Write a Data Management Plan

Veerle Van den Eynden
UK Data Service

UK Data Archive 2016
Overview

- Data policies and funder requirements
- Tools and templates
- Topics
- Exercise
Why data management planning

A data management and sharing plan helps researchers consider: when research is being designed and planned, how data will be managed during the research process and shared afterwards with the wider research community.

Research benefits

- think what to do with research data, how collect, how look after
- keep track of research data (e.g. staff leaving)
- identify support, resources, services needed
- plan storage, short & long-term
- plan security, ethical aspects
- be prepared for data requests (FoI, funder)
Why data management planning

• Many research funders require planning for data management and data sharing in research applications
• Expect to cost sustainable data management and sharing into research

• Overview of requirements:
  • Digital Curation Centre, Funders’ data plan requirements
  • Knight, G. (2012) Funder Requirements for Data Management and Sharing. London School of Hygiene and Tropical Medicine, London.
Research funder data policies (RCUK)

- Publicly funded research data are a public good, produced in the public interest, that should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- in accordance with relevant standards and community best practice
- metadata to make research data discoverable
- legal, ethical, commercial constraints on release of research data
- recognition for collecting & analysing data; limited privileged use
- acknowledge sources of data, intellectual contributions, terms & conditions
- use public funds to support the management and sharing of publicly-funded research data

Research Councils UK Common Principles on Data Policy (2011)
Guidance on best practice in the management of research data (2015)
Concordat on Open Research Data (draft, 2015)
Research funder data policies (RCUK)

Research Councils:

- Data sharing policy mandating or encouraging data sharing
- Data management / sharing planning required
- Award holders responsible for managing & sharing data, except EPSRC
- Fund data sharing support services and infrastructure
  
  e.g. UK Data Service (ESRC)
  
  NERC data centres (NERC)
  
  MRC Data Support Service (MRC)
  
  Atlas Petabyte Storage (STFC)
  
  Archaeology Data Service (AHRC)
<table>
<thead>
<tr>
<th>Funder</th>
<th>Plan required?</th>
<th>Required at application</th>
<th>Data topics in DMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts and Humanities Research Council (AHRC)</td>
<td>Yes</td>
<td>Technical plan</td>
<td>Standards, preservation, continued access and use</td>
</tr>
<tr>
<td>Biotechnology and Biological Sciences Research Council (BBSRC)</td>
<td>Yes</td>
<td>Data management and sharing plan</td>
<td>Type, format, standards, sharing methods, restrictions, sharing timeframe</td>
</tr>
<tr>
<td>Cancer Research UK (CRUK)</td>
<td>Yes</td>
<td>Data sharing plan</td>
<td>Volume, format, standards, metadata, documentation, sharing method, timescale, preservation, restrictions</td>
</tr>
<tr>
<td>Department for International Development (DFID)</td>
<td>Yes</td>
<td>Access and data management plan</td>
<td>Repositories, limits, timescale, responsibilities, resources, access strategy</td>
</tr>
<tr>
<td>Engineering and Physical Sciences Research Council (EPSRC)</td>
<td>No</td>
<td>Policy framework at institutional level (from 2015)</td>
<td></td>
</tr>
<tr>
<td>Economic and Social Research Council (ESRC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Volume, type, quality, archiving plans, difficulties sharing, consent sharing, IPR, responsibilities</td>
</tr>
<tr>
<td>Medical Research Council (MRC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Collection methods, documentation, standards, preservation, curation, security, confidentiality, sharing and access, timescale, responsibilities</td>
</tr>
<tr>
<td>Natural Environment Research Council (NERC)</td>
<td>Yes</td>
<td>Outline data management plan</td>
<td>Data management procedures, created data</td>
</tr>
<tr>
<td>Science and Technology Facilities Council (STFC)</td>
<td>Yes</td>
<td>Data management plan</td>
<td>Type, preservation, metadata, value, sharing, timescale, resources needed</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>Yes</td>
<td>Data management and sharing plan</td>
<td>What data, when share, where share, how access, limits, how preserve, what resources</td>
</tr>
</tbody>
</table>
Research funder data policies (EU)

European open access policies: Horizon 2020, European Research Council (ERC)

- communication & recommendation on access to / preservation of scientific information (July 2012) (publications & research data)
- pilot on open access to research data, primarily data underlying (open access) scientific publications for H2020
- FAQ open access to publications & data in Horizon 2020
- data management guidelines for Horizon 2020 (~ policies)
- DMP is WP deliverable after 6 months of project start
## Overview of USA funder requirements for data management and sharing plans

<table>
<thead>
<tr>
<th>Funder</th>
<th>Data topics in DMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Science Foundation (NSF), includes Social, Behavioral and Economic Sciences Directorate; other directorates have different requirements</td>
<td>Expected data, retention, how share data, how manage data, legal/ethical restrictions on access, metadata, data storage &amp; preservation, data format &amp; dissemination, roles &amp; responsibilities</td>
</tr>
<tr>
<td>National Institutes of Health (NIH); data sharing plan required if funding is over $500,000</td>
<td>Data sharing</td>
</tr>
<tr>
<td>Gordon and Betty Moore Foundation (GBMF)</td>
<td>Data description, data management, data sharing</td>
</tr>
<tr>
<td>Institute of Museum and Library Services (IMLS)</td>
<td>Data description, data restrictions, documentation, IPR, metadata, storage, access, archiving and sharing</td>
</tr>
<tr>
<td>National Endowment for the Humanities (NEH)</td>
<td>Expected data, roles &amp; responsibilities, data retention, data formats &amp; dissemination, storage &amp; preservation</td>
</tr>
<tr>
<td>National Oceanic and Atmospheric Administration (NOAA)</td>
<td>Data description, stewardship, documentation, data sharing, contact, storage, protection, archiving &amp; preservation</td>
</tr>
<tr>
<td>Bill and Melinda Gates Foundation; data access plan if funding over $500,000</td>
<td>Expected data, data access, timeframe for sharing, storage and dissemination</td>
</tr>
<tr>
<td>National Institute of Justice (data archiving plan)</td>
<td>Data management and archiving process, confidentiality protections, tasks associated with data preparation and archiving, costing</td>
</tr>
<tr>
<td>Institute of Education Sciences (data sharing plan)</td>
<td>Expected data, data management, confidentiality of private information, roles and responsibilities, schedule for data sharing, format, documentation, how to share, limitations to sharing</td>
</tr>
</tbody>
</table>

Source: California Digital Library 2013
Journal / publisher data policies

- data underpinning publication accessible
  - upon request from author
  - as supplement with publication
  - in public repository
  - in mandated repository (e.g. PANGAEA – Elsevier)
- citation via unique DOIs
- e.g. BioMed Central open data statement
- global registries of data repositories:
  - databib.org/
  - re3data

![Growth in number of data repositories over time](image.png)

*Source: Databib 2012*
ESRC research data policy

Publicly-funded research data are a public good, produced in the public interest, which shall be made openly available and accessible with as few restrictions as possible in a timely and responsible manner that meets a high ethical standard and does not violate privacy or harm intellectual property (ESRC Research Data Policy, 2015)

• ESRC grant include a **data management plan** with their application, as an attachment to the Je-S form

• ESRC award holders deposit their research data in the **ReShare** repository (managed by UK Data Service) within three months of the end of their grant, to preserve them and to make them available for new research.

Researchers who collect the data initially should be aware that ESRC expects that others will also use it, so consent should be obtained on this basis and the original researcher must take into account the long-term use and preservation of data. (ESRC Framework for Research Ethics, 2012)
ESRC data management plan

Assessment of existing data
Information on new data
Quality assurance of data
Backup and security of data
Difficulties in data sharing and measures to overcome these
Consent, anonymisation, re-use strategies
Copyright / Intellectual Property Ownership
Responsibilities
Management and curation

ESRC DMP guidance
NERC outline data management plan

- Project information
- Organisation
- Roles and responsibilities
- Data generation activities
- Data management approach
- Metadata and documentation
- Data quality
- Exceptions or additional services

NERC DMP guidance
Horizon 2020 data management plan

- Data set reference and name
- Data set description
- Standards and metadata
- Backup and security of data
- Data sharing
- Archiving and preservations (incl storage and backup)

Horizon 2020 DM guidelines
Tools and templates

• Funder template for DMP
  • ESRC DMP requirements in data policy and DMP guidance
  • MRC DMP guidance and template
  • AHRC technical plan requirements
  • NERC DMP guidance and template

• DCC’s DMPonline tool

“We back up our data on sticky notes because sticky notes never crash.”
DM checklist

- Are you using standardised and consistent procedures to collect, process, check, validate and verify data?
- Are your structured data self-explanatory in terms of variable names, codes and abbreviations used?
- Which descriptions and contextual documentation can explain what your data mean, how they were collected and the methods used to create them?
- How will you label and organise data, records and files?
- Will you apply consistency in how data are catalogued, transcribed and organised, e.g. standard templates or input forms?
- Which data formats will you use? Do formats and software enable sharing and long-term validity of data, such as non-proprietary software and software based on open standards?
- When converting data across formats, do you check that no data or internal metadata have been lost or changed?
- Are your digital and non-digital data, and any copies, held in a safe and secure location?
- Do you need to securely store personal or sensitive data?
- If data are collected with mobile devices, how will you transfer and store the data?
- If data are held in various places, how will you keep track of versions?
- Are your files backed up sufficiently and regularly and are back-ups stored safely?
- Do you know what the master version of your data files is?
- Do your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data?
- Are you gaining (written) consent from respondents to share data beyond your research?
- Do you need to anonymise data, e.g. to remove identifying information or personal data, during research or in preparation for sharing?
- Have you established who owns the copyright of your data? Might there be joint copyright?
- Who has access to which data during and after research? Are various access regulations needed?
- Who is responsible for which part of data management?
- Do you need extra resources to manage data, such as people, time or hardware?

https://www.ukdataservice.ac.uk/manage-data/plan/checklist
DM plan
PUCCUK project

Design research:
- design survey
- design interview questions & schedule

Prepare fieldwork

Collect data:
- survey online
- do interviews
- gather media info into files

Analyze data:
- analyze survey
- code interviews & media info to analyze

Processing data:
- survey to SPSS & validate
- transcribe interviews
- transcriptions 
- check accuracy

Check transcriptions for suitability

Transcribed interviews to interviewees for approval

Storage area set up for all files
- check university backup policy
- check server
- set up SPSS database
- ind. labels, codes questions text
- planned methods written out

Design consent form
- include data sharing
- contact interviewees
- contact transcriber & agree terms

Test audio recorder
- attention file format
- WAV

Regular transfer audio files to server
- regular copy of survey responses table to server (daily/weekly)
- whilst survey is online
- attention to file naming

During interviews:
- discuss data sharing
- agree whether anonymising needed
- take relevant notes as context
- store consent forms safely
Key planning issues

- Know your legal, ethical and other obligations towards research participants, colleagues, research funders and institutions
- Know your institution’s policies and services: storage and backup strategy, research integrity framework, IPR policy, institutional data repository
- Assign roles and responsibilities to relevant parties
- Incorporate data management into research cycle
- Implement and review management of data during project meetings and review
Roles & responsibilities

- Project director: design, oversee research
- Research staff: design research, collect, process and analyze data, where keep data, who has access
- Laboratory or technical staff: generate metadata and documentation
- Database designer
- External contractors: data collection, data entry, transcribe, process, analysis; agree standard protocols
- Support staff: manage and administer research and funding, ethical review and assess IPR
- Institutional IT services: data storage, security and backup services
- External data centres: facilitate data sharing.
Cost research data management

- Cost RDM into research applications / research budgets / DMPs
- List and identify resources needed to make research data shareable beyond primary research team - above planned standard research procedures and practices
- Resources = people, skills, equipment, infrastructure, tools to manage, document, organise, store and provide access to data
- Early planning can reduce costs
- No ‘easy rules’
  - extra costs depend on standard research management practices
  - extra costs depend on long-term storage / preservation / publishing plans - repository may carry those costs
    e.g. UK Data Archive, funded by ESRC, this covers all data processing / curation / preservation / dissemination costs
- Budget for duration of research project
- Overhead costs – institutional infrastructure
How cost RDM?

**STEP 1**
- check data management activities in table and tick what applies to your proposed research; we propose 18 essential RDM activities

**STEP 2**
- for each selected activity, estimate / calculate additional time and/or resources needed and cost this

**STEP 3**
- add data management costs to your research application; coordinate resourcing and costing with your institution, research office and institutional IT services
Our data management costing tool

- Developed in discussion with researchers, research funders, research managers and administrators
- [www.data-archive.ac.uk/media/247429/costingtool.pdf](http://www.data-archive.ac.uk/media/247429/costingtool.pdf)
<table>
<thead>
<tr>
<th>ACTIVITY</th>
<th>COMMENTS AND SUGGESTIONS</th>
<th>COST</th>
</tr>
</thead>
</table>
| Data description | - Are data in a spreadsheet or database clearly marked with variable and value labels, code descriptions, missing value descriptions, etc.?  
- Are labels consistent?  
- Do textual data like interview transcripts need description of context, e.g. included as a heading page? | - if data description is carried out as part of data creation, data input or data transcription – low or no additional cost  
- if needed to be added afterwards – higher cost  
- codebooks for datasets can often be easily exported from software packages | |
| Data cleaning | - Do quantitative data need to be cleaned, checked or verified before sharing, e.g. check validity of codes used, check for anomalous values?  
- Will data match documentation, e.g. same number of variables, cases, records, files?  
- Does textual information in data need to be spell-checked? | - if carried out as part of data entry and preparation before data analysis – low or no additional cost  
- if needed afterwards – higher cost | |
| Documentation | - Do you have documentation for the data that describes the context and methodology of how data were gathered, created, processed and quality controlled? | - often essential contextual and methods documentation will be written up in publications and reports  
- if all data creation steps are well documented and documentation is kept well organised during research – low or no additional cost  
- if documentation to be written or compiled specifically afterwards – higher cost | |
| Metadata | - Do structured metadata need to be created when data are shared via a data centre or archive, e.g. completing a deposit form for the UK Data Archive? | - completing a UK Data Archive deposit form may take one to two hours  
- other data centres will have their own metadata forms | |
| Formatting and organising | - Are your data files, spreadsheets, interview transcripts, records etc. all in a uniform format or style?  
- Are files, records and items in the collection clearly named with unique file names and well organised? | - if planned beforehand by developing templates and data entry forms for individual data files (transcripts, spreadsheets, databases) and by constructing clear file structures – low or no additional cost  
- if needed afterwards – higher cost  
- free software exists for batch file renaming to harmonise file names | |
| Transcription | - Will you transcribe qualitative data (e.g. | - if part of research practice – very low or no additional cost |
DM topics

- File formats
- Data documentation
- Quality control
- Storage, backup and security
- Ethical and legal
  - Consent
  - Anonymisation
  - Access control
- Copyright and IP
- Responsibilities
Our managing and sharing data resources

- **UKDS Prepare and manage data web guidance**
- **Sharing social data in multidisciplinary, multi-stakeholder research Best practice guide for researchers**
- **Training programme**
Example DMPs

Exercise: ESRC DMP questions

- existing data sources that will be used by the research project, with references
- analysis of the gaps identified between the currently available and required data for the research
- information on the data that will be produced or accessed by the research project:
  - data volume, data type
  - data quality, formats, standards documentation and metadata
  - methodologies for data collection and/or processing
  - source and trustworthiness of third party data
- planned quality assurance and back-up procedures [security/storage]
- plans for management and curation of primary or third party data
- expected difficulties in data sharing, along with measures to overcome these difficulties, explicitly stating which data may be difficult to share and why
- explicit mention of consent, confidentiality, anonymisation and other ethical considerations and, in particular, strategies taken to not preclude further reuse of data
- copyright and intellectual property ownership of the data
- responsibilities for data management and curation within research teams at all participating institutions