Principles of Research Data Management and Open Research

S. Venkataraman, PhD and Ryan O’Connor
Research Data Specialists
Digital Curation Centre

s.venkataraman@ed.ac.uk

28th November 2019, Queen’s University, Belfast

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

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30</td>
<td>Introduction and welcome</td>
</tr>
<tr>
<td>09:45</td>
<td>Intro to RDM, Open Research and FAIR</td>
</tr>
<tr>
<td>10:45</td>
<td>Break</td>
</tr>
<tr>
<td>11:15</td>
<td>RDM exercise</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30</td>
<td>Intro to Data Management Plans</td>
</tr>
<tr>
<td>14:30</td>
<td>DMP exercise</td>
</tr>
<tr>
<td>15:00</td>
<td>Break</td>
</tr>
<tr>
<td>15:30</td>
<td>DMP exercise cont’d</td>
</tr>
<tr>
<td>16:15</td>
<td>Closing remarks</td>
</tr>
<tr>
<td>16:30</td>
<td>End</td>
</tr>
</tbody>
</table>
About the DCC

• Established in 2004
• Based in Edinburgh and Glasgow
• Works at national and international levels
• One of leading organisations in the world specialising in training, consultancy, policy making and advocacy in digital data management best practice and services provision
• Involved in many international consortia and schools
Learning outcomes

• Be familiar with the curation lifecycle
• Understand the standardisation methods and principles available to add value to your data
• Learn about resources to aid your workflows
• Increase/encourage your level of openness
• Implement and review DMPs
Language is a barrier...

Respondents mentioned 40 terms which were unclear to them in European Commission DMP

“Researchers are not familiar with the following terms/phrases: Metadata, standards for metadata/data, ontologies, mapping with ontologies, interoperability, ... All the ICT jargon”

“With the help from Swedish National Data Service we could clarify many questions. Without this help we would not be able to finish the DMP.”

Is there a reproducibility crisis?

The curation lifecycle

Create

Preserve

Document

Share

Use

Store

dcc.ac.uk
...and open research

- Change the typical lifecycle
- Publish earlier and release more
- Papers + Data + Methods + Code...
- Support reproducibility
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
### The Old Weather Project

Data for research, not from research

---

dcc.ac.uk
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
Validation of results

“It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that have arguably led to riots, poverty and lost jobs.”

www.guardian.co.uk/politics/2013/apr/18/uncovered-error-george-osborne-austerity
Cut down on academic fraud

Stapel – 55 publications – “fictitious data”
Sharing leads to breakthroughs!

"It was unbelievable. It's 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 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
How do you share data effectively?

• Use appropriate repositories, this catalogue is a good place to start
  http://www.re3data.org

• Document and describe it enough for others to understand, use and cite
  http://www.dcc.ac.uk/resources/how-guides/cite-datasets

• Licence it so others can reuse
  www.dcc.ac.uk/resources/how-guides/license-research-data
FOSTER Open Science toolkit

https://www.fosteropenscience.eu/toolkit
Research Data Alliance

https://www.rd-alliance.org

dcc.ac.uk
Who has heard of this before…?

Findable

Accessible

Interoperable

Reusable

Image CC-BY-SA by SangyaPundir
Familiarity with FAIR principles

The majority of researchers surveyed as part of a recent study on open data had never heard of FAIR, regardless of their field. Of the 748 researchers that responded to this question, 144 said they were familiar with the principles. Circles are sized by number of respondents.

- I am familiar with the FAIR principles
- I have previously heard of the FAIR principles but I'm not familiar with them
- I've never heard of the FAIR principles before now

Source: State of Openness

Brock, J. "A love letter to your future self": What scientists need to know about FAIR data *Nature Index* 11 Feb 2019
Compliance with FAIR principles

Of the participants who were familiar with FAIR, about a third said that their data management practices were very compliant with the principles. That proportion is similar across scientists at different stages of their career.

Source: State of Opencare: Data sharing and research management best practices

Brock, J. "A love letter to your future self": What scientists need to know about FAIR data Nature Index 11 Feb 2019
Which of the FAIR principles do you think most needs better definition?
Interoperability is the least understood FAIR principle. Some 42% of the 187 respondents who answered this question felt that it needed further clarification.

- F (Findable): 13%
- A (Accessible): 19%
- I (Interoperable): 42%
- R (Reusable): 26%

Source: State of Openness
Brock, J. "A love letter to your future self": What scientists need to know about FAIR data Nature Index 11 Feb 2019
European perspective...
Research data: institutional crown jewels?

Give us back our crown jewels

Our taxes fund the collection of public data - yet we pay again to access it. Make the data freely available to stimulate innovation, argue Charles Arthur and Michael Cross

Charles Arthur and Michael Cross
The Guardian, Thursday 9 March 2006
Article history

http://www.flickr.com/photos/lifes_too_short_to_drink_cheap_wine/4754234186
What FAIR means: 15 principles

**Findable:**
F1. (meta)data are assigned a globally unique and persistent identifier;
F2. data are described with rich metadata;
F3. metadata clearly and explicitly include the identifier of the data it describes;
F4. (meta)data are registered or indexed in a searchable resource;

**Accessible:**
A1. (meta)data are retrievable by their identifier using a standardized communications protocol;
A1.1 the protocol is open, free, and universally implementable;
A1.2. the protocol allows for an authentication and authorization procedure, where necessary;
A2. metadata are accessible, even when the data are no longer available;

**Interoperable:**
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles;
I3. (meta)data include qualified references to other (meta)data;

**Reusable:**
R1. meta(data) are richly described with a plurality of accurate and relevant attributes;
R1.1. (meta)data are released with a clear and accessible data usage license;
R1.2. (meta)data are associated with detailed provenance;
R1.3. (meta)data meet domain-relevant community standards;

doi: 10.1038/sdata.2016.18

COMPREHENSIVE DESCRIPTIONS CAN BE FOUND AT https://www.go-fair.org/fair-principles/
Common misconceptions

- FAIR data does not have to be open
- The principles do not specify particular technologies or implementations e.g. semantic web
- FAIR is not a standard to be followed or strict criteria – it’s a spectrum / continuum
- It doesn’t only apply to the life sciences
All research data

the wild

Managed data

FAIR data

Open data
Increasing that which is FAIR & open

Managed data

FAIR data

Open data

the wild
as open as possible, as closed as necessary
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
Create

Preserve

Document

Share

Use

Store
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</td>
<td>SPSS portable format (.por)</td>
<td>proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)</td>
</tr>
<tr>
<td>variable labels, code labels, and defined missing values</td>
<td>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></td>
</tr>
<tr>
<td>Tabular data with minimal metadata</td>
<td>comma-separated values (.csv)</td>
<td>delimitext (.txt) with characters not present in data used as delimiters</td>
</tr>
<tr>
<td>column headings, variable names</td>
<td>tab-delimited file (.tab)</td>
<td>widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)</td>
</tr>
<tr>
<td>Geospatial data</td>
<td>ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional)</td>
<td>ESRI Geodatabase format (.mdb)</td>
</tr>
<tr>
<td>vector and raster data</td>
<td>geo-referenced TIFF (.tif, .tiff)</td>
<td>MapInfo Interchange Format (.mif) for vector data</td>
</tr>
<tr>
<td></td>
<td>CAD data (.dwg)</td>
<td>Keyhole Mark-up Language (.kml)</td>
</tr>
<tr>
<td></td>
<td>tabular GIS attribute data</td>
<td>Adobe Illustrator (.ai), CAD data (.dxf or .svg)</td>
</tr>
<tr>
<td></td>
<td>Geography Markup Language (.gml)</td>
<td>binary formats of GIS and CAD packages</td>
</tr>
<tr>
<td>Textual data</td>
<td>Rich Text Format (.rtf)</td>
<td>Hypertext Mark-up Language (.html)</td>
</tr>
<tr>
<td></td>
<td>plain text, ASCII (.txt)</td>
<td>widely-used formats: MS Word (.doc/.docx)</td>
</tr>
<tr>
<td></td>
<td>eXtensible Mark-up Language (.xml)</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GIF (.gif)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIFF other versions (.tif, .tiff)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RAW image format (.raw)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photoshop files (.psd)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BMP (.bmp)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PNG (.png)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>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</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Audio Interchange File Format (.aif)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Waveform Audio Format (.wav)</td>
</tr>
<tr>
<td>Video data</td>
<td>MPEG-4 (.mp4)</td>
<td>AVCHD video (.avchd)</td>
</tr>
<tr>
<td></td>
<td>OGG video (.ogv, .ogg)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>motion JPEG 2000 (.mj2)</td>
<td></td>
</tr>
<tr>
<td>Documentation and scripts</td>
<td>Rich Text Format (.rtf)</td>
<td>plain text (.txt)</td>
</tr>
<tr>
<td></td>
<td>PDF/UA, PDF/A or PDF (.pdf)</td>
<td>widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx)</td>
</tr>
<tr>
<td></td>
<td>XHTML or HTML (.xhtml, .htm)</td>
<td>XML marked-up text (.xml) according to an appropriate DTD or</td>
</tr>
<tr>
<td></td>
<td>OpenDocument Text (.odt)</td>
<td>schema, e.g. XHMTL 1.0</td>
</tr>
</tbody>
</table>

[Link](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

- 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
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
What are 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/

https://rdamsc.dcc.ac.uk/
Controlled vocabularies

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

Example courtesy of Ken Haug, European Bioinformatics Institute (EMBL-EBI)
...and ontologies?
e.g. SNOMED CT (clinical terms) or MeSH

- Defined terms + taxonomy
- Useful for selecting keywords to tag datasets
- You can find many ontologies in the [BARTOC catalogue](https://www.bartoc.eu) and elsewhere

**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 \( n \)
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
Collaborative platforms e.g. OSF

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

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.

https://osf.io
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
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
Primary and secondary data

Reuse

Create

Preserve

Document

Share

Use

Store

dcc.ac.uk
License research data openly

Part of How To Attribute Creative Commons Photos by Foter, licensed CC BY SA 3.0
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 your 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://ufal.github.io/public-license-selector/
Deposit in a data repository

The Re3data catalogue can be searched to find a home for data

www.re3data.org
Criteria for selecting a repository

- Better to use a domain specific repository if available

- Check they match particular data needs e.g. formats accepted, mixture of Open and Restricted Access.

- Do they assign a persistent and globally unique identifier for sustainable citations and to links back to particular researchers and grants?

- Look for certification as a ‘Trustworthy Digital Repository’ with an explicit ambition to keep the data available in long term.

Icons to note open access, licenses, PIDs, certificates…

dcc.ac.uk
What is a Persistent Identifier (PID)?

*a long-lasting reference to a document, file or other object*

- PIDS come in various forms e.g. ORCID, DOI, ISBN...
- Typically they’re actionable i.e. type it into web browser to access
- Many repositories will assign them on deposit
PID Graphs – the next level

• If you have a collection of PIDs describing different objects, these can be joined together in a graph to form relationships
• These graphs can aid in workflows and provenance
Citing research data: why?

Building a Culture of Data Citation

**CREATE**
- Australian researcher creates a research dataset and a publication related to the dataset
- Dataset is stored in a publicly accessible repository
- Researcher uses ANDS services to mint a Digital Object Identifier (DOI) for the dataset

**REWARD**
- Researcher future funding and promotion influenced by dataset citation metrics
- Funding and research groups review publication and dataset citation metrics

**MEASURE**
- Citation metrics services (e.g., Thomson Reuters Data Citation Index, Scopus) accumulate citation references to the dataset and publication

**USE**
- Research community use the DOI to access the dataset and carry out related research
- Research community generate new publications using the DOI to reference the dataset

**do**
- doi is used in data citation

http://ands.org.au/cite-data
Questions?
Exercise - 45 min (+ 30 min discussion)

Imagine you are a biologist who is doing microscopy experiments imaging tissue specimens. The data captured by the imaging is 100s of GB in size and is then cleaned and analysed to produce derivatives of the original captured data. Some of these derivatives may eventually be published. In preparation for publication, the data will also be segmented and annotated using standard ontologies. Documentation will also include metadata standards that will sufficiently describe the experimental procedure to allow reproducibility. Publication of the data is mandatory due to funder policy and must be deposited in a repository within 3 years of data production and must use an open licence without restrictions on reuse.

Now…please split into groups and see if you can answer the following questions using the tools and guidelines that have been described:

- What **file format(s)** should data be captured/preserved in?
- Which **metadata standard(s)** should be used?
- What **ontology(ies)** should be used?
- Which **licence(s)** should be used?
- Which **repository** would be the best fit for these data?
- Do you foresee any problems with the data?

(Hint: not all the questions can be answered definitively! – but why not?)
Introduction to Data Management Plans

S. Venkataraman, PhD and Ryan O’Connor
Research Data Specialists
Digital Curation Centre

s.venkataraman@ed.ac.uk

28th November 2019, Queen’s University, Belfast

This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License
What is a data management plan (DMP)?

A brief plan written at the start of a project to define:
• how the data will be created?
• how it will be documented?
• who will access it?
• where it will be stored?
• who will back it up?
• whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data.
Why make DMPs?

Making plans
They sound dull, but data management plans are essential, and funders must insist on them.

Data is the high-yielding crop of scientific research, but managing it well is the trick. Researchers need to plan how they will store, access, use, and dispose of their data. For complex data, this means understanding how the data will be structured and how it will be signified in some meaningful way. By the time the data is ready to be shared, funders need to ensure that it is securely stored and that it is free of any intellectual property claims.

Making project data freely available is vital for open science.

For the record
Making plans before the data is created is key. It involves outlining the purpose of the data, ensuring it is stored in a way that allows others to access and use it. It also involves considering how the data will be shared and what kind of access will be granted. Without a plan, the data may become difficult to access or may be lost.

https://www.nature.com/articles/d41586-018-03071-1
doi: 10.1038/d41586-018-03071-1

dcc.ac.uk
Digital Curation Centre
Why make DMPs?

(Not for the research funder, but for life we make data management plans)

- Make informed decisions to anticipate and avoid problems
- Avoid duplication, data loss and security breaches
- Develop procedures early on for consistency
- Ensure data are accurate, complete, reliable and secure
- Save time and effort to make your life easier!
Don’t undervalue research data
DCC Checklist for a DMP

The DCC assessed existing funder requirements, DMP templates and other best practice to see what should be included in plans. This was synthesised down into common themes and questions.

• 13 questions on what’s asked across the board
• Prompts / pointers to help researchers get started
• Guidance on how to answer

Common themes in DMPs

1. Description of data to be collected / created
   (i.e. content, type, format, volume...)

2. Standards / methodologies for data collection & management

3. Ethics and Intellectual Property
   (highlight any restrictions on data sharing e.g. embargoes, confidentiality)

4. Plans for data sharing and access
   (i.e. how, when, to whom)

5. Strategy for long-term preservation
Planning trick 1: think backwards

What data organisation would a re-user like?

Design how you will organise data in the project (folder structure, file naming convention, …)
Planning trick 2: include RDM stakeholders

Commercial partners
Publishers
Data
Availability policy

Institution
RDM policy
Facilities

Researchers
Front office

Back office
data centers

Information and awareness
Training
Storage

Research funders
www.openaire.eu/briefpaper-rdm-infonoads
Planning trick 3: ground your plan in reality

Base plans on available skills, support and good practice for the field – show it’s feasible to implement
What makes a good DMP?

- Clear, detailed information that is relevant to the science
  - adopting recognised standards
  - practices in line with norms for that field
  - use of support services e.g. university storage, subject repositories...
- Realistic approach that is feasible to implement
- Evidence of consultation and seeking advice
- Proper justification of restrictions and costs

Have you taken time to reflect on what to do?
Is the information specific enough?

“we will use suitable formats to ensure that our data can be preserved and sustained over the long term”

• Which standards? Name them!
• Show that you know which are suitable
• Does your chosen repository have preferences?
Are decisions justified?

“data will be made available upon request to bona fide medieval historians”

- Why is it restricted?
- Could other communities not reuse the data?
- Will the research team be around to handle access requests in the future?
A better response…

“We will provide MP3 audio files for online dissemination. While this is not an open format, it is well-established and the most widely supported. High-resolution WAV files will be used for the archival master recordings.”

• Be clear, specific and detailed

• Justify decisions
Example plans

Plans from several funders and disciplines via DCC
www.dcc.ac.uk/resources/data-management-plans/guidance-examples

Scientific DMPs submitted to the NSF (USA) provided by DataOne
https://www.dataone.org/data-management-planning

DMPs published in RIO journal

Share yours! - www.dcc.ac.uk/share-DMPs
Data description examples

The final dataset will include self-reported demographic and behavioural data from interviews with the subjects and laboratory data from urine specimens provided.

From NIH data sharing statements

Every two days, we will subsample E. affinis populations growing under our treatment conditions. We will use a microscope to identify the life stage and sex of the subsampled individuals. We will document the information first in a laboratory notebook and then copy the data into an Excel spreadsheet. The Excel spreadsheet will be saved as a comma separated value (.csv) file.

From DataOne – E. affinis DMP example
Metadata examples

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

From ICPSR Framework for Creating a DMP

We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a .txt document that will be stored with the data file. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. EML is one of the accepted formats used in ecology, and works well for the types of data we will be producing. We will create these metadata using Morpho software, available through KNB. The metadata will fully describe the data files and the context of the measurements.

From DataOne – E. affinis DMP example
Data sharing examples

The videos will be made available via the bristol.ac.uk website (both as streaming media and downloads) HD and SD versions will be provided to accommodate those with lower bandwidth. Videos will also be made available via Vimeo, a platform that is already well used by research students at Bristol. Appropriate metadata will also be provided to the existing Vimeo standard.

All video will also be available for download and re-editing by third parties. To facilitate this Creative Commons licenses will be assigned to each item. In order to ensure this usage is possible, the required permissions will be gathered from participants (using a suitable release form) before recording commences.

From University of Bristol Kitchen Cosmology DMP

We will make the data and associated documentation available to users under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

From NIH data sharing statements
Examples restrictions

Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement.

From NIH data sharing statements

1. Share data privately within 1 year. 
   *Data will be held in Private Repository, but metadata will be public*

2. Release data to public within 2 years. 
   *Encouraged after one year to release data for public access.*

3. Request, in writing, data privacy up to 4 years. 
   *Extensions beyond 3 years will only be granted for compelling cases.*

4. Consult with creators of private CZO datasets prior to use. 
   *Pis required to seek consent before using private data they can access*

From Boulder Creek Critical Zone Observatory DMP
Archiving examples

The investigators will work with staff at the UKDA to determine what to archive and how long the deposited data should be retained. Future long-term use of the data will be ensured by placing a copy of the data into the repository.

From ICPSR Framework for Creating a DMP

Data will be provided in file formats considered appropriate for long-term access, as recommended by the UK Data Service. For example, SPSS Portal format and tab-delimited text for qualitative tabular data and RTF and PDF/A for interview transcripts. Appropriate documentation necessary to understand the data will also be provided. Anonymised data will be held for a minimum of 10 years following project completion, in compliance with LSHTM’s Records Retention and Disposal Schedule. Biological samples (output 3) will be deposited with the UK BioBank for future use.

From Writing a Wellcome Trust Data Management and Sharing Plan
DCC support on DMPs

- Webinars and training materials
- How-to guides and other advisory documents
- Checklist on what to cover in DMPs
- Example DMPs
- DMPonline

www.dcc.ac.uk/resources/data-management-plans
Guidance from elsewhere

Framework for Creating a Data Management Plan

This framework can be used as an outline in assembling data management plans to accompany grant applications. Note that some funders have page limits for data management plans—NSF limits plans to two pages.

Elements of a Data Management Plan

This list of elements is informed by a gap analysis that ICPSR conducted of existing recommendations for data management plans and other forms of guidance made available for researchers generating data. The result of the gap analysis was a comparison of existing forms of guidance. Elements that are highly recommended for inclusion in effective data management plans are noted.

See our bibliography for additional readings germane to the elements of a data management plan.

Data Description (Recommended)

Provide a brief description of the information to be gathered -- the nature, scope, and scale of the data that will be generated or collected.

Why this is important:
A good description of the data to be collected will help reviewers understand the characteristics of the data, their relationship to existing data, and the role of data management.

Example 1:
This project will produce public-use nationally representative survey data for the United States covering Americans' social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life.

Example 2:
This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns of drug use among a nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released in a form that does not identify individuals.

Think about why the questions are being asked – why is it useful to consider that topic?

Look at examples to help you understand what to write

www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html
What is DMPonline?
A web-based tool to help researchers write data management plans

https://dmponline.dcc.ac.uk
Main features in DMPonline

- Templates for different requirements (funder or institution)
- Tailored guidance (funder, institutional, discipline-specific etc)
- Ability to provide examples and suggested answers
- Supports multiple phases (e.g. pre- / during / post-project)
- Granular read / write / share permissions
- Customised exports to a variety of formats
- Shibboleth authentication
Key messages

• Data management is part of good practice whether you plan to make the data open or not
  – it benefits you!

• The process of planning is as important as the DMP. Think about the desired end result and plan for this.

• Approach DMPs in whatever way best fits your project. Don’t just let funder requirements drive things.
Thank you!

For DCC resources see: www.dcc.ac.uk/resources

Follow us on twitter: @digitalcuration and #ukdcc