The expanding scholarly record

Data & software & materials, *oh my!*

UKSG: Open Access Lifecycle Seminar
9 March 2021

Jez Cope: 🐦 @jezcope
Data Services Lead
The British Library

These slides: http://bit.ly/uksg-jc-esr
Live transcript: http://bit.ly/uksg-jc-esr-transcript

Agenda

- RDM & Open Data: a (very) brief history
- Why do we want to be Open, anyway?
- Where is Open Research heading?

Research Data is...

- Text, documents, corpora, images, maps, web pages, spreadsheets, interview transcripts, video, audio, 3D models, ...
- Any source of information you use to support a research conclusion or argument

Brief aside: *I try to consistently treat "data" as a mass noun*

Other people prefer to treat it as plural;
I don't mind what you do as long as you're consistent!

BRITISH LIBRARY

Wait, let's back up a bit...

Some background

- 1967: Social Science Research Council Data Bank founded
  - Renamed UK Data Archive in 2000; part of UK Data Service from 2011

Some background

- 1967: Social Science Research Council Data Bank founded
  - Renamed UK Data Archive in 2000; part of UK Data Service from 2011
- 2005: MRC Research Data Policy published (revised 2015)

Some background

- 1967: Social Science Research Council Data Bank founded
  - Renamed UK Data Archive in 2000; part of UK Data Service from 2011
- 2005: MRC Research Data Policy published (revised 2015)
- 2009–2013: Jisc Managing Research Data (MRD) programme

Some background

- 1967: Social Science Research Council Data Bank founded
  - Renamed UK Data Archive in 2000; part of UK Data Service from 2011
- 2005: MRC Research Data Policy published (revised 2015)
- 2009–2013: Jisc Managing Research Data (MRD) programme
- 2011: EPSRC policy framework on research data published
  - Full compliance by 2015

Some background

- 1967: Social Science Research Council Data Bank founded
  - Renamed UK Data Archive in 2000; part of UK Data Service from 2011
- 2005: MRC Research Data Policy published (revised 2015)
- 2009–2013: Jisc Managing Research Data (MRD) programme
- 2011: EPSRC policy framework on research data published
  - Full compliance by 2015
- 2016: FAIR Principles for Data published
  - *More on this later…*

Aside

Much of this ☝ is becoming prey to link rot...

But it's still available thanks to the work of

- The National Archives
- UK Web Archive
- Internet Archive

Issues over the years

- How can we make people understand this is even a thing?

Issues over the years

- How can we make people understand this is even a thing?
- What should be covered under RDM?
- How should RDM be planned?
- How can good RDM be monitored, encouraged or enforced?

Issues over the years

- How can we make people understand this is even a thing?
- What should be covered under RDM?
- How should RDM be planned?
- How can good RDM be monitored, encouraged or enforced?
- Who is responsible for what parts?
- Who should pay for the costs of good RDM?

Issues over the years

- How can we make people understand this is even a thing?
- What should be covered under RDM?
- How should RDM be planned?
- How can good RDM be monitored, encouraged or enforced?
- Who is responsible for what parts?
- Who should pay for the costs of good RDM?
- Moving from RDM to Open Data where possible
- How Open is Open? How Open should / be?
- What should be Open?

Are we nearly there yet?

Photo by Priscilla Du Preez on Unsplash

Why is it taking so looooong...?

Culture change takes time

Open Access was an overnight success...

Why is it taking so looooong...?

Culture change takes time

Open Access was an overnight success...

...which took 2+ decades to happen

Why is it taking so looooong...?

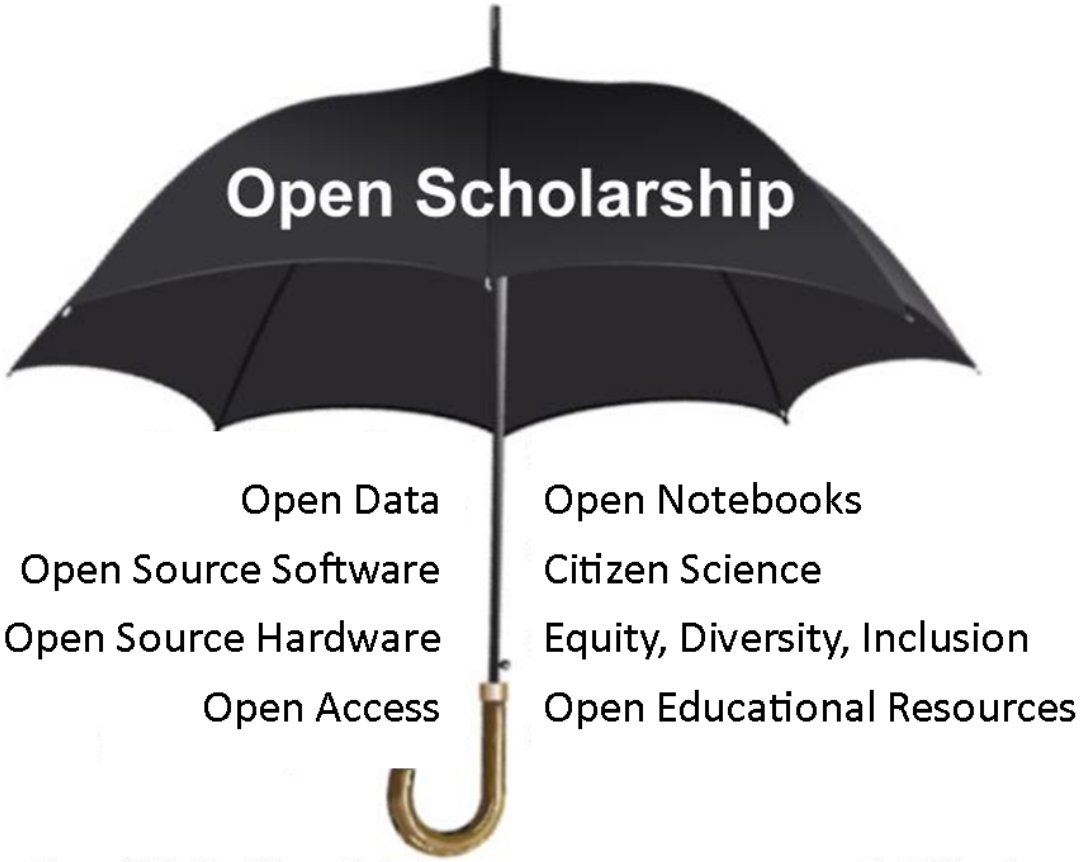
Culture change takes time

Open Access was an overnight success...

...which took 2+ decades to happen

But we *are* making progress!

So where are we now...?



Adapted from: https://www.meetup.com/Berlin-Open-Science-Meetup/

Robin Champieux and Danielle Robinson

Why do we *want* to be Open?

- Transparency?
- Trustworthiness?
- Efficiency?
- Inclusivity?

FAIR principles for data (2016)

- Data should be:
  - Findable
  - Accessible
  - Interoperable
  - Reusable

FAIRsFAIR Project https://fairsfair.eu: *"Fostering Fair Data Practices in Europe - aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle"*

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., … (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(160018). http://dx.doi.org/10.1038/sdata.2016.18

What is scholarship, anyway?

> "An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."
> — *David Donoho in 1995, paraphrasing geologist Jon Claerbout*

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In A. Antoniadis, & G. Oppenheim (Eds.), Wavelets and Statistics (pp. 55–81). : Springer New York. http://dx.doi.org/10.1007/978-1-4612-2544-7_5

Tell the *whole* story

Not just the book/article! Yes, data, but also...

- Methodology
- Software/analysis code
- Intellectual property
- Primary & secondary sources
- Public engagement activities
- Creative outputs: performances, artworks, poetry...

Wellcome Trust call this "Outputs Management"

Research Software

Any software used to conduct research

- *May* include:
  - Commercial software packages
  - Software modules & libraries
  - Even programming languages themselves (e.g. Python, R, Matlab)
  - Workflows
  - Simple analysis scripts

FAIR Research Software

- What does FAIR software look like?
  - FAIR4RS joint RDA/FORCE11/Research Software Alliance working group
- Ongoing work by: Software Sustainability Institute; Society of Research Software Engineers; *and many others*

What else?

- Physical artefacts
  - Museum collections
  - Archives
  - Cell lines
  - Tissue samples
- Methods and protocols
- Instruments
- Projects
- Funders
- ...

PIDs: Tying it all together

Persistent Identifiers (PIDs) provide a linking infrastructure for the scholarly record

- DOI (DataCite, Crossref); ORCID; ISNI; FundREF; RAiD; …

The PID Graph

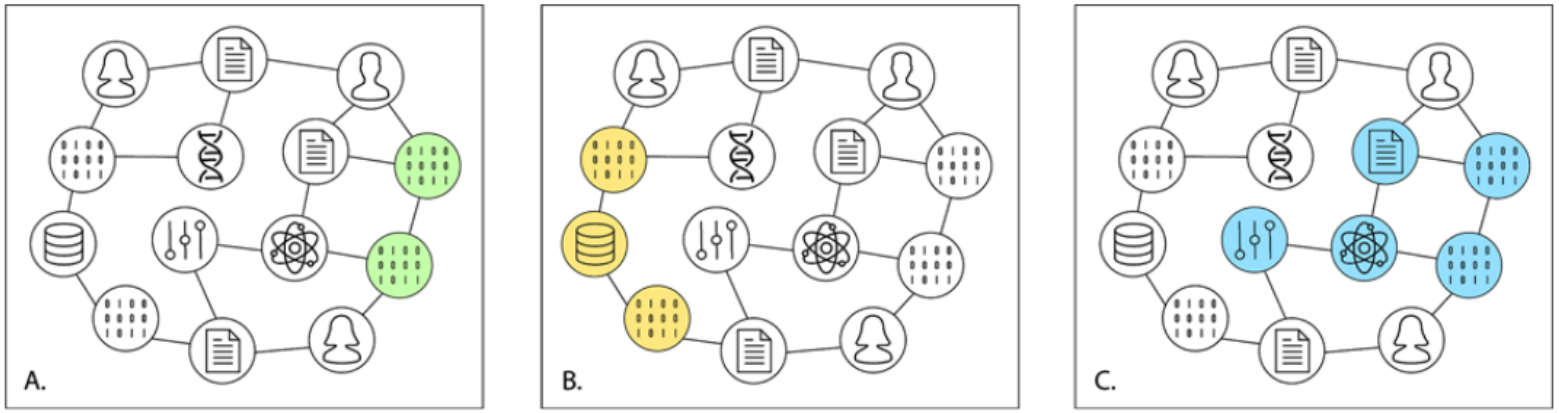


Figure 1 A schematic representation of the PID graph with digital objects connected by PIDs, showing three use cases: A: Different versions of software code, B: Datasets hosted by a particular repository, C: All digital objects connected to a research object.

From Project Freya: https://project-freya.eu

Where can I learn more?

The Turing Way, "an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science https://the-turing-way.netlify.app

Library Carpentry FAIR Research lesson (work in progress) https://github.com/LibraryCarpentry/lc-fair-research/

Research Data Alliance working groups https://www.rd-alliance.org

- RDA 17th (Virtual) Plenary Meeting: 20–23 April 2021

GLAM (Galleries, Libraries, Archives, Museums) Data Science Network: *watch this space...*

Summary

- We're building a resilient scholarly record for the future

Photo by Courtney Hedger on Unsplash