Research Engagement Library

©University of Reading 2023

Introduction

A data management plan (DMP) should be completed for any research project that will involve the collection or creation of data.

Primary data may be collected or created by means of experiment, observation, simulation, and processing or combination of data from existing sources. Secondary data sources may be used as inputs into research, e.g. published literature, archive documents, datasets created by administrative or research data collection activities, and information published by individuals and organisations.

Instructions for completing the Data Management Plan

You can use the template provided to document how you will manage your data and supporting materials such as software code throughout your research project, and how you will preserve them and make them accessible to others in support of your completed thesis and any associated publications.

A DMP is a practical research instrument, which can help you manage your research data effectively and prepare them for long-term preservation and sharing. It is meant to be written iteratively throughout the course of your research, and should be regularly reviewed and updated.

You may not be able to complete all sections of the DMP at a first attempt: while you are in the early stages of research a lot of the practical detail, and some of your key data management decisions, may be as yet undetermined. But you can use the plan to document your data management requirements, identify questions you need answers to and people to ask, and put down markers for future development of the plan.

Guidance is provided below for completing each section of the template. It includes useful links and examples to help you provide relevant information.

You should complete each section of the template. If a section is not relevant to you, simply write N/A and move on to the next section. Do not delete sections from the template.

There is no ideal length for your DMP. It depends on the nature and extent of the data you work with, the complexity of the data management requirements, and the level of detail that is useful to you.

Updating the plan

You are encouraged to update your DMP on a regular basis and to share and discuss it with your Supervisor. Regular DMP reviews allow you to add new relevant information as it arises, to reflect on your data management activities in the light of the plan, and to adjust the plan or your practice as appropriate.

Support

Training on ÓWriting a data management plan for your research projectôis delivered termly through the Reading Researcher Development programme.

For advice on completing the DMP and to request a review of your DMP contact researchdata@reading.ac.uk / 0118 378 6161. Please note that review requests may take up to five working days to be answered.

Guidance

1 Project information

1.1 Project description

In two or three sentences describe the research question(s) you are addressing.

1.2 Organisations

List any organisations in addition to the University that are directly involved in your research, either as funders, or as research partners or collaborators, and describe their role.

Example

I am co-funded by BBSRC and Syngenta under a CASE Studentship.

1.3 Contracts

List any contracts under which your research is being conducted, and indicate where your copy of any relevant contract is held.

If you are being funded by a third party, for example under an industrial sponsorship or CASE agreement, or by an employer, you will need to be aware of how the terms and conditions of sponsorship affect your ownership of, and rights in, any intellectual property (IP) that you create in the course of your research. Data created by you constitute intellectual property, and will be subject to contractual terms and conditions.

Under certain circumstances, for example where University staff have made a significant intellectual contribution to your research or where the University has provided significant financial or material support, you may also be required to assign IP to the University by

Instruments may include hardware, software and paper-based instruments, e.g. data collection forms, lab notebooks. For hardware and software, specification or version/release information should be recorded at the time of data collection. Samples of instruments and collection forms, e.g. survey questions, should be recorded where relevant. If you will be using any experimental facilities, e.g. the ISIS neutron and muon source, or research infrastructure, such as the NERC ARCHER supercomputing service, make a note of this.

The University provides access to a number of secure online services that can be used for the collection of data from research participants. These include the research database and survey platform UoR REDCap, and the survey tools Jisc Online Surveys and Qualtrics.

If you plan to develop any scripts, libraries, plug-ins, software tools or applications as part of your research, briefly describe these. What programming language will you use? How will you handle code dependencies? What methods and tools will you use to develop and manage your code, e.g. version control software such as Git, or a code repository platform such as GitHub or GitLab. Note that the University provides a GitLab service that you can use to maintain and share your code (link below). What computing environment will the code be executed in?

Useful links

UoR REDCap: https://www.reading.ac.uk/research-services/research-data-management/managing-your-data/uor-redcap
Online survey tools: https://www.reading.ac.uk/research-services/research-data-management/managing-your-data/online-survey-tools
University GitLab Git repository server.
https://research.reading.ac.uk/act/knowledgebase/gitlab-git-repository/

Examples

Interviews will be audio-recorded using a digital audio recorder supplied by the Department. Recordings will be transcribed into text and anonymised by myself. Texts will be imported into NVivo for analysis.

NMR spectra will be acquired from biofluid samples in proprietary Topspin format using a Bruker Avance III 700 MHz spectrometer based in the Chemical Analysis Facility. They will be converted to ASCII format for plotting in Excel.

A detailed MySQL database of agricultural productivity, property transactions and population counts in the Berkshire region from the period 1320-1380 will be compiled from manorial records held at the Berkshire Record Office and the National Archives.

The ocean circulation model will be implemented using Fortran 2008. Code will be developed in the University GitLab platform and code files archived at the time of the experiment to preserve the version of the model implemented. The model will be run in the NERC JASMIN data processing environment. Data analysis and visualisation of NetCDF output will be performed using custom scripts written in Python 3.0. Any third-party dependencies will be installed from PyPI and recorded

occasional days thereafter to ensure accuracy of measurements. Original paper records will be scanned and saved. I will check spreadsheet data against paper records.

3 Storage and organisation

3.1 Storage and security

Describe your data storage and security policy.

You should choose a storage and backup solution that will keep your data safe (from loss, theft and corruption) and secure from unauthorised access ï this is especially important if you will be collecting personal or sensitive data. You will need to ensure that the chosen solution has enough storage capacity for your needs.

You should also specify who will have access to your data. As a rule, during the active phase of your research, up until the point you complete your thesis or publish your findings, data should be kept private, and made accessible to others only on a 1rETQq0.000008

you use OneDrive, do not sync any folders containing personal data to your laptop or other devices; instead, store and access the files as required via Office 365 in your browser. Where personal devices are used for the temporary processing of personal data, they should be secured at the very least by password access controls, and preferably by encryption.

If data are collected outside the University network, e.g. using off-network instruments or in a field campaign, you should establish a protocol for backup and transfer to University storage, with backups to the cloud or to separate devices between transfers.

For non-digital data, you may need to take copies and establish a process for transfer of data in digital format (by digitisation or data entry).

Useful links

University Office 365: https://www.reading.ac.uk/digital-technology-services/service-catalogue/office-365

University storage services: https://uor.topdesk.net/tas/public/ssp/content/detail/service?unid=8f3d719246814fa089b9c5d9c8e0e7ff (login required)

Guidance on academic computing resources, including Research Data Storage, Research Cloud computing platform, cloud storage and other services: https://research.reading.ac.uk/act/

Consider what information you or someone else would need to be able to reproduce the data, or to make sense of them and use them. It can be useful to think of documentation in terms of four levels: variable level, file/database level, project level, and metadata level.

Variable level documentation defines your variables, and specifies units of measurement and permitted values (including missing value codes). This information is usually embedded within data files, e.g. as a header, or in column labels. Separate worksheets in a spreadsheet file might contain a list of variables with their full definitions and information about units of measurement and permitted values (these latter could be used for data validation). Variable information may also be recorded as a separate codebook or data dictionary.

File or database-level information describes the components and logical structure of the dataset. This could be as simple as a listing of files with details of their contents, or a database schema. The information could be recorded in a separate readme file.

Project level information describes the research questions and hypotheses the data will be collected to answer or test, the design of the research and the methodologies that will be used, and information about the instruments that will be used to collect and process the data, and records of the research process. There may be standard experimental reporting protocols in your field that you can use to document your methods and instruments. Documentation might include laboratory notebooks, interview schedules, instrument or software specifications and guides, in-line commentary of software code written in the research, interview transcription and anonymisation guidelines, etc.

M 1 1571 0 5eW(s)-3(d)-d/3ut req0.000008871 8871 0 5108.62-3(t)4(g)q0.00(u)-3(i)12(d

Sequencing Experiment (MINSEQE) guidelines (https://www.ebi.ac.uk/fg/annotare/login/).

Useful links

How to make a data dictionary: https://help.osf.io/article/217-how-to-make-a-data-dictionary

Metadata Standards: https://rdamsc.bath.ac.uk/

Life sciences standards: https://fairsharing.org/search?fairsharingRegistry=Standard

ethical approval make any commitment to destroy confidential data by a given time or not to share (anonymised) data collected from research participants. In most cases data can be shared openly if they are anonymised. It is good practice to secure consent for data sharing when you recruit participants, e.g. by including in your consent form a statement such as: ó understand that the data collected from me in this study will be preserved and made available in anonymised form, so that they can be consulted and re-used by othersô

Personal data is any information relating to an identified or identifiable natural person. These data enjoy statutory protection under the General Data Protection Regulation 2016 and the Data Protection Act 2018. Under this legislation any personal data collected by you must be processed fairly and lawfully. Among other things you will be required to issue a Privacy Notice to your research participants, which explains the purpose(s) for which the data are being collected, your lawful basis for processing the data, who the data will be disclosed to, and the rights of the individuals in respect of their personal data. For certain kinds of research, for example involving the processing of sensitive data or human genetic data, you will also need to complete a Data Protection Impact Assessment under the advice of th

no further need to link individuals to data, the linking key can be destroyed, so that the data become fully anonymised.

To make data safe for sharing they will need to be anonymised. Bear in mind that effective anonymisation may involve much more than replacing personal names with pseudonyms, and different techniques are required for quantitative and qualitative data. The UK Data Service provides useful guidance on anonymisation (see below).

You should indicate when personal data will be destroyed. In many cases this is likely to be at the end of the project, if not earlier. But if continued retention of data beyond the end of the project is anticipated, you should state your reason for this, and describe your retention policy. You can retain personal data on a continued basis for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. You do not need to commit to destroy personal data at a set time, but they

means of deposit in a suitable data repository. You are unlikely to need to preserve all the data you collect or create in the course of your research. You will therefore need to select data of value, and dispose of data of little or no value. The following considerations should be borne in mind.

What data will be required to validate your research findings? Test data, results of failed experiments, and data from faulty instruments are obvious candidates for disposal. Data at intermediate stages of processing may also be surplus to requirements, as it is more important to preserve the raw and final data and the record of processing by which they were transformed from one state to the other.

In the case of computer simulations of complex systems, raw output can often run to TB, and individual outputs may be less important than preservation of the model code and input parameters, by which a set of results can be reproduced. Storage, preservation and transfer of data at the TB scale present both technical and financial challenges, to the extent that the cost of meaningful preservation and sharing of such data outputs may be far in excess of any possible benefit.

What is the intrinsic value of the data? Environmental data, for example, are unique to their time and place and have inherent value as part of the historical record. If these are lost they can never be replaced. Experiments can in principle be repeated, and the data reproduced, although the cost of doing so may be high.

Are there any legal/ethical/contractual restrictions on what data can be shared? As a general rule, you would be expected to preserve anonymised data only. For example, you may preserve anonymised transcripts, but dispose of original interview audio recordings. There may also exist reasons to redact data, for example to remove commercially-sensitive information or other information provided in confidence, to obscure the locations of endangered species, or to protect national security.

You should consider the format the data will be preserved in, and any preparation that will be necessary. Suitable preservation formats may be:

> open formats, such as CSV for tabular data, ASCII text (.txt) and PDF/A for text and documentation, XML with an appropriate Document Type Definition (DTD) for structured machine-readable information, JPEG for images, FLAC for audio, and MPEG-4 for video. Included in this category are self-

analysed in a proprietary software, such as MATLAB or SPSS, should be preserved in a format accessible to users without a software licence.

Data should be shared under an open licence that grants broad permission for re-use, unless there are legal, ethical or commercial reasons why data cannot be shared openly. Various open licences exist, but the recommended default for open data is the widely-adopted Creative Commons Attribution 4.0 International License. There are more restrictive standard licence options, such as the Creative Commons Attribution NonCommercial 4.0 International Licence, but these should only be used where the restriction can be justified. When data are shared they should be accompanied by a rights and licence statement, so that terms of use and attribution requirements are clear to any users. This statement should be included in the dataset documentation file. Most repositories will also include rights and licence statements in the online metadata record for a dataset.

Useful links

University guidance on licensing data:

research is undertaken with commercial sponsorship or participation, you may be subject to contractual confidentiality terms or a requirement to provide prior notice of publication. For example, an industrial sponsorship contract will tytne, aq33(u)-3(ir)15(e)-3(m-6(e)6(f)63( )8(

recording/collection equipment, specialist software for which a licence is required, or time at a facility, such as the Diamond Light Source.

If you will require large amounts of storage and computing resource for computational research, either at the University, or at a national facility such as JASMIN, this should be noted here.

If there are likely to be any costs to meet these additional requirements, specify these and state how they will be met.

8.3 Training and information requirements

What training or further information will you need?

If you have been unable to complete any sections of your plan, make a note of them here so that you can find out the information you need and update the plan in due course. Identify any person/organisation you will need to contact and note the questions you need to ask. If you will need training on data management, note your training requirements and the details of any courses you plan or will need to attend.

Examples

> To help me implement my model efficiently, I need training in software development basics and use of version control systems. I will find out about University training in this area.
>
> , ¶ P X Q V X U H Z K L F K L V W K H E H V W V W R U D J HDTVR O X W L R support to discuss this.
>
> I ¶ P X Q V X U H Z K L F K L V W K H E R V W suitable data repository to preserve my data. I will contact the Research Data Manager about this.