DCC

Digital Curation Centre

Principles of Research Data Management and Open Research

S. Venkataraman, PhD
Research Data Specialist
Digital Curation Centre

s.venkataraman@ed.ac.uk

5th December 2019, CODATA/RDA School of Research Data Science, CeNAT, San José, Costa Rica

About the DCC

- Established in 2004
- Based in Edinburgh and Glasgow
- Works at national and international levels
- One of leading organisations in the world specialising in training, consultancy, policy making and advocacy in digital data management best practice and services provision
- Involved in many international consortia and schools
- (We do not curate any data ourselves!)

Learning outcomes

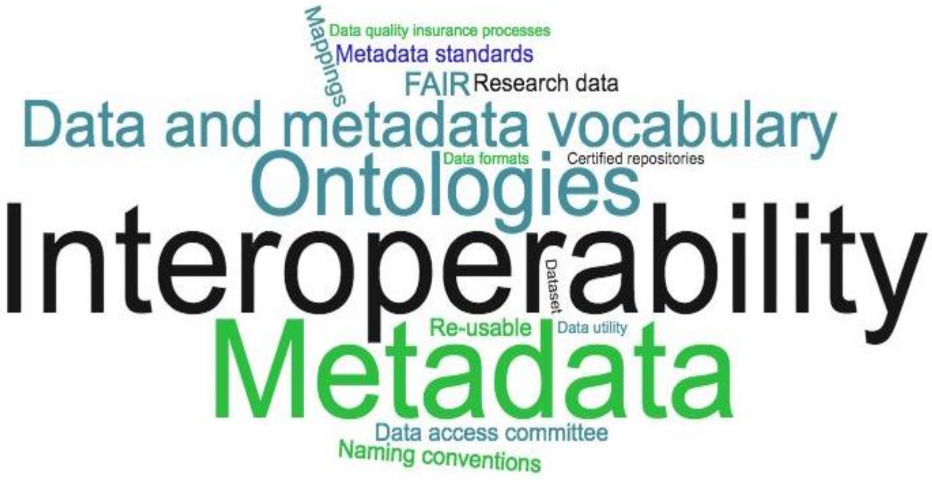
- Be familiar with the curation lifecycle
- Understand the standardisation methods and principles available to add value to your data
- Learn about resources to aid your workflows
- Increase/encourage your level of openness
- Implement and review DMPs

Language is a barrier…

Respondents mentioned 40 terms

which were unclear to them in

European Commission DMP

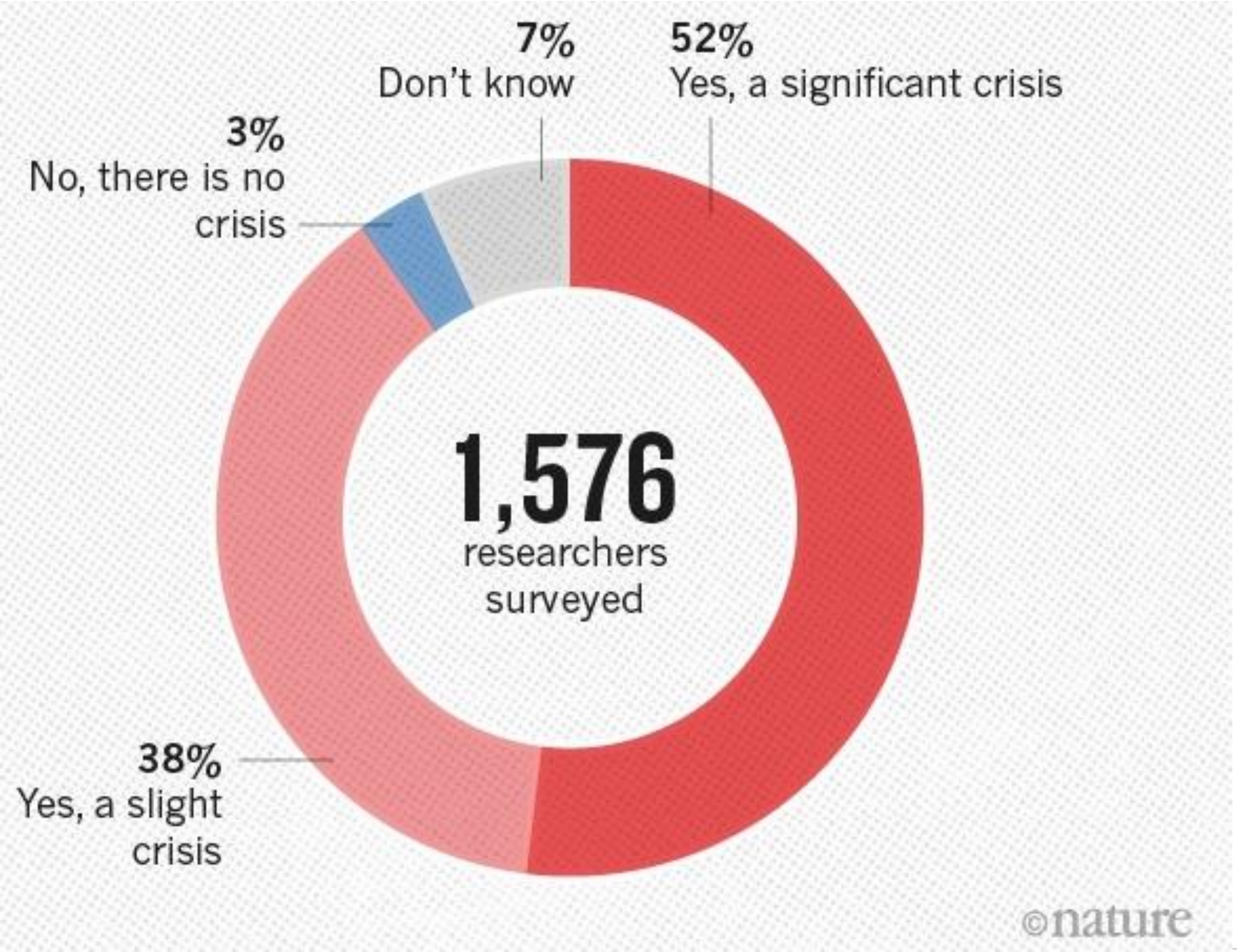


“Researchers are not familiar with the following terms/phrases : Metadata, standards for metadata/data, ontologies, mapping with ontologies, interoperability, ... All the ICT jargon”

“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.”

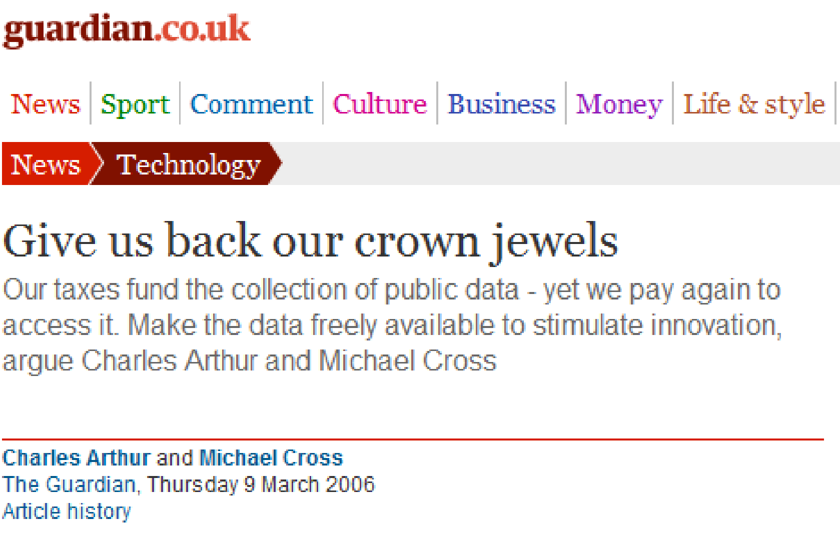
Grootveld et al. (2018). OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans http://doi.org/10.5281/zenodo.1120245

Is there a reproducibility crisis?



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

Research data: institutional crown jewels?



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

http://www.flickr.com/photos/lifes__too_short__to__drink__cheap__wine/4754234186

Why make data available?

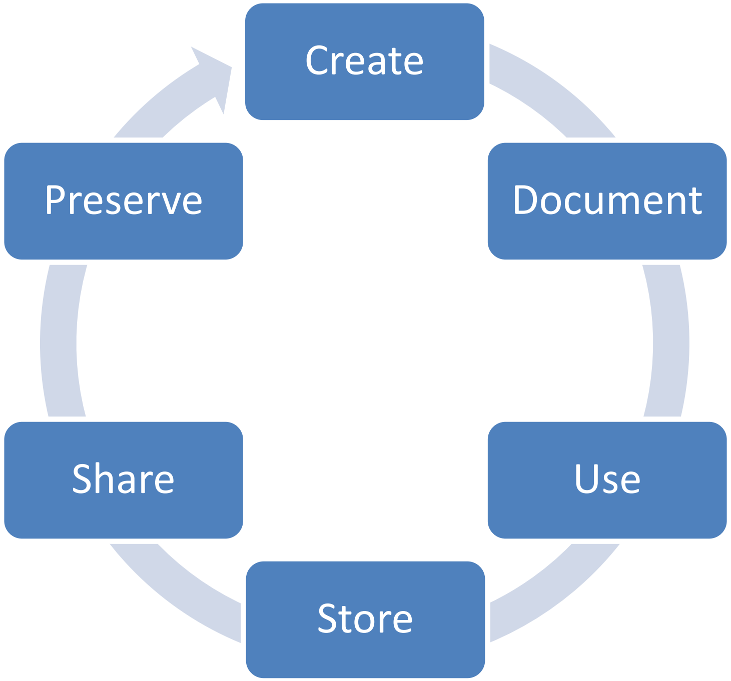
"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

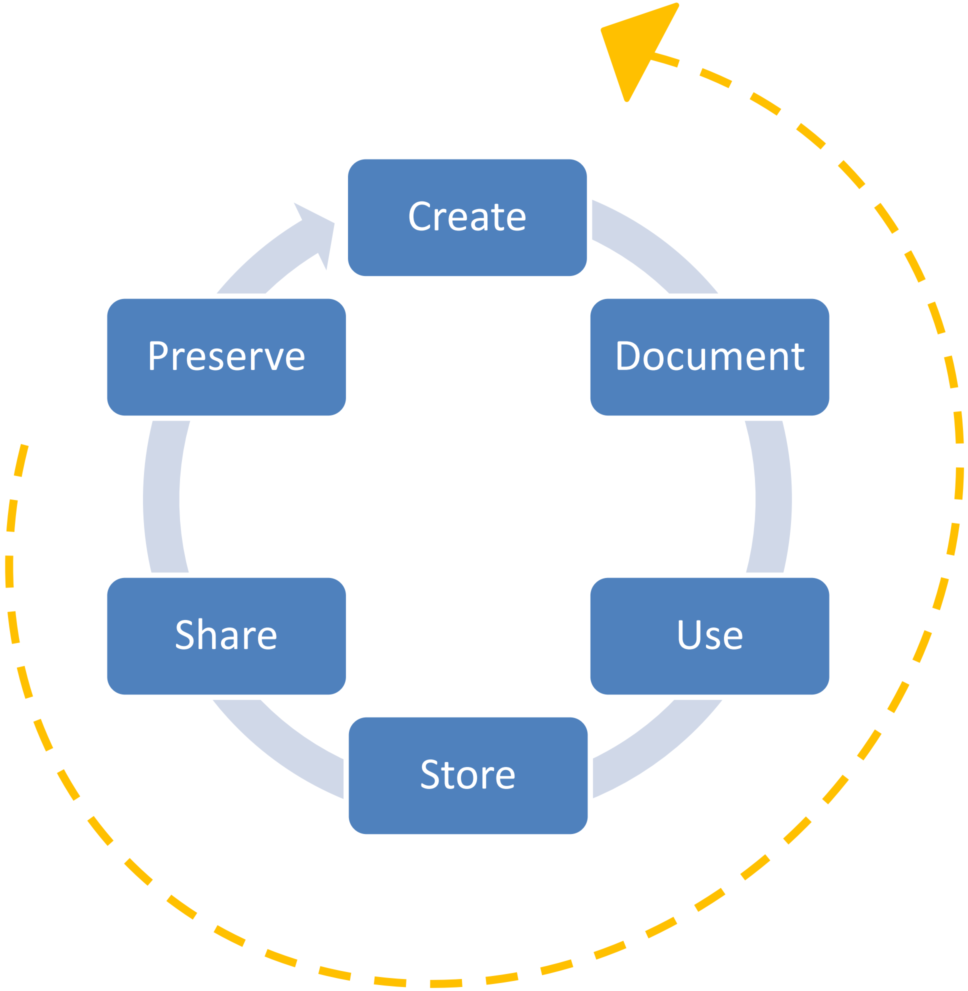
#OpenData
#OpenScience

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 curation lifecycle



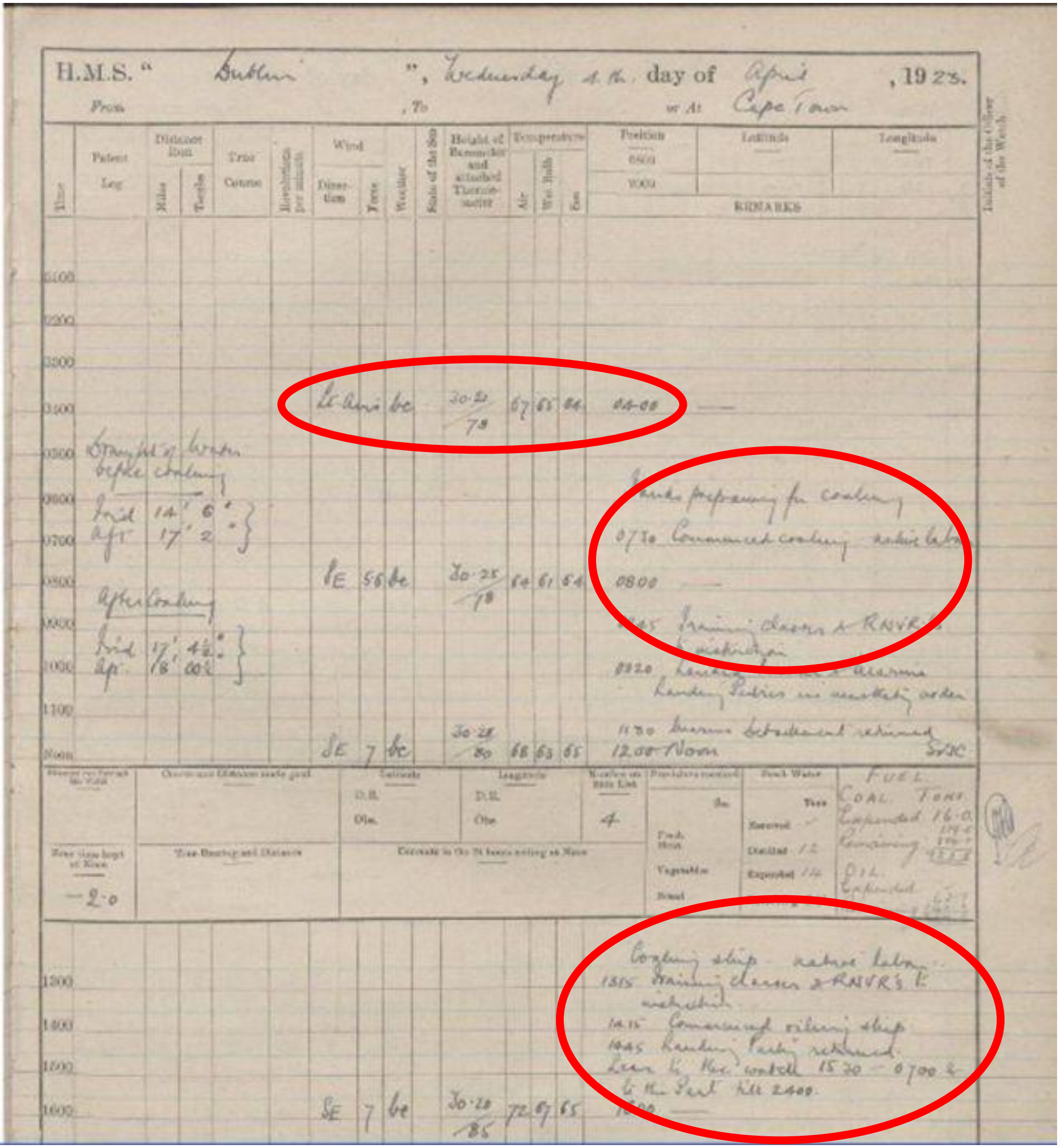
…and open research



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

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

0400 Lt airs bc 30·21/73 67 65 64 04-00 —

Draught of water before coaling
Fwd 14' 6"
Aft 17' 2"

Hands preparing for coaling
0730 Commenced coaling native labour
0800 SE 5-6 bc 30·25/78 64 61 64 08.00 —
0845 Training classes & RNVR's instruction
0920 Landing [illegible] & [illegible] Landing Parties in readiness order

After coaling
Fwd 17' 4½"
Aft 18' 00½"

1130 [illegible] detachment returned
Noon SE 7 bc 30·28/80 68 63 65 12.00 Noon

FUEL
COAL Tons
Expended 16·0
Remaining
OIL
Expended

Coaling ship native labour
1315 Training classes & RNVR's instruction
1415 Commenced oiling ship
1445 Landing Party returned
Leave to the watch 1530 – 0700 & to the Port till 2400.
1600 SE 7 bc 30·20/85 72 67 65 16.00 —

Increased use and economic benefit

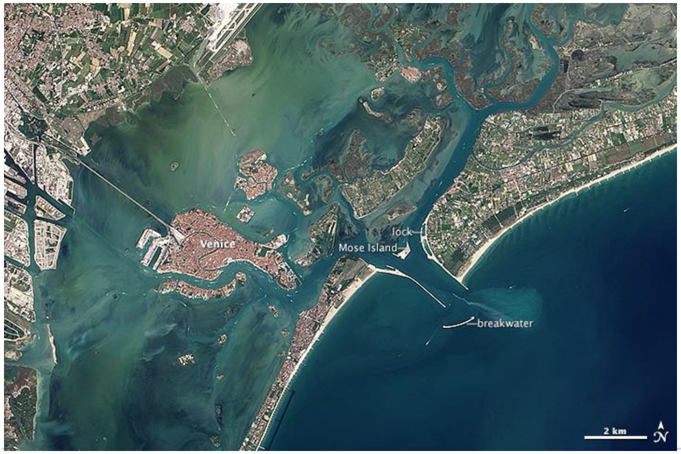
The case of NASA Landsat satellite imagery of the Earth's surface:

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

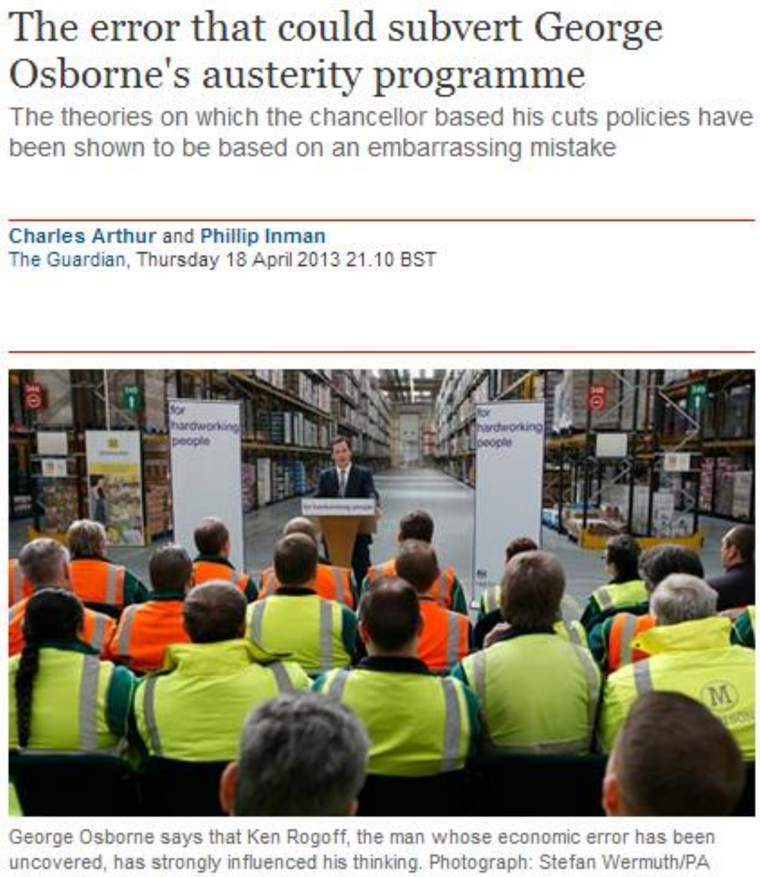
Has stimulated the development of applications from a large number of companies worldwide

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

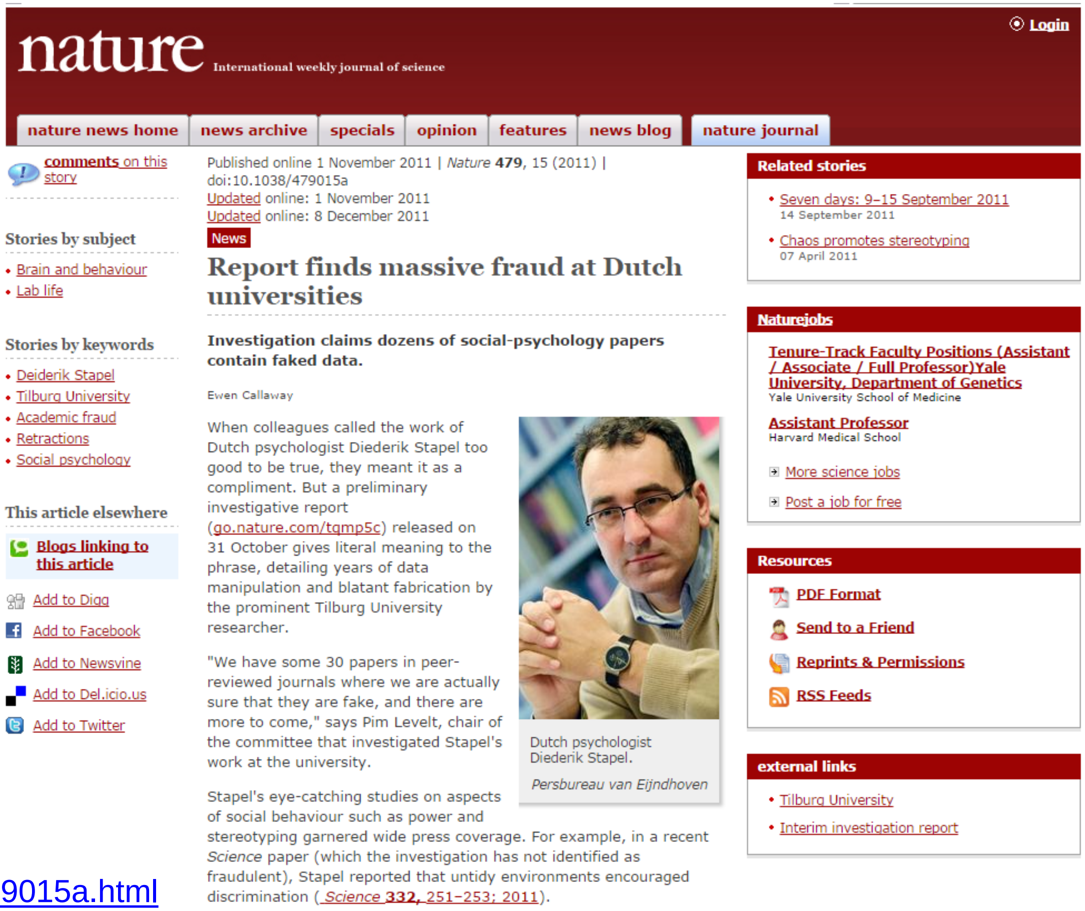
Charles Arthur and Phillip Inman
The Guardian, Thursday 18 April 2013 21.10 BST

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”



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

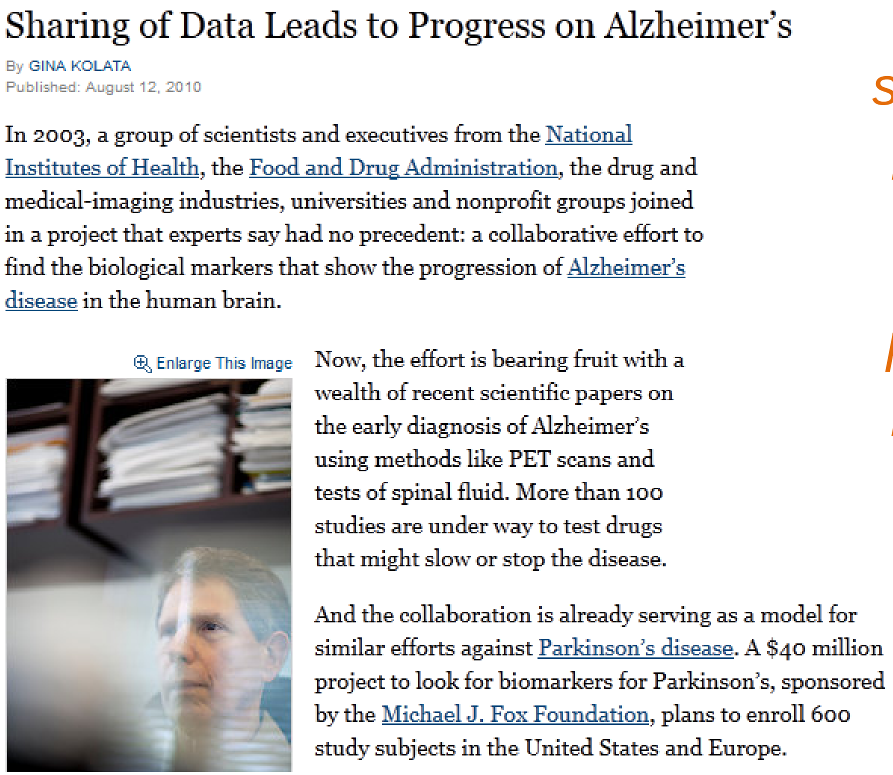
RSS Feeds

external links

- Tilburg University
- Interim investigation report

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

Sharing leads to breakthroughs!



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.

"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

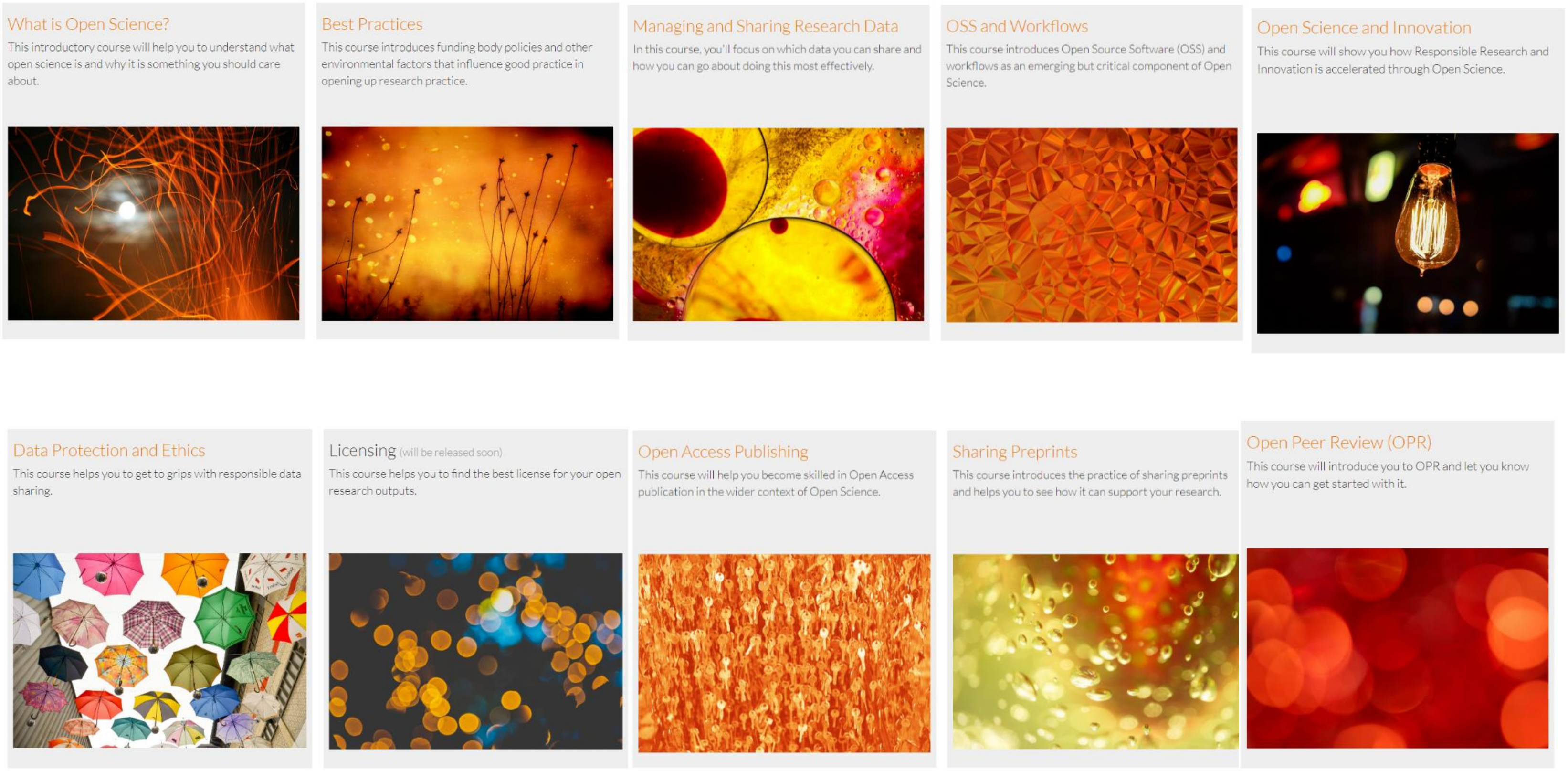
How do you share data effectively?

- Use appropriate repositories, this catalogue is a good place to start
  http://www.re3data.org

- Document and describe it enough for others to understand, use and cite
  http://www.dcc.ac.uk/resources/how-guides/cite-datasets

- Licence it so others can reuse
  www.dcc.ac.uk/resources/how-guides/license-research-data

FOSTER Open Science toolkit



What is Open Science?

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

Best Practices

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

Managing and Sharing Research Data

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

OSS and Workflows

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

Open Science and Innovation

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

Data Protection and Ethics

This course helps you to get to grips with responsible data sharing.

Licensing (will be released soon)

This course helps you to find the best license for your open research outputs.

Open Access Publishing

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

Sharing Preprints

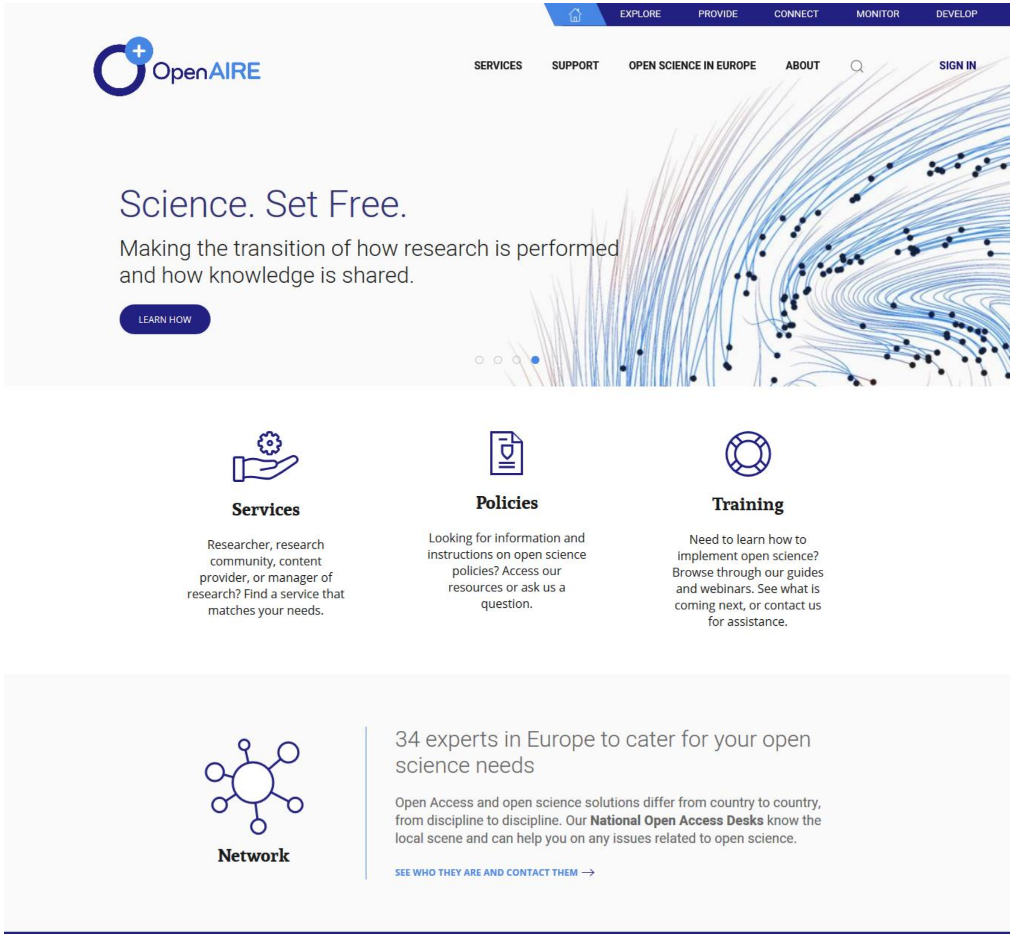
This course introduces the practice of sharing preprints and helps you to see how it can support your research.

Open Peer Review (OPR)

This course will introduce you to OPR and let you know how you can get started with it.

https://www.fosteropenscience.eu/toolkit

OpenAIRE



Science. Set Free.

Making the transition of how research is performed and how knowledge is shared.

Services

Researcher, research community, content provider, or manager of research? Find a service that matches your needs.

Policies

Looking for information and instructions on open science policies? Access our resources or ask us a question.

Training

Need to learn how to implement open science? Browse through our guides and webinars. See what is coming next, or contact us for assistance.

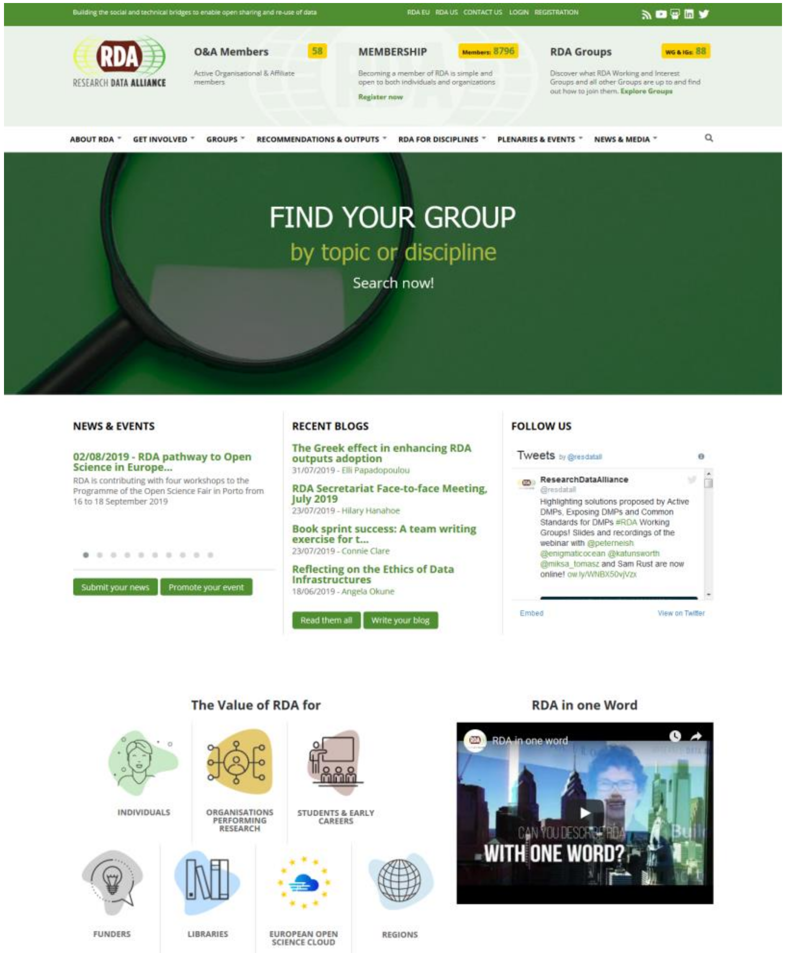
Network

34 experts in Europe to cater for your open science needs

Open Access and open science solutions differ from country to country, from discipline to discipline. Our **National Open Access Desks** know the local scene and can help you on any issues related to open science.

https://www.openaire.eu/

Research Data Alliance



Building the social and technical bridges to enable open sharing and re-use of data

RDA EU RDA US CONTACT US LOGIN REGISTRATION

RESEARCH DATA ALLIANCE

O&A Members 58

Active Organisational & Affiliate members

MEMBERSHIP Members: 8796

Becoming a member of RDA is simple and open to both individuals and organizations

Register now

RDA Groups WG & IGs: 88

Discover what RDA Working and Interest Groups and all other Groups are up to and find out how to join them. **Explore 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...

RDA is contributing with four workshops to the Programme of the Open Science Fair in Porto from 16 to 18 September 2019

Submit your news | Promote your event

RECENT BLOGS

The Greek effect in enhancing RDA outputs adoption
31/07/2019 - Elli Papadopoulou

RDA Secretariat Face-to-face Meeting, July 2019
23/07/2019 - Hilary Hanahoe

Book sprint success: A team writing exercise for t...
23/07/2019 - Connie Clare

Reflecting on the Ethics of Data Infrastructures
18/06/2019 - Angela Okune

Read them all | Write your blog

FOLLOW US

Tweets by @resdatall

ResearchDataAlliance @resdatall

Highlighting solutions proposed by Active DMPs, Exposing DMPs and Common Standards for DMPs #RDA Working Groups! Slides and recordings of the webinar with @peterneish @enigmaticocean @katunsworth @miksa_tomasz and Sam Rust are now online! ow.ly/WNBX50vjVzx

Embed View on Twitter

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

Who has heard of this before…?

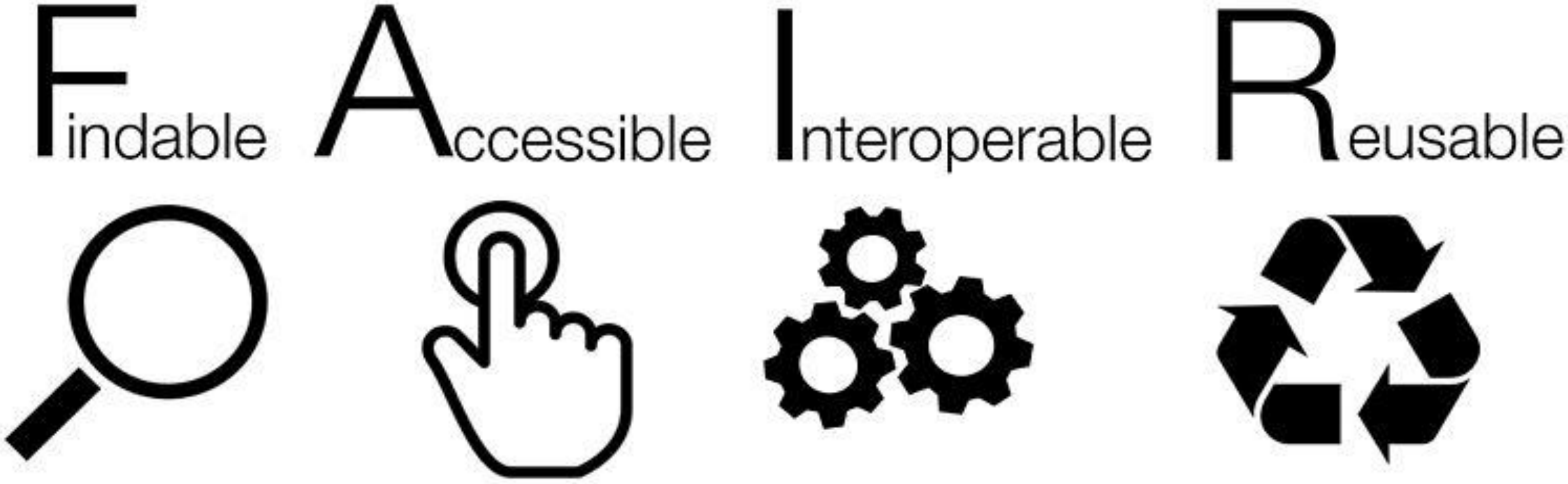
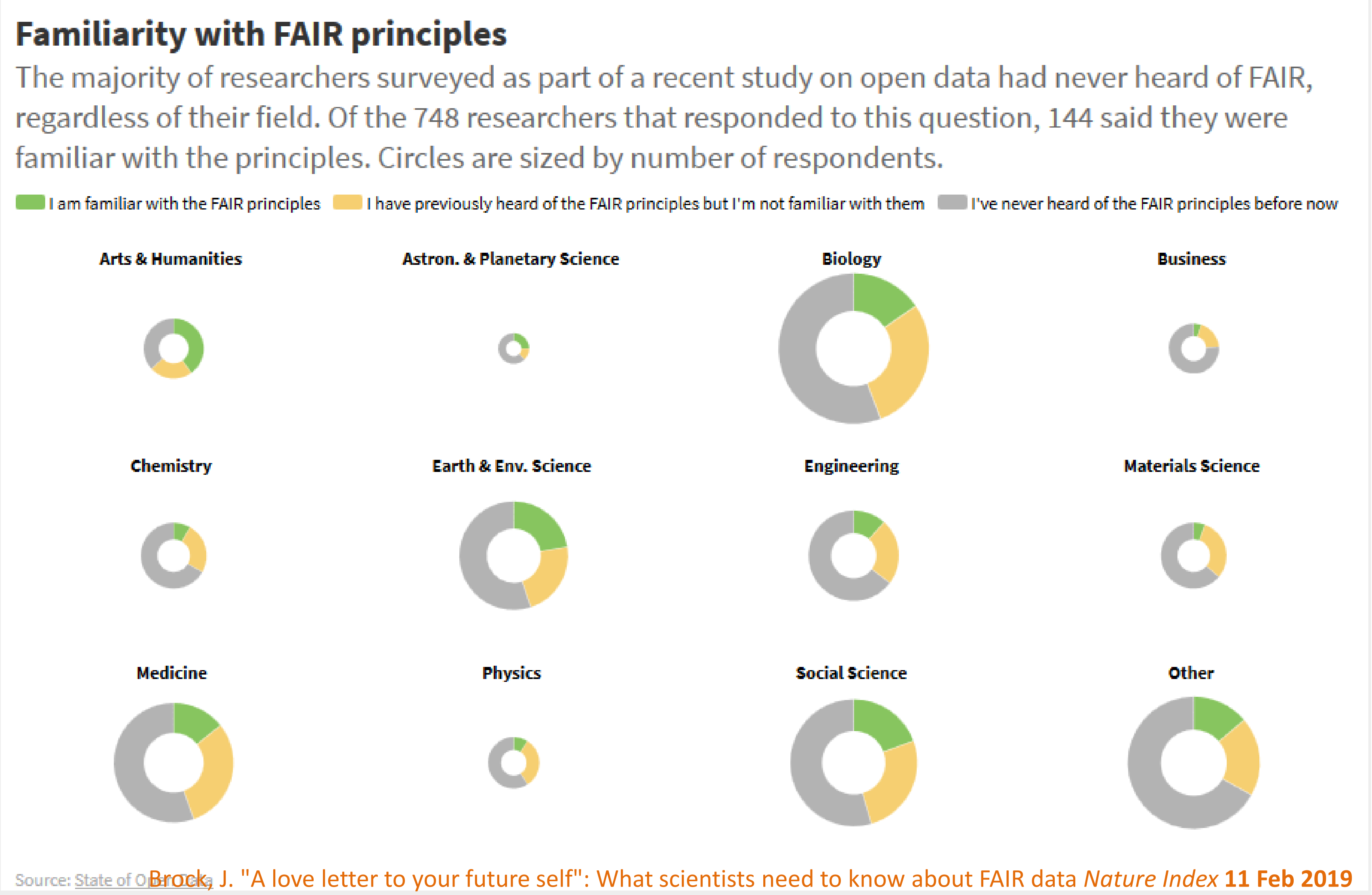


Image CC-BY-SA by SangyaPundir

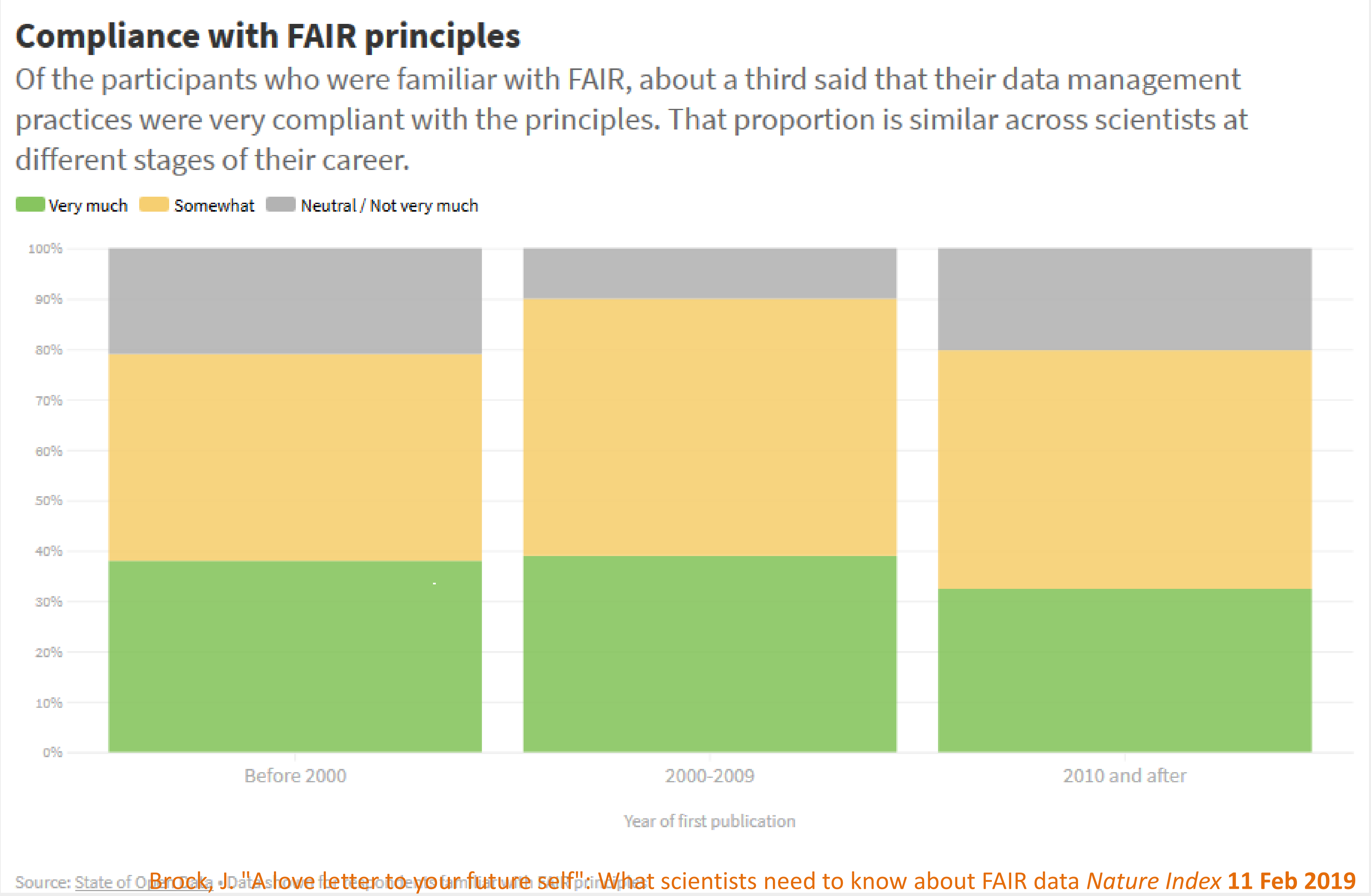


Familiarity with FAIR principles

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

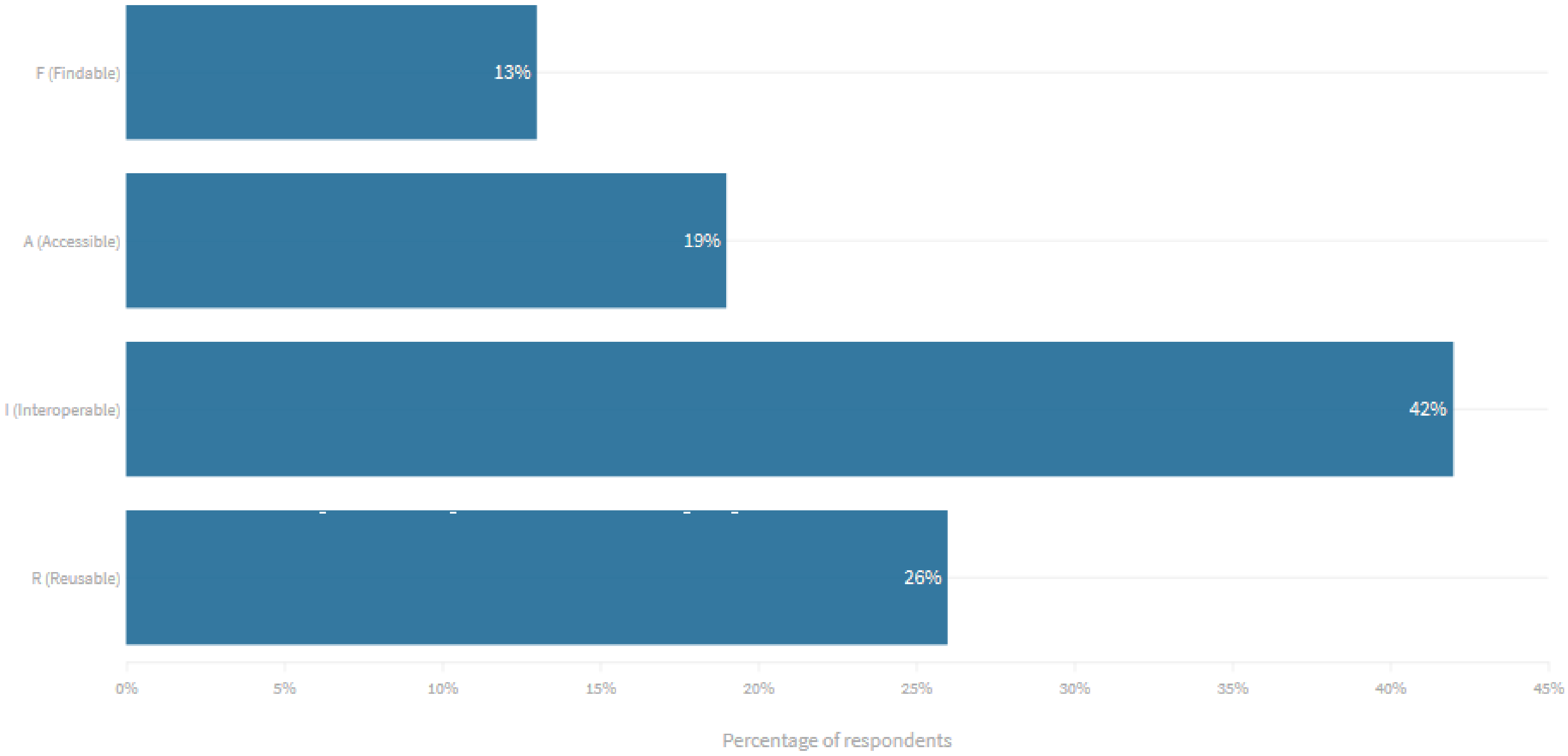
Of the participants who were familiar with FAIR, about a third said that their data management practices were very compliant with the principles. That proportion is similar across scientists at different stages of their career.

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…

European Commission

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

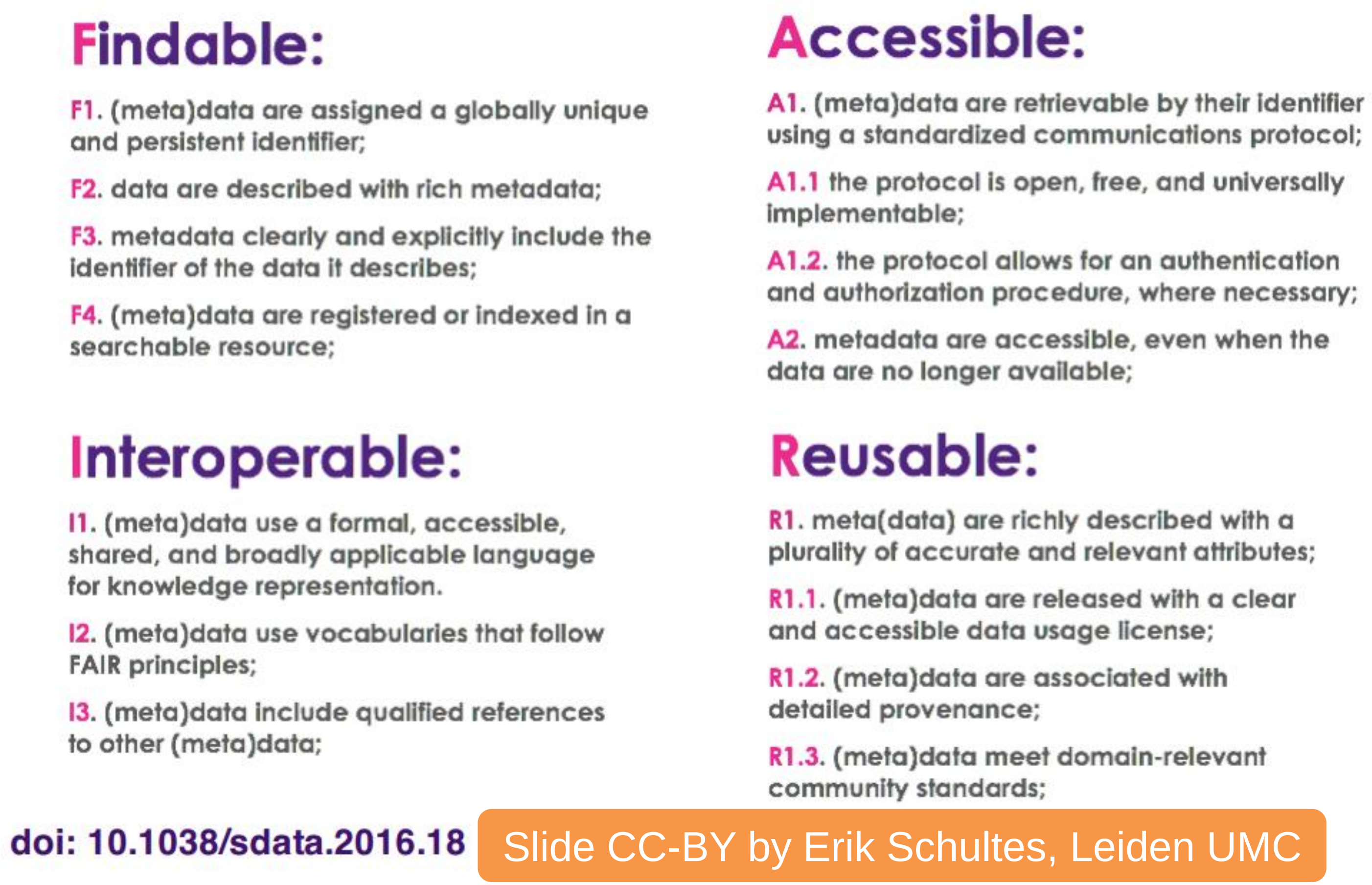
TURNING FAIR INTO REALITY

2018

Research and Innovation

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

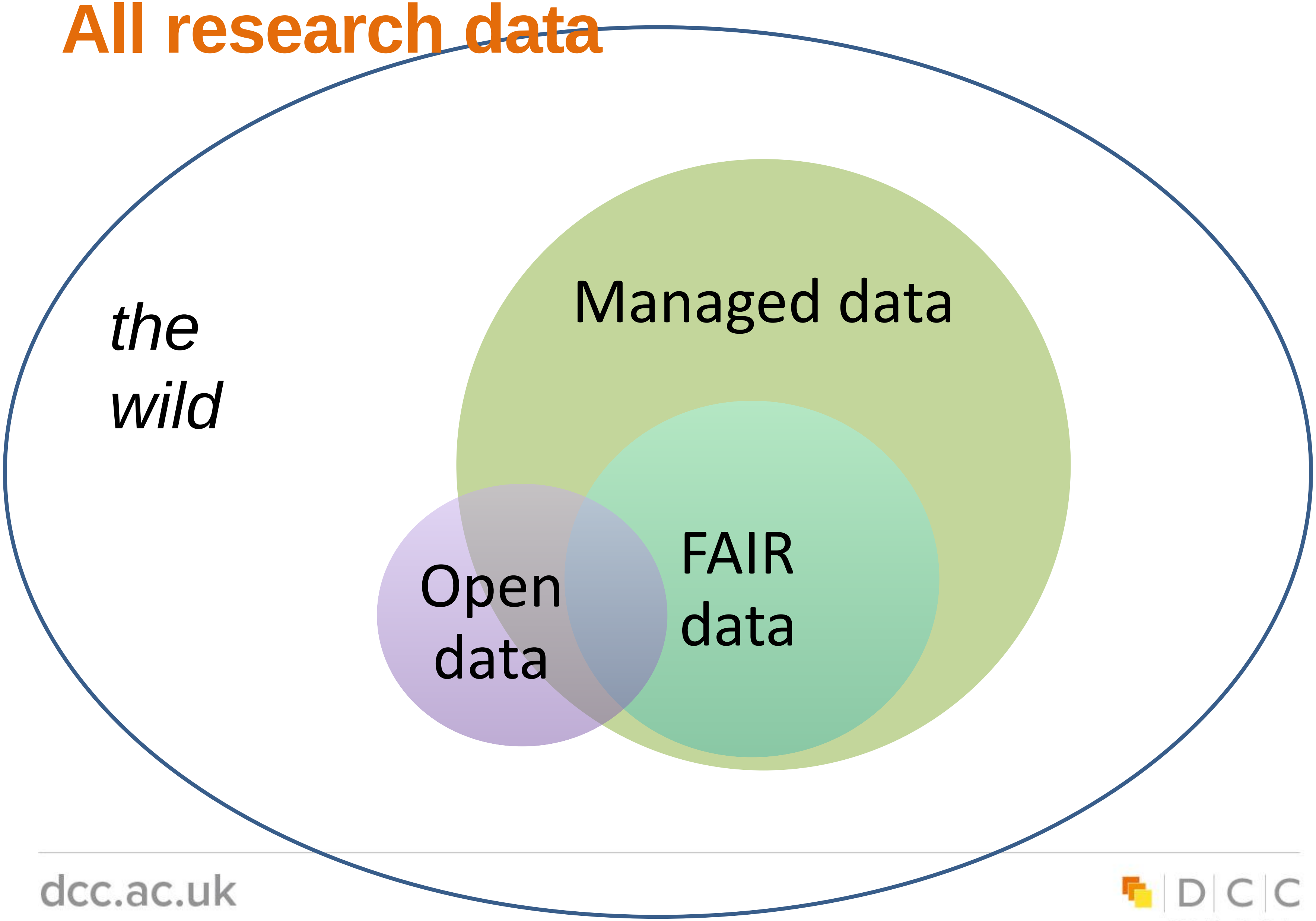
Slide CC-BY by Erik Schultes, Leiden UMC

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

All research data



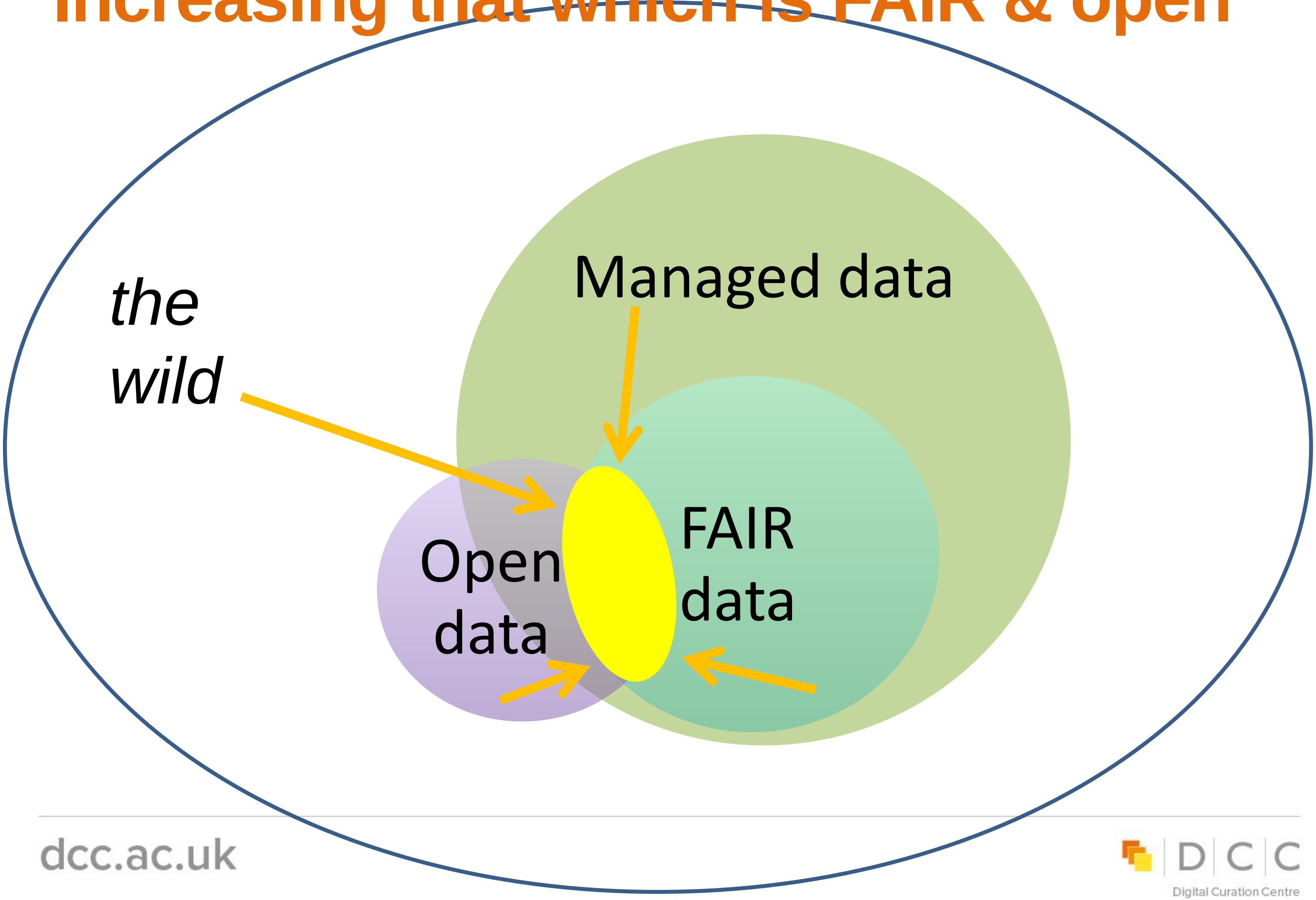
the wild

Managed data

Open data

FAIR data

Increasing that which is 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

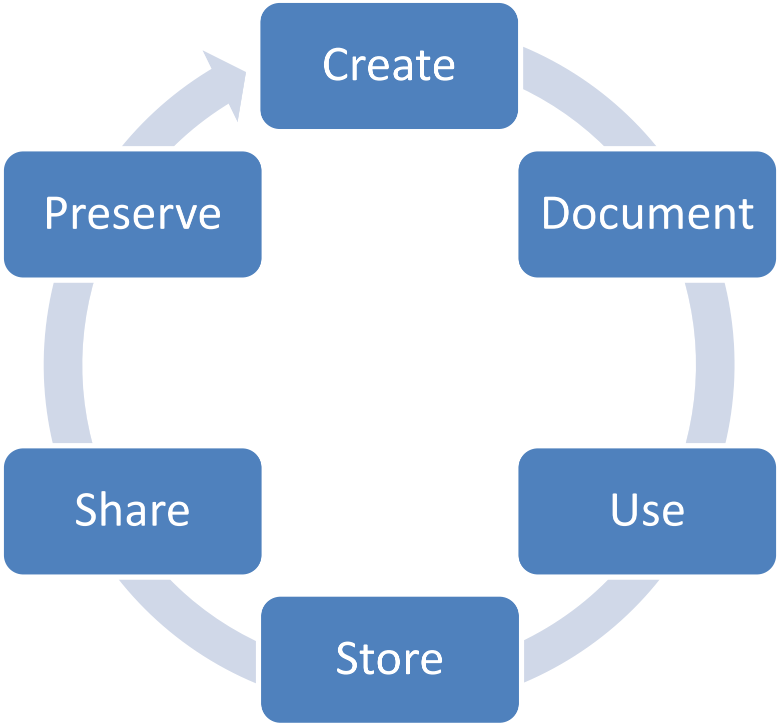
RDM & the Data Lifecycle

Image CC-BY-SA by Janneke Staaks www.flickr.com/photos/jannekestaaks/14411397343

dcc.ac.uk

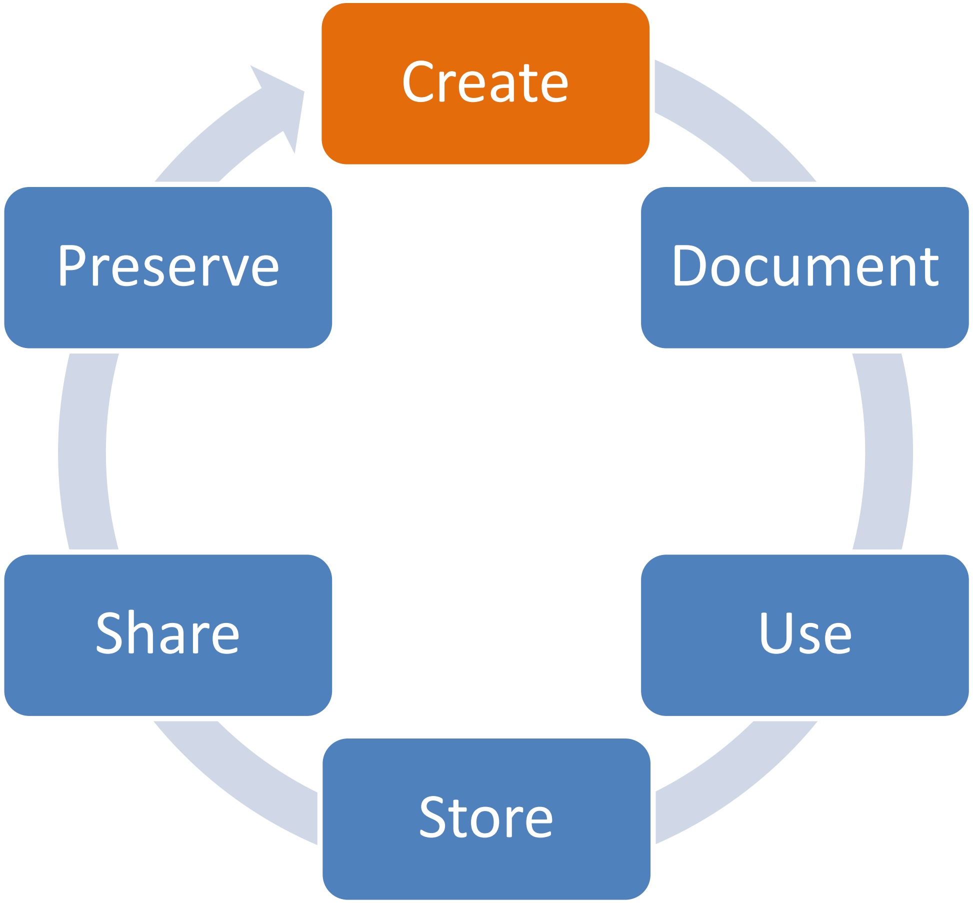
D C C
Digital Curation Centre

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

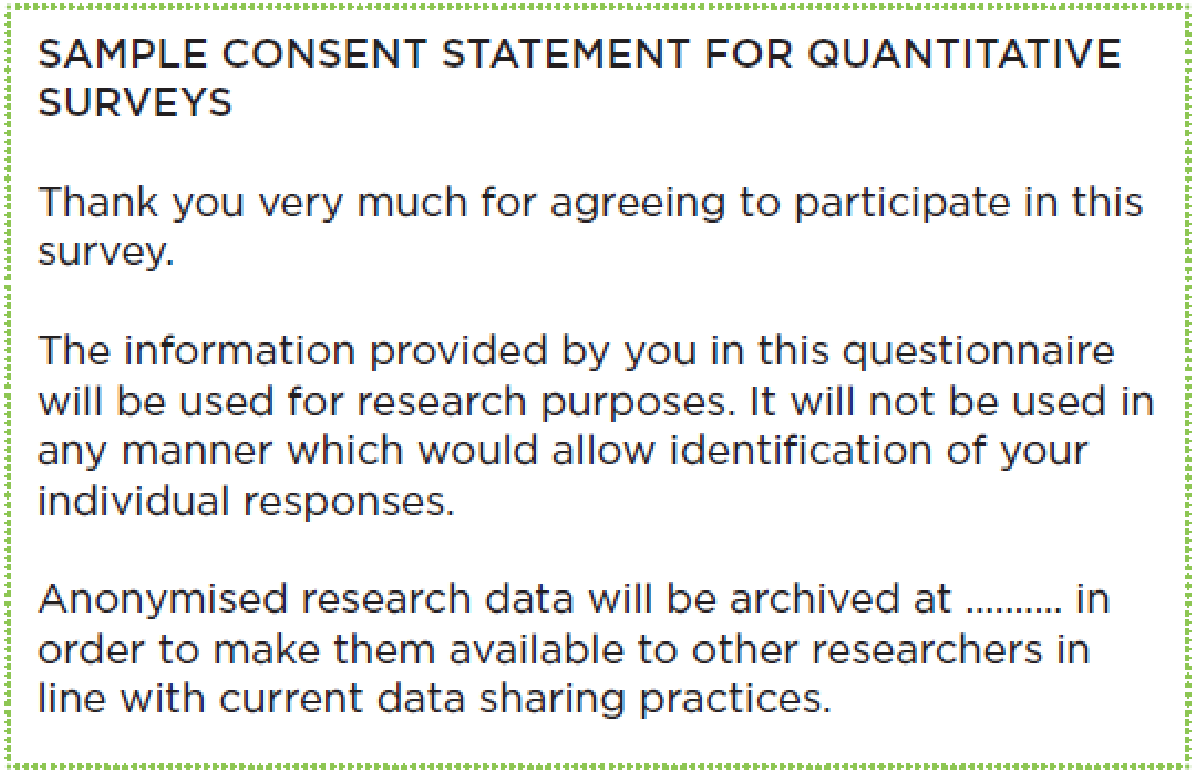

Create
Document
Use
Store
Share
Preserve

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

Data centres may suggest preferred formats for deposit

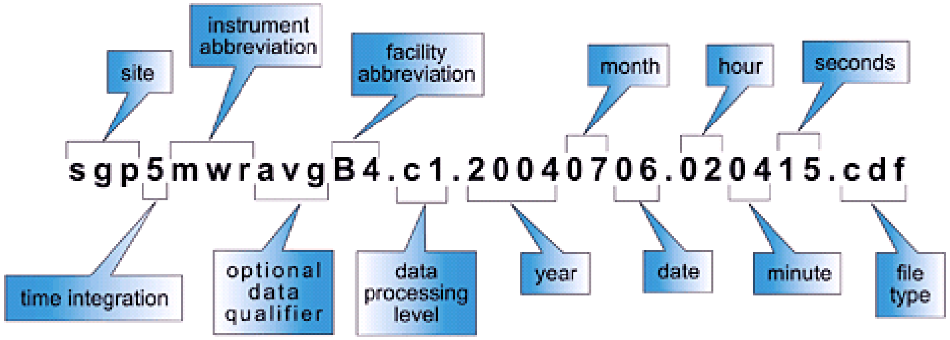
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

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

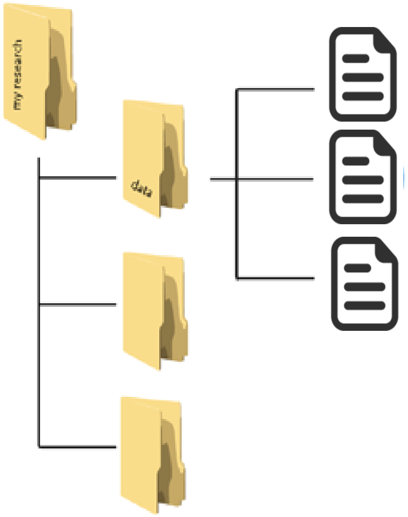
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

How will you organise your data?

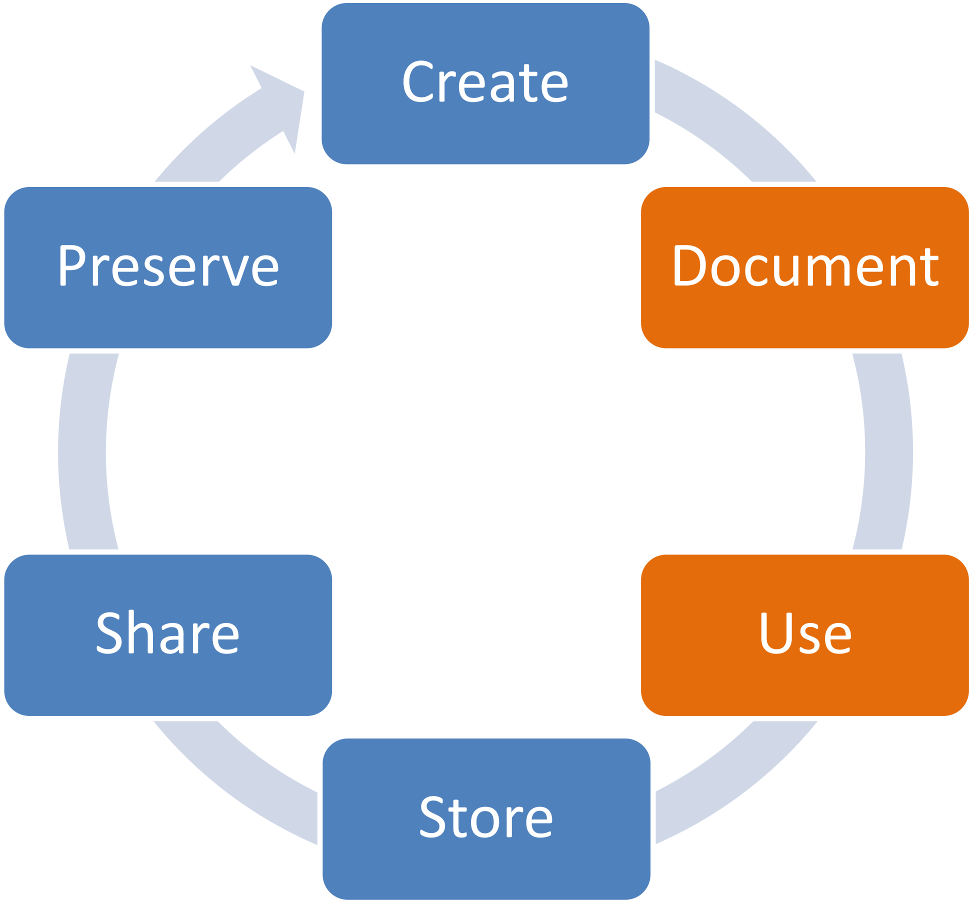
An example netCDF data file name is depicted below:



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



- Keep file and folder names short, but meaningful
- Agree a method for versioning
- Include dates in a set format e.g. YYYYMMDD
- Avoid using non-alphanumeric characters in file names
- Use hyphens or underscores not spaces e.g. day-sheet, day sheet
- Order the elements in the most appropriate way to retrieve the record


Create
Document
Use
Store
Share
Preserve

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?

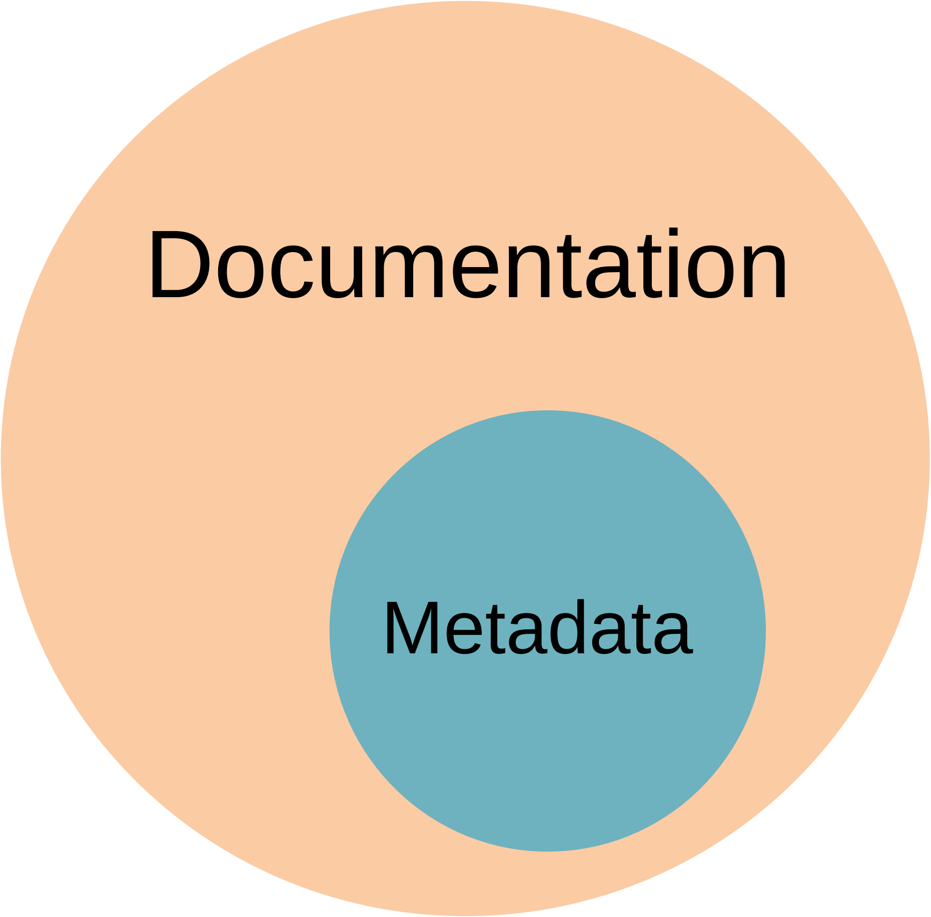
Metadata

- Standardised
- Structured
- Machine and human readable

Metadata helps to cite &

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

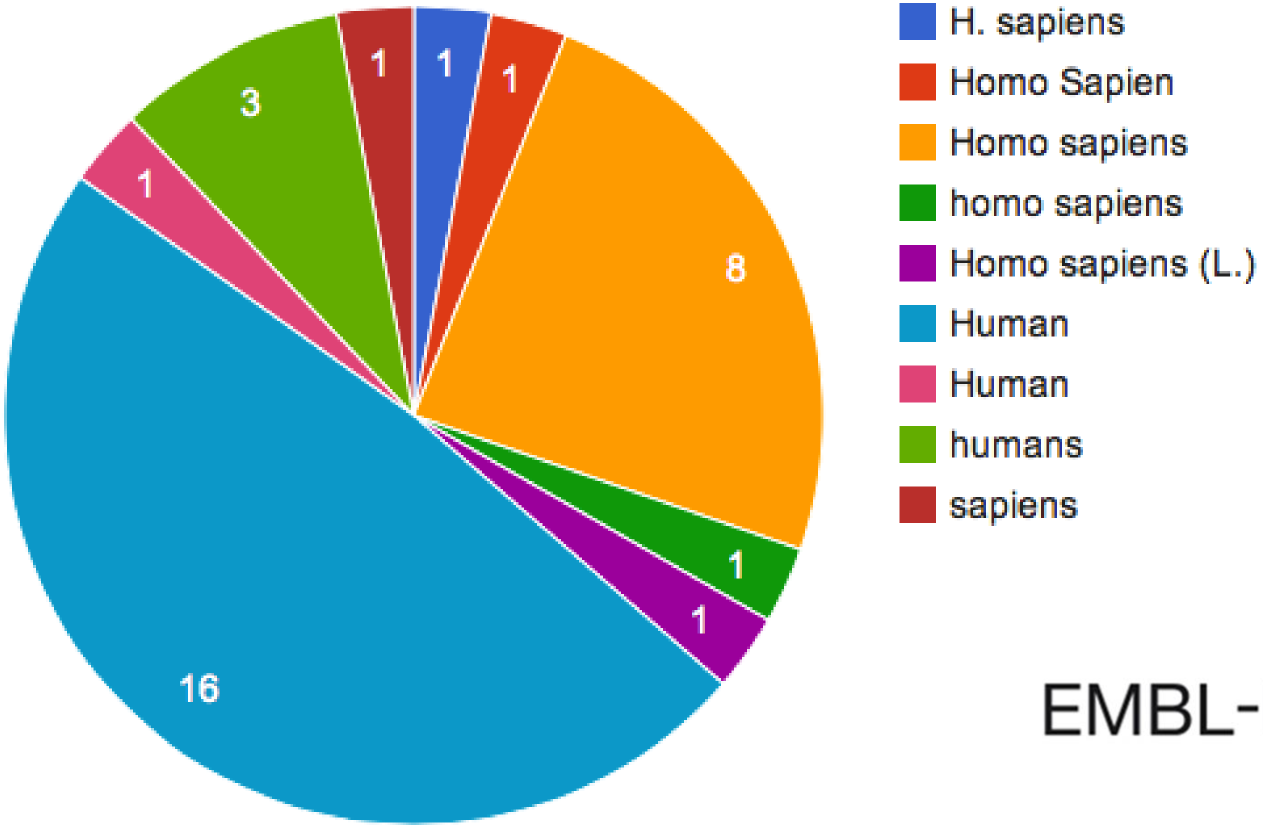
Search for standards in catalogues like:

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……”



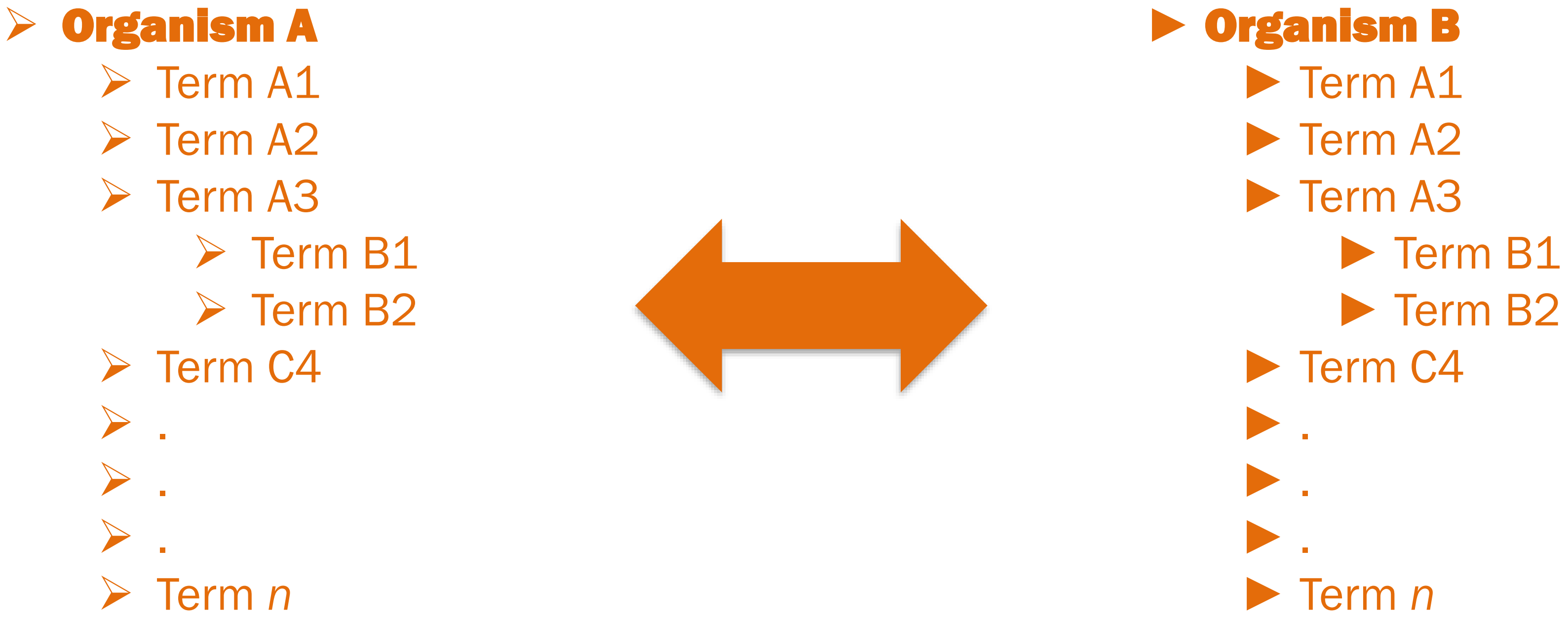
EMBL-EBI

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

…and ontologies?

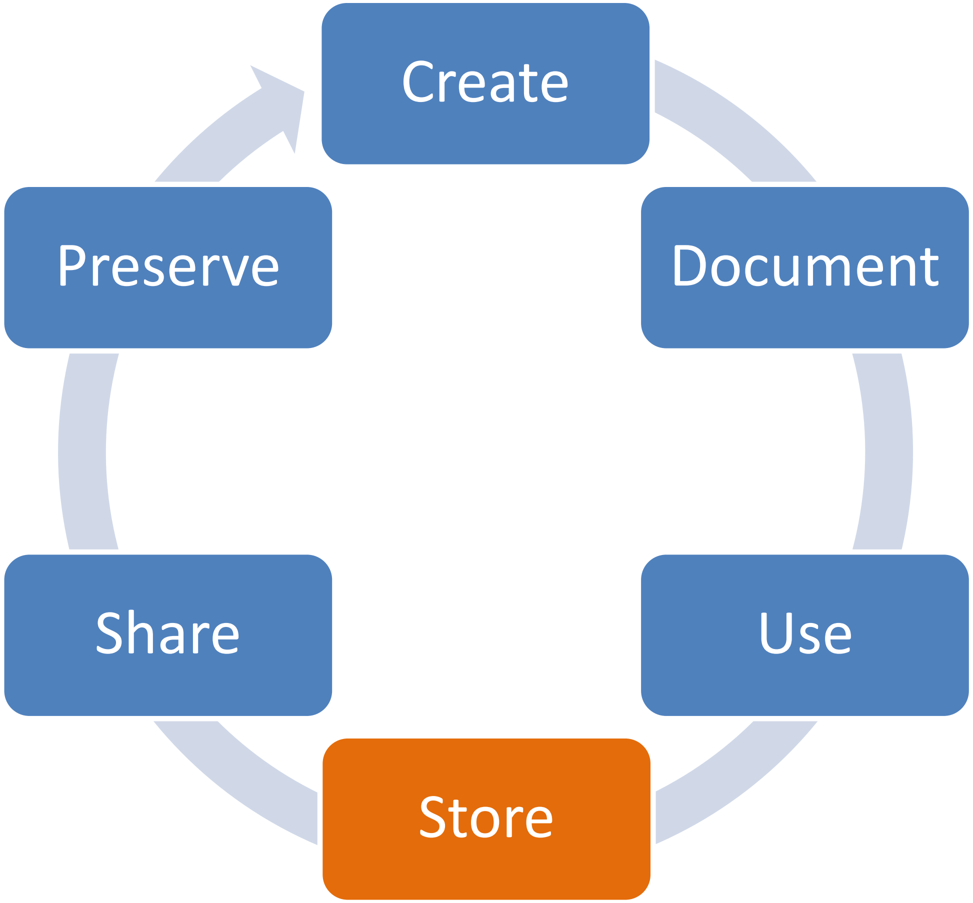
e.g. SNOMED CT (clinical terms) or MeSH

- 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*

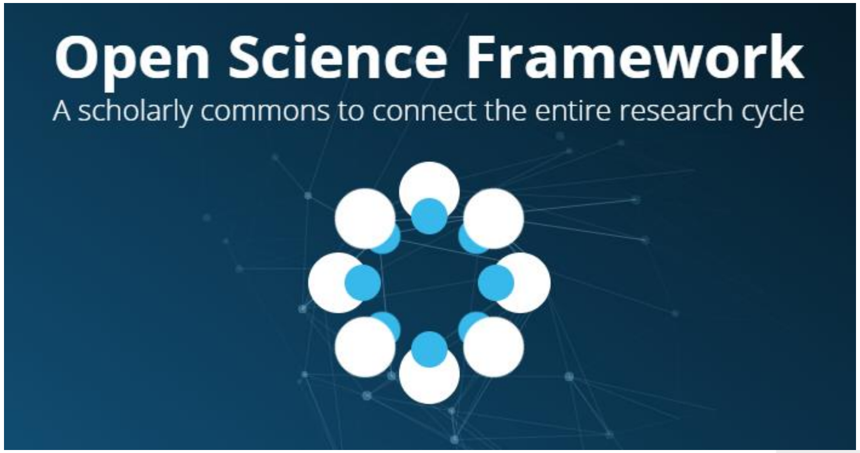

Create
Document
Use
Store
Share
Preserve

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?

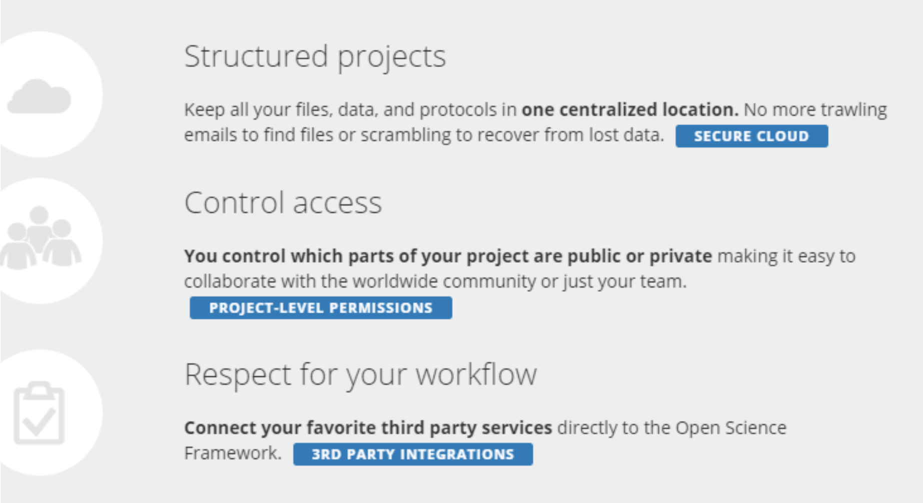
The decision will be based on how sensitive your data are, how robust you need the storage to be, and who needs access to the data and when

Collaborative platforms e.g. OSF



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.

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.

Respect for your workflow

Connect your favorite third party services directly to the Open Science Framework.

https://osf.io

Third-party tools for collaboration

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

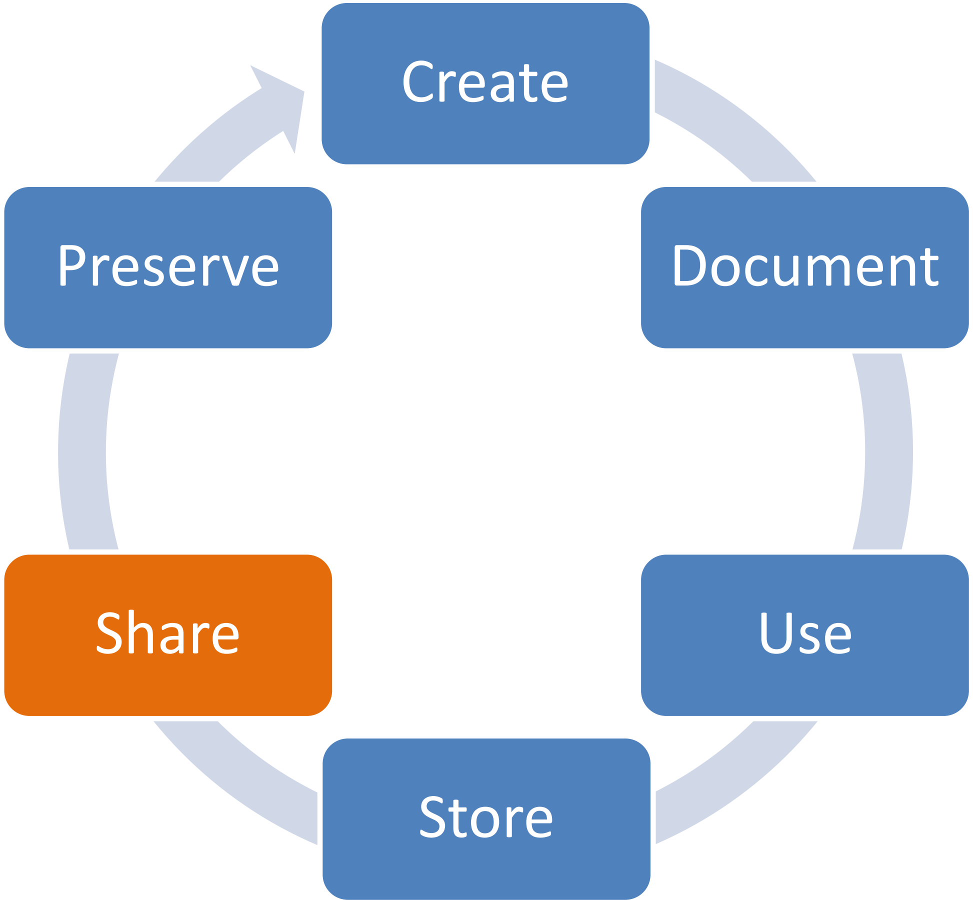
Backup and preservation – not the same thing!

Backups

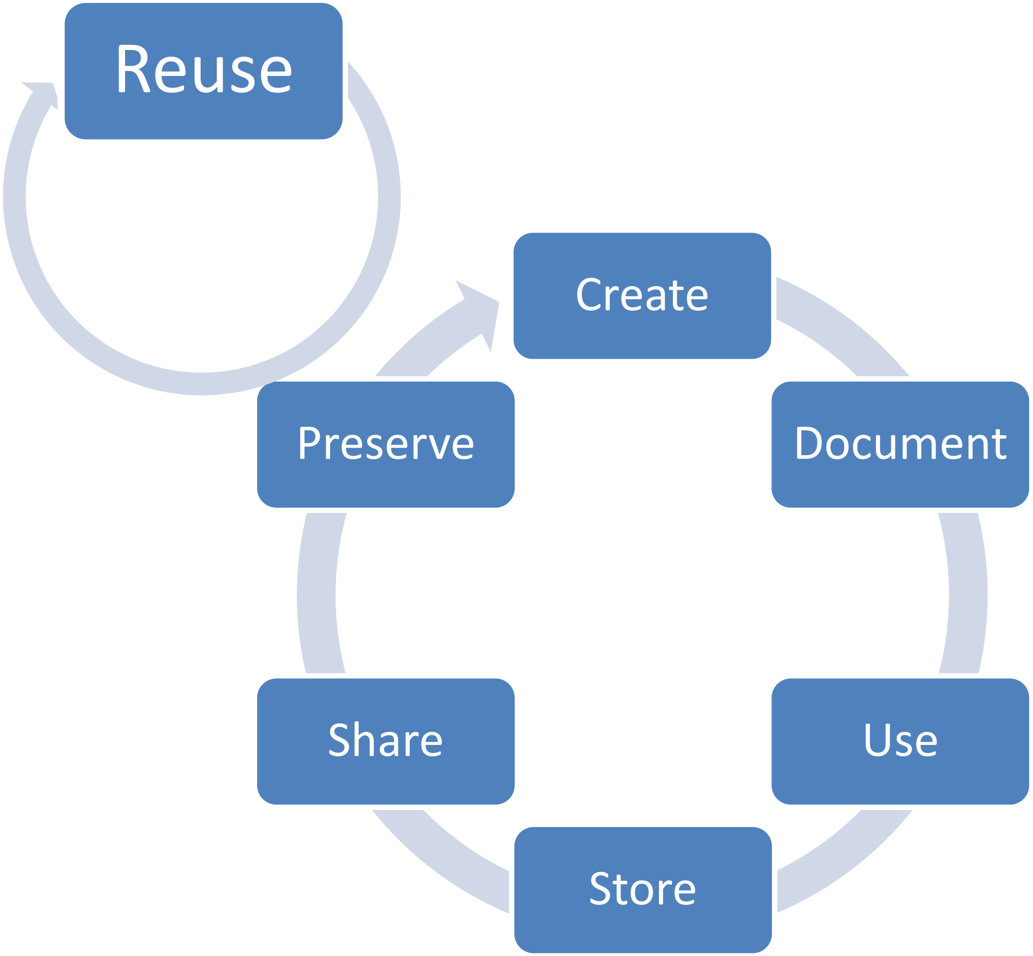
- Used to take periodic snapshots of data in case the current version is destroyed or lost
- Backups are copies of files stored for short or near-long-term
- Often performed on a somewhat frequent schedule

Archiving

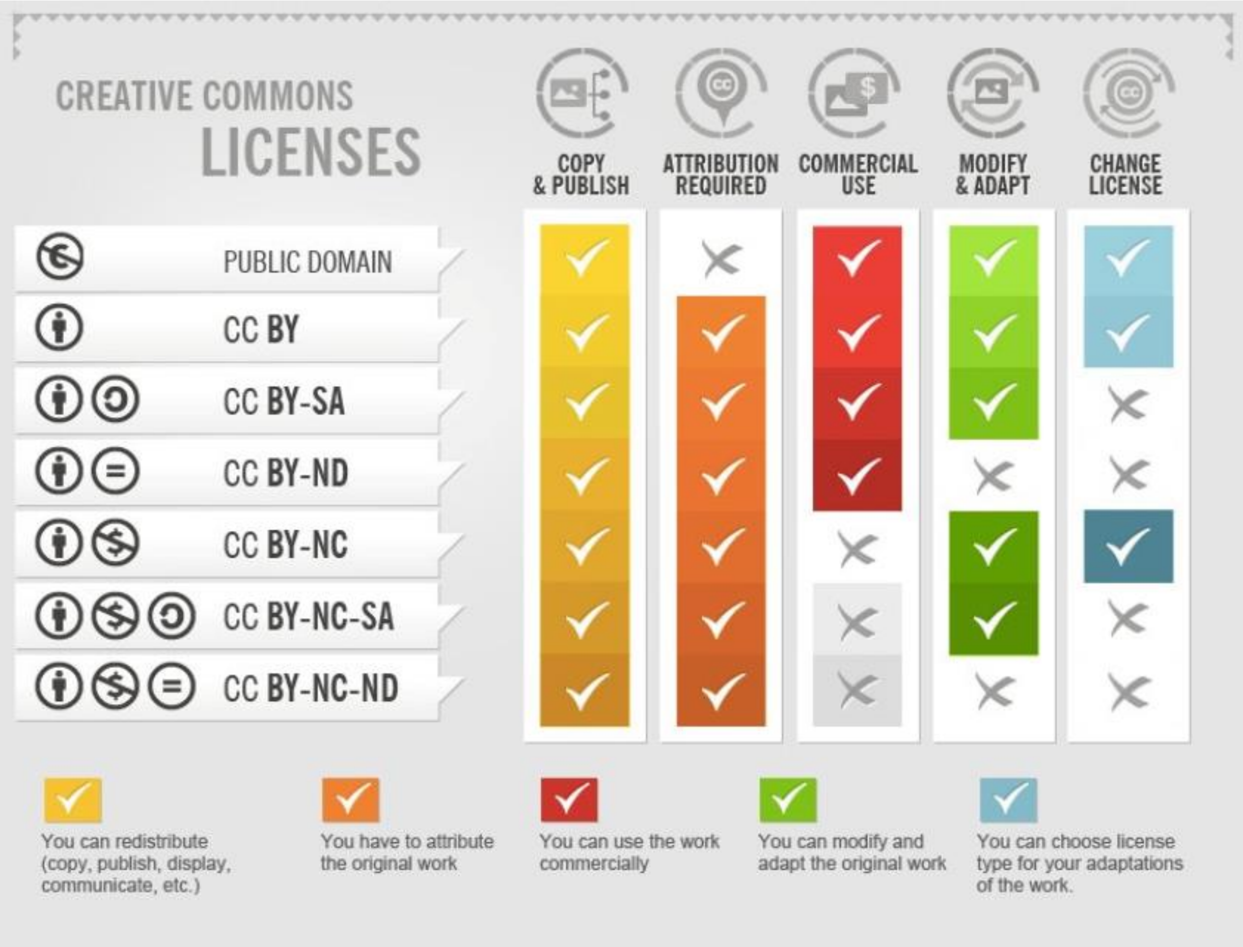
- Used to preserve data for historical reference or potentially during disasters
- Archives are usually the final version, stored for long-term, and generally not copied over
- Often performed at the end of a project or during major milestones


Create
Document
Use
Store
Share
Preserve

Primary and secondary data



License research data openly



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

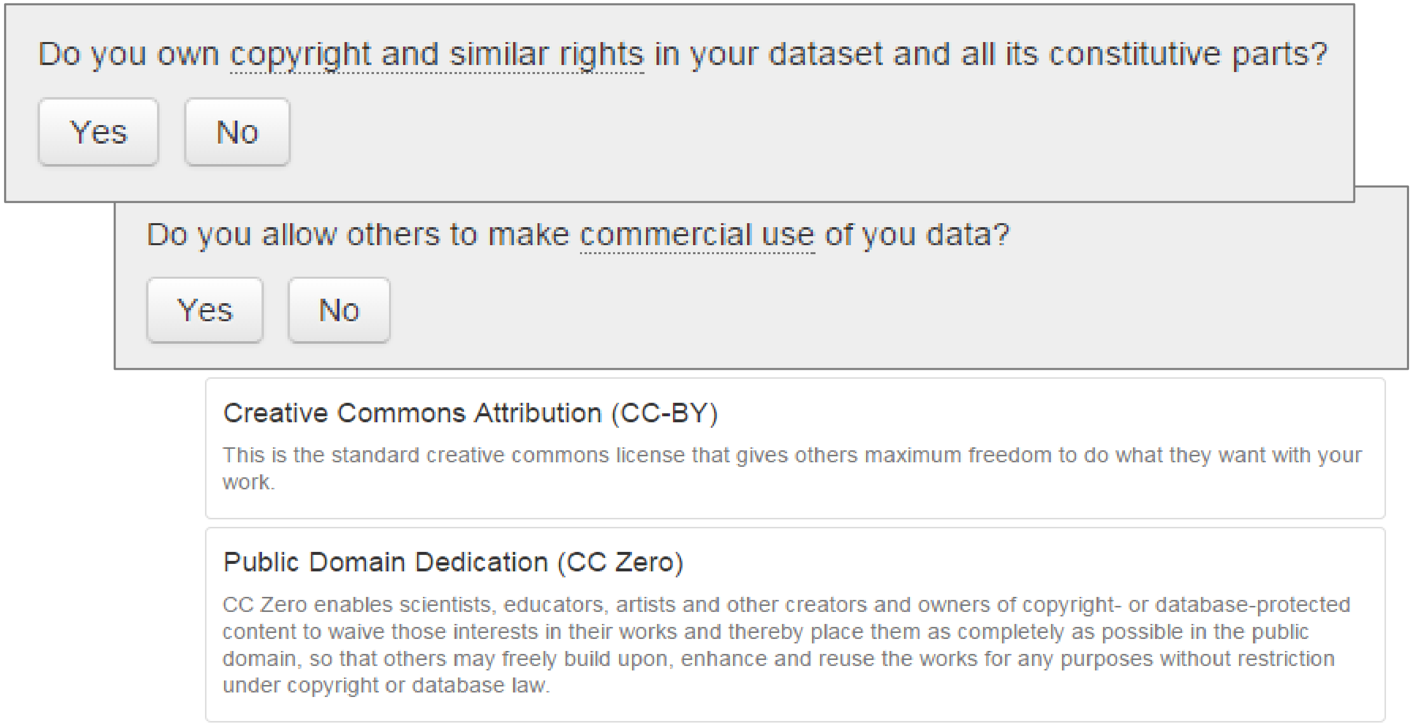
✓ You can modify and adapt the original work

✓ 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

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?

Do you allow others to make commercial use of you data?

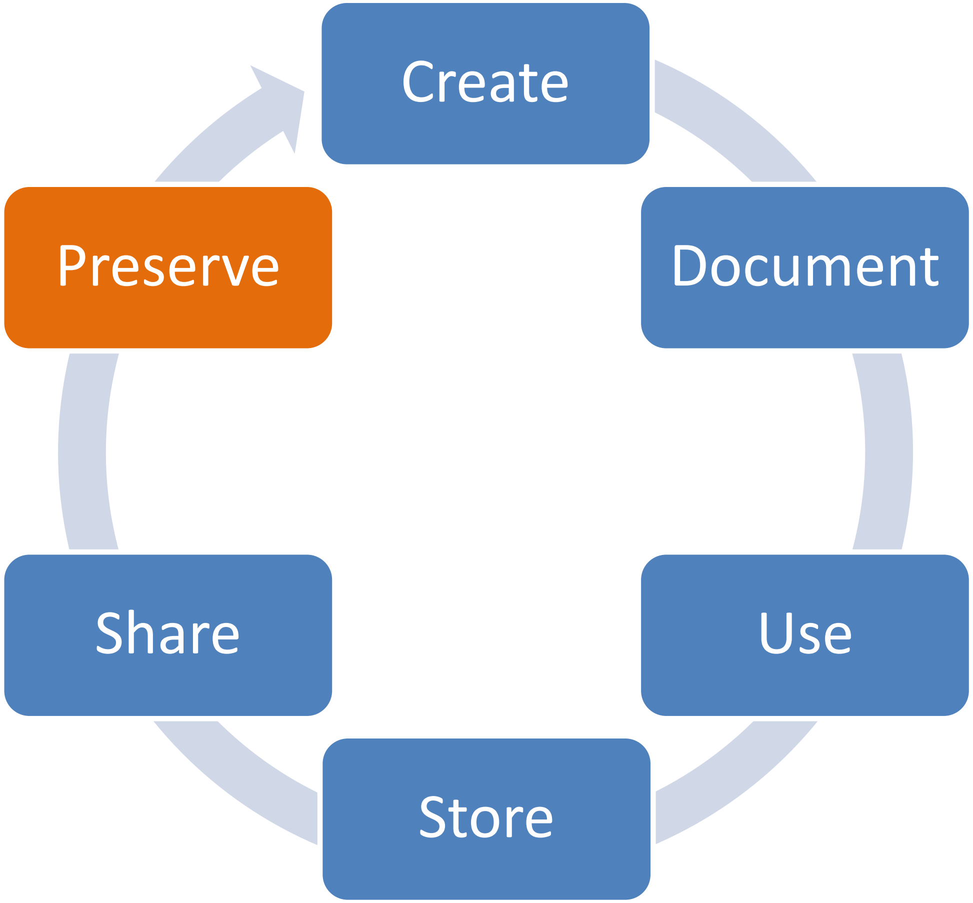
Creative Commons Attribution (CC-BY)

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)

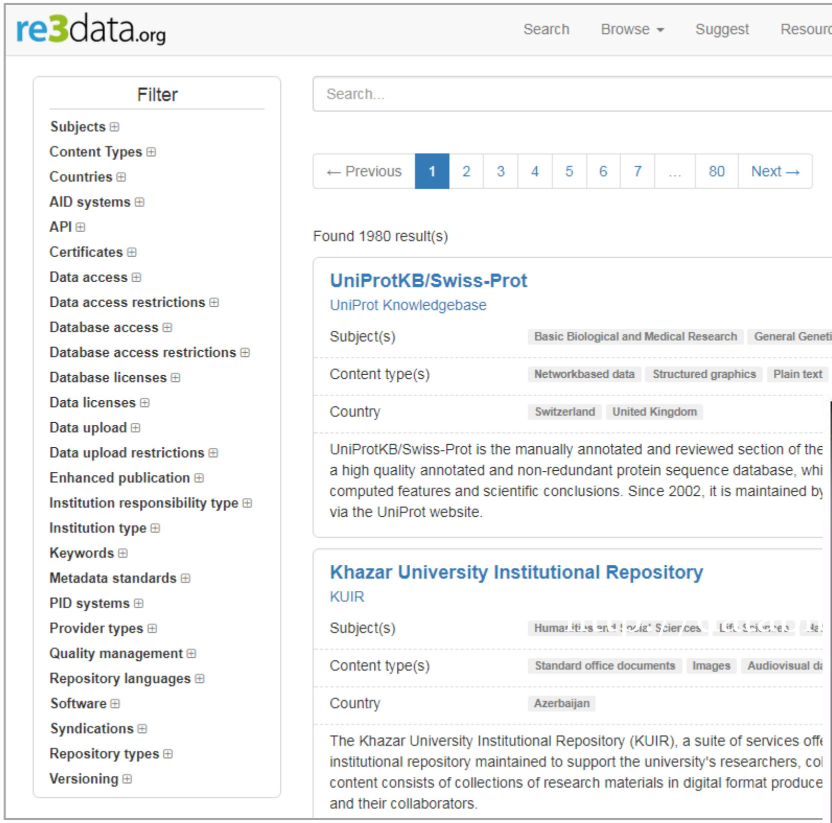
CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

https://ufal.github.io/public-license-selector/


Create
Document
Use
Store
Share
Preserve

Deposit in a data repository

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



re3data.org Search Browse Suggest Resources Contact DataCite

Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- AID systems ⊞
- API ⊞
- Certificates ⊞
- Data access ⊞
- Data access restrictions ⊞
- Database access ⊞
- Database access restrictions ⊞
- Database licenses ⊞
- Data licenses ⊞
- Data upload ⊞
- Data upload restrictions ⊞
- Enhanced publication ⊞
- Institution responsibility type ⊞
- Institution type ⊞
- Keywords ⊞
- Metadata standards ⊞
- PID systems ⊞
- Provider types ⊞
- Quality management ⊞
- Repository languages ⊞
- Software ⊞
- Syndications ⊞
- Repository types ⊞
- Versioning ⊞

Search... Search

Toogle short help

← Previous 1 2 3 4 5 6 7 … 80 Next → Sort by

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 by via the UniProt website.

Khazar University Institutional Repository

KUIR

Subject(s)	Humanities and Social Sciences · Life Sciences
Content type(s)	Standard office documents · Images · Audiovisual da
Country	Azerbaijan

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

www.re3data.org

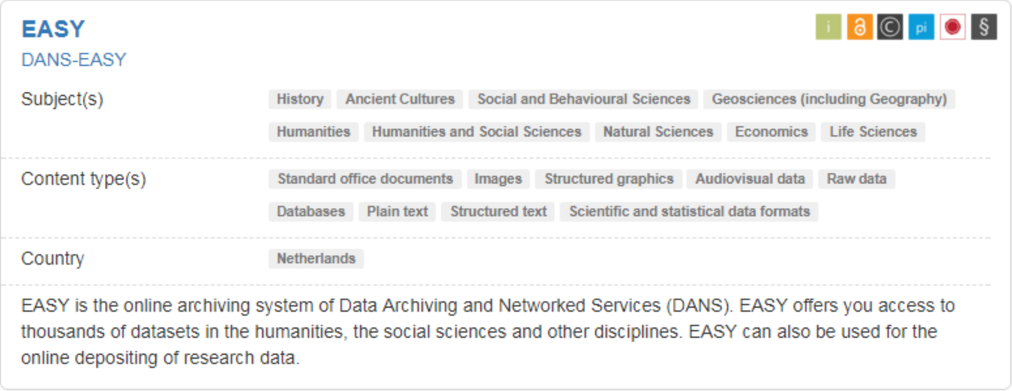
www.fosteropenscience.eu/content/re3data-demo

dcc.ac.uk

DCC Digital Curation Centre

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

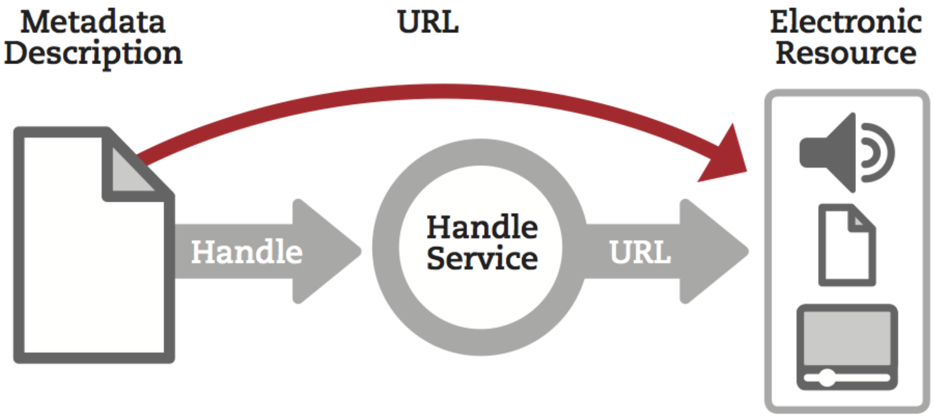
EASY is the online archiving system of Data Archiving and Networked Services (DANS). EASY offers you access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data.

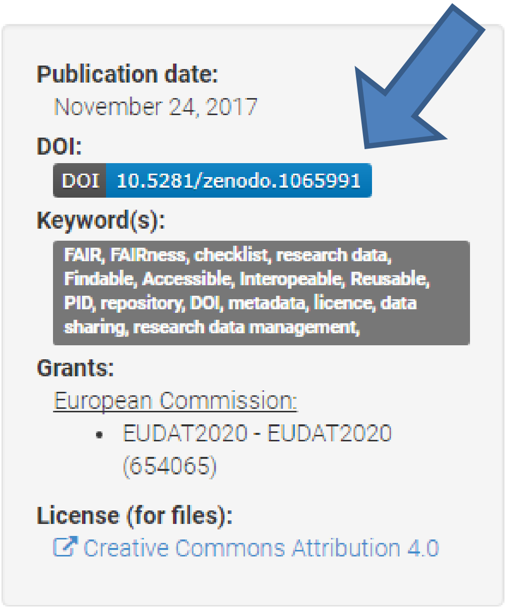
Icons to note open access, licenses, PIDs, certificates…

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

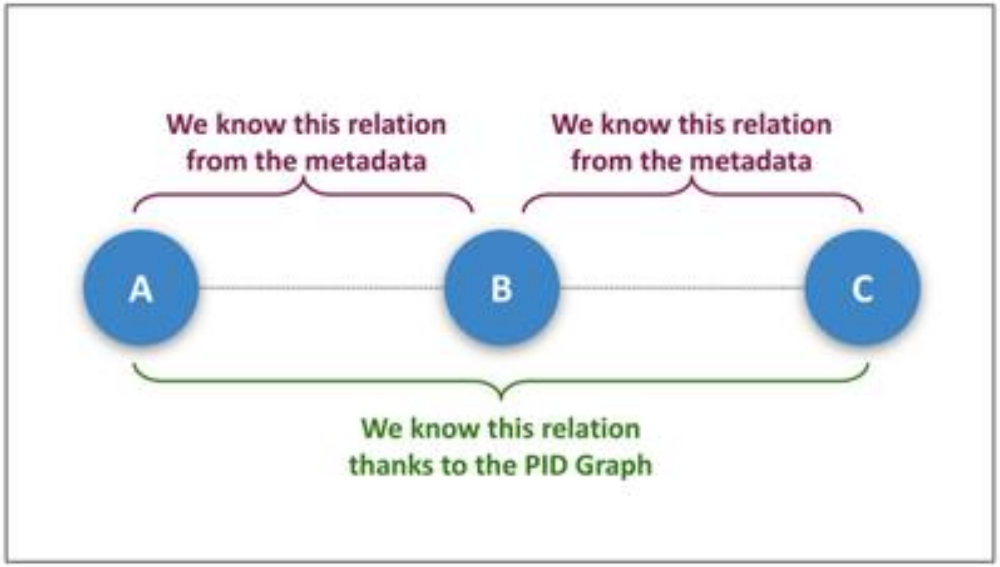
Keyword(s):
FAIR, FAIRness, checklist, research data, Findable, Accessible, Interopeable, Reusable, PID, repository, DOI, metadata, licence, data sharing, research data management,

Grants:
European Commission:
- EUDAT2020 - EUDAT2020 (654065)

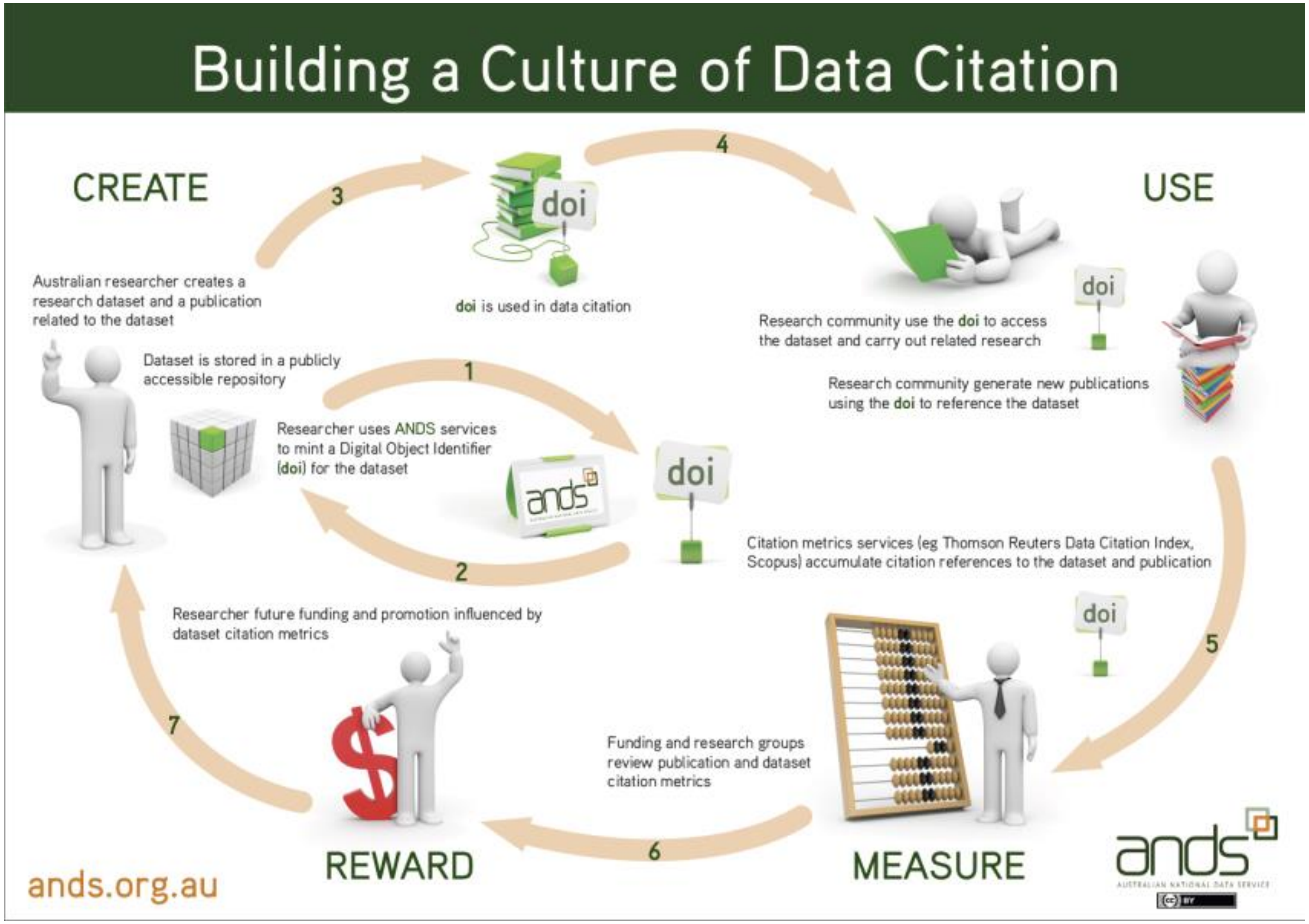
License (for files):
Creative Commons Attribution 4.0

PID Graphs – the next level

- If you have a collection of PIDs describing different objects, these can be joined together in a graph to form relationships
- These graphs can aid in workflows and provenance


We know this relation from the metadata
We know this relation from the metadata
A
B
C
We know this relation thanks to the PID Graph

Citing research data: why?



Building a Culture of Data Citation

Australian researcher creates a research dataset and a publication related to the dataset

Dataset is stored in a publicly accessible repository

Researcher uses ANDS services to mint a Digital Object Identifier (doi) for the dataset

doi is used in data citation

Research community use the doi to access the dataset and carry out related research

Research community generate new publications using the doi to reference the dataset

Citation metrics services (eg Thomson Reuters Data Citation Index, Scopus) accumulate citation references to the dataset and publication

Funding and research groups review publication and dataset citation metrics

Researcher future funding and promotion influenced by dataset citation metrics

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

Questions?

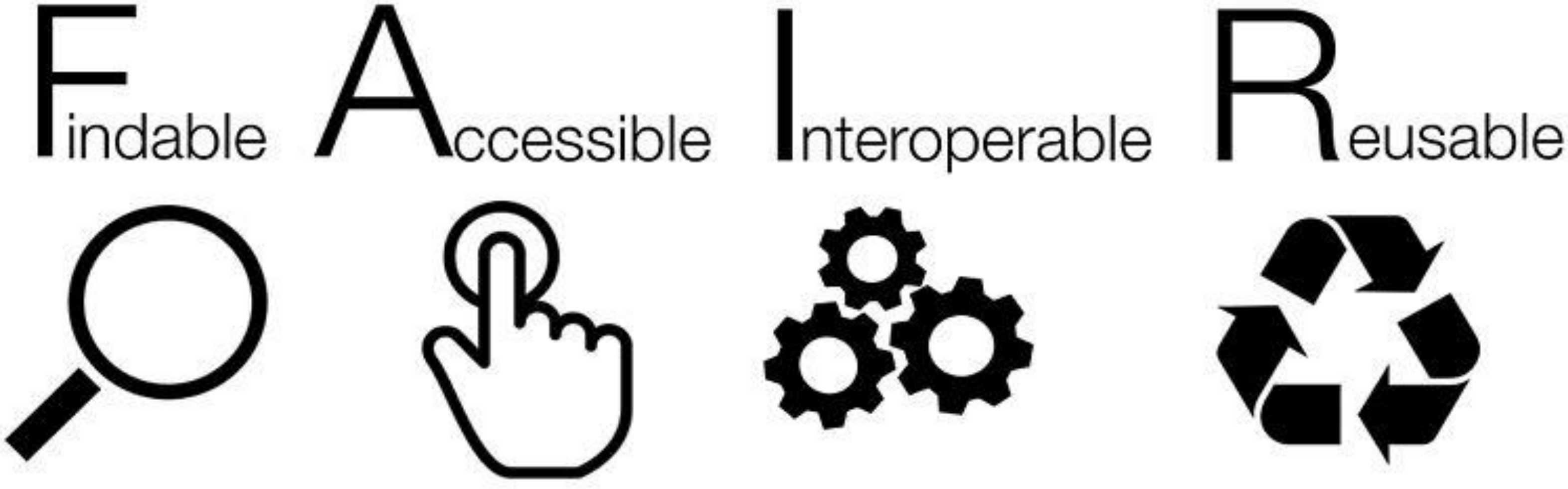


Image CC-BY-SA by SangyaPundir

Digital Curation Centre

Introduction to Data Management Plans

S. Venkataraman, PhD
Research Data Specialist
Digital Curation Centre

s.venkataraman@ed.ac.uk

3rd December 2019, Universidad de Costa Rica

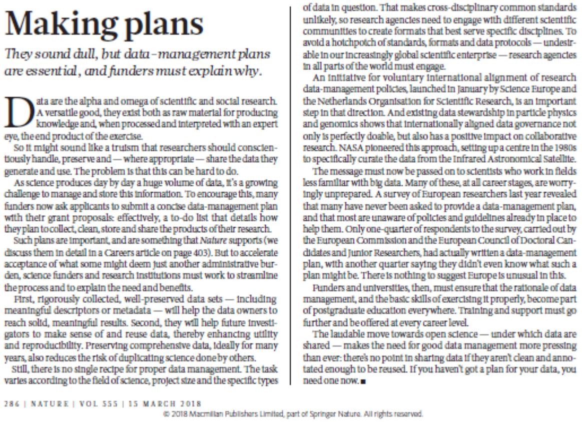
What is a data management plan (DMP)?

A brief plan written at the start of a project to define:

- how the data will be created?
- how it will be documented?
- who will access it?
- where it will be stored?
- who will back it up?
- whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data.

Why make DMPs?



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.

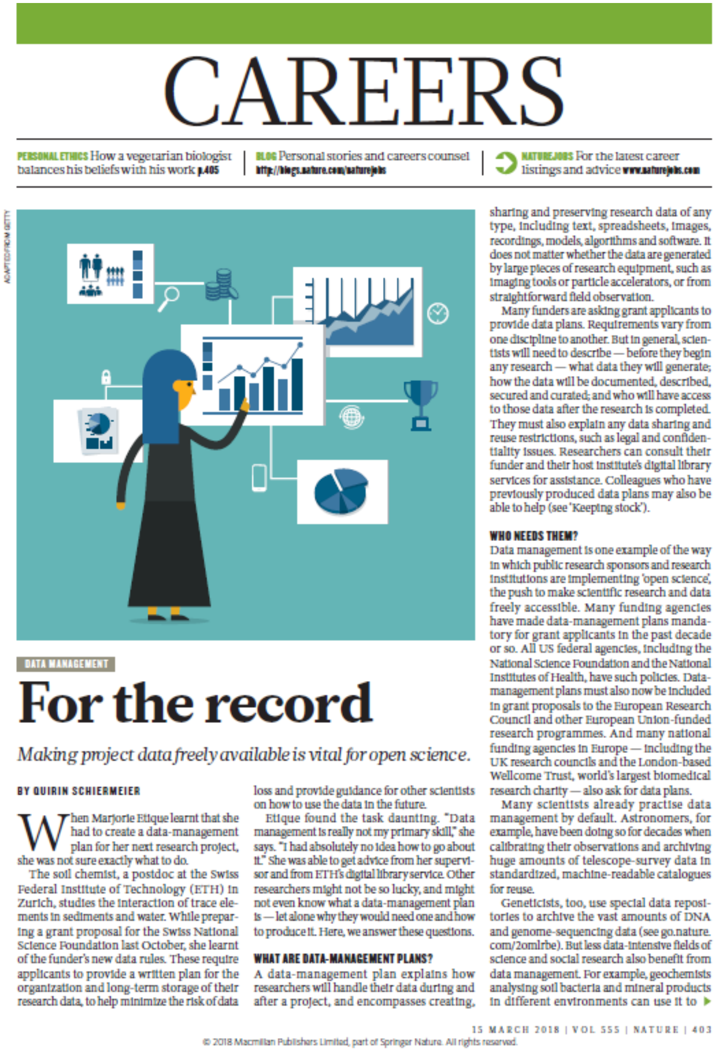
An initiative for voluntary international alignment of research data-management policies, launched in January by Science Europe and the Netherlands Organisation for Scientific Research, is an important step in that direction. And existing data stewardship in particle physics and genomics shows that internationally 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. ■

286 | NATURE | VOL 555 | 15 MARCH 2018



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 ▶

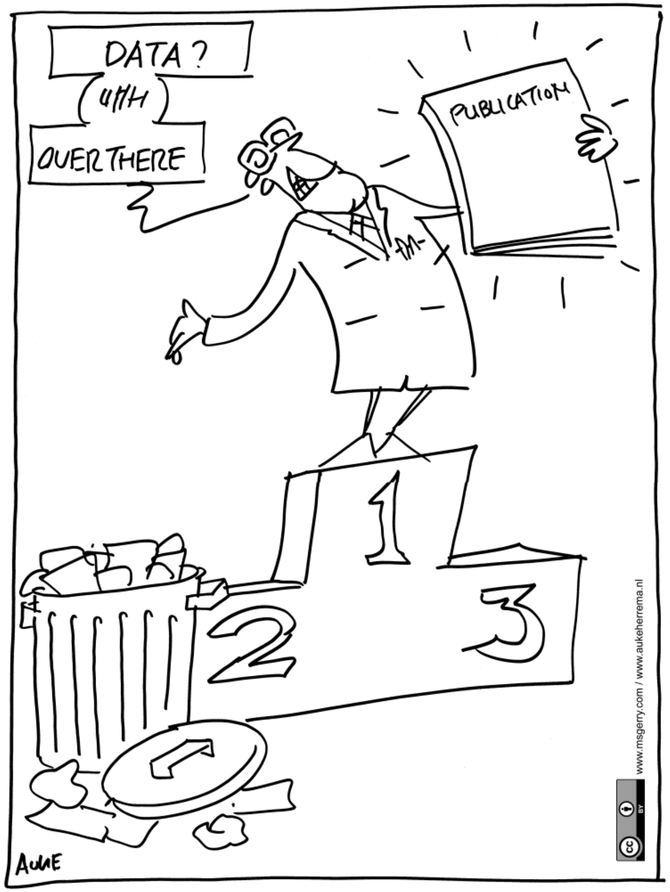
15 MARCH 2018 | VOL 555 | NATURE | 403

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

Why make DMPs?

- 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!

Don’t undervalue research data



PUBLICATIONS AND DATA

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

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

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)

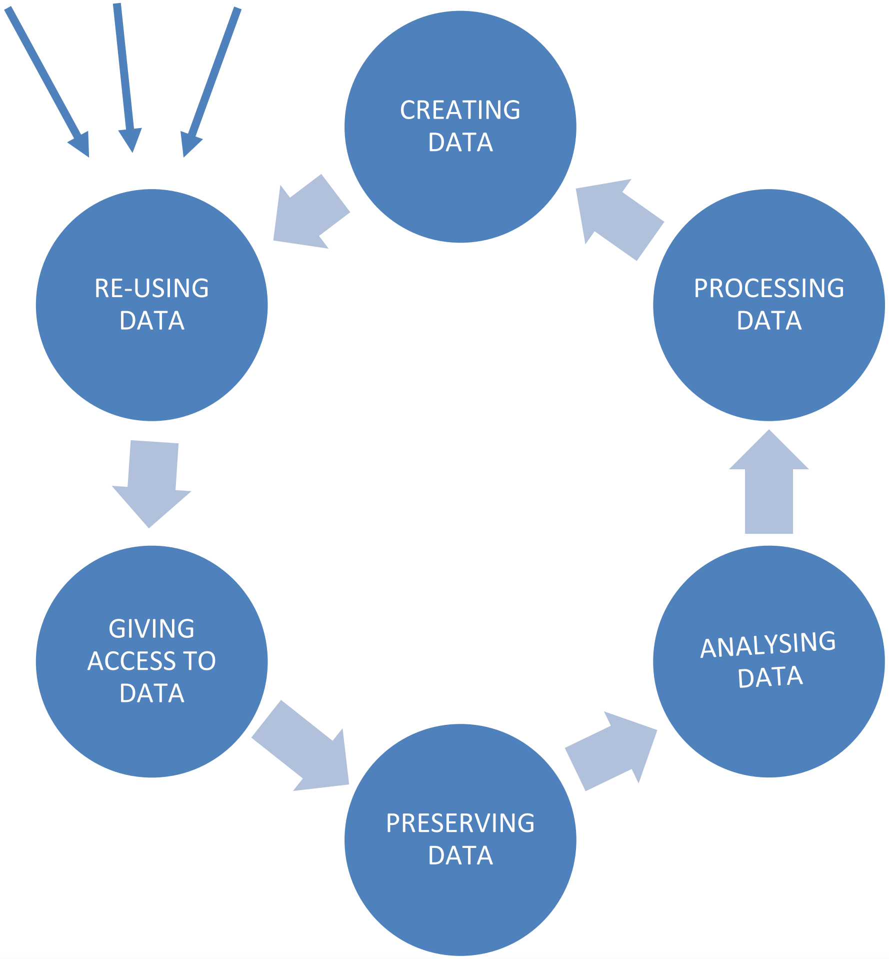
4. Plans for data sharing and access

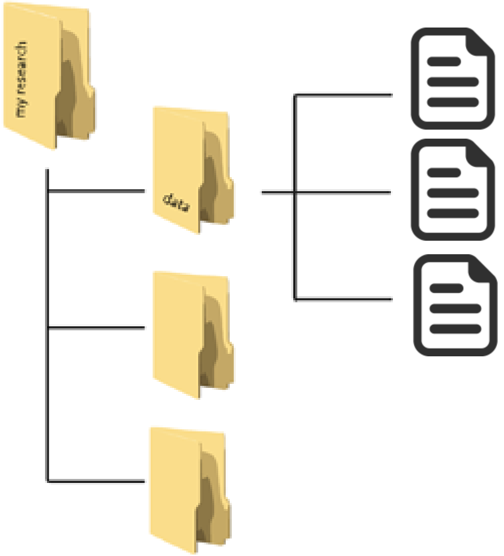
   (i.e. how, when, to whom)

5. Strategy for long-term preservation

Planning trick 1: think backwards

What data organisation would a re-user like?



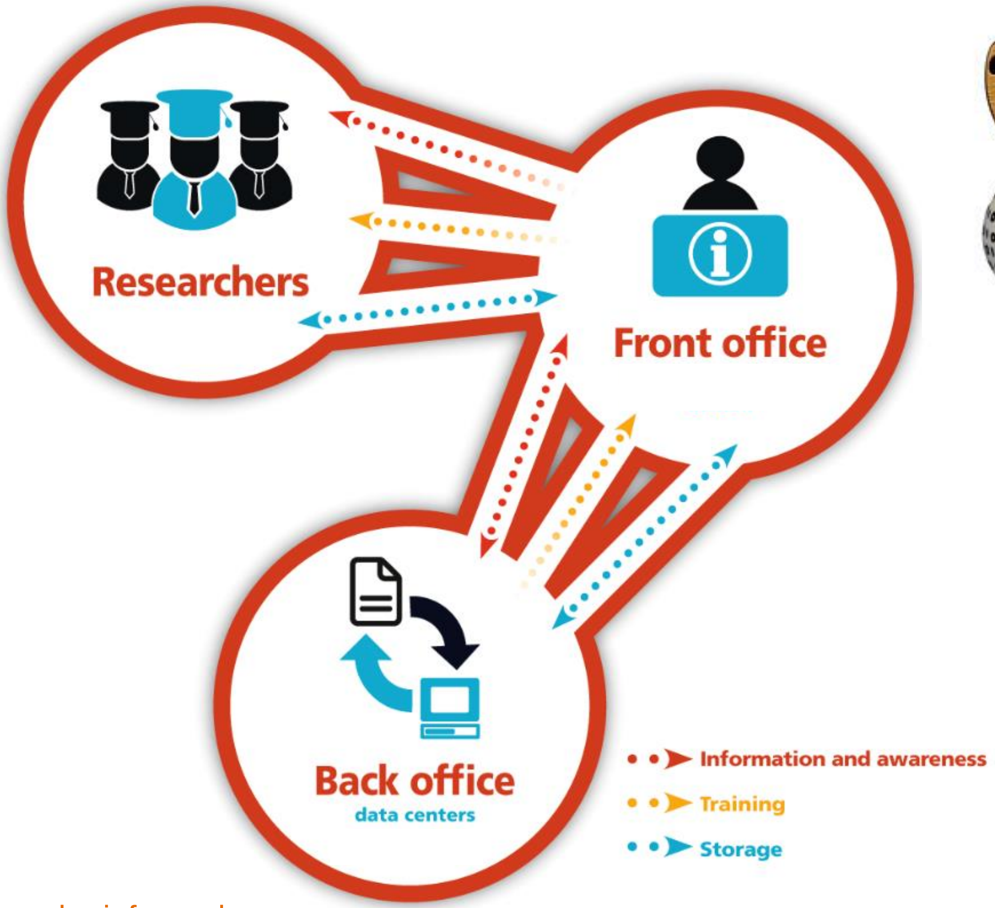


Design how you will organise data in the project (folder structure, file naming convention, …)

Planning trick 2: include RDM stakeholders

Commercial partners

Publishers
Data
Availability
policy



Institution
RDM policy
Facilities

€$£

Research
funders

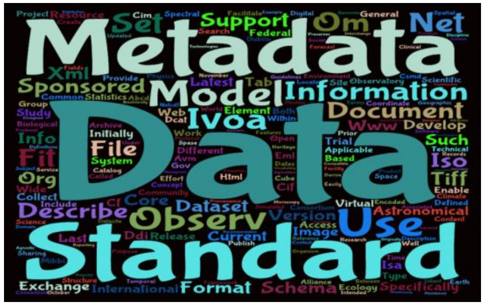
www.openaire.eu/briefpaper-rdm-infonoads

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

Data description examples

The final dataset will include self-reported demographic and behavioural data from interviews with the subjects and laboratory data from urine specimens provided.

From NIH data sharing statements

Every two days, we will subsample E. affinis populations growing under our treatment conditions. We will use a microscope to identify the life stage and sex of the subsampled individuals. We will document the information first in a laboratory notebook and then copy the data into an Excel spreadsheet. The Excel spreadsheet will be saved as a comma separated value (.csv) file.

From DataOne – E. affinis DMP example

Metadata examples

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

From ICPSR Framework for Creating a DMP

We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a .txt document that will be stored with the data file. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. EML is one of the accepted formats used in ecology, and works well for the types of data we will be producing. We will create these metadata using Morpho software, available through KNB. The metadata will fully describe the data files and the context of the measurements.

From DataOne – E. affinis DMP example

Data sharing examples

The videos will be made available via the bristol.ac.uk website (both as streaming media and downloads) HD and SD versions will be provided to accommodate those with lower bandwidth. Videos will also be made available via Vimeo, a platform that is already well used by research students at Bristol. Appropriate metadata will also be provided to the existing Vimeo standard.

All video will also be available for download and re-editing by third parties. To facilitate this Creative Commons licenses will be assigned to each item. In order to ensure this usage is possible, the required permissions will be gathered from participants (using a suitable release form) before recording commences.

From University of Bristol Kitchen Cosmology DMP

We will make the data and associated documentation available to users under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

From NIH data sharing statements

Examples restrictions

Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement.

From NIH data sharing statements

1. Share data privately within 1 year.
   Data will be held in Private Repository, but metadata will be public
2. Release data to public within 2 years.
   Encouraged after one year to release data for public access.
3. Request, in writing, data privacy up to 4 years.
   Extensions beyond 3 years will only be granted for compelling cases.
4. Consult with creators of private CZO datasets prior to use.
   Pis required to seek consent before using private data they can access

From Boulder Creek Critical Zone Observatory DMP

Archiving examples

The investigators will work with staff at the UKDA to determine what to archive and how long the deposited data should be retained. Future long-term use of the data will be ensured by placing a copy of the data into the repository.

From ICPSR Framework for Creating a DMP

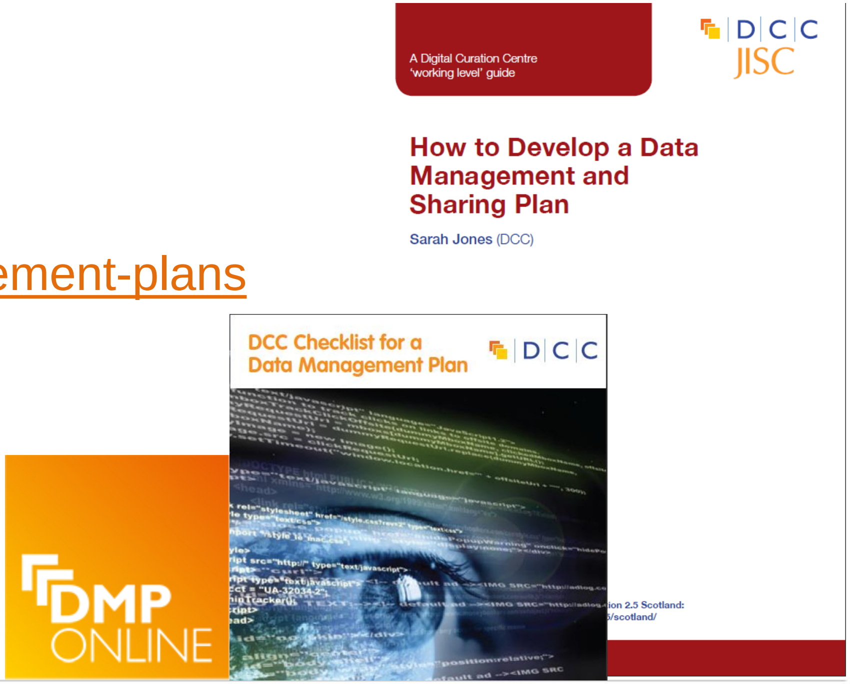
Data will be provided in file formats considered appropriate for long-term access, as recommended by the UK Data Service. For example, SPSS Portal format and tab-delimited text for qualitative tabular data and RTF and PDF/A for interview transcripts. Appropriate documentation necessary to understand the data will also be provided. Anonymised data will be held for a minimum of 10 years following project completion, in compliance with LSHTM's Records Retention and Disposal Schedule. Biological samples (output 3) will be deposited with the UK BioBank for future use.

From Writing a Wellcome Trust Data Management and Sharing Plan

DCC support on DMPs

- Webinars and training materials
- How-to guides and other advisory documents
- Checklist on what to cover in DMPs
- Example DMPs
- DMPonline

www.dcc.ac.uk/resources/data-management-plans



A Digital Curation Centre 'working level' guide

DCC JISC

How to Develop a Data Management and Sharing Plan

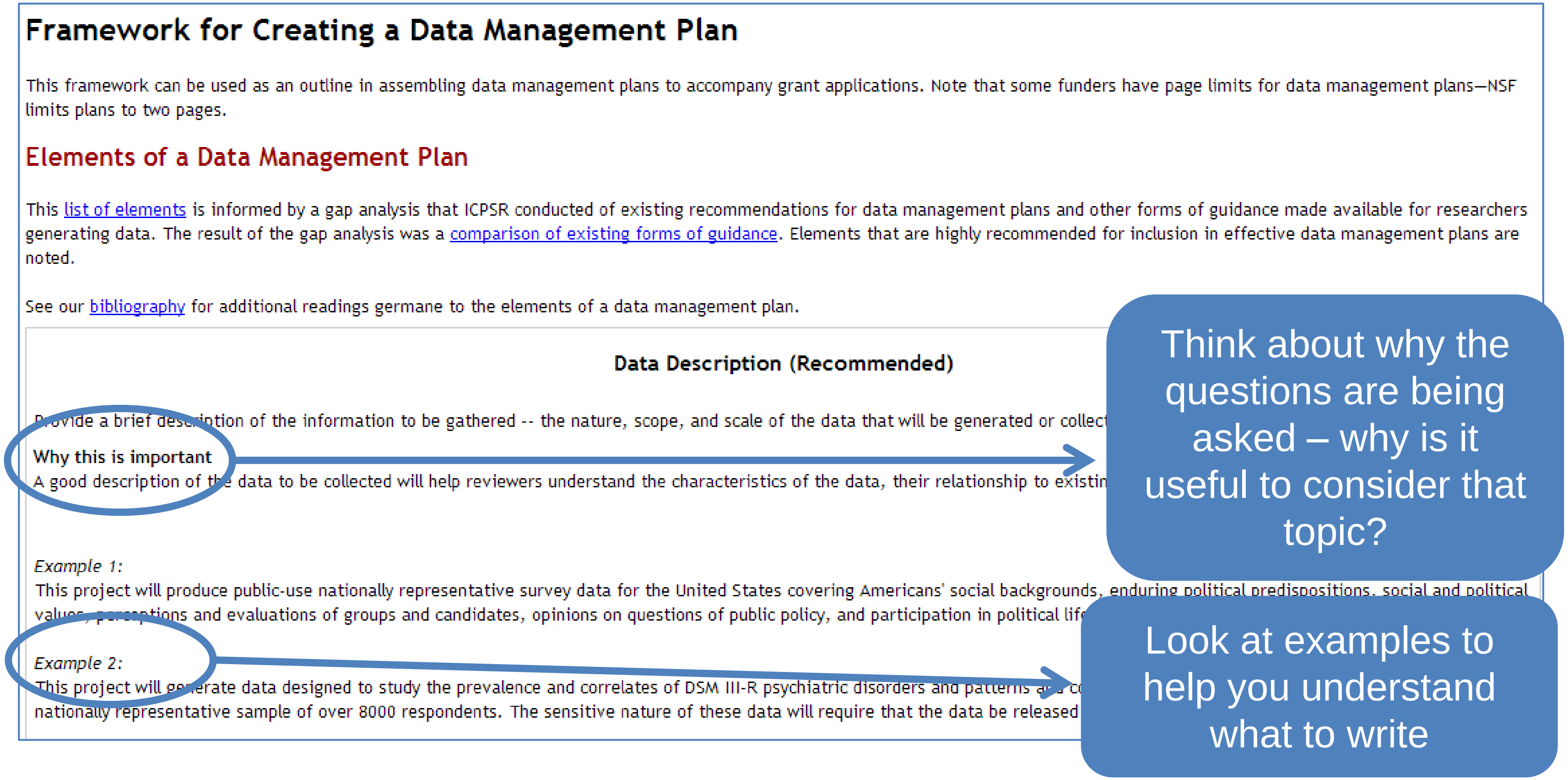
Sarah Jones (DCC)

DCC Checklist for a Data Management Plan

DCC

DMP ONLINE

Guidance from elsewhere



Framework for Creating a Data Management Plan

This framework can be used as an outline in assembling data management plans to accompany grant applications. Note that some funders have page limits for data management plans—NSF limits plans to two pages.

Elements of a Data Management Plan

This list of elements is informed by a gap analysis that ICPSR conducted of existing recommendations for data management plans and other forms of guidance made available for researchers generating data. The result of the gap analysis was a comparison of existing forms of guidance. Elements that are highly recommended for inclusion in effective data management plans are noted.

See our bibliography for additional readings germane to the elements of a data management plan.

Data Description (Recommended)

Provide a brief description of the information to be gathered -- the nature, scope, and scale of the data that will be generated or collect

Why this is important
A good description of the data to be collected will help reviewers understand the characteristics of the data, their relationship to existin

Example 1:
This project will produce public-use nationally representative survey data for the United States covering Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political lif

Example 2:
This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns a
nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released

Think about why the questions are being asked – why is it useful to consider that topic?

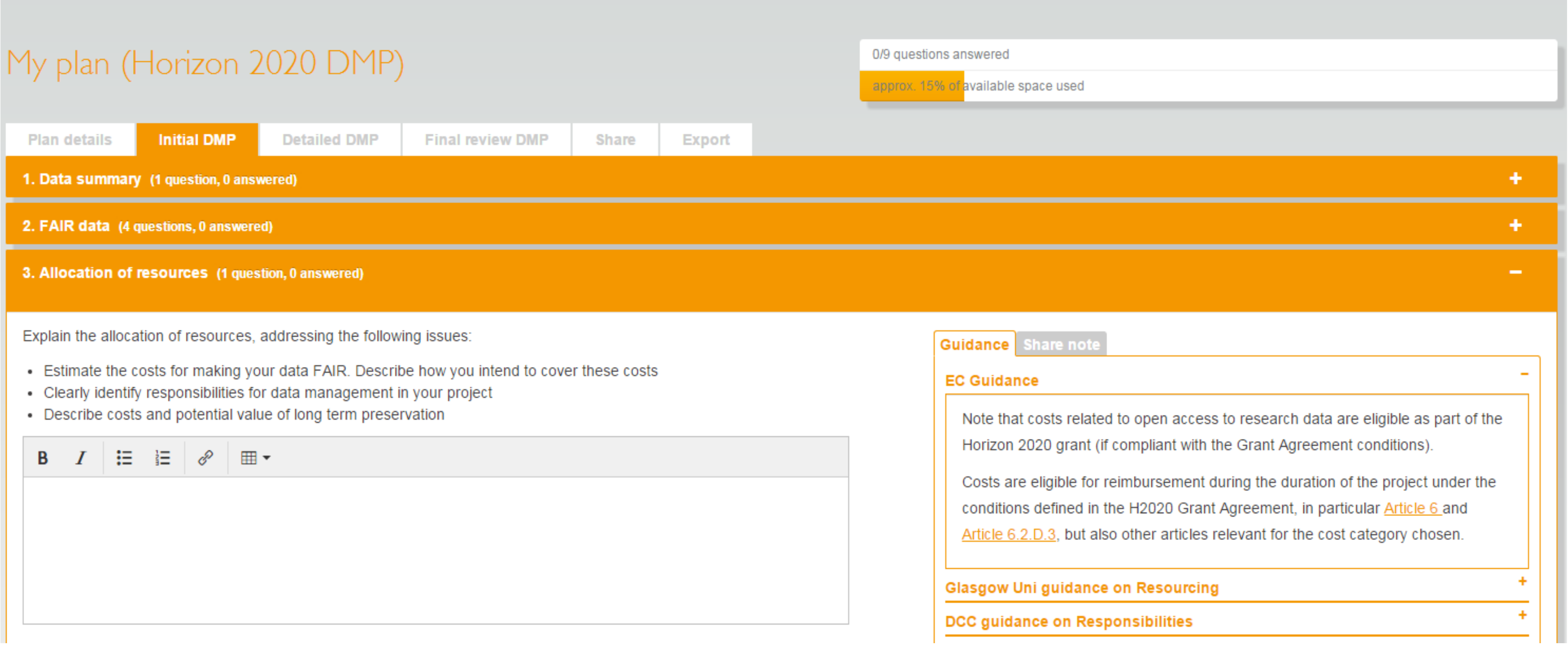
Look at examples to help you understand what to write

www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html

What is DMPonline?

A web-based tool to help researchers write

data management plans



My plan (Horizon 2020 DMP)

0/9 questions answered

approx. 15% of available space used

1. Data summary (1 question, 0 answered) +

2. FAIR data (4 questions, 0 answered) +

3. Allocation of resources (1 question, 0 answered) –

Explain the allocation of resources, addressing the following issues:

- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs
- Clearly identify responsibilities for data management in your project
- Describe costs and potential value of long term preservation

EC Guidance –

Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Costs are eligible for reimbursement during the duration of the project under the conditions defined in the H2020 Grant Agreement, in particular Article 6 and Article 6.2.D.3, but also other articles relevant for the cost category chosen.

Glasgow Uni guidance on Resourcing +

DCC guidance on Responsibilities +

https://dmponline.dcc.ac.uk

Main features in DMPonline

- Templates for different requirements (funder or institution)
- Tailored guidance (funder, institutional, discipline-specific etc)
- Ability to provide examples and suggested answers
- Supports multiple phases (e.g. pre- / during / post-project)
- Granular read / write / share permissions
- Customised exports to a variety of formats
- Shibboleth authentication

Key messages

- Data management is part of good practice whether you plan to make the data open or not
  - **it benefits you!**

- The process of planning is as important as the DMP. Think about the desired end result and plan for this.

- Approach DMPs in whatever way best fits your project. Don’t just let funder requirements drive things.

Questions?

Exercise - 45 min (+ 15 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?)

Thank you!

For DCC resources see:

www.dcc.ac.uk/resources

Follow us on twitter:

@digitalcuration and #ukdcc

Feedback form: https://forms.gle/tELB93RwNzHr2baf6