Creating a Data Management Plan (DMP)

Dr. Chris Emmerson
Research Data Manager
Welcome

- Research Data Service
- RDM landscape
- Data lifecycle
- Data management Planning
  - 6 themes
- Questions
Research Data Service
Research Data Management

Managing and publishing research data is part of good research conduct. Research funders also require data to be made openly available with as few restrictions as possible. The Research Data Service has been created to assist researchers during the three key stages of the data lifecycle: before, during and after a project.

Before a Project

Discover the benefits of writing a data management plan (DMP) and research funder expectations.

During a Project

Find out the options available to store research data during your research project.

After a Project

Training
RDM landscape
RDM landscape

• **Funder requirements**
  – RCUK Common Principles, 2011
  – EPSRC, 2015
  – EU, 2016
  – UKRI Research Concordat on Open Research Data, 2018

• **Newcastle University Policy**

• Key element of good research conduct – not new
Funder expectations

1. Make data openly available where possible
2. Have policies & plans. Preserve data of long-term value
3. Metadata for discovery / reuse. Link to data from publications
4. Be mindful of legal, ethical and commercial constraints
5. Allow limited embargoes to protect the effort of creators
6. Acknowledge sources to recognise IP and abide by T&Cs
7. Ensure cost-effective use of public funds for RDM
Funder expectations

1. Make data openly available where possible
2. Have policies & plans. Preserve data of long-term value
3. Metadata for discovery / reuse. Link to data from publications
4. Be mindful of legal, ethical and commercial constraints
5. Allow limited embargoes to protect the effort of creators
6. Acknowledge sources to recognise IP and abide by T&Cs
7. Ensure cost-effective use of public funds for RDM

• Two main drivers
  – Public good
  – Digital data
Digital data

The rise of digital data

- Volume of data (data deluge) – rapid increase
- Computing power – Moore’s law
- Global economy – waking up to the value of data
Research data lifecycle

RDM is “the active management and appraisal of data over the lifecycle of scholarly and scientific interest”
Research data lifecycle

Plan

Create

Process

Preserve

Share

Reuse

Public Storage

Archival Storage

Active Storage

Dispose

Research Activity
Data lifecycle
Research data lifecycle

Before: Plan, Create
During: Process
After: Preserve, Reuse, Share

Public Storage
Archival Storage
Active Storage

Reuse
Share
Preserve

Active Storage
Public Storage
Archival Storage
Data Management Plans (DMPs)

Create

Reuse

Share

Preserve

Public Storage

Archival Storage

Active Storage

Before

During

Active

Storage

Public

Storage

Archival

Storage

Reuse

Share

Preserve
Planning

• To help you manage your data
• To make informed decisions so you don’t have to figure out things as you go
• To anticipate and avoid problems e.g. data loss or not enough storage
• To ensure your data continue to be usable throughout and beyond your project
• To help you win ethical approval
• To make your life easier!
  – post viva
DMPs

• Who requires them?
  • Funding bodies
  • Newcastle University (encourages)

• When are they required?
  – Most at application stage / project start but:
    • They should become a living document
What to include in the plan?

“I have a plan so cunning, you could put a tail on it and call it a fox.”
What to include?

1

2

3

4

5

6
What to include?

1. Data types, format, standards, scale and method

2. 

3. 

4. 

5. 

6. 
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data types, format, standards, scale and method</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Metadata content and format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What to include?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Data types, format, standards, scale and method</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Metadata content and format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Project storage and data security</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What to include?

1. Data types, format, standards, scale and method
2. Metadata content and format
3. Project storage and data security
4. Data sharing
5. 
6. 
What to include?

1. Data types, format, standards, scale and method
2. Metadata content and format
3. Project storage and data security
4. Data sharing
5. Long-term preservation
What to include?

1. Data types, format, standards, scale and method
2. Metadata content and format
3. Project storage and data security
4. Data sharing
5. Long-term preservation
6. Roles, responsibilities and resourcing
1. Data types, format, standards, scale and method

• State the relationship to other data available
  – Existing data sources
  – Gaps between available data and that required for the research
1. Data types, format, standards, scale and method

- State the relationship to other data available
  - Existing data sources
  - Gaps between available data and that required for the research
1. Data types, format, standards, scale and method

• State the relationship to other data available
  – Existing data sources
  – Gaps between available data and that required for the research
1. Data types, format, standards, scale and method

• Description of data to be produced
  – Observational, experimental, simulation
  – Digital data?

• How the data will be acquired
  – Method

• Detail the file formats and scale of the data
1. Data types, format, standards, scale and methods

**Information on new data:** In summary the project will create the following data:

- ~90 key informant, policy, industry interviews generally lasting up to one hour. These will be audio recorded (~10-15Mb each MP3 format) (where allowed, contemporaneous notes taken where not), transcribed & translated to English where needed and converted to RTF for storage (~150 files (original and translation) (0.3 Mb each);

- ~90 hour long or longer household interviews. These will be audio recorded (~15-20Mb each MP3 format), transcribed (& translated to English where needed) and converted to RTF for storage (~150 files (original and translation) (0.3 Mb each);

- 30 ethnographic studies of household practices. This will generate fieldnotes, audionotes (c30hrs MP3 ~300Mb total), ~ 300 photographs (JPG 1Mb each), ~30 go-along interview recordings (c30hrs total, MP3 ~300Mb), video recordings of practices (~40 episodes each of ~5mins, each at 40Mb MPEG), ~30 household inventories (0.3Mb each RTF files), transcriptions and translations of audio material (~40 RTF 0.3MB each)~30 in-depth studies of digital media usage, including screen shots (~50 JPG 0.4Mb each), mined data from social media (raw feeds, GNIP power track metadata & archive of reports in CSV/RTF format integrable with Nvivo, ~1Gb)

- ~1500 response social survey with demographic variables, responses to questions (stored as CSV data)
2. Metadata content and format

Metadata: data about data

A crucial part of making data user-friendly, shareable, and useable

Do you understand your older work?

• What metadata will be recorded
  – Helps makes data meaningful in the future
  – Study level (research context, design, methods)
  – Data level (labelling and documenting)
2. Metadata content and format

• How the metadata will be created and/or captured
  – Lab notebooks
  – Project database (.CSV)

• What format will be used for the metadata
  – Standards exist
3. Project storage and data security

- Where are you going to store your PhD data?
3. Project storage and data security

Describe the quality assurances of data storage

- Storage options
  - H: drive (4/5GB)
  - Shared school storage (??)
  - Back up procedures, security
3. Project storage and data security

Office 365

- **OneDrive**
  - Providing secure online cloud storage
  - Can access offline
  - One off sharing

- **Groups / Teams**
  - Collaborative tool
  - Inbox, calendar, file library, OneNote
3. Project storage and data security

- De-centralised storage
  - USB devices, external hard disks, laptops, PCs, instrument servers, cloud storage
  - Data fragmented on different devices/servers
  - Often 1 copy in 1 location

- Issues
  - No standard approach, data at risk or loss of data
  - No realisation that the data are the real asset of research
  - Estimating volume of data is unfamiliar
3. Project storage and data security

A Dropbox security breach in 2012 has affected more than 68 million account holders, according to security experts.

Last week, Dropbox reset all passwords that had remained unchanged since mid-2012 “as a preventive measure”.

In 2012, Dropbox had said hacks on "other websites" had affected customers who used their Dropbox accounts through these other services.

---

**Cloud Storage**

Credit: Patty - photos - Flickr

---

Apple scammers have their heads in the iCloud

iPad and iPhone users warned not to fall for fake emails and texts aimed at tricking them into handing over their login details.

---

Hacker collects 272m email addresses and passwords, some from Gmail

Security firm announces it has persuaded fraudster to give up database of email addresses along with passwords users use to log in to websites.
3. Project storage and data security

• Legal requirements – **GDPR, Data Protection Act 2018**
  – Research purposes
  – Don’t cause damage or distress to participants
  – Don’t identify research participants
  – Update consent and information sheets

• Anonymising data
  – Can’t identify subjects
  – Code sheet held separately
  – Resources at the end
3. Project storage and data security

• Structured:
  – Hierarchical file structure
  – Balance between breadth and depth
  – Name folders after work – not individuals
  – Review your files and folders regularly

• Best practice for files:
  • Be consistent
  • Number the versions in an appropriate scale – 01/001
  • Version control table for important documents
  • Year at start of the file YYYY_MM_DD
  • Try to avoid spaces/gaps
3. Project storage and data security

- What would you do?
  - Lost
  - Stolen
  - Damaged

- Have back-ups

CASH REWARD
for returning my lost backpack

- Black [AK] Burton Rucksack
- Lost on Friday 15 July at 6 pm in the Panton Arms pub 43, Panton St, Cambridge
- Containing a laptop (white MacBook), a black external hard drive and scientific research documents

Thank you!!
4. Data sharing

- Encourages scientific enquiry and debate
- Promotes innovation and potential new data uses
- Leads to new collaborations between data users and data creators
- Maximises transparency and accountability
- Enables scrutiny of research findings
- Encourages the improvement and validation of research methods
- Reduces the cost of duplicating data collection
- Increases the impact and visibility of research
- Provides credit to the researcher as a research output in its own right
- Provides great resources for education and training
4. Data sharing

sharing is caring
4. Data sharing

- Analysis of over 100 years of data about Australian bushfire deaths
- Significantly changed government and emergency services thinking in this area
- Enabled new understanding about decisions made by victims caught in bushfires
- Directly influenced the ‘stay and defend or leave early response advice’

#dataimpact campaign
4. Data sharing

• Obligations to share
  – Funder/ institution / personal

• Details of data sharing
  – When
  – Where (data repository)
  – Who can access?

• Restrictions to data access
  – Ethical / commercial / IP
  – Period of exclusive use
4. Data sharing

Findable, Accessible, Interoperable, Reusable

Before, During, After
4. Data sharing

- Deposit into a research data repository that provides a robust identify (e.g. DOI)
  - deposit in a funder repository
  - deposit in a specialist repository
  - deposit in an institutional repository
  - deposit in a general repository
  - submitting to a journal to support a publication
  - dissemination via a project or institutional website
  - informal peer-to-peer exchange
4. Data sharing

- Persistent identifier (DOI) creates discovery metadata
- Searchable through google as well as the repository
- DOI included in research outputs:
  - Papers
  - Posters
  - Reports
  - Thesis
4. Data sharing

- British Antarctic Survey
- NERC Science of the Environment
- European Nucleotide Archive
- re3data.org
- Findable
- Zenodo
- DataCite
- UK Data Archive

A digital repository for Newcastle University's research data

View the data catalogue
Research data management guidance
Deposit research data
Submit deposited data to archive

Before
During
After
4. Data sharing

Institutional Research Data Repository
April 2019
4. Data sharing

• Accessible and applicable language in describing the data

• Outline access criteria for the datasets (metadata)

• “as open as possible as closed as necessary”

• Justify any restrictions
  – Ethical / commercial / IP
  – Period of exclusive use
4. Data sharing

Sharing is encouraged, but not suitable for all data:

• Ethical duty
• Informed consent
• Enable researcher to gain maximum benefit from findings

<table>
<thead>
<tr>
<th>For</th>
<th>Against</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meet funder obligations</td>
<td>Commercial implications</td>
</tr>
<tr>
<td>Encourage research uptake</td>
<td>Data can’t be anonymised</td>
</tr>
<tr>
<td>Higher citation rate</td>
<td>3rd party constraints</td>
</tr>
<tr>
<td>Encourage validation of results</td>
<td>Reduced participant involvement</td>
</tr>
</tbody>
</table>

3rd party constraints

Data can’t be anonymised

Commercial implications

Meet funder obligations

Encourage research uptake

Higher citation rate

Encourage validation of results

For

Against
5. Long-term preservation

• What data will be preserved
  – Not all data

• Details of data preservation
  – How long (10 years)

• Data transformations
  – File format migration
  – Data prepared inline with repository
5. Long-term preservation

- Clear licence, allow data usage
  - Creative commons

- Accurate information on provenance

- Supporting documentation
  - Helps makes data meaningful in the future
  - Study level (research context, design, methods)
  - Data level (labelling and documenting)

- Open file formats
## 5. Long-term preservation

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Recommended formats</th>
<th>Acceptable formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data</td>
<td>Comma-separated values (.csv), tab-delimited file (.tab)</td>
<td>SPSS (.sav), MS Access (.widely-used formats: MS Excel (.xls/.xlsx)</td>
</tr>
<tr>
<td></td>
<td>PSS (.por)</td>
<td>mdb/.accdb)</td>
</tr>
<tr>
<td>Textual data</td>
<td>Rich Text Format (.rtf), plain text (.txt)</td>
<td>Hypertext Mark-up language (.html)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Widely-used formats: MS Word (.doc/.docx)</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>PNG (.png)</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>Video data</td>
<td>MPEG-4 (.mp4)</td>
<td>AVCHD video (.avchd)</td>
</tr>
<tr>
<td>Documentation and</td>
<td>Rich Text Format (.rtf)</td>
<td>Plain text (.txt)</td>
</tr>
<tr>
<td>scripts</td>
<td>PDF/A (.pdf/a)</td>
<td>Widely-used formats: MS Word (.doc/docx), MS Excel (.xls/.xlsx)</td>
</tr>
</tbody>
</table>
6. Roles, responsibilities and resourcing

- Outline the roles and responsibilities for the implementing the DMP
  - Who’s responsible for data management and updating DMP
- What resources will be required
  - Storage cost, hardware/software support, training
Review

• A DMP defines:

1. What data will be created and how? (Does the data exist)
2. How the data will be structured and stored?
3. How will you share and preserve the data?
Summary

• Good practice in RDM is essential for high quality research

• Planning is your first step – DMPs

• Consider the nature of the data at each stage of the lifecycle

• Your data - your responsibility

• Support exists within the University
RDM Support

Find out more
http://research.ncl.ac.uk/rdm/

Contact us
rdm@ncl.ac.uk

Chris Emmerson
Research Data Manager
“So many boxes”

http://youtu.be/N2zK3sAtr-4
Resources

• Online guidance: http://research.ncl.ac.uk/rdm/
• Pilot repository: https://research.ncl.ac.uk/rdm/afteraproject/rdr/
• Guides:
  – DCC’s How to develop a data management and sharing plan
  – UKDA’s Managing and sharing data
• Online DMP tool reflecting funders’, institutional and disciplinary requirements (UoN version coming soon)
• Tutorials:
  – University of Edinburgh: MANTRA research data management training course
  – University of Bristol: Research data bootcamp
Resources

• Consent:
  – [UK Data Archive](#)

• Anonymisation
  – [UK Data Archive](#)

• General Repositories
  – [Datacite](#)
  – [Re3Data](#)
  – [Figshare](#)
  – [Zenodo](#)
  – [Research Data Discovery Service (Alpha)](#)
  – [Dryad](#) (life sciences)