An Introduction to Research Data Management and Open Science

S. Venkataraman
Research Data Specialist
Digital Curation Centre

The Hong Kong University of Science and Technology, 18th March 2019

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License
What is RDM?

Image CC-BY-SA by Janneke Staaks www.flickr.com/photos/jannekestaaks/14411397343
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
Is there a reproducibility crisis?

Creating data

Image CC-SA-ND by Bill Dickinson www.flickr.com/photos/skynoir/8270436894
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)
<table>
<thead>
<tr>
<th>Type of data</th>
<th>Recommended formats</th>
<th>Acceptable formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data with extensive metadata variable labels, code labels, and defined missing values</td>
<td>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</td>
<td>proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)</td>
</tr>
<tr>
<td>Tabular data with minimal metadata column headings, variable names</td>
<td>comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements</td>
<td>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)</td>
</tr>
<tr>
<td>Geospatial data vector and raster data</td>
<td>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)</td>
<td>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</td>
</tr>
<tr>
<td>Textual data</td>
<td>Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema</td>
<td>Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti</td>
</tr>
<tr>
<td>Image data</td>
<td>TIFF 6.0 uncompressed (.tif)</td>
<td>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)</td>
</tr>
<tr>
<td>Audio data</td>
<td>Free Lossless Audio Codec (FLAC) (.flac)</td>
<td>MPEG-1 Audio Layer 3 (.mp3) if original created in this format Audio Interchange File Format (.aif) Waveform Audio Format (.wav)</td>
</tr>
<tr>
<td>Video data</td>
<td>MPEG-4 (.mp4) OGG video (.ogv, .ogg) motion JPEG 2000 (.mj2)</td>
<td>AVCHD video (.avchd)</td>
</tr>
<tr>
<td>Documentation and scripts</td>
<td>Rich Text Format (.rtf) PDF/UA, PDF/A or PDF (.pdf) XHTML or HTML (.xhtml, .htm) OpenDocument Text (.odt)</td>
<td>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</td>
</tr>
</tbody>
</table>

[https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats](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](http://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
What is 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
Controlled vocabularies

“MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in…….”

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

dcc.ac.uk
Why are ontologies important?

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

Include ontologies as well
Defined terms + taxonomy

Useful for selecting keywords to tag datasets

- **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
License research data openly

DCC how-to guide: [www.dcc.ac.uk/resources/how-guides/license-research-data](http://www.dcc.ac.uk/resources/how-guides/license-research-data)
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?

- Yes
- No

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

- Yes
- No

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://www.eudat.eu/services/userdoc/license-selector
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
Useful tools for documentation

E-lab notebooks, wikis, etc

- Record experiment procedures and results
- Share protocols

http://openwetware.org
Managing data

Image CC-SA-ND by Bill Dickinson www.flickr.com/photos/skynoir/8270436894

dcc.ac.uk
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
How to keep your data secure?

Develop a practical solution that fits your circumstances

– Store your data on managed servers
– Restrict access to collaborators or smaller subset
– Encrypt mobile devices carrying sensitive information
– Keep anti-virus software up-to-date
– Use secure data services for long-term sharing

www.wsj.com/articles/SB10001424052748703843804575534122591921594
Collaborative platforms e.g. OSF

Open Science Framework
A scholarly commons to connect the entire research cycle

https://osf.io

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.
Data-specific platforms e.g. OMERO

What is OMERO?

From the microscope to publication, OMERO handles all your images in a secure central repository. You can view, organize, analyze and share your data from anywhere you have internet access. Work with your images from a desktop app (Windows, Mac or Linux), from the web or from 3rd party software. Over 140 image file formats supported, including all major microscope formats.

http://www.openmicroscopy.org/site/products/omero

dcc.ac.uk
Third-party tools for collaboration

ownCloud
• Open source product with Dropbox-like functionality
• Used by many universities and service providers to offer ‘approved’ solution

https://owncloud.org

Using Dropbox and other cloud services
University RDM services e.g. Edinburgh

- DataStore
- Compute & Data Facility (HPC)
- DataSync
- Wiki service
- Subversion
- Electronic Lab Notebook
- DataShare repository
- DataVault
- Pure (research info)
- Secure data service

www.ed.ac.uk/information-services/research-support/research-data-service
One copy = risk of data loss
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
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 sharing and being Open
Why make data available?

Making plans

They sound dull, but data-management plans are essential, and funders must expressly require them.

Data aren’t just “raw” raw materials, but products that can be reconstituted and reconstituted. They produce ideas that can be reconstituted and reconstituted. And people who produce ideas can be reconstituted and reconstituted. They’re often the only tangible outcome of projects that we fund. So it’s not surprising that data management is a big business. Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.

Data management is big business.
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!

“IT was unbelievable. It's not science the way most of us have practiced in our careers. But we all realized 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 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
FOSTER Open Science toolkit

https://www.fosteropenscience.eu/toolkit

dcc.ac.uk
Open Peer Review module example

Open Peer Review
This module will introduce you to Open Peer Reviewing and let you know how you can get started with it.

Introduction
This module introduces you to open peer review (OPR), an emerging practice which is gaining momentum as part of Open Science.
Upon completing this module, you will:
- understand what OPR means and how it supports Open Science;
- be aware of OPR workflows and which aspects of the review process can be conducted openly;
- know how to write a constructive and responsible open peer review;
- know about useful tools and services that can support you putting OPR into practice.

OPR in three minutes
In this short video, Tony Ross-Hellauer introduces the concept of open peer review and explains why transparency is strongly needed in the peer review process.

What is Open Peer Review? Tony Ross-Hellauer

Quiz - Are you an Open Peer Reviewer?
Transparency can be added to peer review through:
- Accessible evaluation reports
- Platforms that allow interaction
- Revealed identities of reviewers

What are the benefits of open peer review?
- It is not biased
- My results can be published more quickly
- My review is a citable research output
Specialisation pathways

2-4 hours of content
- The reproducible research practitioner
- The responsible data sharer
- The Open Access Author
- The open peer reviewer
- The open innovator

For more information, see [www.fosteropenscience.eu/learning-paths](http://www.fosteropenscience.eu/learning-paths)
Case study approach

Using the EC Open Science Monitor approach to share practical examples of activity from the Life Sciences, Social Sciences and Humanities.

Open Access
Open Peer Review
Open Source Licensing
Open Research Data
Open Innovation

Why?
The analysis of big data in a performant and reproducible manner is an increasing pressing issue in many scientific fields including and mostly in life science disciplines. This problem has been fuelled by the combined reliance on increasingly complex data analysis methods and the exponential growth of biological datasets. When considering the installation, deployment and maintenance of bioinformatic pipelines, an even more challenging picture emerges due to the lack of community standards. Moreover, the effect of limited standards on reproducibility is amplified by the very diverse ranges of computational platforms and configurations on which these applications are expected to be applied (workstations, clusters, HPC, clouds, etc.). The Nextflow open source technology provides a simple but yet effective solutions to many of these problems.
How do you share data effectively?

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

- Document and describe it enough for others to understand, use and cite

  \url{http://www.dcc.ac.uk/resources/how-guides/cite-datasets}

- Licence it so others can reuse

  \url{www.dcc.ac.uk/resources/how-guides/license-research-data}
### Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

<table>
<thead>
<tr>
<th></th>
<th>LEVEL 0</th>
<th>LEVEL 1</th>
<th>LEVEL 2</th>
<th>LEVEL 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citation standards</strong></td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td><strong>Data transparency</strong></td>
<td>Journal encourages data sharing—or says nothing.</td>
<td>Article states whether data are available and, if so, where to access them.</td>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td><strong>Analytic methods (code) transparency</strong></td>
<td>Journal encourages code sharing—or says nothing.</td>
<td>Article states whether code is available and, if so, where to access them.</td>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td><strong>Research materials transparency</strong></td>
<td>Journal encourages materials sharing—or says nothing.</td>
<td>Article states whether materials are available and, if so, where to access them.</td>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
<td>Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td><strong>Design and analysis transparency</strong></td>
<td>Journal encourages design and analysis transparency or says nothing.</td>
<td>Journal articulates design transparency standards.</td>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
<td>Journal requires and enforces adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td><strong>Preregistration of studies</strong></td>
<td>Journal says nothing.</td>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
<td>Journal requires preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
<td>Journal requires preregistration of studies and provides link in article to meeting requirements.</td>
</tr>
<tr>
<td><strong>Preregistration of analysis plans</strong></td>
<td>Journal says nothing.</td>
<td>Journal encourages prereanalysis plans and provides link in article to registered analysis plan if it exists.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
<td>Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.</td>
</tr>
<tr>
<td><strong>Replication</strong></td>
<td>Journal discourages submission of replication studies—or says nothing.</td>
<td>Journal encourages submission of replication studies.</td>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
<td>Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.</td>
</tr>
</tbody>
</table>

[http://science.sciencemag.org/content/348/6242/1422](http://science.sciencemag.org/content/348/6242/1422)
Thank you!

In collaboration with: