Managing research data and Horizon 2020

Sarah Jones
Digital Curation Centre, Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC

Consorcio Madroño conference on Data Management Plans and Horizon 2020, ETSI Industriales, Madrid, 25th February 2015

Funded by: Jisc

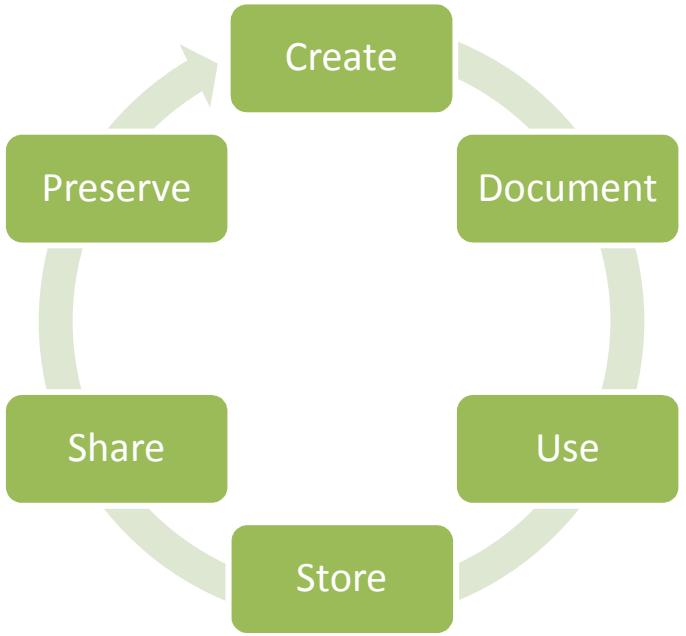
WHY MANAGE RESEARCH DATA?

Benefits and drivers

What is Research Data Management?

The active management of data throughout the lifecycle

- Data Management Planning
- Creating data
- Documenting data
- Accessing / using data
- Storage and backup
- Selecting what to keep
- Sharing data
- Data licensing and citation
- Preserving data
- …



Why manage research data?

Direct benefits for you

- To make your research easier!
- Stop yourself drowning in irrelevant stuff
- Have data organised so you know which versions are most up-to-date
- Make sure you can understand and reuse your data again later

Research integrity

- To avoid accusations of fraud or bad science
- Evidence findings and enable validation of research methods
- Codes of practice on good research conduct
- Many research funders worldwide now require Data Management and Sharing Plans

Potential to share

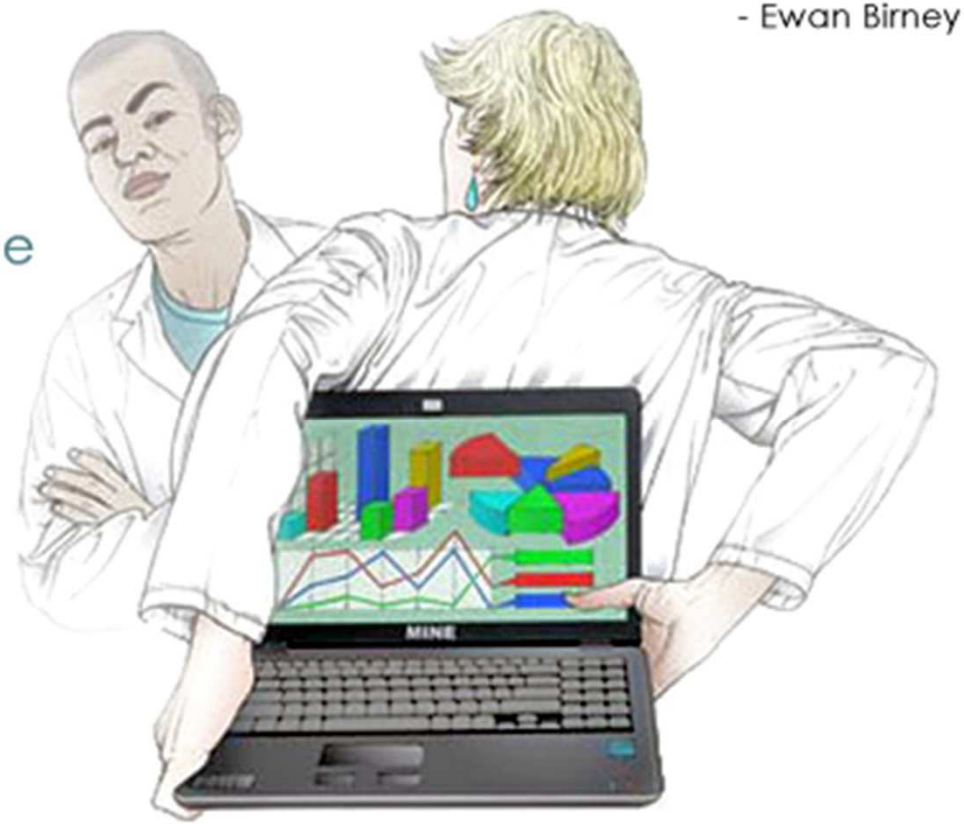
- So others can reuse and build on your data
- To gain credit – several studies have shown higher citation rates when data are shared
- For greater visibility, impact and new research collaborations
- Promote innovation and allow research in your field to advance faster

It’s part of good research practice

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

- Ewan Birney

#OpenData
#OpenScience

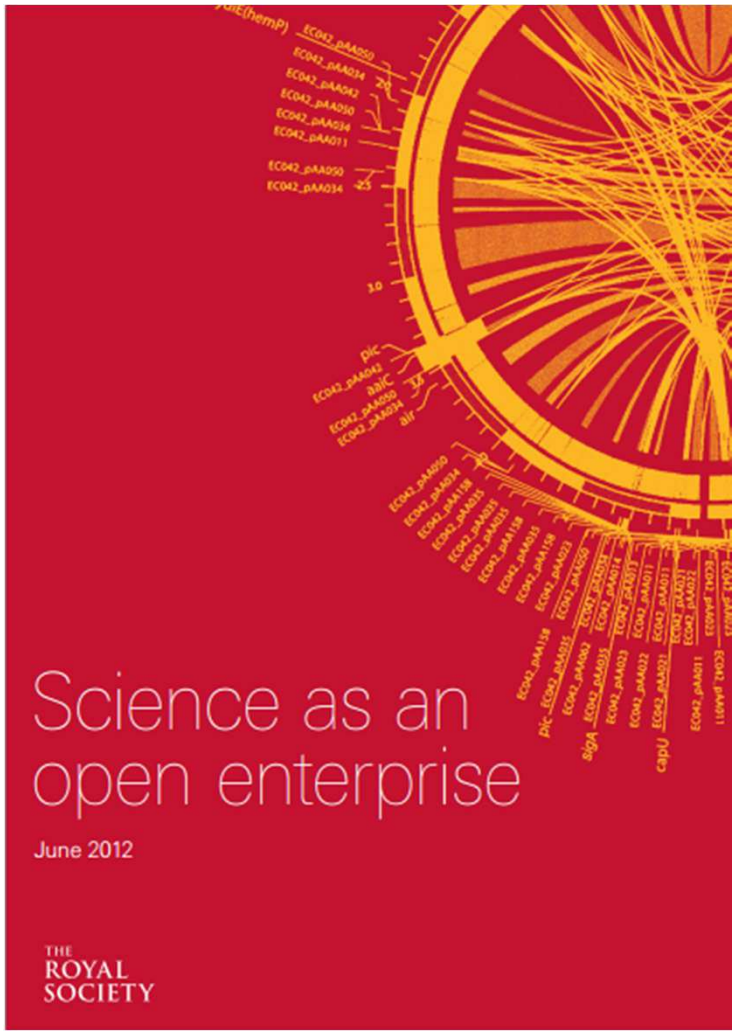


Original image via doi:10.1038/461145a. "Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science." - *Nature* 461, 145

Science as an open enterprise

"Much of the remarkable growth of scientific understanding in recent centuries is due to open practices; open communication and deliberation sit at the heart of scientific practice."

The Royal Society report calls for 'intelligent openness' whereby data are accessible, intelligible, assessable and usable.



https://royalsociety.org/policy/projects/science-public-enterprise/Report

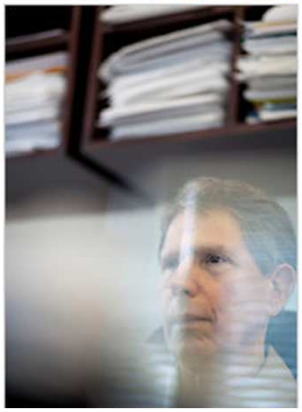
More scientific breakthroughs

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 tc find the biological markers that show the progression of Alzheimer's disease in the human brain.

Enlarge This Image



Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

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

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

But there are also opportunity costs

THE OPPORTUNITY COST OF MY #OPENSCIENCE WAS 35 HOURS + $690

By Emilio Bruna
http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690

For his most recent paper:

1. Double checking the main dataset and reformatting to submit to Dryad: **5 hours**
2. Creating complementary file and preparing metadata: **3 hours**
3. Submission of these two files and the metadata to Dryad: **45 minutes**
4. Preparing a map of the locations: **1 hour**
5. Submission of map to Figshare: **15 minutes**
6. Cleaning up and documenting the code, uploading it to GitHub: **25 hours**
7. Cost of archiving in Dryad: **US$90**
8. Page Charges: **$600**

So what needs to change?

Conclusions from Emilio Bruna:

- Develop a better system of incentives from the community for archiving data and code
- Teach our students how to do this NOW - it's much easier if you develop good habits early
- Minimise the actual and opportunity costs

We need to stop telling people "You should" and get better at telling people "Here's how"

HOW TO MANAGE DATA?

10 things to think about

1. What file formats are most appropriate?

- Do you have a choice or do the instruments you use only export in certain formats?
- What is common in your field? Try to use something that is accepted and widespread.
- Does your data centre recommend formats? If so it's best to use these.

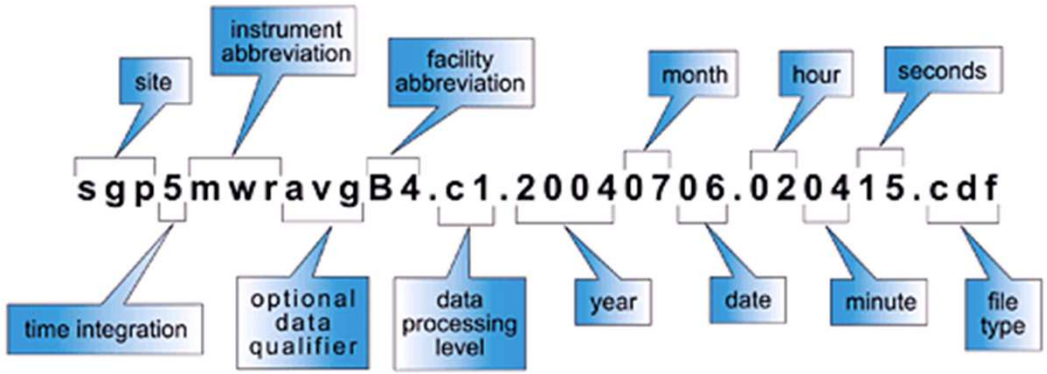
If you want your data to be re-used and sustainable in the long-term, you typically want to opt for open, non-proprietary formats.

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Further examples: http://www.data-archive.ac.uk/create-manage/format/formats-table

2. How will you name your files?

An example netCDF data file name is depicted below:



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

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

www.jiscdigitalmedia.ac.uk/guide/choosing-a-file-name

3. Can others understand the data?

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

- Have you documented what you did and how?
- Did you develop code to run analyses? If so, this should be kept and shared too.
- Is it clear what each bit of your dataset means? Make sure the units are labelled and abbreviations explained.
- Record metadata so others can find your work e.g. title, date, creator(s), subject, format, rights…,

4. What metadata standards will you use?

Use relevant standards for interoperability

RDA
RESEARCH DATA ALLIANCE

www.dcc.ac.uk/resources/metadata-standards

5. 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

6. Who will do the backup?

Use managed services where possible (e.g. University filestores rather than local or external hard drives), so backup is done automatically

3… 2… 1… backup!

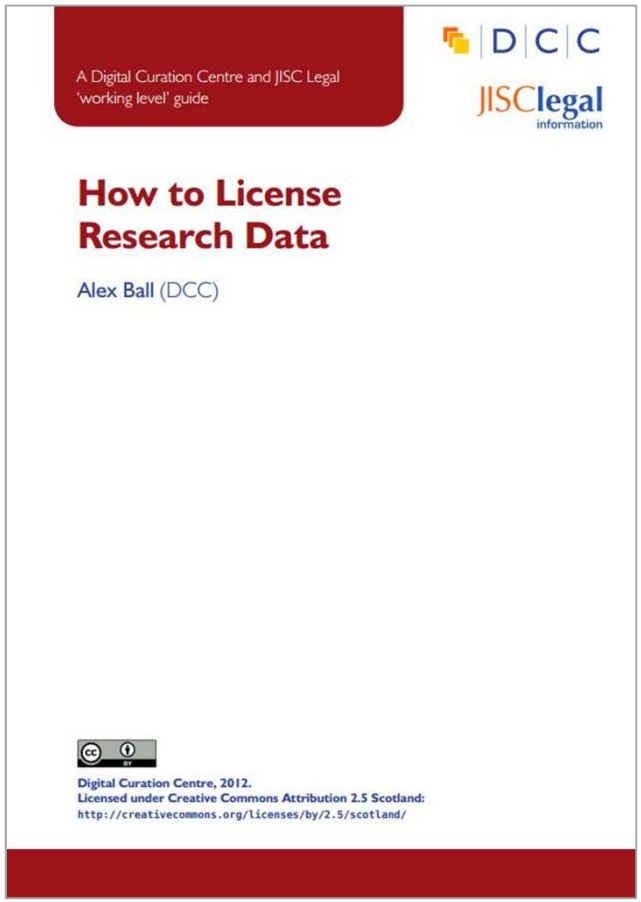
- at least 3 copies of a file
- on at least 2 different media
- with at least 1 offsite

Ask central IT team for advice

7. Can you publish / share your data?

- Who owns the data?
- Have you got consent for sharing?
- Do any licences you've signed permit sharing?
- Is the data in suitable formats?
- Is there enough documentation?

8. How will you license your data?



- Can your data be made available openly? e.g. CC0 or CC-BY
- Do you need to place certain restrictions on who can use the data or how?

This DCC how-to guide outlines pros and cons of each approach and gives practical advice on how to implement your licence.

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

9. Which data need to be kept?

Five steps to follow

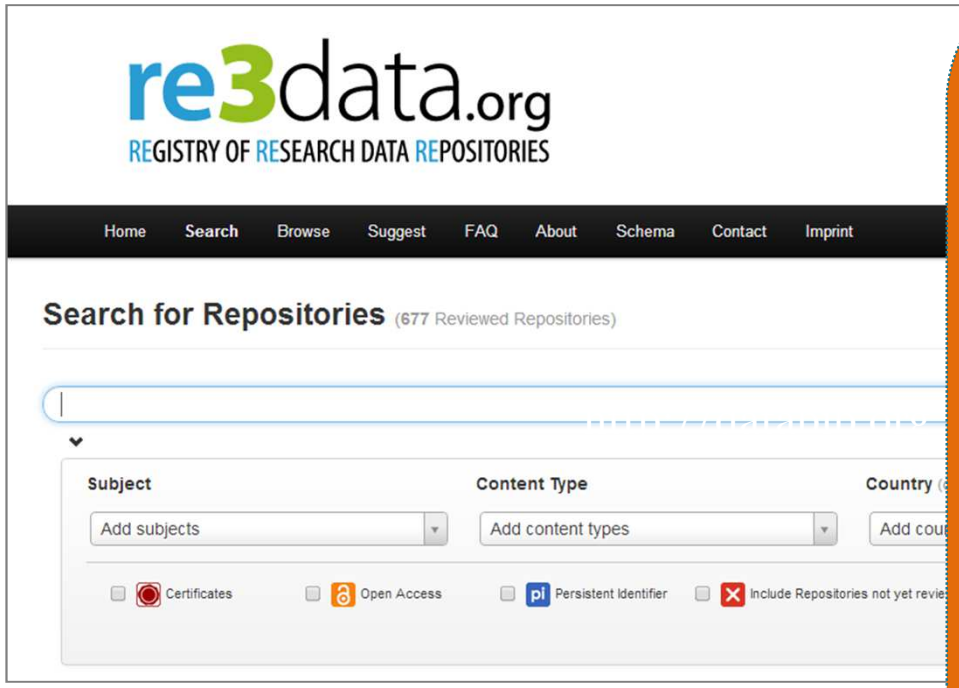
1. **Could** this data be re-used
2. **Must** it be kept as evidence or for legal reasons
3. **Should** it be kept for its potential value
4. **Consider costs** – do benefits outweigh cost?
5. **Evaluate criteria** to decide what to keep

5 steps to decide what data to keep

www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

10. Who will share & preserve the data?

- Does your publisher or funder suggest a repository?
- Are there data centres or community databases for your discipline?
- Does your university offer support for long-term preservation?



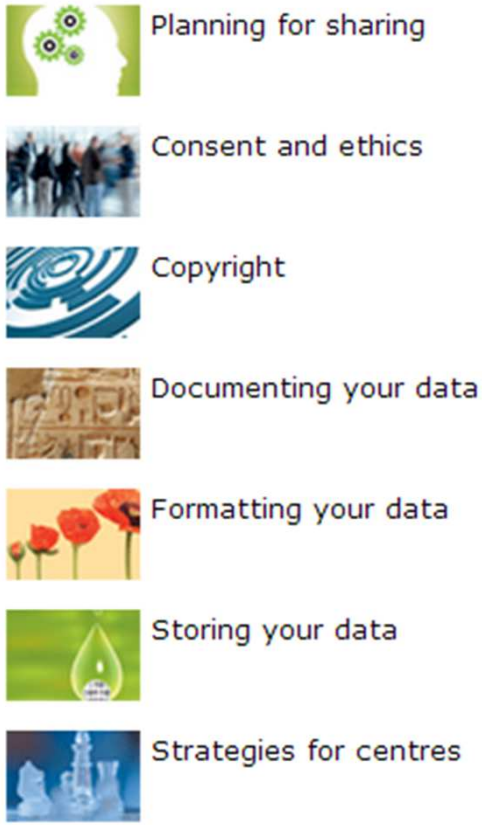
http://service.re3data.org/search

Zenodo

- Joint effort by OpenAIRE-CERN
- Multidisciplinary repository
- Multiple data types
  - Publications
  - Long tail of research data
- Citable data (DOI)
- Links funding, publications, data & software

www.zenodo.org

Managing and sharing data: a best practice guide

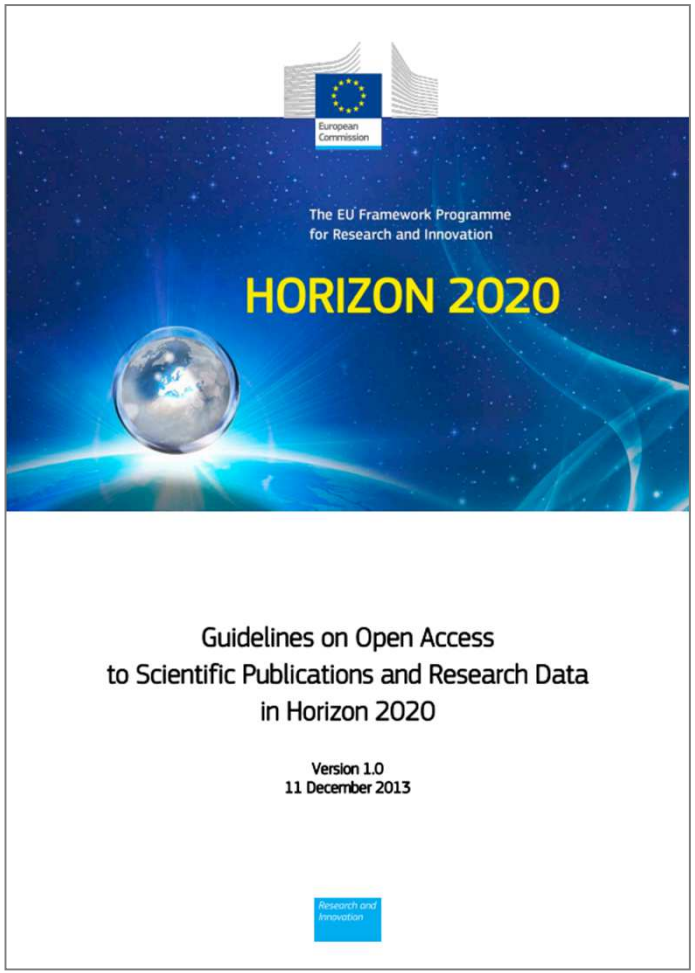


http://data-archive.ac.uk/media/2894/managingsharing.pdf

HORIZON 2020 OPEN DATA PILOT

Guidance and support to meet the EC requirements

Why open access and open data?



"The European Commission's vision is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full."

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

What is open data?

"Open data and content can be freely used, modified and shared by anyone for any purpose"

http://opendefinition.org

Tim Berners-Lee's proposal for five star open data - http://5stardata.info

- ★ make your stuff available on the Web (whatever format) under an open licence
- ★★ make it available as structured data (e.g. Excel instead of a scan of a table)
- ★★★ use non-proprietary formats (e.g. CSV instead of Excel)
- ★★★★ use URIs to denote things, so that people can point at your stuff
- ★★★★★ link your data to other data to provide context

How to make data open?

https://okfn.org

1. Choose your dataset(s)

   What can you may open? You may need to revisit this step if you encounter problems later.

2. Apply an open license

   Determine what IP exists. Apply a suitable licence e.g. CC-BY

3. Make the data available

   Provide the data in a suitable format. Use repositories.

4. Make it discoverable

   Post on the web, register in catalogues…

Support on Data Management Plans

What to cover in DMPs:

1. Description of data to be collected / created
2. Standards and methodologies for data collection & management
3. Any issues or restrictions due to ethics and Intellectual Property
4. Plans for data sharing and access
5. Strategy for long-term preservation

Example DMPs, guidance, tools and support at:

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

OpenAIRE

Open Access Infrastructure for research in Europe

- aggregates data on OA publications
- mines & enriches it content by linking thing together
- provides services & APIs e.g.

  to generate publication lists

www.openaire.eu

http://vimeo.com/108790101

EUDAT

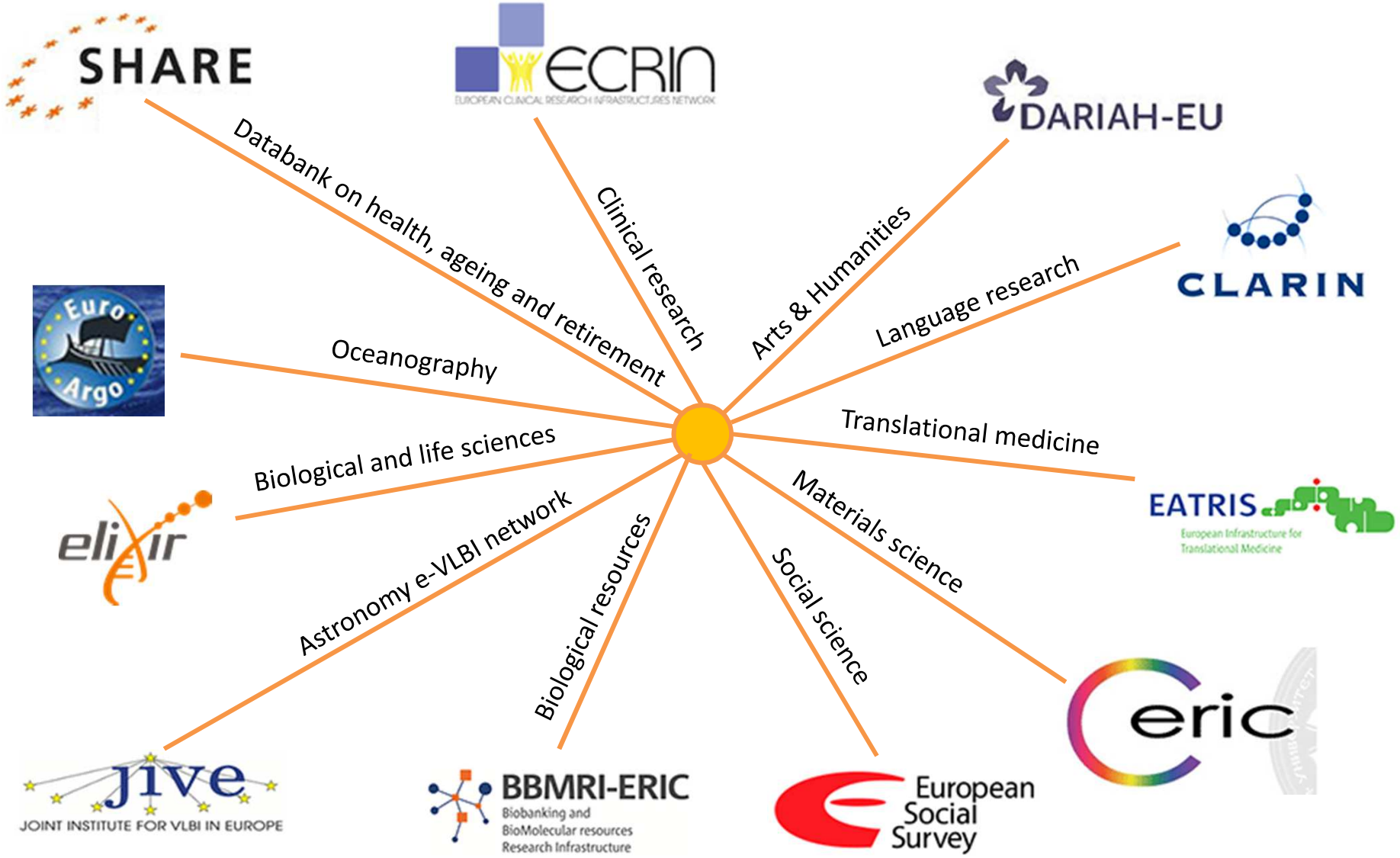
Data Infrastructure project offering various services:

- B2DROP: sync & exchange data
- B2SHARE: store & share
- B2STAGE: get data to computation
- B2SAFE: replicate data safely
- B2FIND: find data

They also offer a licensing wizard:
http://ufal.github.io/lindat-license-selector

www.eudat.eu

Discipline-specific infrastructure



FOSTER open science

- Open access and open data training across Europe
- Portal with access to reusable learning objects
- e-learning courses forthcoming
- Check out the training programme
  www.fosteropenscience.eu/events

Sharing European Research Outcomes: Raising Awareness on Open Access, Open Research Data and Open Science
May 13 (Madrid) – 14 (Valencia) – 28 (Gijón)

FOSTER

Gracias por su atención

DCC guidance, tools & case studies:

www.dcc.ac.uk/resources

Follow us on twitter:

@digitalcuration and #ukdcc