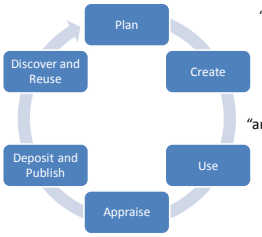
because good research needs good data

Good Practice – What RDM means in the real world

Kevin Ashley
Digital Curation Centre
www.dcc.ac.uk
@kevingashley
Kevin.ashley@ed.ac.uk

What is research data management?



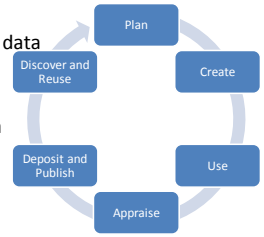
"the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

"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

What is involved in RDM?

- 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…

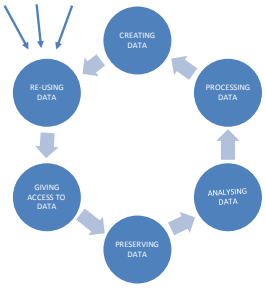


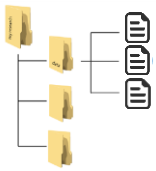
Data creation

- 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/
- Ensure consent forms, licences and partnership agreements don't restrict opportunities to share data
  - http://www.dcc.ac.uk/resources/how-guides/license-research-data
  - http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

Planning trick 1: think backwards

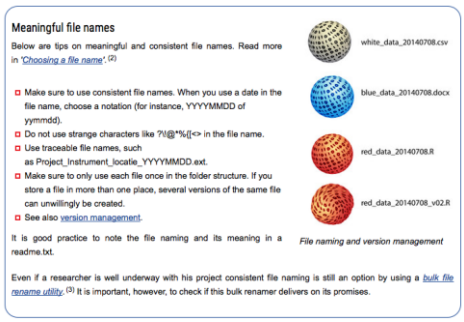
What data organisation would a re-user like?





Design a data organisation for the project (folder structure, file naming convention, …)

Data organisation



Meaningful file names

Below are tips on meaningful and consistent file names. Read more in '*Choosing a file name*'. (2)

- Make sure to use consistent file names. When you use a date in the file name, choose a notation (for instance, YYYYMMDD of yymmdd).
- Do not use strange characters like ?\@*%[]<> in the file name.
- Use traceable file names, such as Project_Instrument_locatie_YYYYMMDD.ext.
- Make sure to only use each file once in the folder structure. If you store a file in more than one place, several versions of the same file can unwillingly be created.
- See also version management.

It is good practice to note the file naming and its meaning in a readme.txt.

Even if a researcher is well underway with his project consistent file naming is still an option by using a *bulk file rename utility*. (3) It is important, however, to check if this bulk renamer delivers on its promises.

white_data_20140708.csv
blue_data_20140708.docx
red_data_20140708.R
red_data_20140708_v02.R

File naming and version management

http://datasupport.researchdata.nl/en/start-de-cursus/ii-onderzoeksfase/organising-data

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

Documentation and metadata

Metadata: basic info e.g. title, author, dates, access rights…
Documentation: context, workflows, methods, code, data dictionary…

Use standards wherever possible for interoperability

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

Data creation: documentation

- Collect together all the information users would need to find and understand the data
- Create metadata as you go, it's more time-consuming and less effective to do it at the end of a project
- Use standards where possible
  - » Data Documentation Initiative http://www.ddialliance.org/
  - » DCC Metadata Catalogue http://www.dcc.ac.uk/resources/metadata-standards
- Name, structure and version files clearly

Metadata devil: http://www.truthdig.com/cartoon/item/nsa_its_just_metadata_20130812

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?

Storage and backup

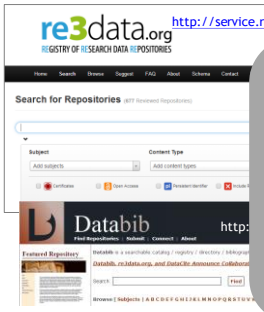
- Use managed services where possible e.g. shared drives rather than local or external hard drives
- Consider the security implications of where you store data and how you transfer it
- 3… 2… 1… backup!
  - » at least 3 copies of a file
  - » on at least 2 different media
  - » with at least 1 offsite

Pile of flash drives: www.flashdrivepros.com
Dalian University fire: www.weirdworldnews.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

Data repositories



http://service.re3data.org/search

Zenodo

- OpenAIRE-CERN joint effort
- Multidisciplinary repository
- Multiple data types
  - Publications
  - Long tail of research data
- Citable data (DOI)
- Links to funder, publications, data & software

www.zenodo.org

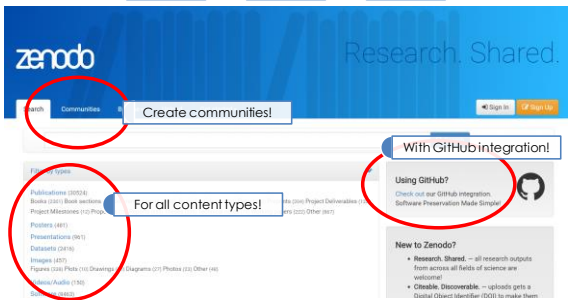
Data repositories

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

http://service.re3data.org/search

OpenAIRE services: zenodo.org

Upload → Describe → Publish



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!

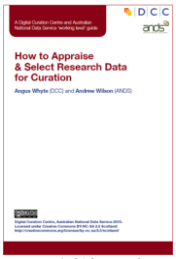
DCC guidelines - repositories

- Is the repository reputable?
- Will it accept the data you want to deposit?
- Will data be safe in legal terms?
- Will the repository sustain data value?
- Will the repository support analysis and track data usage?

The Data Deluge is upon us

Sensor's ability to produce data outstrips IT's ability to process it

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.

Outline

Why select, rather than 'file and forget'!

Take five steps to inform your choise…

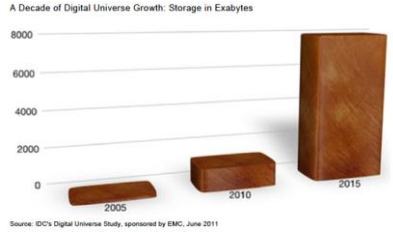
① Think. What could be reused for what purpose?
② Recognise compliance risks
③ (Gu)estimate long-term value
④ Judge the cost factors
⑤ Decide – what action needed

The onus is on you, but it's a partnership

So what tools and practical help do you need?

Why not keep it all?

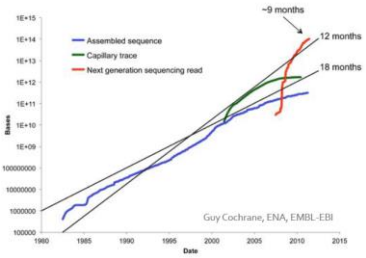
Globally, data volumes are doubling every two years



*John Gantz and David Reinsel 2011 **Extracting Value from Chaos**. www.emc.com/digital_universe.*

Data volumes escalate

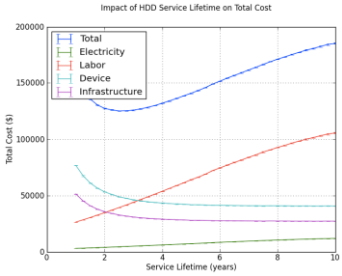
Volumes rising faster in data-intensive research domains
e.g. DNA sequence data is doubling every 6-8 months



2016-04-13 *"ELIXIR and Open Data" View from an ELIXIR Node" Borend Mons, ELIXIR Launch event, 18th Dec 2013* 22

Storage mgmt costs rise long-term

Hardware costs decline, but power and staff costs keep rising



2016-04-13 *David Rosenthal blog.dshr.org/2012/05/lets-just-keep-everything-forever-in.html* 23

The storage is cheap fallacy

- Decreasing hardware costs offset by data volume growth
- Backup and mirroring multiplies cost of preserved data
- Discovery becomes harder as chaff outweighs the wheat
- Curation of unused data is a waste of resources

2016-04-13 24

Storage Strategies

Good practice

Weigh up risks,
Value, and costs

Select, share, safeguard what you can afford to, or dispose of it

- ▷**F**indable
- ▷**A**ccessible
- ▷**I**nteroperable
- ▷**R**eusable

FAIR Principles
www.force11.org/group/fairgroup

Bad practice

Keep everything until…
lost by natural wastage

Fragmented
Risking unauthorised **disclosure** or **loss**
- •Bit rot
- •Media degradation
- •Obsolescence
- •(software, device,
- • format, media)
- •Fire, flood, theft
- •Organisation failure

When should selection begin?

Appraisal should begin as early as possible!

Periodically for longitudinal and reference datasets

What questions must be answered?

1. What must be kept to manage compliance risk?
2. What data could be re-used?
3. What data has value and should be kept?
4. Given costs what will or won't be kept?
5. How will it be kept and shared, on what terms?

What 'must' be kept?

Some data may be part of research record, evidence for e.g. …

- Supporting patent applications or IP
- Evidence of investigations or inquiries
- Health & Safety (Lab book)

Compliance also about data that won't be kept, or may only be shared with approved researchers…

Step 2 What *could* it be reused for?

Step back and reflect – typical reuse purposes

1. Verification
2. Further analysis
3. Reputation building
4. Resource development
5. Further publications inc. data articles
6. Learning and teaching materials
7. Private reference

Then relative to these, which data **must** be kept and which data and related materials will have significant value?

e.g. High Energy Physics community

Levels of data to preserve	Reuse purpose
1) Additional documentation (e.g. wikis, news forums)	Publication-related information search
2) Data in a simplified format	Outreach, simple training analyses
3) Analysis level software and the data format	Full scientific analysis based on existing reconstruction
4) Reconstruction and simulation software and basic level data	Full potential of the experimental data

Adapted from: DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics, May 2012 . http://arxiv.org/abs/1205.4667

Step 3 What data *should* have value

Indicators that data have value

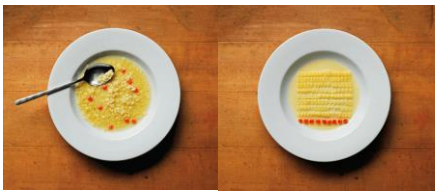
1. Quality of the data and its description
   complete, accurate, reliable, valid, representative etc
2. Demand high
   known users, integration potential, reputation, recommendation, appeal
3. Replication difficulty
   difficult, costly, or impossible to reproduce
4. Low barriers
   legal/ ethical, copyright non-restrictive terms and conditions
5. Rarity
   unique copy or other copies at risk

Which related material does data depend on for its value?

Step 4 Cost factors

Consider these when deciding what to keep because

- Costs incurred during project may add to the data's value
- Need to make sure post-project costs are covered



What action needs to be taken to ensure preservation is costed?

Step 5 Your data appraisal

Establish a clear idea of what data needs packaged at end

1. Title, contributors, description, access rights *
2. Reuse purpose(s)
3. Value for purpose
4. Risk of budget shortfall
5. Keep it or not? *
6. Reasons for disposal *
7. Actions to prepare for preservation or disposal

* What anyone outside the project most needs to know
(but the rest will help)

Licensing research data

Horizon 2020 Open Access guidelines point to:

or

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

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

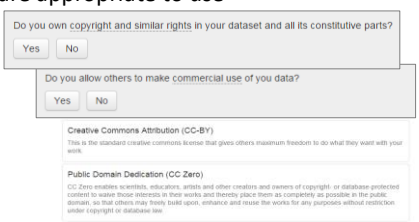
ND No Derivatives
Severely restricts use

These clauses are not open licenses

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

EUDAT licensing tool

Answer questions to determine which licence(s) are appropriate to use



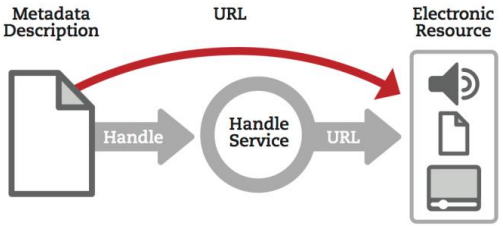
http://ufal.github.io/public-license-selector

What are persistent identifiers?

- They are an alphanumeric code identifying a resource, organisation or individual
- They must be
  - Unique
  - Persistent
- Ideally they should be actionable too

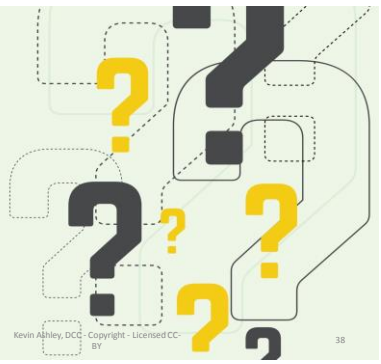
DataCite · EZID Long-term identifiers made easy · ORCID · isni

How do persistent identifiers work



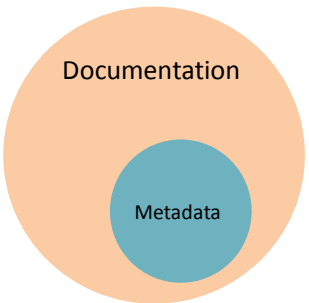
What is metadata?

Data about data



What is the difference?

- Metadata
  - Standardised
  - Structured
  - Machine and human readable



What is the minimum required?

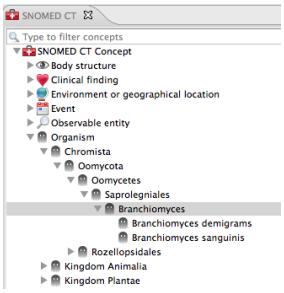
- Repository requirements
- Could be lead by DataCite
- Citation/disambiguation
  - Identifier
  - Creator
  - Title
  - Publisher
  - Publication Year
- Licencing/access conditions

Aiding discoverability

- Catalogue or discovery metadata
- Structured so that search engines can uncover it.
- Must be exposed in machine-readable form eg XML
- OAI-PMH?

Controlled vocabularies

- E.g. SNOMED CT (clinical terms) or MeSH
- Include ontologies as well
  - Defined terms + taxonomy
- Useful for selecting keywords to tag datasets



Ensuring the utility of the data

- The what, why and how data creation must be understood
- Data dictionaries
- Columns/rows labelled
- Variable ranges defined

Metadata standards

- These can be general – such as Dublin Core
- Or discipline specific
  - Data Documentation Initiative (DDI)
    - Social Sciences
  - Ecological Metadata Language (EML)
    - Ecology
  - Flexible Image Transport System (FITS)
    - Astronomy

Readme files

We recommend that a ReadMe be a plain text file containing the following:

- for each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication
- for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units
- any data processing steps, especially if not described in the publication, that may affect interpretation of results
- a description of what associated datasets are stored elsewhere, if applicable
- whom to contact with questions

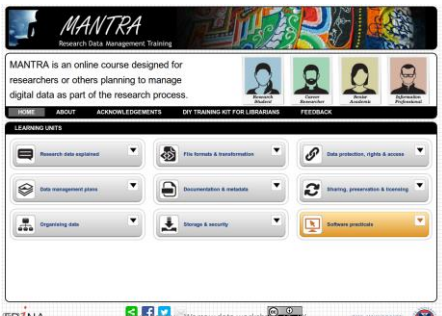
Documentation - Final thoughts

- The level of documentation required is likely to be proportional to the complexity of the data
- Ensuring values and terms are correctly defined is important
- Tools exist to simplify the creation of metadata
- Data which can't be published in digital form can still be made visible

RDM and sharing : a best practice guide

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

Acquire research data skills



Finally

- We can't cover it all today
- There's lots online that you can use to improve your skills
- Practice makes perfect – so does talking with colleagues