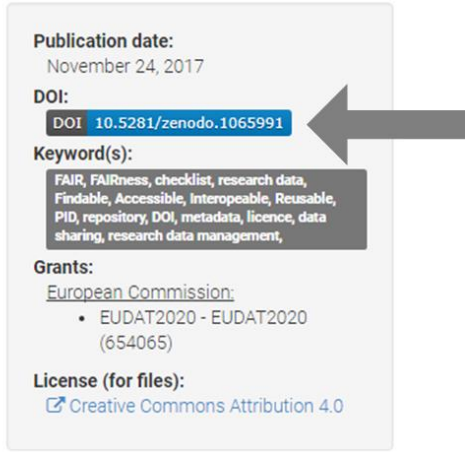
www.re3data.org

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.





www.re3data.org

PID systems – always use

Data Schools

DIGITAL OBJECTS

ARK

RESEARCHERS & ORGANISATIONS

OTHER ACTIVITIES

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?

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

Best practice

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

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

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

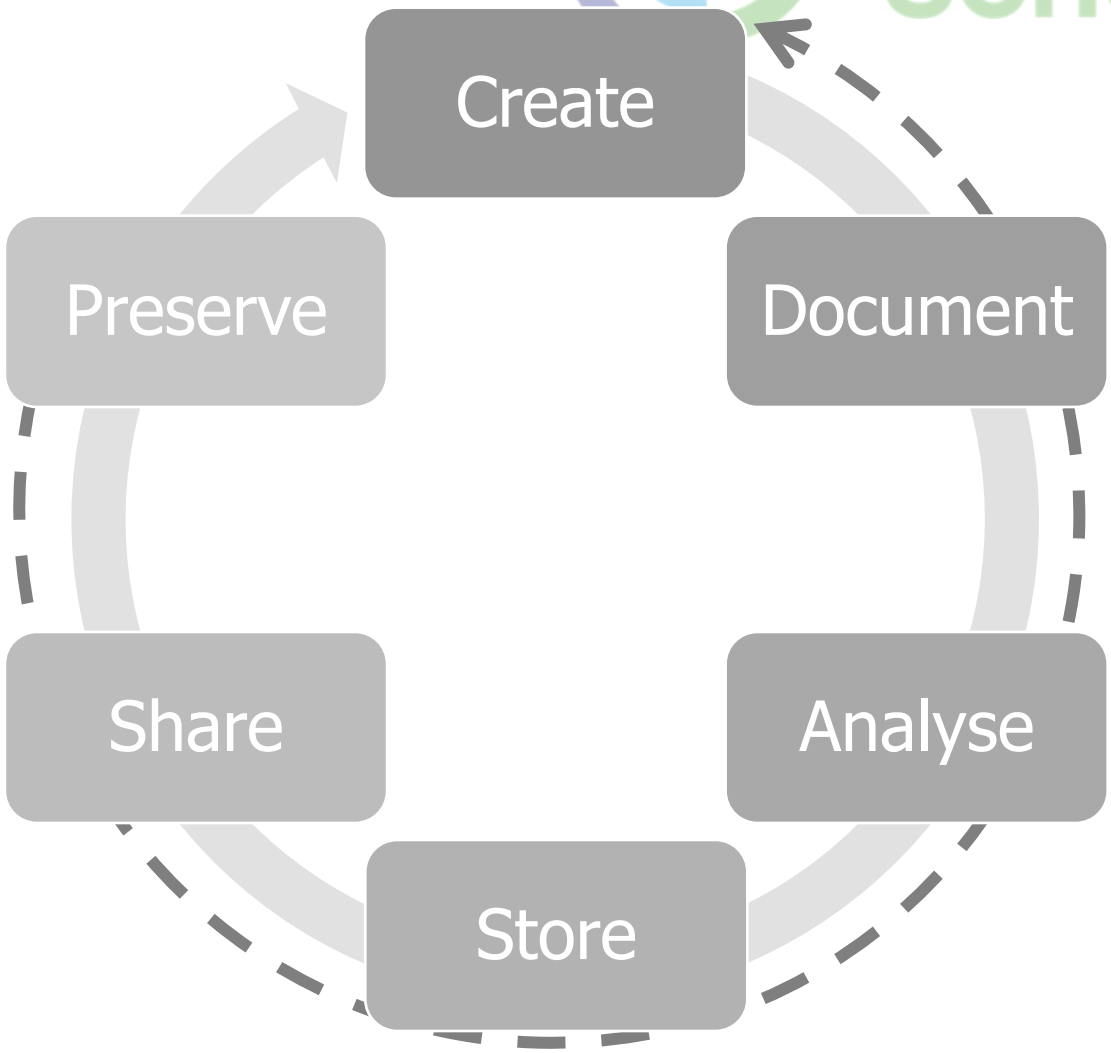
Charles Arthur and **Michael Cross**

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."

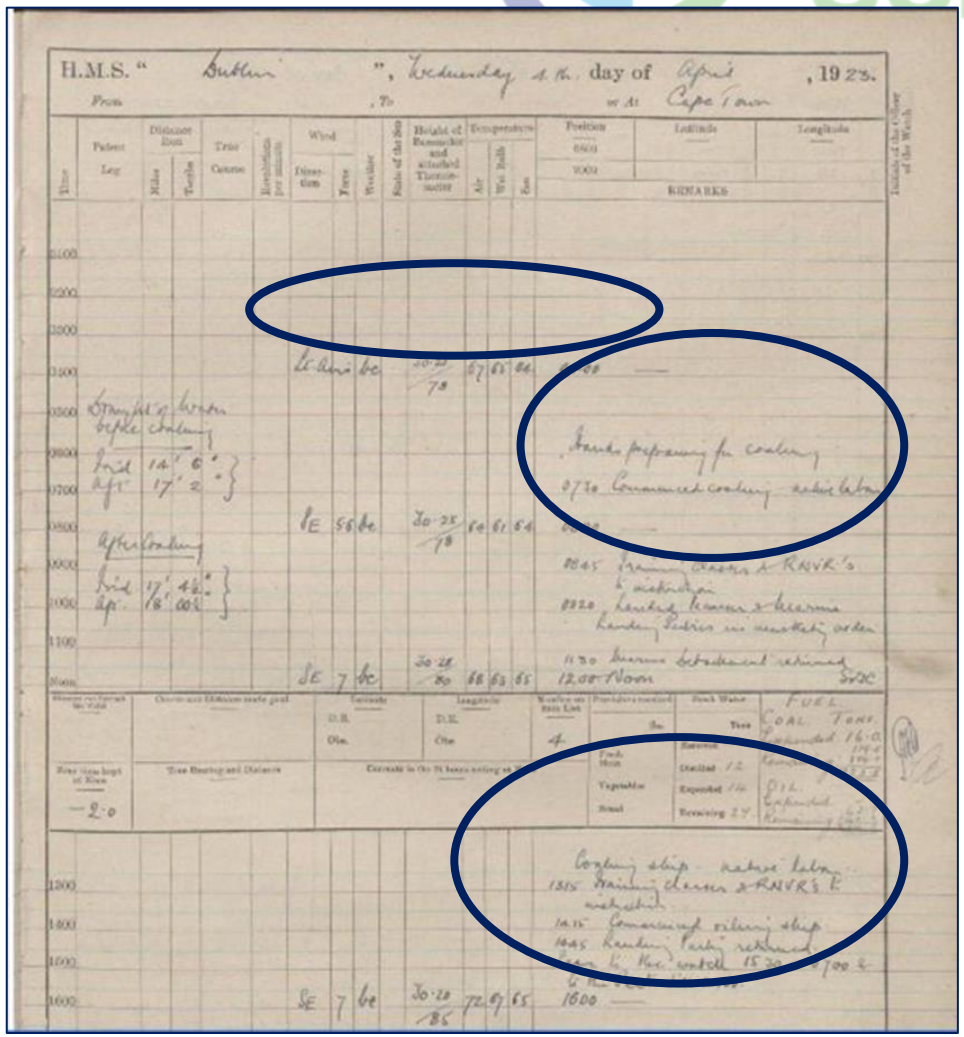
- Ewan Birney

#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

The Old Weather Project

Data for research, not from research



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

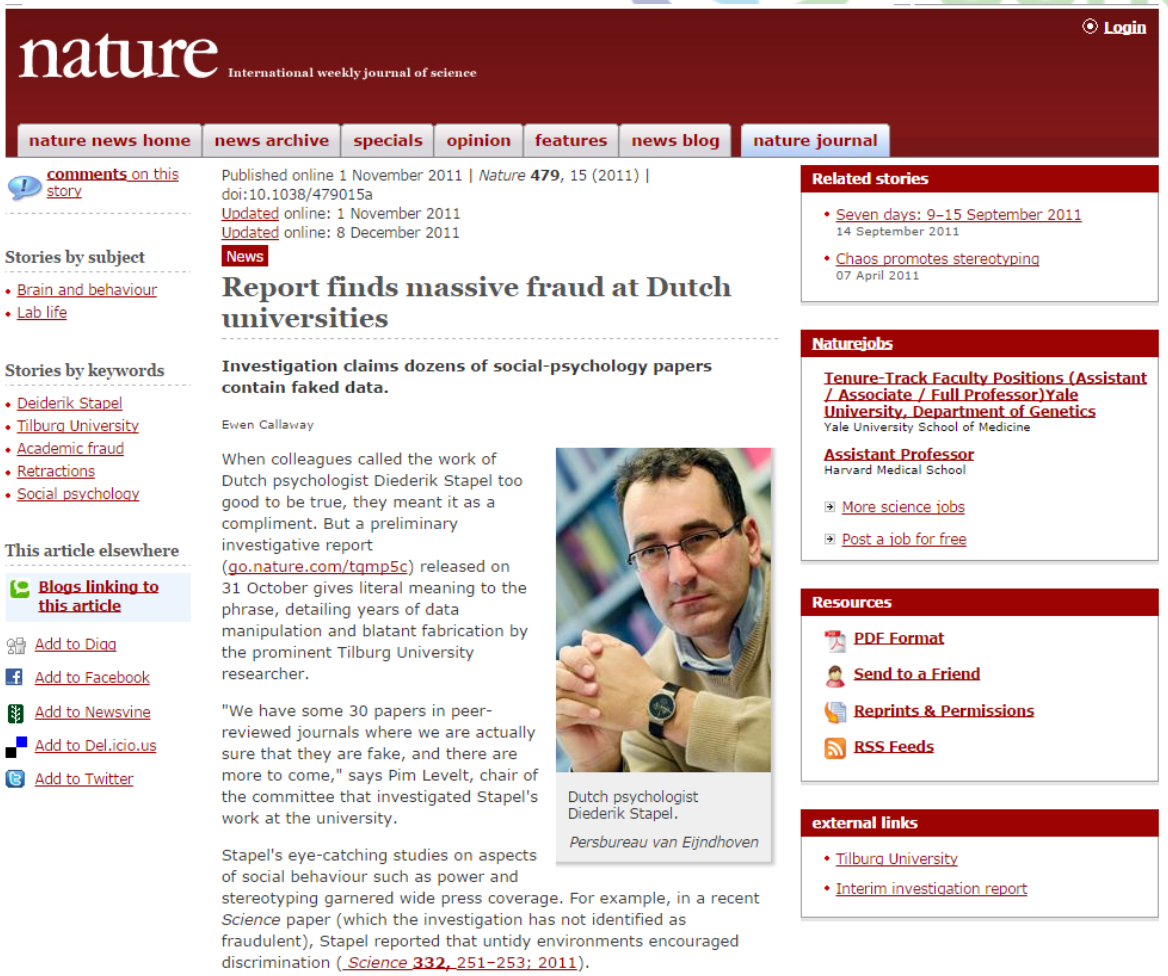
Charles Arthur and **Phillip Inman**
The Guardian, Thursday 18 April 2013 21.10 BST

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

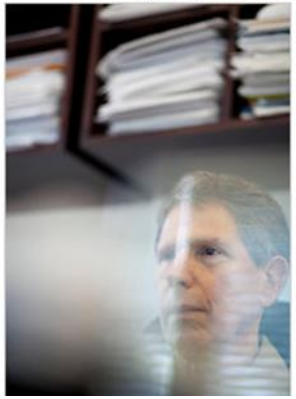
http:///www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0

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.

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.

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:

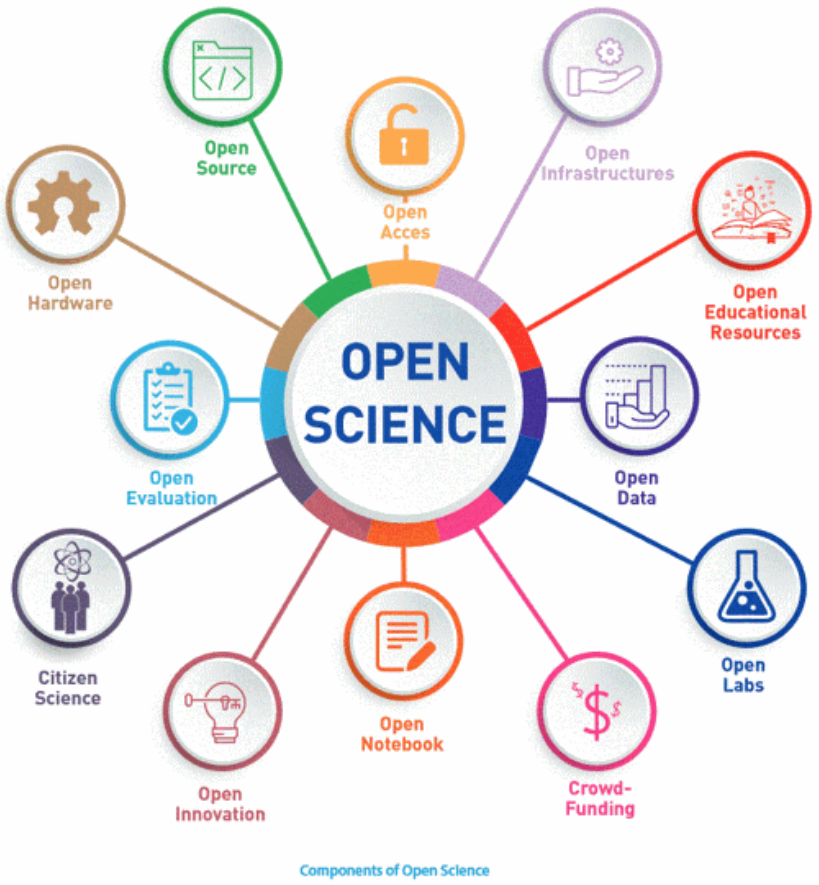
  http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- License it so others can reuse:

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

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Open Science by UNESCO

- Collaboration
- Reproducibility
- Transparency
- Trust



The picture was taken from the UNESCO Open Science brochure.

Who has heard of this before...?

Findable **A**ccessible **I**nteroperable **R**eusable

- Metadata
- PIDs
- Repositories

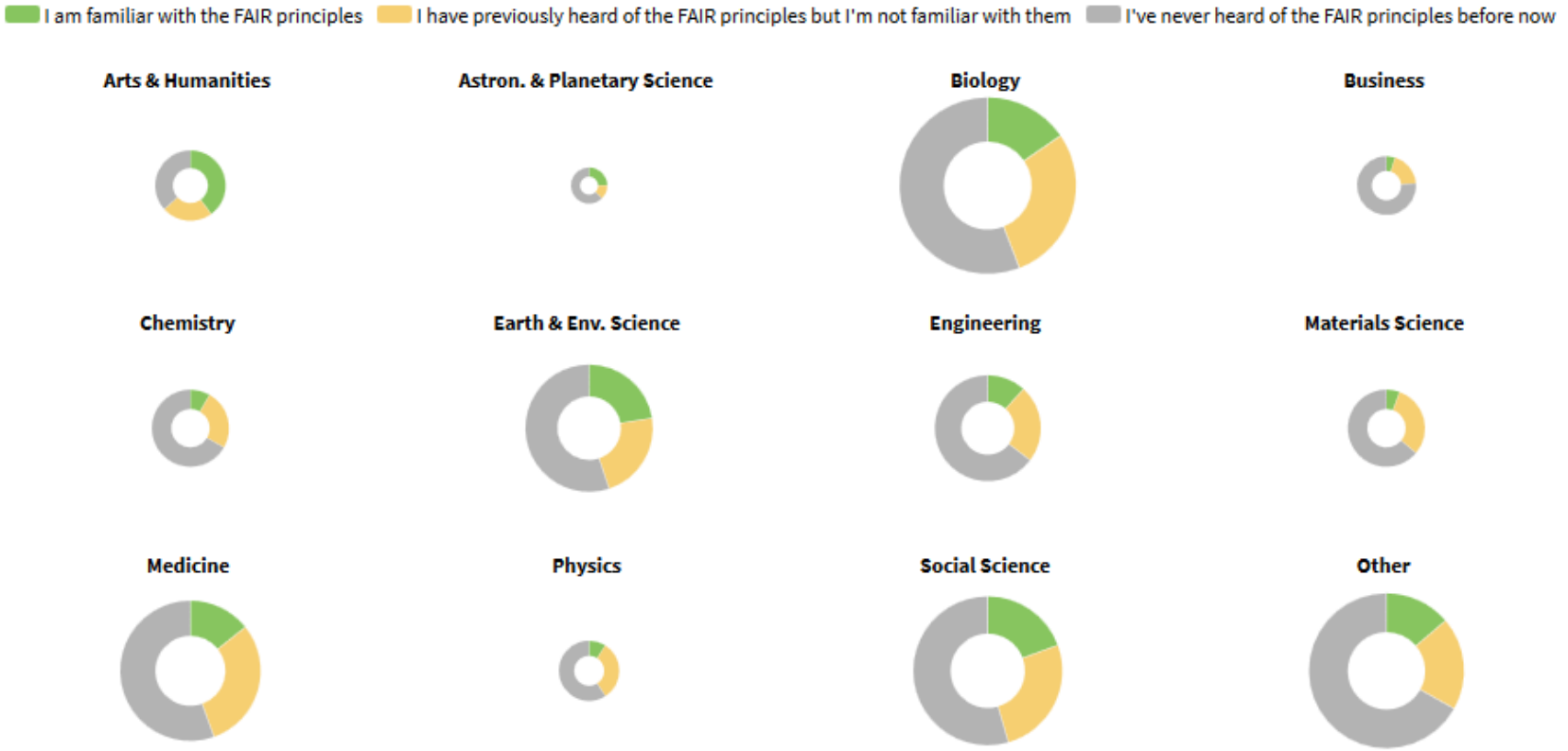
- Metadata
- Open file formats and software

- Metadata
- Ontologies
- Repositories

- Metadata
- Licences

Familiarity with FAIR principles

The majority of researchers surveyed as part of a recent study on open data had never heard of FAIR, regardless of their field. Of the 748 researchers that responded to this question, 144 said they were familiar with the principles. Circles are sized by number of respondents.

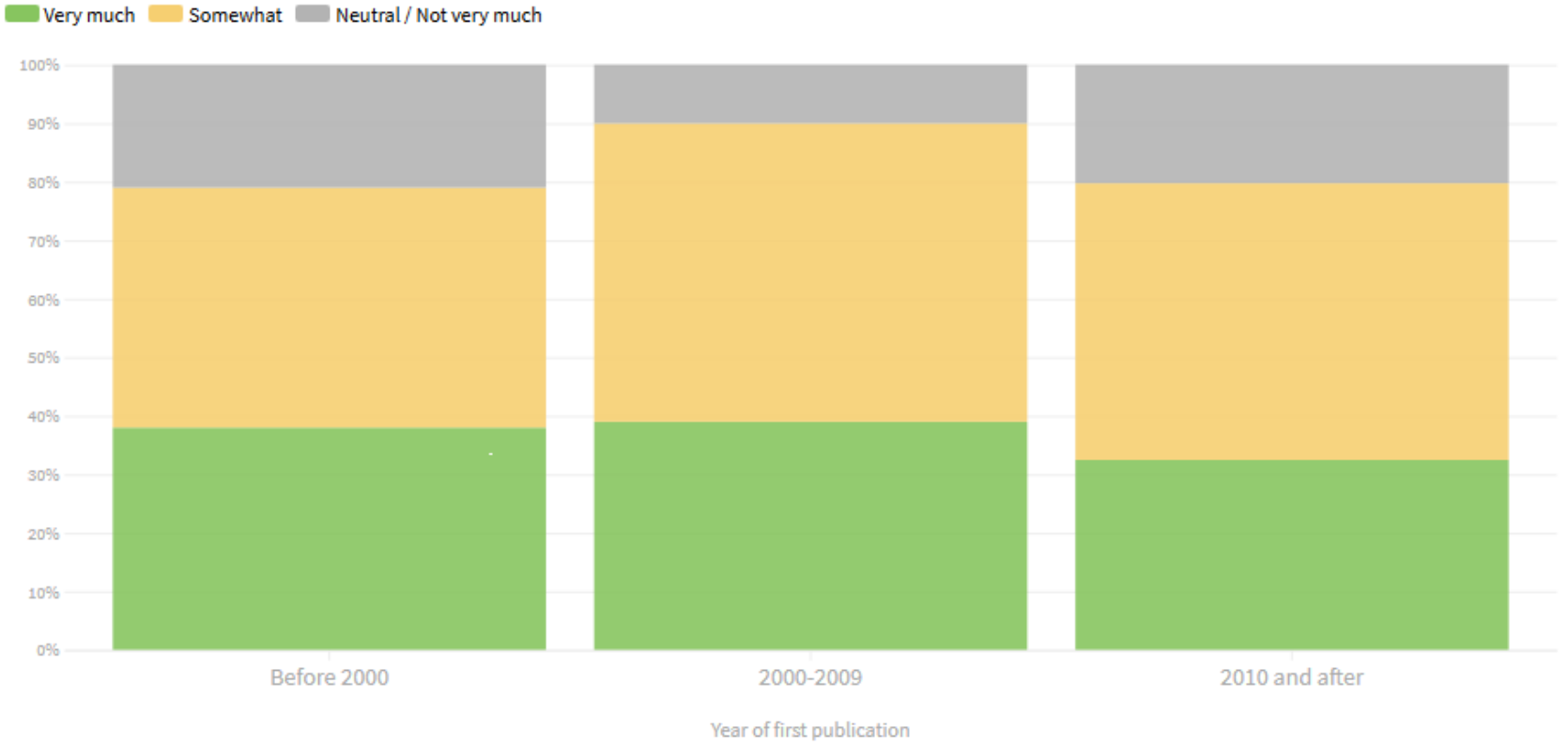


Source: State of Open Data

Brock, J. "A love letter to your future self": What scientists need to know about FAIR data *Nature Index* **11 Feb 2019**

Compliance with FAIR principles

Of the participants who were familiar with FAIR, about a third said that their data management practices were very compliant with the principles. That proportion is similar across scientists at different stages of their career.

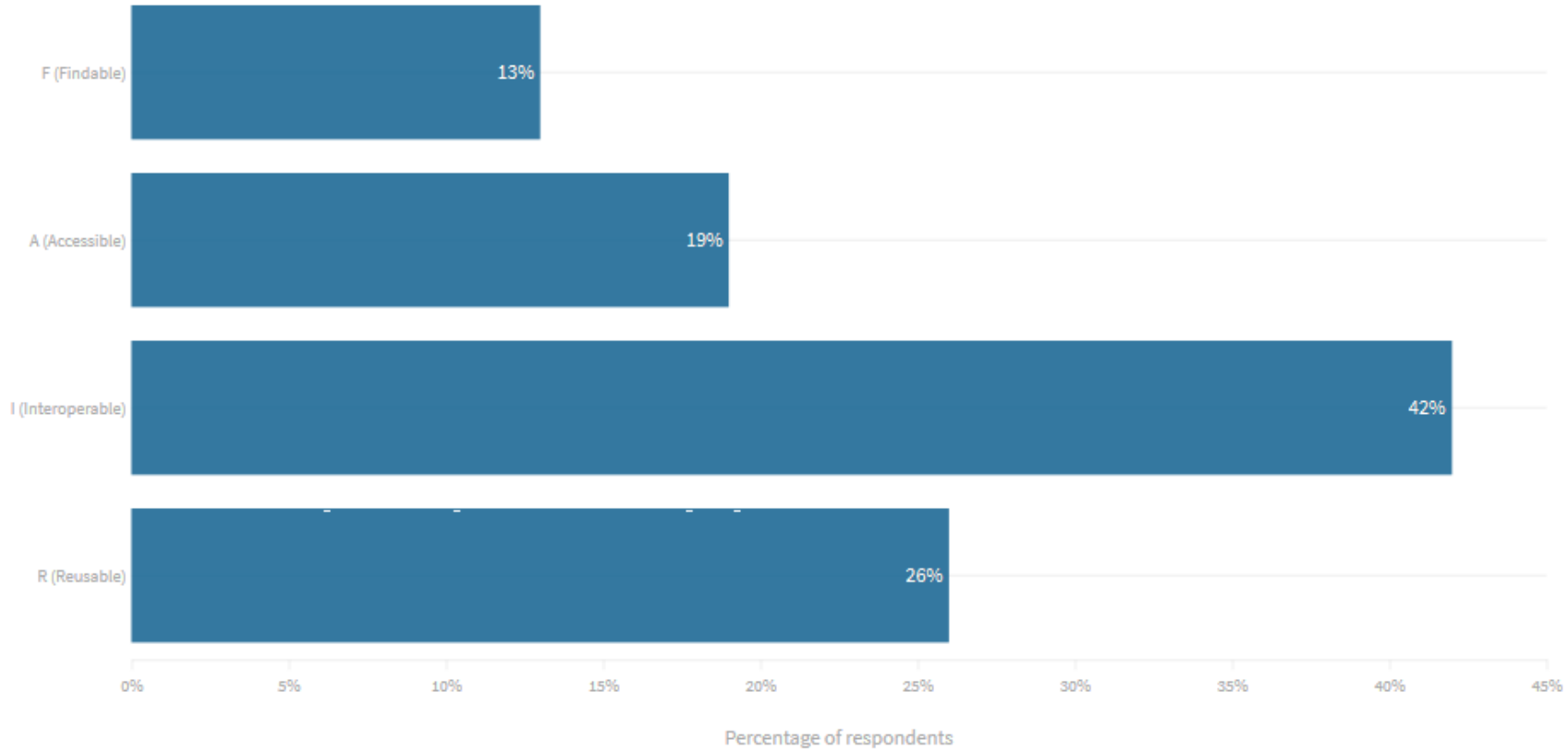


Source: State of Open Data • Data shown for respondents familiar with FAIR principles

Brock, J. "A love letter to your future self": What scientists need to know about FAIR data *Nature Index* **11 Feb 2019**

Which of the FAIR principles do you think most needs better definition?

Interoperability is the least understood FAIR principle. Some 42% of the 187 respondents who answered this question felt that it needed further clarification.



Source: State of Open Data

Brock, J. "A love letter to your future self": What scientists need to know about FAIR data *Nature Index* **11 Feb 2019**

European perspective...

https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

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;

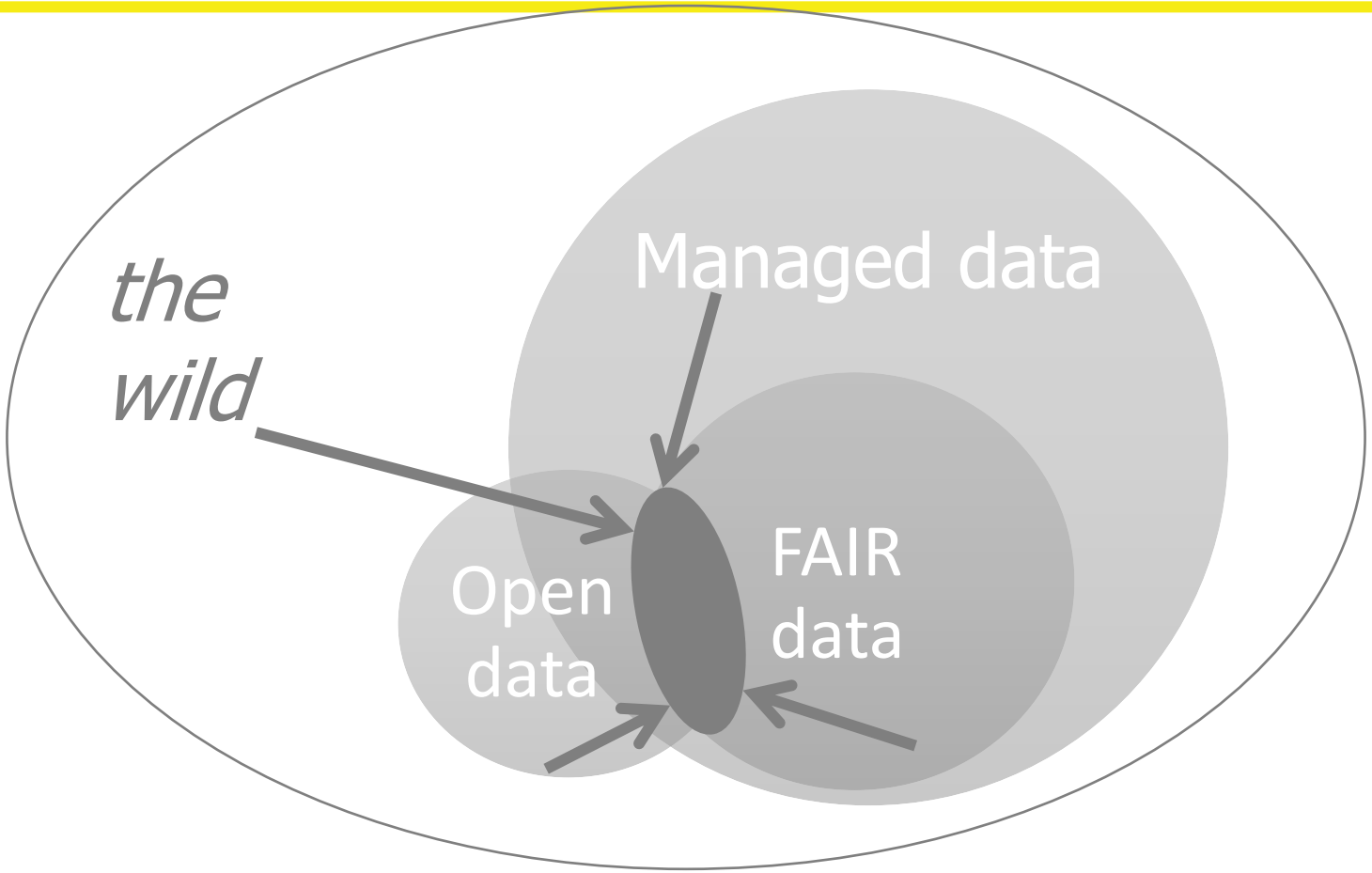
doi: 10.1038/sdata.2016.18

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

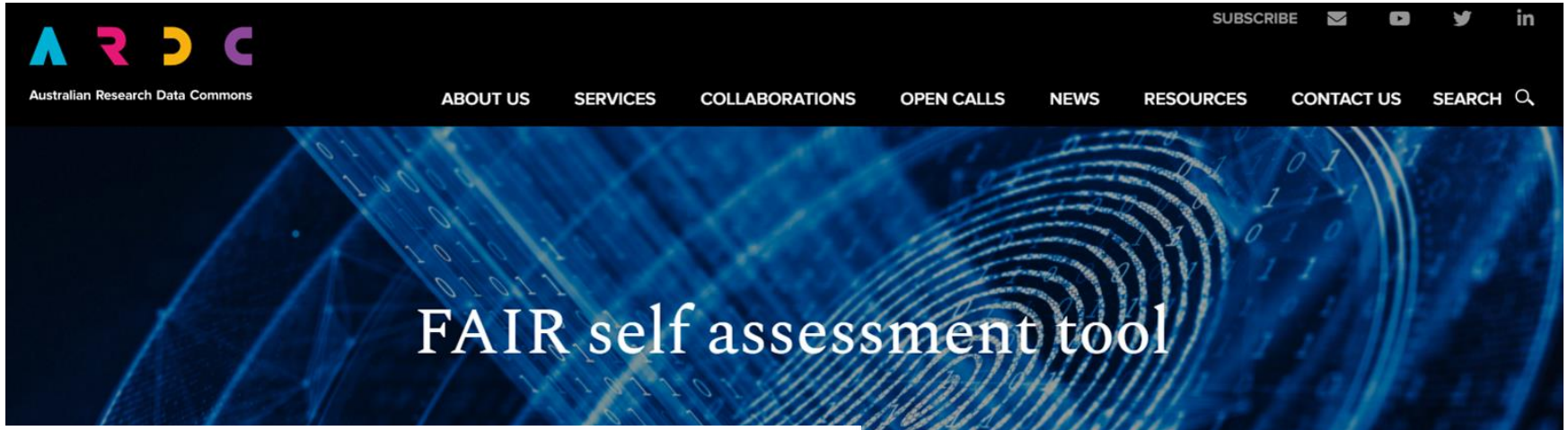


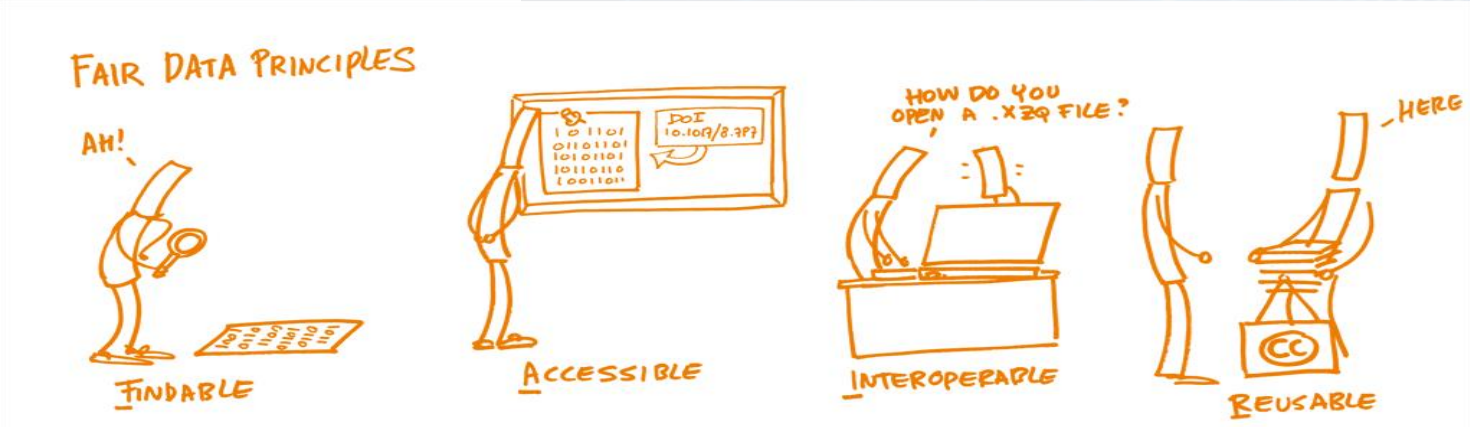
FAIR ≠ Open

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





ssment tool. Using this tool you will be able to assess
nce its FAIRness (where applicable).

e, Accessible, Interoperable and Reusable (FAIR). Once
'green bar' indicator based on your answers in that
ll 'FAIRness' indicator is provided.

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

OpenAIRE RESEARCH GRAPH
Putting research into context. Making the connections.
explore.openaire.eu

USAGE ANALYTICS
Make your content count
provide.openaire.eu

zenodo
a catch-all-repository
www.zenodo.org

OpenAIRE | MONITOR
Simplify research tracking & monitoring
monitor.openaire.eu

amnesia
Anonymize your datasets
amnesia.openaire.eu

OpenAIRE | CONNECT
Build an Open Research Gateway for your community
connect.openaire.eu

OA BROKER
Routing information across content providers
provide.openaire.eu

ScholeXplorer
Interlinking data and publications
scholexplorer.openaire.eu

METADATA VALIDATOR
Helping you apply the OpenAIRE Guidelines
provide.openaire.eu

OpenAIRE | EXPLORE
Discover and share research
explore.openaire.eu

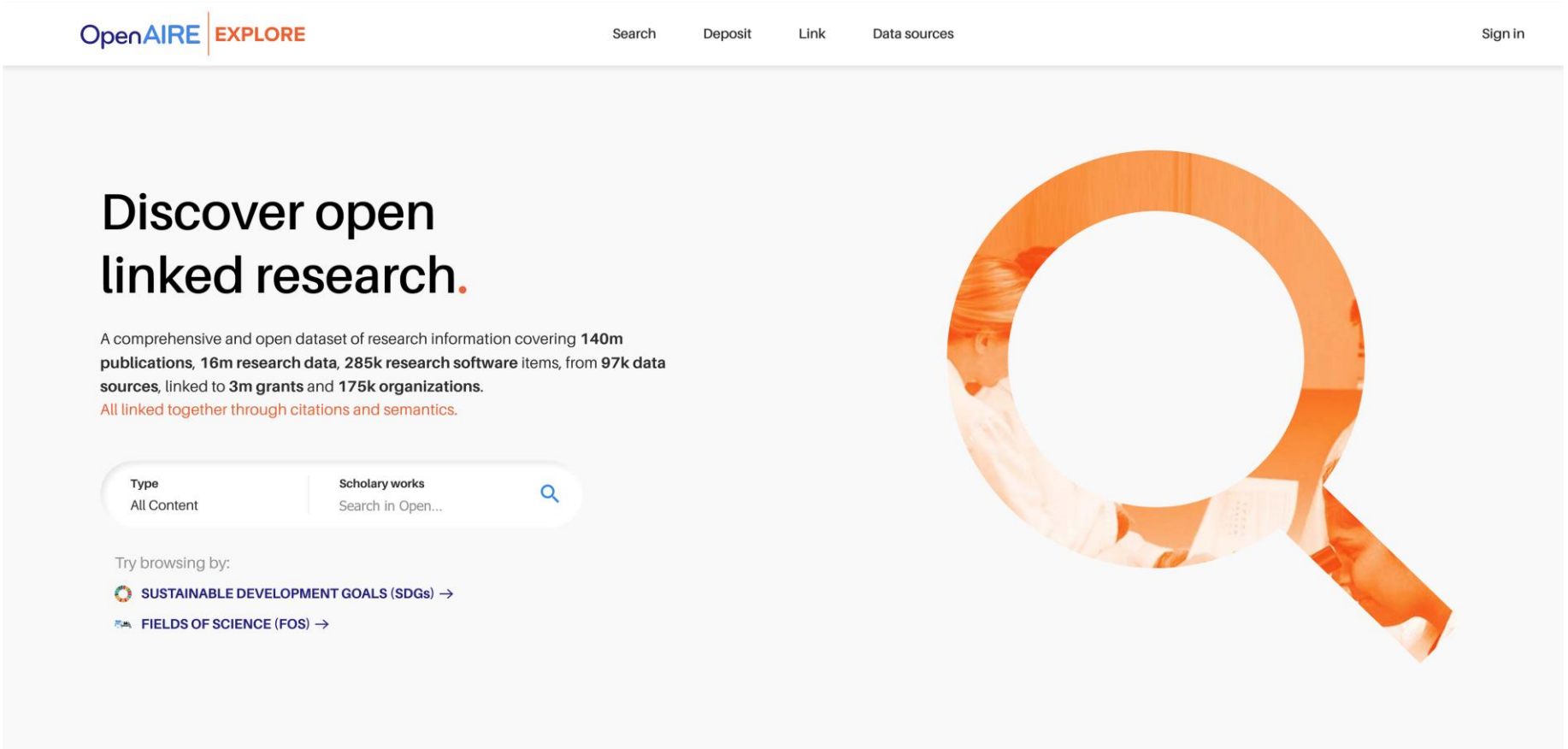
OpenAIRE | DEVELOP
APIs to access the OpenAIRE Research Graph
develop.openaire.eu

argos
Plan and follow your data
openaire.eu/argos

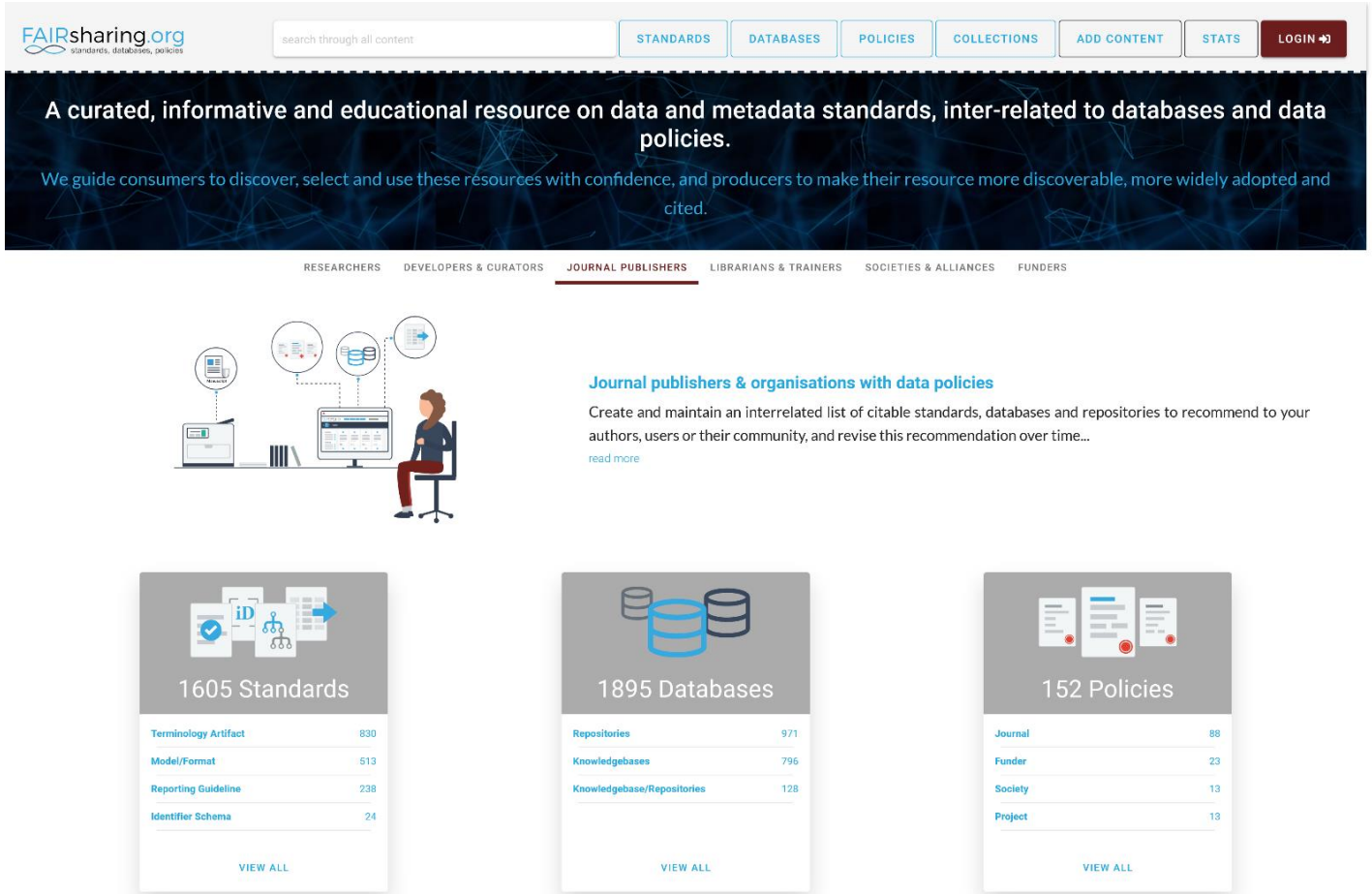
Scholarly communication services for all actors

https://www.openaire.eu/

OpenAIRE EXPLORE



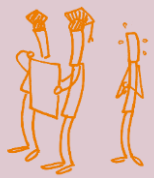
https://explore.openaire.eu/



https://www.fairsharing.org

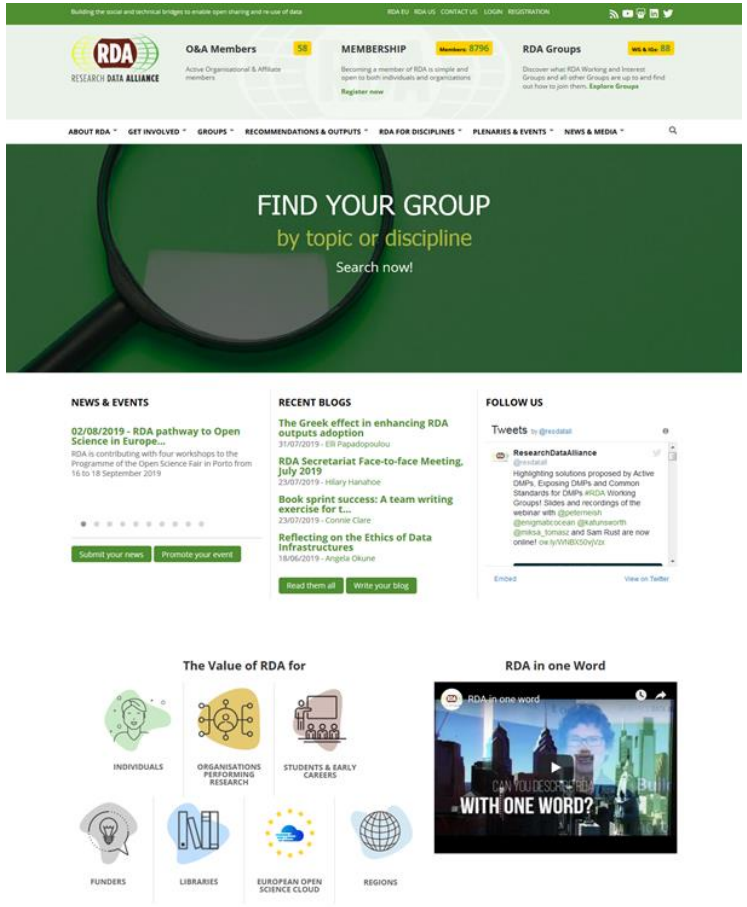
FOSTER Open Science

Data Schools

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



https://www.rd-alliance.org

Data Schools

Thank you!

Questions?