DCC because good research needs good data

Data Management Planning

What is a data management plan?

A brief plan written at the start of your project, updated as it progresses, to define:

- » how will your data be created?
- » how will it be documented?
- » who will access it?
- » where will it be stored?
- » who will back it up?
- » whether (and how) will it be shared & preserved?

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

Why develop a DMP?

▷ They can help you to:

- » Make informed decisions that anticipate & avoid problems
- » Avoid duplication, data loss and security breaches
- » Provide guidelines for everyone working on the project
- » Make it more likely that your data has future value
- » To comply with funder requirements…

Which UK funders require a DMP?

Some other funders that require DMPs

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

DMP template for postgraduates

▷ Defining your data

▷ Looking after your data

▷ Sharing your data

▷ Archiving your data

▷ Executing your plan

http://opus.bath.ac.uk/30772

DataTrain

Training and DMP template

Used by archaeology and social anthropology students

Focuses on submission of data with thesis

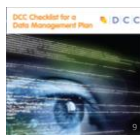## DCC Checklist for a DMP

▷ 13 questions on what's asked across the board

▷ Prompts / pointers to help researchers get started

▷ Guidance on how to answer

www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

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...

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

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 and... nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released th...

Look at examples to help you understand what to write

www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html

Help from the DCC

How to Develop a Data Management and Sharing Plan

Sarah Jones (DCC)
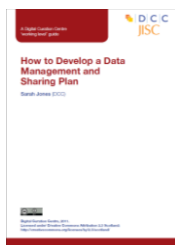
A web-based tool to help researchers write data management plans

https://dmponline.dcc.ac.uk

www.dcc.ac.uk/resources/how-guides/develop-data-plan

Tips on writing DMPs

▷ Keep it simple, short and specific

▷ Seek advice - consult and collaborate

▷ Base plans on available skills and support

▷ Make sure implementation is feasible

▷ Justify any resources or restrictions needed

▷ Don't overthink it

www.youtube.com/watch?v=7OJtiA53-Fk

Example plans

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

Rural Economy & Land Use (RELU) programme examples [webpage]
Two examples from a joint ESRC, BBSRC and NERC programme

Health sciences DMP [PDF, 11 pages]
Example DMP produced by the DATUM for Health RDM training project

Psychology DMP [PDF, 11 pages]
A very detailed, fictional psychology DMP produced by the DMTpsych RDM training project, based on a seminal psychology experiment

UCSD Example Data Management Plans [webpage]
Over 20 example plans submitted to the National Science Foundation (NSF) in the United States by academics at UC San Diego

Colorado School of Mines examples [webpage]
A variety of US example DMPs from Mines and elsewhere

NSF data management plans [webpage]
5 DMPs submitted to the NSF, shared by the DataOne initiative

Biology and chemistry DMPs [webpage]
Three example DMPs from the USA shared by NECDMC, an instructional tool for teaching RDM to undergraduates, graduate students, and researchers in the health sciences, sciences and engineering.

- https://dmponline.dcc.ac.uk/help#PlanningHelp

Data standards and documentation

Why is this important?
Creating data in formats preferred for archiving helps to ensure that they will be usable in the future. Good descriptive metadata are essential for effective data use.

e.g. "Quantitative survey data files generated will be processed as SPSS system files with DDI XML documentation. 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."

n.b. EXPLANATIONS AND EXAMPLES COURTESY OF NIH, ICPSR and University of Bristol

Documentation and standards

Metadata: basic info e.g. title, author, dates, access rights...
Documentation: methods, code, data dictionary, context...

Use standards wherever possible for interoperability

www.dcc.ac.uk/resources/metadata-standards

Data storage and backup

Why is this important?
Data should be stored securely to avoid unauthorised access and automatically backed up to prevent data loss.

e.g. "Bristol's Research Data Storage Facility (RDSF) will be used to store the data during the project. Backup procedures are robust (overnight backup, copies held remotely on tape) and secured access is in place.
Recordings made in the field will be copied to the RDSF via a secure web connection, by the PI, as soon as possible, but some delays are expected due to Sri Lankan facilities. Therefore an external, portable hard drive (and an identical copy, for redundancy purpose) will be used for backup in the field."

Plans for data sharing

Why is this important?
Sharing data helps to advance science and to maximize the research investment. Research funders and universities probably expects you to share data wherever possible.

e.g. "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. "

Seek consent for data sharing & preservation

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.

If you don't ask, data centres won't be able to accept your data – regardless of any conditions on the original grant or your desire for the data to be shared.

License your data for reuse

Outlines pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

	NC	Non-Commercial
		What counts as commercial?
	SA	Share Alike
		Reduces interoperability
	ND	No Derivatives
		Severely restricts use

www.dcc.ac.uk/resources/
how-guides/license-research-data

Long-term preservation plans

Why is this important?
Digital data need to be actively managed over time to ensure that they will always be available and usable. Depositing data resources with a trusted digital archive can ensure that they are curated and handled according to good practices in digital preservation.

> e.g. "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."

Who should pay for RDM?

Funding Research Data Management
"A conversation with the funders"

The DCC held a special event on this topic in the UK, but there's still a long way to go

www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf-special-event-funding-research-data-management

What to charge and how?

Direct costs

» In-project costs that must be incurred before the end of the grant
» Potentially hardware, staff, expenses, costs of preparing data for deposit...
» Could include charges levied by repositories (pay once store forever)
» May also include costs ordinarily recovered indirectly (e.g. storage) if the requirement is exceptional and exceeds the norm

Indirect costs

» The general cost of providing RDM services and infrastructure
» Designated data services should be used if provided
» Outsourcing to a third-party is also an option

Remember to make a clear justification for any costs

Key messages

▷ All costs are eligible, but:
  » Direct costs must be incurred before a grant ends
  » Nothing can be double funded (recovered indirectly and as a direct)

▷ Researchers are expected to use designated data repositories.

▷ There is no rule of thumb to measure the proportion of a grant that may acceptably be spent on research data management.

▷ A clear justification of resources is needed for each specific case.

Be specific for each grant

▷ A flat rate charging structure for RDM services (e.g. 10% of each grant) is not appropriate.

▷ The value of a research grant is not a good proxy for the volume or complexity of data it may generate.

▷ Base costs on each specific case and make a clear justification for them. They should be auditable.

Making data FAIR

▷ **F**indable
– *Assign persistent IDs, provide rich metadata, register in a searchable resource,...*

▷ **A**ccessible
– *Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...*

▷ **I**nteroperable
– *Use formal, broadly applicable languages, use standard vocabularies, qualified references...*

▷ **R**eusable
– *Rich, accurate metadata, clear licences, provenance, use of community standards...*

www.force11.org/group/fairgroup/fairprinciples

More elaborate DMP

Scientific research data should be easily:

1. Discoverable
Are the data discoverable and identifiable by a standard mechanism e.g. DOIs?

2. Accessible
Are the data accessible and under what conditions e.g. licenses, embargoes?

3. Assessable and intelligible
Are the data and software assessable and intelligible to third parties for peer-review? E.g. can judgements be made about their reliability and the competence of those who created them?

4. Useable beyond the original purpose for which it was collected
Are the data properly curated and stored together with the minimum software and documentation to be useful by third parties in the long-term?

5. Interoperable to specific quality standards
Are the data and software interoperable, allowing data exchange? E.g. were common formats and standards for metadata used?

DMPonline

A web-based tool to help researchers write DMPs
Includes templates for funders, organisations, disciplines

https://dmponline.dcc.ac.uk
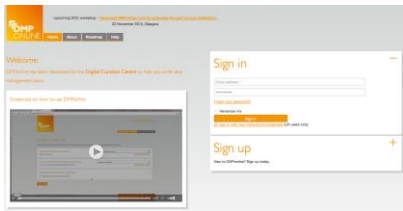
How the tool works

Click to write a generic DMP

Or choose your funder to get their specific template

Pick your uni to add local guidance and to get their template if no funder applies

Choose any additional optional guidance

Site and funder-specific guidance

Deliver the DMP and keep it up to date

▷ EC: "Since DMPs are expected to mature during the project, more developed versions of the plan can be included as additional deliverables at later stages. (…) New versions of the DMP should be created whenever important changes to the project occur due to inclusion of new data sets, changes in consortium policies or external factors."

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

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

TIME TO WORK!

Exercise: Writing a DMP

In pairs or small groups, look at the basic DMP questions and consider:

- Can you answer these for yourself?

- For those you can, sketch out some answers

- Where might you seek help for those you can't?

- Which might be subject to change as research progresses?

We're going to tackle this in a number of steps…

Writing your own data management plan

1. In small groups (2-3), work through the five questions in the 'Defining your data' & 'Looking after your data' sections of the workbook (15 minutes)

2. Each group will feedback some of the responses they've considered based on their own research activity

Defining your data

➢ Define what data will be created and how - this should be communicated to the whole research team

➢ Develop procedures for consistency and data quality

➢ Choose appropriate software and formats - some are better for long-term preservation and reuse

➢ Ensure consent forms, licences and partnership agreements allow for options to share data if desired

Step 2

Next, work through the four questions in the

'Looking after your data'

section of the workbook (15 minutes)

_____

_____

_____

_____

_____

_____

_____

Looking after your data

▷ Make use of University secure storage and automatic back-ups rather than manual if you can, if unavoidable plan a realistic update approach that reflects updates, changes etc.

▷ Keep a copy off-site

▷ Collect together all the information users would need to understand and reuse the data

▷ Name, structure and version files clearly

▷ Create metadata at the time - it's hard to do later

▷ Use disciplinary standards where possible

_____

_____

_____

_____

_____

_____

Step 3

Now, work through the three questions in the

'Sharing your data'

section of the workbook (15 minutes)

_____

_____

_____

_____

_____

_____

_____

Sharing your data

▷ Clarify your rights – does your University or funder have any ownership rights?

▷ Clarify rights relating to use of third party data

▷ Restrict access to those who need to read/edit data during active phase of research

▷ Consider the data security implications of where you store data and collaborate on working documents and from which devices you access files (e.g., DropBox)

▷ Choose appropriate methods to transfer / share data (encrypted media rather than email)

▷ How will you protect any data that can't be shared

Step 4

Lastly, work through the four questions in the

'Archiving your data'

section of the workbook (15 minutes)

Archiving your data

▪ Decide what data should be kept – some, all, none?

▪ Who will keep the data?

▪ Will the data be publicly accessible or available on request (e.g., for validation purposes)

▪ Make sure that the archive normalisation procedures meet your needs

▪ Consider whether you will need any specific software/tools to visualise or manipulate the data

▪ How long will the data need to be retained? When/how will reappraisal occur?

▪ Are there any costs that need to be considered? If so, how will these be met?

Exercise: Assessing a DMP

In pairs or small groups, read the DMP provided and discuss:

▷ Is there enough information provided in the common areas?
  » Data description
  » Standards and methodologies
  » Restrictions (e.g. ethics, IP, etc)
  » Plans for sharing
  » Preservation

▷ If not, what would you need help to complete?

▷ Where would you go to get this help?

Share your example DMPs!

EUDAT OpenAIRE

Share your DMPs

Please share your example DMPs so we can add them to the DCC list of examples for others to learn from.

There are increasing requirements for Data Management and Sharing Plans worldwide. Various universities have started to collate examples to help researchers understand how to respond and the level of detail to provide. The DCC collates examples, and together with the DMPTool we are promoting public sharing of DMPs. Some titles such as the Research Ideas and Outcomes (RIO) journal also publish DMPs.

If you are willing to share your examples, please share a link and any other details as desired. Many thanks!

Project name / DMP title:

Link to DMP:

Subject area:

Main data type(s):

Funder (if relevant):

Contact name:

Email address:

Comments:

Send us links to your DMPs

We will add them to the DCC list

Aim to cover wide range of disciplines and funders

www.dcc.ac.uk/share-DMPs

Tips for writing DMPs

EUDAT OpenAIRE

▷ Seek advice - consult and collaborate

▷ Consider good practice for your field

▷ Base plans on available skills & support

▷ Make sure implementation is feasible

▷ Think about things early…

Key messages

▷ The principles of good research conduct hold for all of us, across disciplinary boundaries.

▷ Data management is all in a day's work.

▷ Planning and reflection are more important than the plan – but write the DMP and keep it up to date.

▷ Planning data management is team work.

▷ Think about the desired end result and plan for this.

▷ Decisions made early affect what you can do later.

Lessons from History

Helmuth (Karl Bernhard Graf) von Moltke the Elder

What we learn from history

• "No battle plan survives contact with the enemy"

▷   - Helmuth von Moltke the elder

• "... I have always found that plans are useless, but planning is essental"

▷   - Dwight D. Eisenhower

• The data management plan is an evolving object, along with the research