An introduction to Open Research Data
Data management: a shift within the scientific process
My office desktop is a little cluttered.

Image: Christina Bonnington/Wired
“Can you send me your dataset?”

“Ok wait, I hope I can still find all the data”
“Ok, but I don’t know where all the datasets are”
“Ok, but I need to clean the data first”
“Ok, but I first need to transform it to a readable format”
“Woops, I don’t have it anymore”

- What will be the name of your dataset?
- What?
- Do you have metadata on the dataset?
- Metadata?
- What standard did you used?
- A data standard?
1. Navigating & using data is a challenge
2. Digital data are fragile
3. Research data are undervalued
MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.
4. Data may be valued... by publishers
verified and reproduced.

2.4 Safeguards

- Researchers comply with codes and regulations relevant to their discipline.

- Researchers handle research subjects, be they human, animal, cultural, biological, environmental or physical, with respect and care, and in accordance with legal and ethical provisions.

- Researchers have due regard for the health, safety and welfare of the community, of collaborators and others connected with their research.

- Research protocols take account of, and are sensitive to, relevant differences in age, gender, culture, religion, ethnic origin and social class.

- Researchers recognise and manage potential harms and risks relating to their research.

2.5 Data Practices and Management

- Researchers, research institutions and organisations ensure appropriate stewardship and curation of all data and research materials, including unpublished ones, with secure preservation for a reasonable period.

- Researchers, research institutions and organisations ensure access to data is as open as possible, as closed as necessary, and where appropriate in line with the FAIR Principles (Findable, Accessible, Interoperable and Re-usable) for data management.

- Researchers, research institutions and organisations provide transparency about how to access or make use of their data and research materials.

- Researchers, research institutions and organisations acknowledge data as legitimate and citable products of research.

- Researchers, research institutions and organisations ensure that any contracts or agreements relating to research outputs include equitable and fair provision for the management of their use, ownership, and/or their protection under intellectual property rights.
EC: The Open Research Data Pilot (H2020)

Flexible ORD pilot:
From limited to default in 2017

- Foster Open Science
- Avoid duplication of research and loss of resources

Data Management Planning

Open Access to research data (or partially opt-out)
What is "research data" by the way?

• It can be a lot of things and their nature also depends on the field.
• For instance:
  • numerical, textual, audiovisual, multimedia
  • spreadsheets/tabular data, field notes, databases, images, audio recordings, marked up texts, surveys, instrument, readings
  • experimental, observational, simulation, derived/compiled data
  • digital or non-digital data
  • primary or secondary data
  • raw, processed or analyzed data.

H2020: deposit everything needed to validate results presented in scientific publication (data, metadata, documentation)
Main objectives of Open Research Data

• Allow access and re-use of research data collected on public fundings
• Allow scientific verifiability of the data and improve their quality
• Allow reproducibility of research results
• Value the data outside of the traditional scientific process
• Prevent fraud
• Prevent redundancy and waste
• Allow new opportunities for (inter)national collaborations
• Improve data conservation
A SHIFT FROM THIS...
CREATING DATA
RE-USING DATA
GIVING ACCESS TO DATA
ANALYSING DATA
PRESERVING DATA
PROCESSING DATA

...TO THIS
Data management plan
What is a data management plan?

- A living document outlining how data will be handled during and after a project
- Increasingly required by research funders/institutions
- A good practice in itself!
What topics does a DMP cover?

• **Description of data** to be collected/created (content, type, format, volume ...)

• Methodologies, standards for collecting/creating data & **data documentation (metadata)**

• **Ethics & intellectual property** issues (informed consent, anonymization, any restrictions on sharing – e.g. confidentiality, copyright, embargoes – usage licenses...)

• Plans for **data sharing** & access (how, when, with whom...) or **restriction**

• Strategy for long-term **preservation** (what to keep, for how long, where...)
Example of data documentation

“Descriptive metadata of data items will be captured in XML files in accordance with the Darwin Core schema, which is an international metadata standard for biodiversity data. In addition, datasets will be accompanied by a separate readme.txt file providing study-level documentation including the field methods used for data collection.”
Can you send me your dataset?

“Ok wait, I hope I can still find all the data”
“Ok, but I don’t know where all the datasets are”
“Ok, but I need to clean the data first”
“Ok, but I first need to transform it to a readable format”
“Woops, I don’t have it anymore”

- What will be the name your dataset?
- What?
- Do you have metadata on the dataset?
- Metadata?
- What standard did you used?
- A data standard?

“Here it is!”
Examples of Data Management Plan

- The EC
- DMPTool website
- Digital Curation Centre website
- DMPOnline website
Content of a Data Management Plan according to the EC

Data summary  FAIR Data principles  Resources  Data security  Ethical aspects
A look at FAIR data principles

- Findable
  - How to discover your data?
  - How to understand your data?

- Accessible
  - Where to find your data?
  - Can people access your data?

- Interoperable
  - Can other people use your data?

- Re-usable
  - What is permitted?
  - Can the data be open?
A look at FAIR data principles

- Licensing
- Standards
- Vocabulary
- Methodologies
- Software
- Documentation
- Data repository
- Metadata
- Persistent identifier
- Naming convention
- Keywords
- Versioning
- Findable
- Accessible
- Interoperable
- Re-usable
- Naming convention
- Keywords
- Versioning
- Licensing

Metadata standards
A tool that may help you: DMPOnline
### My plans

The table below lists the plans that you have created, and any that have been shared with you by others. These can be edited, shared, exported or deleted at anytime.

<table>
<thead>
<tr>
<th>Name</th>
<th>Owner</th>
<th>Shared?</th>
<th>Last edited</th>
<th>Select an action</th>
</tr>
</thead>
<tbody>
<tr>
<td>My project (NERC Template)</td>
<td>Valerie McCutcheon</td>
<td>Yes (with 1 people)</td>
<td>17-12-2013</td>
<td>Edit Export</td>
</tr>
<tr>
<td>Withdrawal of services for young people</td>
<td>Selina Shaw</td>
<td>Yes (with 2 people)</td>
<td>20-11-2013</td>
<td>Edit Share Export Delete</td>
</tr>
</tbody>
</table>

[Create plan](#)
Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

If applying for funding, select your research funder.
Otherwise leave blank.

Funder:
- Cancer Research UK
- Economic and Social Research Council
- European Commission (Horizon 2020)
- Medical Research Council
- National Science Foundation (USA)
- Natural Environment Research Council
- Science and Technology Facilities Council
- Wellcome Trust
Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

- If applying for funding, select your research funder.
  - Medical Research Council

- To see institutional questions and/or guidance, select your organisation.
  - University of Glasgow

- Tick to select any other sources of guidance you wish to see.
  - Generic guidance from the Digital Curation Centre

Create plan
All applicants submitting funding proposals to the MRC are required to include a Data Management Plan as an integral part of the application. MRC Institutes and Units are required to submit one as part of the Quinquennial Review (QQR) report.

For more information see the [MRC guidance on data management plans](https://mdcguide.dcc.ac.uk/projects/my-project-mrc-template-~88?show_form=yes)

A [DMP template](https://mdcguide.dcc.ac.uk/projects/my-project-mrc-template-~88?show_form=yes) is also provided.

<table>
<thead>
<tr>
<th>Sections</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Proposal name</td>
<td>- 0. Enter the proposal name</td>
</tr>
<tr>
<td>1. Description of Data</td>
<td>- 1.1 Type of Study&lt;br&gt;- 1.2 Types of Data&lt;br&gt;- 1.3 Format and scale of the data</td>
</tr>
<tr>
<td>2. Data collection / generation</td>
<td>- 2.1 Methodologies for data collection / generation&lt;br&gt;- 2.2 Data quality and standards</td>
</tr>
<tr>
<td>3. Data management, documentation and curation</td>
<td>- 3.1 Managing, storing and curating data&lt;br&gt;- 3.2 Metadata standards and data documentation&lt;br&gt;- 3.3 Data preservation strategy and standards</td>
</tr>
<tr>
<td>4. Data security and confidentiality of potentially disclosive personal information</td>
<td>- 4.1 Formal information/data security standards&lt;br&gt;- 4.2 Main risks to data security</td>
</tr>
<tr>
<td>5. Data sharing and access</td>
<td>- 5.1 Suitability for sharing&lt;br&gt;- 5.2 Discovery by potential users of the research data&lt;br&gt;- 5.3 Governance of access&lt;br&gt;- 5.4 The study team's exclusive use of the data&lt;br&gt;- 5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions&lt;br&gt;- 5.6 Regulation of responsibilities of users</td>
</tr>
<tr>
<td>6. Responsibilities</td>
<td>- 6. Responsibilities</td>
</tr>
<tr>
<td>7. Relevant policies</td>
<td>- 7. Relevant institutional, departmental or study policies on data sharing and data security</td>
</tr>
<tr>
<td>8. Author and contact details</td>
<td>- 8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their telephone &amp; email contact details</td>
</tr>
</tbody>
</table>
### MRC Data Management Plan

<table>
<thead>
<tr>
<th>Plan Details</th>
<th>MRC Data Management Plan</th>
<th>Share</th>
<th>Export</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Proposal name</td>
<td>1 question, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Description of Data</td>
<td>3 questions, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Data collection / generation</td>
<td>2 questions, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Data management, documentation and curation</td>
<td>3 questions, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Data security and confidentiality of potentially disclosive personal information</td>
<td>2 questions, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Data sharing and access</td>
<td>6 questions, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Responsibilities</td>
<td>1 question, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Relevant policies</td>
<td>1 question, 0 answered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Author and contact details</td>
<td>1 question, 0 answered</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The clinical data collected from this project will be documented using CDISC metadata standards.

### MRC Guidance

How the data will be collected/generated and which community data standards (if any) will be used at this stage.

### Glasgow guidance on Data Capture Methods

Advice on filenaming, documentation and organising your data is available on the University's [data management webpages](#).

### DCC guidance on Data Capture Methods

[Link to DCC guidance](#)
You can give other people access to your plan here. There are three permission levels.

Users with "read only" access can only read the plan.
Editors can contribute to the plan.
Co-owners can also contribute to the plan, but additionally can edit the plan details and control access to the plan.

Add each collaborator in turn by entering their email address below, choosing a permission level and clicking "Add collaborator".

Those you invite will receive an email notification that they have access to this plan, inviting them to register with DMPonline if they don't already have an account. A notification is also issued when a user's permission level is changed.

Collaborators

<table>
<thead>
<tr>
<th>Email address</th>
<th>Permissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarah Jones</td>
<td>Owner</td>
</tr>
</tbody>
</table>

Add collaborator

Email

Permissions: Read only

Add collaborator
## Withdrawal of services for young people

**ESRC Data Management and Sharing Plan**

### Existing data

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Explain the existing data sources that will be used by the research project (with references)</td>
<td>The ESDS archive has been systematically searched using a series of search terms related to and derivative of ‘public service withdrawal’, ‘impacts’, and/or ‘children’ and ‘young people’. Our overall assessment is that there are no datasets that will adequately address the aims of this project. The following datasets are tangentially related to this project.</td>
</tr>
<tr>
<td>- Present an analysis of the gaps identified between the currently available and required data for the research</td>
<td>Given the contemporary nature of the proposed project, we know of no datasets that cover users’ (especially young people’s) views and experiences of austerity measures and service withdrawal. The proposed project will therefore capture new and unprecedented data, for which there is an evident demand among national and regional stakeholders, decisionmakers and service-providers (see ‘Pathways to Impact’ attachment). Moreover, whilst several datasets incorporate longitudinal data, none includes data gleaned from oral history and multigenerational family interview methods central to the proposed project. The proposed project therefore represents an extension to the methods and data quality of the tangentially-related projects listed in section 1. Finally, the novel use of ‘impact’ activities to generate data (e.g. workshops, mapping software) exceeds the scope of all extant and even tangentially-related datasets.</td>
</tr>
</tbody>
</table>

### Information on new data

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>- State the data volume, type (e.g. qualitative or quantitative)</td>
<td>The project shall generate new quantitative data (Bristol Online Survey outputs, SPSS data and outputs), qualitative data (digital audio files, audio transcripts, digital photographic and video data, workshop outputs, NVivo files), and mapping data (TIFF files). Metadata, in the form of pdfs and</td>
</tr>
</tbody>
</table>
ESRC Data Management and Sharing Plan

Existing data

Explain the existing data sources that will be used by the research project (with references)

The EDSO archive has been systematically searched using a series of search terms related to and derivative of ‘public service withdrawal’, ‘impacts’, and/or ‘children’ and ‘young people’. Our overall assessment is that there are no datasets that will adequately address the aims of this project. The following datasets are tangentially related to this project.

Present an analysis of the gaps identified between the currently available and required data for the research

Given the contemporary nature of the proposed project, we know of no datasets that cover users’ (especially young people’s) views and experiences of austerity measures and service withdrawal. The proposed project will therefore capture new and unprecedented data, for which there is an evident demand among national and regional stakeholders, decisionmakers and service-providers (see ‘Pathways to ‘impact’ attachment’).

Moreover, whilst several datasets incorporate longitudinal data, none includes data gleaned from oral history and multigenerational family interview methods central to the proposed project. The proposed project therefore represents an extension to the methods and data quality of the tangentially-related projects listed in section 1. Finally, the novel use of ‘impact’ activities to generate data (e.g. workshops, mapping software) exceeds the scope of all extant and even tangentially-related datasets.

Information on new data

State the data volume, type (e.g. qualitative or quantitative data), quality and formats

The project shall generate new quantitative data (Bristol Online Survey outputs, SPSS data and outputs), qualitative data (digital audio files, audio transcripts, digital photographic and video data, workshop outputs, NVivo files), and mapping data (TIF files). Metadata, in the form of pdfs and Excel spreadsheets, shall be used to facilitate the management and archiving of these data. Data shall be stored in password-protected folders on the host institution’s secure servers. Data transfer between the PI and Co-I shall take place via face-to-face meetings.

Outline the standards and methodologies for data collection (including documentation)

There will be four methods of data capture.

1. Online survey - the survey will be constructed using BOS, and data shall be initially captured via this interface. These data shall be exported (as CSF files) into SPSS. Data shall be stored and analysed as SPSS spreadsheets.

2. Interview and oral history methods. Qualitative data shall be captured using high spec digital voice recorders with external microphones. Data shall be transcribed and saved as MS Word files, stored on NVivo files, and ultimately archived as pdfs.
Useful resources on Data Management Planning

UK funder requirements for Data Management Plans [webpage]
A summary of requirements from UK Research Councils and other charitable or health funders.

US funder requirements for Data Management Plans [webpage]
A summary of requirements from the NSF, NIH and other key funders in the USA.

DCC Checklist for a Data Management Plan [PDF, 3 pages]
A list of 13 questions and associated guidance, that represent the main issues to come up in Data Management and Sharing Plans. The Checklist is used as a generic template in DMPonline, and is presented when no funder or organisational requirements are applicable for the user.

How to develop a Data Management and Sharing Plan [PDF, 8 pages]
A guide by the Digital Curation Centre that outlines typical funder requirements for DMPs and the types of considerations to make when responding.

Example Data Management Plans

Technical plan submitted to the AHRC [PDF, 7 pages]
A DMP submitted by a researcher from the University of Bristol, also including comments from the reviewers

Two social science DMPs [PDF, 7 pages]
Example plans from researchers at the University of Leeds, shared as part of the Leeds RoaDMaP training materials

Health sciences DMP [PDF, 11 pages]
Example DMP produced by the DATUM for Health RDM training project
Limitation of the tool (KU Leuven)

• Copying DMP is not possible: no quick re-use of a previously filled out DMP
• Linking different DMP is not possible: no referring to related DMPs
• Not easy export to integrate with other systems: “send information in this section to field X in application Y”
• Not easy to retrieve info from other systems: “retrieve information for this section from field X in application”
• No conditional logic: if the answer to “are you working with personal data?” is negative, then your are directed to the next relevant section, skipping others.
Upcoming in-depth analysis

- How to develop an institutional strategy that fully address the wide range of needs of the scientific community and the requirements of the funds providers?
- Should we use a tool to store the data? Should we develop our own? Should we adapt our own repository?
- How to fully address the question of data sharing (security, ethic, IPR)?
- How to preserve data efficiently?
- How to increase awareness within the scientific community? How to create a common culture of data sharing within disciplines that are not used to it?
- What are the tools and resources available to help at each part of the data management process? What have other institutions chosen to do and why?
Thank you!

François Paquot
francois.paquot@ulg.ac.be
Ressources

Slides adapted from:


All of these presentations are available [here](#)