University of Reading

Data Management Plan Guidance for Postgraduate Researchers

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 ‘Writing a Data Management Plan for your research project’ is delivered termly through the Reading Researcher Development programme. For more information see http://www.reading.ac.uk/gs-reading-researcher-development-programme.aspx.

For advice on completing the DMP and to request a review of your DMP contact Robert Darby, Research Data Manager, at r.m.darby@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 means of an Assignment Agreement. If you sign such an agreement, you should record relevant details here.

Your Supervisor should inform you of any contractual issues concerning IP rights at the start of the project and as they arise during your research.

Always keep a copy of any contract related to your research, and check its IP and publication clauses. If you cannot find your contract, you can ask your School Postgraduate Research Administrator for help: http://www.reading.ac.uk/gs-meet-the-team.aspx.

Useful link

- University Code of Practice on Intellectual Property, clauses 5.5-5.14 (Students). http://www.reading.ac.uk/reas-PP_IntellectualPropertyOwnership.aspx

Example

- *The industrial sponsorship agreement for my PhD, between the University and Syngenta, is filed under University project code F123456. I have stored a copy of the contract in my project folder, under /Documentation/Contract.*

2 Data collection

2.1 Secondary data sources

Identify key secondary data sources you will use as inputs into your research.

Identify the data provider and the precise dataset where possible, ideally with a DOI reference, or a data collection or archive, with URL reference. Note the licence terms or any conditions placed on use of the data.

You only need to identify data sources: you do not need to include here information about published literature.

Examples

- *ONS Opinions Survey, Census Religion Module, 2009 (http://doi.org/10.5255/UKDA-SN-8078-1) held in the UK Data Service. Data are accessible via the Secure Lab only. Access requirements: ONS Accredited Researcher status (including the requirement to attend and pass the Safe User of Research data Environments (SURE) training course); submission of a project proposal; and completion of a Secure Access User Agreement signed by the University. Copies of data cannot be made and any data outputs must be approved for release by the UKDS.*
- *Landsat 8 longitudinal satellite data covering the Antarctic region, available from US Geological Survey EROS (https://landsat.usgs.gov/). Data are public domain and can be freely used, reproduced and distributed. Citation is requested (https://eros.usgs.gov/about-us/data-citation).*
- *Soil survey in England, Scotland and Wales carried out during 2013 and 2014. Available from Environmental Information Data Centre at https://doi.org/10.5285/17bebd7e-d342-49fd-b631-841ff148ecb0. Data are available under Open Government Licence.*

2.2 Primary data collection

List the types, formats and quantities of data that you will collect or create.

Briefly itemise and describe the types of data you will collect or create in order to answer your research questions. You need only describe the categories of data or the key variables, sufficient to identify their essential characteristics, e.g. telephone interviews, EEG brain activity measurements, weather system model outputs.

For each data type, state what physical and digital file formats data will be stored in. Bear in mind that data may be collected, processed, analysed and preserved in different formats. For long-term preservation of data collected in proprietary file formats, e.g. specific to a data collection instrument, such as a digital camera or an NMR spectrometer, or to proprietary software, such as MATLAB, you should plan to save data to an open format or one accessible without a software licence where possible, such as CSV for tabular data, or JPEG for images. Widely-used formats are acceptable, e.g. MS Word. The UK Data Service provides guidance on recommended file formats for commonly-used file types: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.

For each data type, try to quantify the amount of data you expect to collect, e.g. in numbers of experiments/observations/interviews, duration of collection period and frequency of data collection, or overall numbers of bytes (MB, GB, TB). You may be able estimate total data volume by calculating the numbers of files of each type you intend to collect, and the average size of each file. If you plan to collect a substantial volume of data (anything from 10s GB upwards), it is important to have some idea of the volume of data you will be dealing with, as you may need to consider practical issues related to storage, transfer and long-term preservation and sharing of such data.

Examples

- *I will collect 30 one-hour interviews with farmers to find out about their use of and attitudes to pesticides. These will be stored as MP3 recordings of about 60MB per recording, and transcribed into Microsoft Word (.docx) for analysis and preservation.*
- *I will generate weather system simulation data in NetCDF format. I will perform ~50 model runs with different input parameters. Initial conditions data will be 10 GB approx. per run. Each run will generate approximately 40-50 GB of data. I therefore anticipate a total storage requirement up to 3 TB. The final preservable output will be in the region of 50-100 GB.*
- *EEG measurements will be taken from 15 subjects over a course of approximately 2 hours per subject. Data will be collected from the instrument as ASCII text files, imported into MATLAB for analysis, and exported as CSV for preservation. Total data volume is expected to be ~15-20 GB.*
- *NMR spectra will be collected from 10 samples in proprietary Bruker Topspin format. Files will be stored in ASCII text for further analysis and preservation.*
- *Daily measurements of key plant characteristics, along with date and time of measurement and other environmental variables will be recorded for 60 plants*

over 40 days using paper forms and input into Excel for storage and analysis. Data will be preserved in CSV format.

- *Up to 1000 2-D X-ray images of DNA crystal samples at approximately 1 GB per dataset will be generated. Data will be screened and discarded at the machine as they are generated. Raw data will be processed into the much smaller CIF format for further analysis and preservation. Total actual storage requirement is not expected to exceed 150 GB.*

2.3 Instruments and methods

Describe your proposed data collection methods, and the instruments and software you will use to collect and process the data, including details of any software you will create.

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/res-redcap.aspx
- Online survey tools: https://www.reading.ac.uk/res-survey-tools.aspx
- 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 in code commentary. Analysis scripts will be archived to ensure reproducibility of final results.*
- *To clean and process instrument data, I will write a re-usable C++ library, developed and documented using CWEB.*

2.4 Quality control

Describe the data quality controls you will use.

Consider how you will maintain consistency and accuracy of data throughout the data collection and processing workflow. How will you reduce the risk of introducing errors in the data, and mitigate the impact of errors when they occur? Various quality control strategies can be used:

- Standardise and document your workflows, so that another person could follow your instructions and achieve the same result as you, for example, by writing a step-by-step protocol for data collection, or guidelines for formatting and anonymisation of interview transcriptions. Follow established procedures where relevant, such as laboratory Standard Operating Procedures. Remember to preserve and comment scripts used for data processing.
- Define your data structures and data collection forms or templates in advance. For example, set up a spreadsheet with variables clearly labelled in column headings, including units of measurement. Include in the document a separate worksheet with instructions for data entry. This should provide a full definition of variables, and information about permitted values for given variables (including missing value codes).
- Establish error control processes. Make use of any data validation functions in your software, e.g. Excel allows you to specify permitted values for a cell or range of cells. Methods such as double entry of data and random sample checking can reduce the incidence of error. Review data to check they make sense. Data visualisation can help to identify suspicious outliers and anomalies: a trendline with an obvious spike in it may highlight an incorrect value.

Examples

- *I will develop and pilot an interview schedule, so that interviews follow a standard format. Recordings will be transcribed following guidelines specifying the transcription format, tags to be used and anonymisation rules.*
- *Sensor instruments will be calibrated prior to data collection and recalibrated monthly to reduce drift error. Data will be saved in ASCII format and imported into MATLAB for cleaning and analysis. MATLAB scripts will archived to ensure reproducibility of results from raw data.*
- *Laboratory experiments will follow Standard Operating Procedures. Lab notebook entries will be regularly reviewed and signed by the Supervisor. All experiments will be repeated and results analysed for significance. Results that cannot be replicated will be discarded.*
- *Daily measurements of plant characteristics during the field trial will be recorded by the technician on a paper template and later entered into a spreadsheet. Measurements will be undertaken by the technician and myself on day 1 and on 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 – 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 need-to-know basis. It is in your interest to protect the intellectual property you create, so that you can be the first to derive advantage from it in developing and recording your research findings in your thesis and other publications.

If you will collect personal data or any other sensitive information, you may have a statutory or contractual obligation to restrict access to data, as well as an ethical obligation to protect the confidentiality of your data sources.

By default you should use allocated University storage in your OneDrive account or on the University network as your primary storage location. These are University-warranted highly resilient services providing access control, automatic file replication and backup. Raw data, documentation and master versions of files should be kept here. Personal and sensitive data can be held in these locations. OneDrive provides 1+ TB of storage per account. You may also have allocated capacity in a research group or project fileshare on the University network: your Supervisor will be able to advise if this is the case.

The online data collection services provided by the University (UoR REDCap, Jisc Online Surveys, Qualtrics) also offer secure storage and backup for data they hold, and are suitable for collecting and maintaining identifiable information. Data held in UoR REDCap are stored and backed up on University premises. Jisc Online Surveys and Qualtrics provide secure and resilient data storage, but they are third-party services, and export of data for storage on University infrastructure is advisable.

If data are acquired using specialist infrastructure, such as the ISIS neutron and muon source, or the JASMIN supercomputing environment, raw data may be stored in the facility infrastructure, and data copied or extracted locally as required.

Personal devices such as laptops and other cloud services (e.g. Dropbox, GoogleDrive) may be used as a temporary workspace or for sharing data with colleagues, but with these caveats: cloud services other than OneDrive should not be used to store personal/sensitive data; and personal data should not be stored on personal devices: if 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/internal/its/services/office365.aspx
- University storage services: log in to IT Service Catalogue and select File storage. http://www.reading.ac.uk/internal/its/services/sercat2017.aspx
- Guidance on academic computing resources, including Research Data Storage, Research Cloud computing platform, cloud storage and other services: https://research.reading.ac.uk/act/
- Encryption Policy (including practical guidance on encryption): download from https://www.reading.ac.uk/internal/imps/policiesdocs/imps-policies.aspx

Examples

- *All data and documentation will be stored in my University OneDrive account. My personal devices will be synchronised with OneDrive to ensure these are stored and backed up securely. Data will be shared with my Supervisor on request.*
- *Raw X-ray image data, which may be up to 1 TB in volume, will be stored at the Diamond Light Source (DLS) beamline facility, where they will be archived to tape and retained for a minimum of 7 years. Data held at DLS will be accessible only by myself and system administrators. Processed data created at DLS, which will be much smaller in volume, will be transferred by VPN to an access-controlled*

area in my research group's University network collaborative fileshare, where they will be accessible to my Supervisor and myself.

- *Participants' consent forms and completed questionnaires will be stored in a locked filing cabinet in my Supervisor's office. Questionnaire data will be entered into the study database, to be stored in my personal drive on the University network.*

3.2 Organisation

Describe how your data will be organised.

There are three elements to data organisation: filing system, file naming, and version control.

- Filing system: you should establish a folder structure for your data and supporting documentation. There are many different ways this might be organised; the important thing is that the structure is logical, legible, and meaningful for its purpose. For example, try not to create too many layers in your folder hierarchy. If different files have different access permissions, establish these at a high level. Then organise files into folders according to task (e.g. work package, experiment), then a significant defining property (e.g. location, sample number, run, company name) or type of data (e.g. raw, processed, final).
- File naming: you should establish conventions for naming files, using significant properties of the data to allow the contents of a file to be easily and unambiguously identified. Some properties you might use include: data collection method or instrument, data type, location, subject, and date. You don't have to force all files into a rigid convention, but if you adopt some basic standards they will help you find and organise files. For example, by always writing dates in YYYMMDD format, you will be able to sort files chronologically. Avoid spaces in file names; you can use _ or – to separate elements, or run them together using CamelCase.
- Version control: You should have a system for keeping track of different versions of files if they will be modified frequently or by different people. For most purposes you can use version numbers or dates in your file name. At larger scales, you may need to use a version control system such as Git or Mercurial.

Useful links

- File naming convention worksheet: https://authors.library.caltech.edu/103626/
- MIT Libraries on data management and file organisation: http://bit.ly/2dMsNVv

Examples

- *A screenshot of my folder structure is attached to the end of this document.*
- *Experimental data will be stored in folders using the following convention: [experiment]-[date]/[reagent]-[replicate number].*
- *Folders will be organised by Work Package, with WP0 holding project documentation. Within each WP folder, I will create folders for Data, Methodology, WP documentation, and, where relevant, Participants. Within the Data folder I will create sub-folders for Raw data (which will be read-only), Data analysis, and Final*

data. Where a WP contains a Participants folder access will be restricted to myself and my Supervisor.

- *I have set up folders for each instrument. I will run scripts to write files from the instrument to related folders using the collection date and automatic numbering in the format [instrument name]-YYYY-DD-MM-[autonumber].*
- *The version number will be appended to the filename using the convention _001. As new versions are saved the version number will be incrementally increased.*
- *Code will be written and stored in a GitHub repository, which provides version control.*

4 Documentation and metadata

Describe the documentation and metadata you will record about your data.

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.
- Metadata level information is a structured description of an information item such as a dataset consisting of a set of defined elements. It is usually created when a dataset is deposited into a data repository or described in a data catalogue, and will be composed of information generated at the first three levels of documentation. The metadata description enables a dataset to be discovered online and provides key information to enable the data to be understood and

used. Core metadata properties are typically: Creator(s), Title, Publisher, Publication Year, Resource Type, Unique Identifier, e.g. DOI. Additional properties may be included to facilitate discovery and use, such as description, keywords, temporal and geographical references, rights and licence information, and links to related publications.

You will not need to create a metadata record for your data until you have completed data collection and analysis, and are in the final stages of your research or preparing a publication. At this stage you should be thinking of depositing your data in a relevant repository. But if you have identified a specific repository that you plan to deposit data in, it is worth familiarising yourself with their metadata requirements, so that you have all the information you need when the time comes. For example, if you are conducting microarray or next-generation sequencing experiments and plan to deposit data in ArrayExpress, you should be prepared to record your experiment using the Minimum Information About a Microarray Experiment (MIAME) or Minimum Information About a Sequencing Experiment (MINSEQE) guidelines (https://www.ebi.ac.uk/arrayexpress/submit/overview.html).

Useful links

- Metadata Standards: http://rd-alliance.github.io/metadata-directory/standards/.
- Life sciences standards: https://fairsharing.org/standards/.
- Giraldo O, Garcia A, Corcho O. (2018). ‘A guideline for reporting experimental protocols in life sciences’. PeerJ 6:e4795 https://doi.org/10.7717/peerj.4795

Examples

- *Model input and output files will be in NetCDF format. Each file contains a header section with information about variables contained in the body of the file. Input and output files will be saved to separate folders for each run and automatically numbered, and input parameters and model specifications for each run will be separately recorded. Model code will be managed in GitHub and will be fully commented.*
- *I will develop a study protocol, describing my hypotheses, study design and methods, which will be publicly registered before I begin data collection using the Open Science Framework Registry (https://osf.io/registries/).*
- *Field trial data will be recorded in Excel spreadsheets. Column headings will include the variable name and unit of measurement. A separate worksheet will provide a full listing and definition of variables. I will set up Excel data validation rules where possible to ensure values entered are in the correct ranges.*
- *Quantitative survey data and interview transcriptions will ultimately be deposited in the ReShare repository at the UK Data Service, and will be documented to V. 2.5 of the Data Documentation Initiative (DDI) Codebook standard, which is used by the UK Data Service and social science data archives worldwide. Anonymisation protocols for quantitative and qualitative data will be fully documented.*

5 Ethics and Data Protection

Describe how you will manage ethical issues and compliance with data protection law relating to research participants, where relevant.

You have an ethical obligation to protect the confidentiality of personal information provided to you by research participants, and you must also comply with data protection law if you collect and process personal data. You must consider how you will meet these obligations in the way you manage your research data, and how you will make data collected from research participants safe for sharing, by means of anonymisation or controlled access procedures.

Any research involving human subjects will need to receive approval from your School's or the University's Research Ethics Committee. You should not in your application for 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: 'I 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 others'.

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 the University Information Management & Policy Services Officer.

You must ensure that personal data are kept secure and are not disclosed to unauthorised persons. You should use a locked storage container such as a filing cabinet in a locked office for paper-based personal data; for digital data, password-protected or, preferably, encrypted storage. This particularly applies in the case of special category sensitive personal data, which include information about an individual's: race; ethnic origin; politics; religion; trade union membership; genetics; biometrics (where used for ID purposes); health; sex life; or sexual orientation. Such personal data should be encrypted, and not stored or shared by means of cloud services other than a University OneDrive account, or transferred via unencrypted channels (e.g. via email). You can transfer data to a location on the University network using VPN, which provides an encrypted channel.

Avoid storing personal data on portable devices such as laptops or external hard drives. If you must do so, ensure the device is encrypted. If 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.

Remember that consent forms are personal data and also an important part of the research record. They must be retained by you and/or your School for a minimum period of five years from the completion of the research, and for as long as the personal data are held.

You should also think about what data you actually need and plan your research accordingly. Do not collect more personal data than necessary. Data can often be pseudonymised for purposes of processing and analysis, with the personally-identifying information and their linked IDs stored separately from the working dataset. This helps to minimise the risk of inappropriate disclosure. When the study is complete and if there is 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 should be managed under a retention schedule that specifies periodic reviews, so that they can be securely destroyed when no longer needed.

Useful links

- University Research Ethics guidance: http://www.reading.ac.uk/REClethicshomepage.aspx
- University Data Protection, Remote Working and Encryption Policies: http://www.reading.ac.uk/imps-policies.aspx
- University Data Protection and Research, including a Data Protection Checklist for researchers, and sample information sheets and consent forms: http://www.reading.ac.uk/imps-d-p-dataprotectionandresearch.aspx
- UK Data Service Legal and ethical issues, including guidance on consent with model consent form, and anonymisation: https://www.ukdataservice.ac.uk/manage-data/legal-ethical

Examples

- *Data will consist of mental health screening questionnaires and cognitive test results. Signed e-consent forms, screening questionnaires and cognitive test results will be collected and stored in a UoR REDCap project database accessible* *by myself and my supervisor only. Participants will be assigned a unique ID number in the REDCap database. Identifying details will be stored in the REDCap database in a separate demographics instrument. On completion of the study an anonymised dataset will be exported from REDCap, with tagged identifier fields removed and ID numbers hashed to ensure that individual data records cannot be relinked to participants. This anonymised dataset will be archived to a suitable data repository for long-term preservation. PDFs of e-consent forms will be exported from REDCap and stored for five years after the project in my supervisor's University OneDrive account. When all relevant data have been exported from the REDCap, the REDCap project will be deleted.*

- *Interviews will be copied from the audio recording device to my laptop, and from there transferred to my University OneDrive account. The interviews will be removed from the audio recorder and laptop once transferred into secure storage. Transcription and de-identification of interviews will be undertaken by a professional transcribing service. Audio files will be securely transferred by an encrypted channel. The service contract will require the transcribing service to comply with data protection laws and maintain the confidentiality of information supplied. I will review transcriptions to check that data have been satisfactorily de-identified. Original recordings will be destroyed on completion of the research and transcripts will be anonymised prior to archiving.*

6 Intellectual Property Rights

State how you will you manage copyright and Intellectual Property Rights (IPR) issues.

There may be two kinds of Intellectual Property Right (IPR) in a collection of data: copyright, which attaches to the 'creative' work of selecting and arranging facts in a specific presentation; and sui generis database right, which protects any 'substantial investment in obtaining, verifying or presenting the contents' of a database. These rights are protected in law.

If you will derive new data from secondary sources, you may need to check that you will have permission to share these data. This would particularly apply if your derived data extract or re-use all or a substantial part of the contents of the source database. You do, however, have a right to extract insubstantial parts of the data for any purpose. If your data are derived by significant processing and do not reproduce the source data, they are likely to qualify as new data beyond the scope of any legal protection enjoyed by the source database.

By default, as a student, you are likely to own your IP, unless it has been otherwise assigned by e.g. a contract of industrial sponsorship, or an IP assignment agreement (see section 1.3). If your research is carried out under contract, you should check the terms of the contract to establish who has ownership of your data. Research contracts will have IP clauses which deal with ownership of IP arising under the contract. Under industrial sponsorship contracts IP created by the student generally belongs to either the student or the University. In the latter case, ownership by the University does not prevent you from public disclosure of the data by deposit in a data repository, as University policy promotes an open research data culture.

These contracts also have Publication clauses, which generally grant other parties the right to be notified of and have the opportunity to approve or delay any intended publication. This right exists irrespective of who owns the IP created under the contract. Deposit of data in a data repository for long-term preservation and sharing will constitute effective publication, so any notice requirements must be met. The standard notice period is 30 days.

If your data will be created jointly with others, for example by members of a research group working together, there will be other parties' IP rights to consider. But bear in mind the definitions of IPR above: in most cases, neither Supervisors nor data collectors acting under your instruction are creators of the data, and they do not have IPR in the data. Your Supervisor may have substantial input into the design and conduct of your research, but they will not be the one collecting the data and selecting and presenting them as a collection of facts. In most cases also, technicians or other individuals collecting data under your instruction are not creators of the data, as they generally have no creative input into how the data are selected and presented.

Qualitative research may involve collection of IP from participants: for example, interviewees own copyright in their spoken words, and some research may involve participants creating data, e.g. photographs and artworks. In this case you should ask participants to transfer copyright in their data to you, or agree to joint copyright, or grant you a licence to use their data, e.g. by including in the consent form a statement such as, 'I agree to assign the copyright I hold in any materials related to this project to [name of researcher]'. It is easier if copyright can be assigned to you, as your use of the data is then not restricted in any way; but requesting a licence grant may be fairer in some participatory research scenarios or where participants set a high value on their data.

Useful links

- Carroll MW (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol13(8): e1002235. https://doi.org/10.1371/journal.pbio.1002235

Examples

- *Under the terms of my industrial sponsorship contract, ownership of IP created by me is vested in the University, and there is no reason why the data cannot be made publicly available when I complete my thesis. In accordance with the publication clauses of my sponsorship contract, I will give my industrial supervisor at least 30 days' written notice before making the data publicly available in my chosen data repository.*
- *IPR in data created by myself will belong to me. Research participants will be asked to keep a photographic diary during the experiment period. I will ask participants to either transfer copyright in their photographs to me, or to agree to joint ownership of copyright.*

- *Data will be created collectively with students and staff in my research group through experimental data collection at beamline facilities and data processing. As individual contributions are difficult to identify, it has been agreed with my Supervisor that IPR will be jointly shared by the University and students involved in data collection and processing.*

7 Preservation and sharing

7.1 Data selection

Identify the data you will preserve and share at the end of your project.

Data that support the findings reported in your thesis and in any publications that result from the research should be preserved, and made publicly accessible where possible, by 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-describing formats encoded in text files, where the file contains a header with information about the variables reported in the body of the file: examples include the NetCDF format used in climate system models, and the FASTA format for representing nucleotide or peptide sequences;
- widely-used proprietary formats, such as MS Excel and MS Access for tabular data and databases, MS Word for text, TIFF 6.0 uncompressed for images, and MP3 or WAV for audio.

For example, raw instrument data in a proprietary format may be preserved, but also or alternatively converted into an ASCII/CSV format, to be more widely accessible; data 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: http://www.reading.ac.uk/res-licensing-data.aspx.

Examples

- *Full results of all completed experiments will be retained, and preserved in CSV format.*
- *I will preserve anonymised interview transcripts in MS Word format, but will destroy original audio recordings to remove the risk of accidental disclosure.*
- *Simulation code and input parameters, recorded in NetCDF format, will be retained, but intermediate files will be destroyed as it will be more cost-effective to recreate them if necessary than to preserve them.*
- *For 1D and 2D NMR measurements, original Free Induction Decay (FID) files will be converted to CSV for plotting and preservation. Original FID files will not be preserved.*
- *High-resolution microscopy images will be preserved in TIFF 6.0 uncompressed format.*

- *Raw survey data will be preserved in CSV format. Statistical analysis results will be exported in SPSS portable format (.por) for preservation.*
- *The dataset will be licensed under a Creative Commons Attribution 4.0 International License.*

7.2 Data repositories

Identify any data repositories that will be used to preserve and share your data.

Data repositories are sustainably-managed services that exist to preserve and provide access to data over the long-term. They actively preserve and curate data, publish information online in the form of metadata records, so that other people can discover the data, and assign a unique persistent identifier, such as a DOI, to each dataset, so that it can be uniquely and persistently identified, cited and linked to from other sources, such as your thesis and research publications. Mostly these services are free to use.

You should plan to use a subject or data type-specific repository where a suitable one exists. These are authoritative community resources for a particular field of research or type of data; they include some services that are run by funders of research. They provide subject-specialist data curation, e.g. by applying disciplinary metadata standards, by curating data in standard formats, and by providing tools for data manipulation and visualization. Examples include:

- the ReShare research data repository at the UK Data Service, the primary resource for social science data, funded by ESRC. https://reshare.ukdataservice.ac.uk/
- the several NERC data centres, which host environmental data collected in NERC-funded research, including climate and weather, oceanographic, terrestrial and freshwater, geoscientific, and polar data. http://www.nerc.ac.uk/research/sites/data/
- The databases of the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), among which are ArrayExpress, a database of functional genomics experiments, including microarray and RNS sequence expression data, and Metabolights, a cross-species archive for metabolomics. See https://www.ebi.ac.uk/services/all
- Cambridge Crystallographic Data Centre, a reference database of crystal structures. https://www.ccdc.cam.ac.uk/

For guidance on choosing a suitable repository, visit http://www.reading.ac.uk/res-data-repository.aspx.

Not all areas of research are well-served by community data services. Coverage tends to be best in the environmental sciences, molecular biology and chemistry, the social sciences and archaeology.

In the absence of a suitable community resource for your data, you can deposit them in the University’s data repository, the University of Reading Research Data Archive. Visit the Archive webpage at http://www.reading.ac.uk/res-research-data-archive.aspx for guidance including a Data Deposit Checklist, instructions for depositing data, information on recommended data formats, and a link to the Archive.

You may also use a general-purpose data sharing service if you wish. Examples include Zenodo (https://zenodo.org/), figshare (https://figshare.com/), and Dryad for scientific and medical data (https://datadryad.org/). Zenodo and figshare are effectively self-publishing services, and do not offer the subject-specialist data curation and quality control that may be provided by other services.

If there is an inherent risk of personal identification in data, e.g. genetic data, these may be shared under controlled access conditions, such as requiring a prospective user of the data to register with the data service, demonstrate bona fides and sign a confidentiality agreement. There are data repositories that can manage controlled access to data, e.g. the UK Data Service for social science microdata (https://www.ukdataservice.ac.uk/get-data/how-to-access/conditions) and the European Genome-phenome Archive for human sequence and genotype experiment data (https://www.ebi.ac.uk/ega/home). The University Archive can also hold restricted-access datasets. Contact the Research Data Manager if you need assistance with this.

You should plan to deposit your final dataset so that it can be cited by DOI in the final version of your thesis deposited in CentAUR, and in any research publications. You should reference the data at appropriate places in your thesis, and include a full data citation in the references list, including the following metadata: Dataset Creator(s), Title, Year of Publication, Publisher, Resource Type, Unique Identifier (i.e. DOI). Many data repositories, including the University's Research Data Archive, provide a citation that can be copied and pasted into a reference list.

Useful links

- University guidance on choosing a suitable data repository: http://www.reading.ac.uk/res-data-repository.aspx.

Examples

- *Anonymised survey and interview data and supporting documentation will be deposited in the ReShare data repository at the UK Data Service.*
- *Crystal structures will be deposited in the Cambridge Crystallographic Data Centre.*
- *Data from field experiments will be deposited in the University of Reading Research Data Archive. Metagenomics data from soil samples will be deposited in EBI Metagenomics.*

7.3 Code and research software

Identify any code or research software that will be preserved or maintained after the project and how this will be done.

If you will write any computer code to generate, process, analyse or visualise research data, or if you plan to develop software as a research output in its own right, you should consider long-term preservation and maintenance.

In most cases scripts and segments for code written e.g. for purposes of data processing, statistical analysis or data visualisation can be archived alongside data.

Where the research software is more substantial or has been written in the context of an ongoing project or established community, a development-oriented approach to publication and maintenance of code may be appropriate. This might be the case, for example, where the code written contributes to a published model. Options might include contributing code to a language-specific network such as PyPI, RubyGems.org or Hackage; or sharing a public code repository using a platform such as GitHub or the University GitLab.

As with other intellectual property, software code should be shared under licence. Analysis code archived alongside data can be shared under the dataset licence (see 7.1 above). Where the full source code of software developed or modified in the research is distributed, it will be more appropriate to apply a software licence. You are encouraged to share code under an Open Source licence, which grants broad permission for re-use, unless there are legal, ethical or commercial reasons why this cannot be done. A variety of Open Source licences exists, including broadly permissive licences (such as Apache 2.0 License), and licences which afford specific protections (e.g. GNU GPL 3.0, which requires any modifications to be distributed on the same terms). Refer to the links below for further guidance.

Note that if you have modified existing source code then your licensing options may be constrained by any licence applied to the original code.

Useful links

- University guidance on managing and publishing research software and source code: http://www.reading.ac.uk/res-research-software.aspx.
- Open Source licences for software: https://opensource.org/licenses.
- Choose an Open Source license: https://choosealicense.com/.

Examples

- *Scripts in R (for statistical analysis) and Python (for data visualisation) will be archived alongside data files, as specified above.*
- *Model code will be hosted and developed during the project using GitHub. On completion of the research, the code repository will be made public under a GNU GPL licence, with releases shared through CRAN. I will encourage contributors to submit issue reports and pull requests, but will retain control of the official codebase.*

7.4 Timeframe for data sharing

Specify when data will be made available, and whether there will be any restrictions on access.

Data should be made available no later than publication of any findings that rely on them, so that results placed on public record can be validated against the underlying evidence. Data supporting your thesis should be made available by the date on which the final electronic version deposited in CentAUR is publicly released. Many data repositories, including the University's Research Data Archive, allow you to deposit data under embargo pending their release date.

Data should be made publicly accessible wherever possible, and most data repositories provide unrestricted access to the data they host by default. Most research data can be shared publicly, although they may need prior redaction to remove personal and confidential information. If commercial exploitation of research results is anticipated, public release of data can be delayed until IP protection has been confirmed.

If data have been obtained or derived from existing sources you may need to be aware of any licence conditions that affect whether and how the data can be shared. If your 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 typically require you to give your industrial sponsor at least 30 days' advance notice of any intended publication.

Examples

- *Data will be made available no later than public release of the final electronic version of my thesis deposited in CentAUR, or publication of any research outputs on which I am named, whichever is earlier. I do not anticipate any restrictions on access to data.*
- *Data will be embargoed for 12 months to enable patent applications to be filed, and will then be publicly released.*
- *Analytical data outputs derived from the ONS Opinions Survey microdata must be submitted to the UK Data Service to be approved for public release.*
- *In accordance with the terms of my industrial sponsorship contract public release of data will be notified to the industrial sponsor for approval at least 30 days before the date of intended publication. Data will be redacted to abide by the commercial confidentiality terms of the sponsorship agreement.*

8 Implementation

8.1 Responsibilities

Who will be responsible for data management activities?

You will have overall responsibility for data management in your project, but you should also identify any persons or organisations that will be responsible for delivering specific data management activities. This might include colleagues, e.g. laboratory technicians, or members of a research group; or service providers that may have responsibility for specific aspects of data collection or management, e.g. survey companies, laboratory service providers, professional transcription services.

If other people or organisations are undertaking data management functions, you will need to consider how they can be properly instructed and trained if necessary, and what supervision or quality controls you will put in place to ensure responsibilities are executed to the required standards. Any third parties processing personal data on your behalf should do so under a service contract setting out their data protection responsibilities. The University's IMPS Officer can advise on this requirement. Relevant information can be documented in the Data Collection section of your DMP, and the DMP, or relevant parts of it, can be shared as necessary.

8.2 Resource requirements

Specify any resources this plan will require.

Will you need any resources beyond standard University provision for purposes of collection and data management? This may include hardware, such as data 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.*
- *I'm unsure which is the best storage solution for my needs, I will contact IT support to discuss this.*
- *I'm unsure which is the most suitable data repository to preserve my data. I will contact the Research Data Manager about this.*