because good research needs good data

Introduction to Research Data Management

Open University 09th June 2015

Jonathan Rans

Digital Curation Centre

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License.

Who we are

- The DCC (Est. 2004) is:
  - A national-level centre of expertise in digital preservation with a particular focus on Research Data Management (RDM)
  - Working closely with a number of UK institutions to boost RDM capability across the HE sector
  - Also involved in a variety of national and international collaborations

What will we cover?

1. Definitions we work to
2. Why take a formal approach to managing your data?
3. What does Research Data Management encompass?

Definitions

Definitions of research data?

- “Research data, unlike other types of information is collected, observed, or created, for purposes of analysis to produce original research results.”

- “Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created.“

- “Evidence which is used or created to generate new knowledge and interpretations. ‘Evidence’ may be intersubjective or subjective; physical or emotional; persistent or ephemeral; personal or public; explicit or tacit; and is consciously or unconsciously referenced by the researcher at some point during the course of their research.”

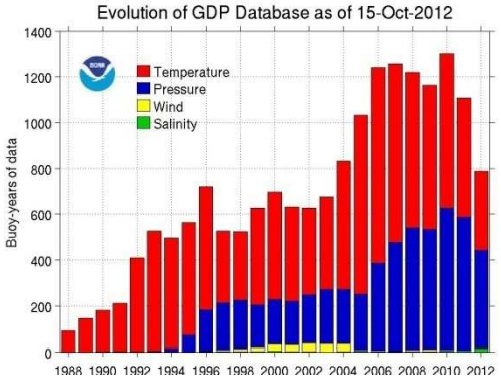
THE UNIVERSITY OF EDINBURGH

EPSRC

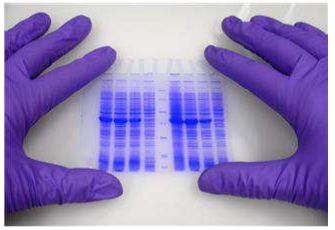
KAPTUR

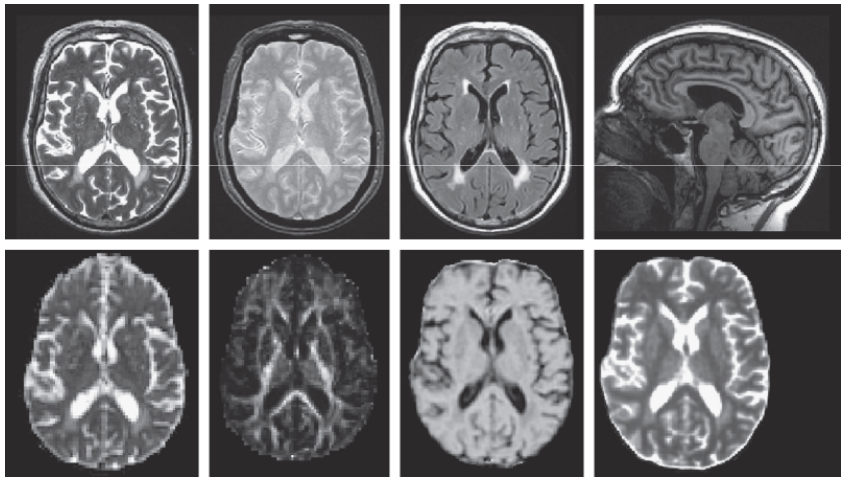
managing visual arts research data

So, what might this include?

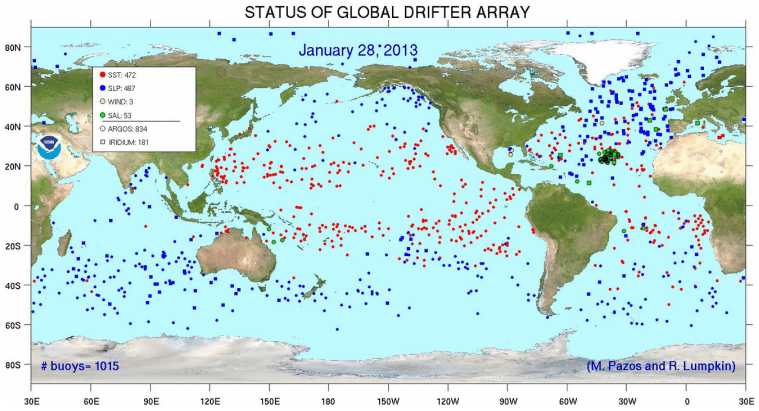


http://www.aoml.noaa.gov/phod/dac/array_growth.html





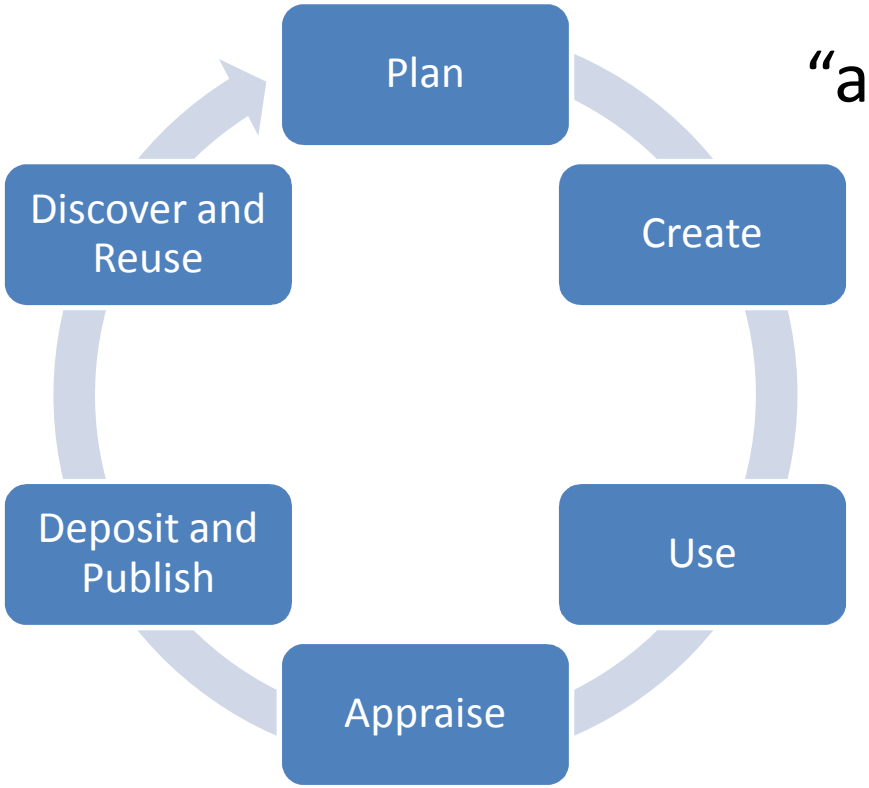
http://www.sbirc.ed.ac.uk/documents/lbc_protocol.pdf



http://www.aoml.noaa.gov/phod/graphics/dacdata/globpop.gif

Anything & everything produced in the course of research

What is research data management?



"an explicit process covering the creation and stewardship of research materials to enable their use for as long as they retain value."

Data management is part of good research practice

Why manage research data?

Why is RDM an issue?

- Digital technology now used very widely in research, and is enabling new research and scientific paradigms
- Research funders and publishers know that digital research data can be expensive to produce but inexpensive to share, making reuse more feasible and desirable
- The challenge is to ensure digital research findings can be reproduced and cited

Why manage research data?

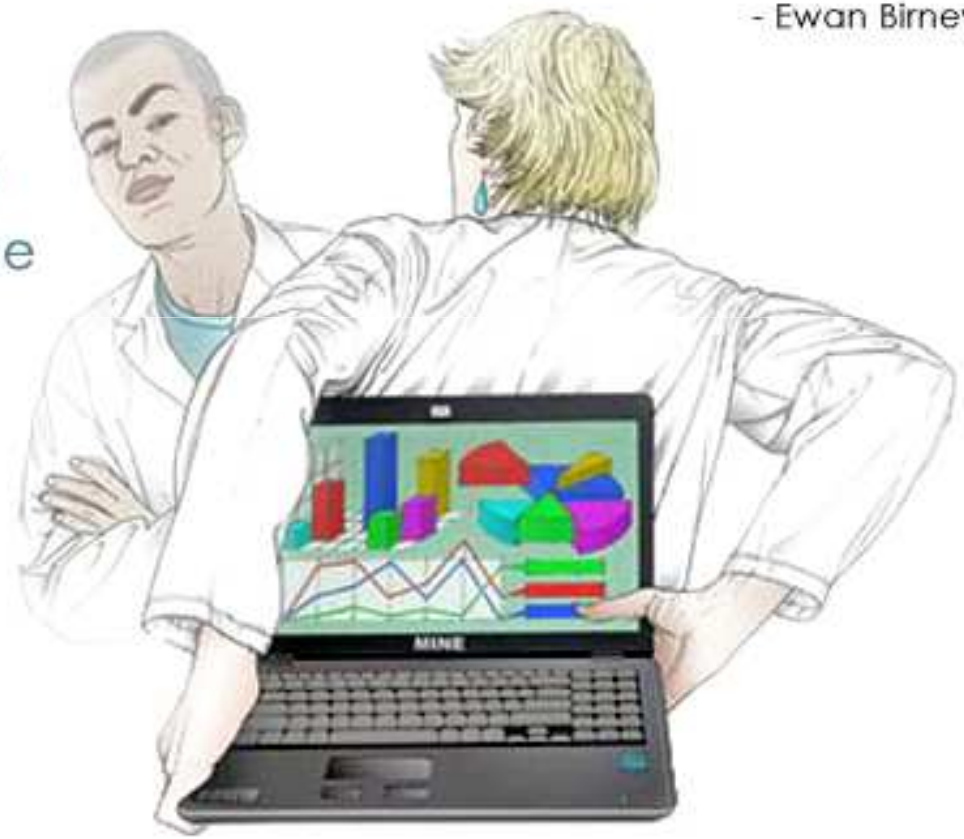
- To make research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To share data so others can use and learn from it
- To get credit for producing the data
- Because somebody else said to do so

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

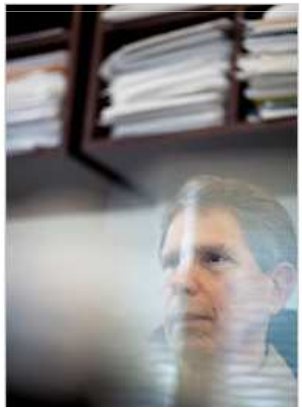
Sharing leads to breakthroughs

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.

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.

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

"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

Benefits of data sharing (3)

“There is evidence that studies that make their data available do indeed receive more citations than similar studies that do not.”

Piwowar H. and Vision T.J 2013 "Data reuse and the open data citation advantage“ https://peerj.com/preprints/1.pdf

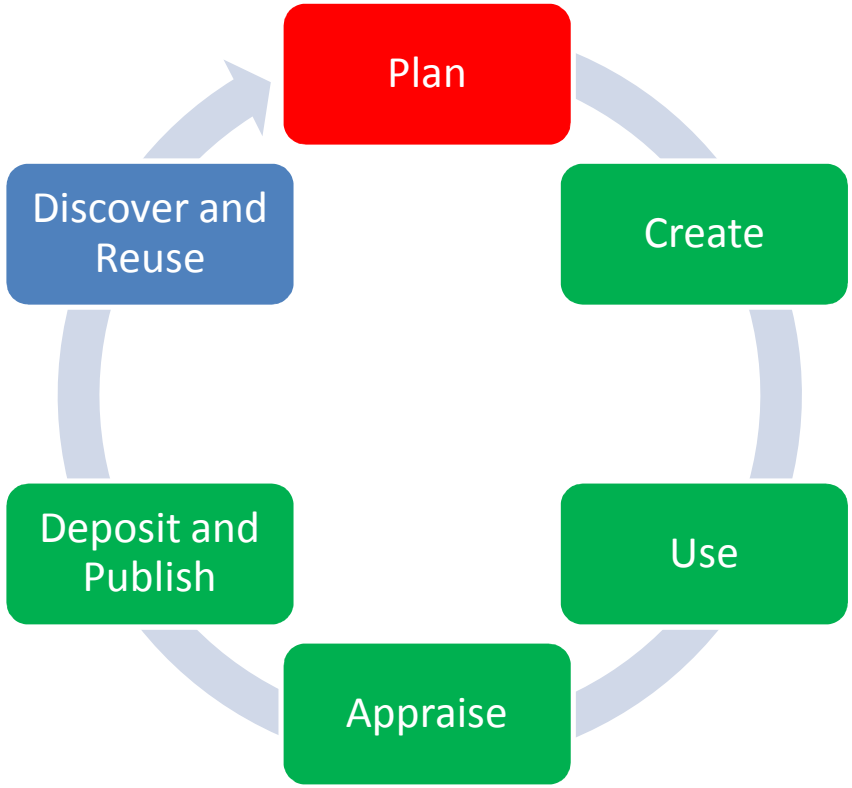
… more citations

9% - 30% increase

How can you manage your data?

Stages of RDM

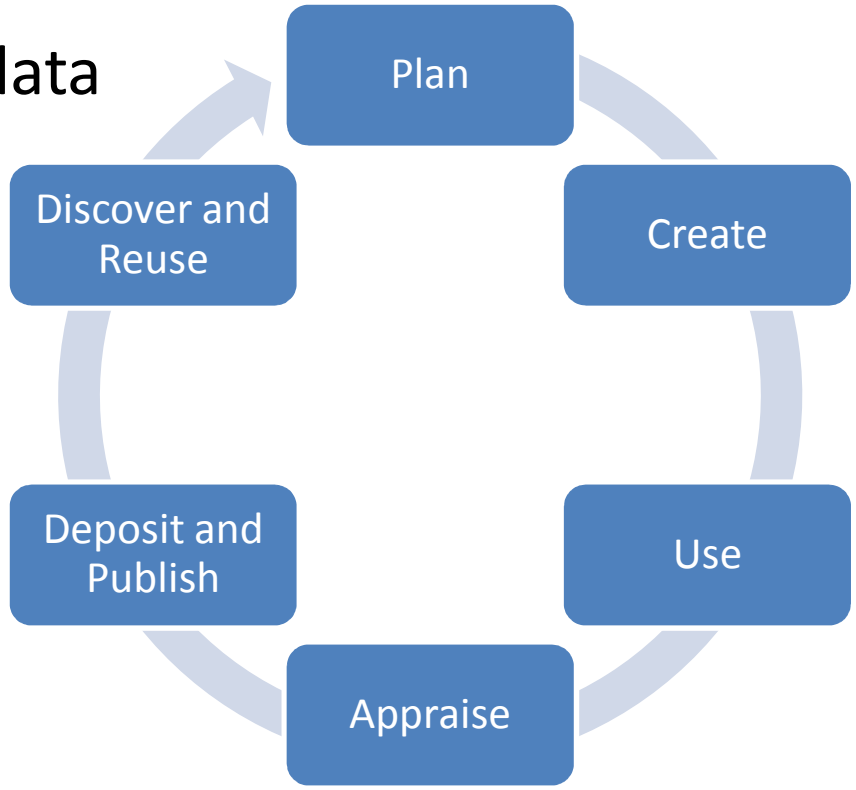
- Pre-award
- Project phase
- Post-project



Lifecycle adapted from Edinburgh University Mantra materials:
http://www.slideshare.net/edinadocumentationofficer/mantra-poster2#

The Research Data Lifecycle

- Data Management Planning
- Data creation
- Annotating / documenting data
- Analysis, use, versioning
- Storage and backup
- Publishing papers and data
- Preparing for deposit
- Archiving and sharing
- Licensing
- Citing…

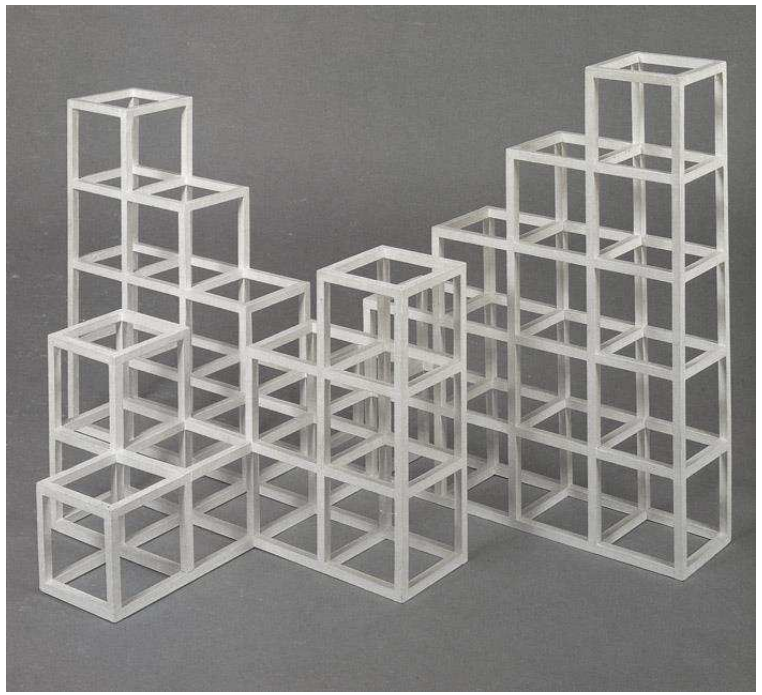


Experimental design

- Ensure consent forms, licences and partnership agreements don't restrict what you can do with your data
  - http://www.dcc.ac.uk/resources/how-guides/license-research-data
  - http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

Data creation

- Adopt file naming conventions:
  - http://www.jiscdigitalmedia.ac.uk/guide/choosing-a-file-name/
- Design a good project folder structure
  - http://research-data-toolkit.herts.ac.uk/document/research-project-file-plan/
- Develop a method for describing new versions of your files.



Some formats are better for long-term

It's preferable to opt for formats that are:

- Uncompressed
- Non-proprietary
- Open, documented
- Standard representation (ASCII, Unicode)

Data centres may have preferred formats for deposit e.g.

Type	Recommended	Non-preferred
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

Describe your data

- Broadly speaking metadata is description of your data.
  - Disambiguation – dataset title, creator, date of accession, etc.
  - Discovery – e.g. tags
  - Reuse – Controlled vocabularies, metadata schema, etc.

Create description as you work!

Where to store data?

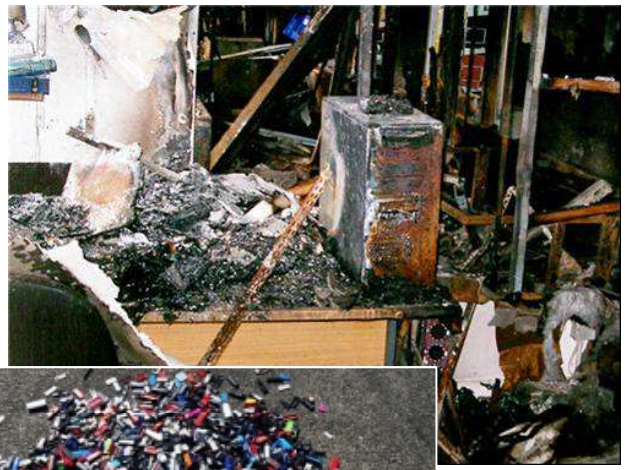
- Your own device (PC, flash drive, etc.)
  - And if you lose it? Or it breaks?
- Departmental drive or university filestore
  - Should be more robust with automated back-up
- “Cloud” storage
  - Do they care as much about your data as you do?

Data security

- Aim to develop a practical solution that fits your circumstances
- Encrypt and password protect mobile devices carrying sensitive information

Storage and backup

- Use managed services where possible e.g. shared drives rather than local or external hard drives
- 3… 2… 1… backup!
  - at least 3 copies of a file
  - on at least 2 different media
  - with at least 1 offsite



Appraisal and deposit

How to Appraise & Select Research Data for Curation
Angus Whyte, Digital Curation Centre, and Andrew Wilson, Australian National Data Service (2010)

1. **Relevance to Mission** – including any legal/funder requirement to retain the data beyond its immediate use.
2. **Scientific or Historical Value** – significance and relationship to publications etc.
3. **Uniqueness** – can it be found elsewhere / if we don't preserve it, who will?
4. **Potential for Redistribution** – quality / IP / ethical concerns are addressed.
5. **Non-Replicability** – either impossible to replicate (e.g. atmospheric or social science data) or not financially viable.
6. **Economic Case** – costs of managing and preserving the resource stack up well against potential future benefits.
7. **Full Documentation** – surrounding / contextual information necessary to facilitate future discovery, access, and reuse is adequate.

Why hand data over for preservation?

- To preserve the data themselves “Data rot”
  - Bitwise preservation
  - Format migration
- To preserve contextual information
  - Often held in a researcher’s head
  - Notes often aren’t detailed enough
- Protecting digital objects requires specialist skills and particular information to be captured
- The aim is to enable the reuse of data

Not everything can, or should be preserved!

Thank-you for listening!

Jonathan Rans
J.Rans@ed.ac.uk
@JNRans

Image Credits

Research Dictionary: http://www.saquiresearch.com

Himalayas: www.cntraveller.com

Sol Lewitt modular structure: www.saatchigallery.com

Harvey Rutt, Southampton. Recovered from: http://www.computerweekly.com

Consent: http://www.mycustomer.com

Pile of flash drives: www.flashdrivepros.com

Dalian University fire: www.weirdworldnews.org

Field researcher: Chris Rainier from:
http://www.wsj.com/articles/SB10001424052748703843804575534122591921594