data tree

datatree.org.uk
@_datatree

Data management questions and answers

What is a data management plan and what should I include in one?

A data management plan (DMP) is a structured document detailing how data will be collected, managed, preserved and shared in a research project. It addresses data management throughout the research lifecycle and should include information on the following:

- What data will be collected or created (including both primary and secondary data);
- The instruments and methods that will be used to collect and process data (including any software that may be created);
- The quality control processes that will be applied to maintain consistency and accuracy of data and reduce the incidence and impact of error;
- How data will be stored and kept secure during the active phase of the project;
- How data will be organised, so that they can used efficiently;
- What information will be recorded about the data, so that they can be validated, interpreted and used by yourself and others;
- How any ethical and data protection issues relating to any research participants will be handled, where this is relevant;
- How intellectual property rights in the data will be handled;
- How and when data and supporting materials (such as software code) will be preserved and shared at the end of the project;
- Who will be responsible for doing what, and what resources will be required.

Various data management plan templates exist, for example those provided by DMPonline (https://dmponline.dcc.ac.uk/) or the Digital Curation Centre's Checklist for a Data Management Plan (http://www.dcc.ac.uk/resources/data-management-plans/checklist).

Why should I write a data management plan?

Reasons to write a DMP:

- Make life easier
    - Well-organised data management increases your efficiency, and saves time and effort in the long run;
- Protect yourself and others
    - Reduce the risk of costly/embarrassing/damaging accidents – losing data, disclosing confidential data;
- Preserve the integrity of your research
    - Well-documented data demonstrate the authenticity of your research and the reliability of your findings;
- Plan ahead for sharing
    - Public sharing of data supporting research findings is integral to the practice of scholarly communication, and must be anticipated and planned for from the outset
    - Data that are preserved and accessible in the long-term can be re-used to your benefit and others'.

How much data will I collect?

Thinking about how much data you will collect can help you with research planning, e.g. if you need to use resources, such as experimental instruments or facilities, or if you need to ensure you have sufficient storage for your needs – especially important if you will be collecting substantial volumes of data.

You should be able to quantify, e.g. based on sample or average output volume per experimental unit, and then scaling up by the expected number of units. Your research materials can be divided into experimental units in different ways, e.g. for modelling, the units may be the total input and output volumes for a single model run; for experiments, they may be the raw and analysed files per experiment; for sensor observations, they may be the volume of data per 24 hour period.

You will probably find that you start your research with a very vague idea of how much data you will collect and a large margin of error, but as you proceed through the research, you will refine your estimate. As you become more experienced in research you will be able to better estimate the volume of data you are likely to generate and use in a project.

Where should I store my data?

Always observe the 3-2-1 rule:

- Have at least three copies of your data;
- Store the copies on two different media;
- Keep one backup copy offsite.

Institutional network storage services in universities and at experimental and computing facilities will meet these requirements, typically by replication in separate on-site data centres and offsite tape backup with a minimum recovery period such as 3 months. Most cloud services will also replicate files in multiple server locations and provide a file recovery period (e.g. for Microsoft OneDrive this is 93 days).

Cloud services (e.g. Google Drive, Dropbox, OneDrive) can be useful if you need to share data with team members/partners/collaborators, but they are commercial services and do not have any institutional commitment to preserve the integrity of your data, so should be used with caution. They may not be suitable for storage of personal and sensitive data. Many universities now provide access to institutionally-warranted cloud services: these will be more or less the same as the standard cloud services, but may provide additional storage and security warranties.

In general, you should use your institutional services as the primary storage location(s) for your data, as these will be managed to provide guaranteed data security and storage resilience. Personal devices and cloud accounts can be used as working storage or for sharing data, providing appropriate security is in place, but you would be advised to store raw data and master files with your institution.

What formats will my data be stored in?

Data may be collected, processed, analysed, and preserved in different formats, so you may need to think about your data at different stages of the workflow. For example, variables obtained from raw observational data may be converted to a format suitable for input into a computer model. At the end of the project, the researcher may wish to preserve the raw observational data, as well as the model inputs and outputs.

For long-term preservation, consideration should be given to storing data in open or widely-used formats. 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 GeoTIFF format for embedding georeferencing information in TIFF files;
- 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.

How can I share my data safely with my Supervisor or trusted peers?

Local network services will provide means for shared access to a storage location, such as a project or research group fileshare. If you have allocated capacity in shared storage on your institutional network, then the fileshare administrator should be able to grant read-only access to other authorised network users, such as your supervisor.

You can also use a cloud service to share data with trusted peers, providing it is safe and secure in proportion to any risk in the data. Bear in mind that until such time as the results of your research are made public, you should be under no obligation to make your data public, and it can be in your interest to keep them private or to share them only in confidence with trusted peers. For example, if you have an option to share your data by creating a public URL or by making the data accessible only to nominated users, it will be safer to use the latter option.

If I wanted to reproduce my research findings in 5 years' time, what information would I need to record now?

It can be useful to think in terms of four levels of documentation:

1. 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;
2. 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;
3. Project level information describes the research questions and hypotheses the data have been collected to answer or test, the design of the research and the methodologies used, information about the instruments used to collect and process the data (which might include the full commented code of any scripts or software that have been written), and records of the research process. Documentation might include a description of the research questions and methods, instrument or software specifications and guides, field notes, etc.;
4. Metadata level information is contained in a structured description of an 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.

How do I know my data are accurate and reliable? How will other people know?

Consider how you will **maintain** and **document** 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, and preserving and commenting any scripts you write;
- 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. Recalibrate instruments at fixed periods to correct for drift. 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.

Who owns my data?

Students:

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. 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 generally does not prevent you from public disclosure of the data by deposit in a data repository, as Universities promote an open research data culture in policies.

Employees:

It is standard for employment contracts to state that IP created by employees in the course of their employment belongs to the employer, unless any contract specifies otherwise.

How can I ensure my data are preserved and remain accessible to myself and others in the long term?

The best way to do this is to deposit your data in a **data repository** where they will be preserved and made accessible to others under a suitable licence.

A data repository is an organisational service that exists to preserve and provide access to research data. The organisation managing the repository may be a data centre serving a research domain or community (such as a NERC data centre, or the EC-funded Zenodo repository, which exists to preserve EC-funded research but is freely open to all), a University of other research organisation, or a general-purpose data sharing service (such as figshare or Dryad).

A data repository must meet certain criteria to qualify as a **trustworthy data repository**. It must be a sustainable service that is committed to preserving and maintaining access to data in the long term. A trustworthy data repository will:

- Actively preserve data, e.g. replicating, maintaining bit integrity, migrating to preservation formats;
- Publish metadata to established standards and enable online discovery;
- Assign persistent unique identifiers (e.g. DOIs) to datasets;
- Issue a licence notice for the data, making the terms of use and attribution requirements clear;
- Manages access to data so that they can be used by other people.

You can search for data repositories at https://www.re3data.org/.

Why should I share my data?

There are two main reasons why you should do this:

1. So that the reported results of your research can be validated or tested or corrected by others. If you put scientific findings on public record, then it is incumbent on you to provide the evidence that substantiates these findings, to allow others to scrutinise the evidence and your interpretations. If you are not transparent with the evidential basis for your findings, then these findings are not demonstrably reliable or trustworthy;
2. So that, where data may have value to others, they are made available and usable. Research data may have many potential uses, for example: to other researchers as inputs into new research; to policy-makers as evidence in support of a policy; and to industry, in the development of new products and services.

You should always bear in mind that though you may think of the data as being yours, they would not exist without public funding. Universities and other research organisations are funded by public money, both by direct Government vote and by winning grants from public and charitable funders of research. Many public funders adopt the view that data and other outputs generated by the research they fund are a public good, paid for with public money, and produced in the interest of increasing the sum of public knowledge, and contributing to social and economic benefit. Therefore, there is an ethical obligation to make the data and other outputs of such research publicly accessible.

I don’t want to share my data until it’s absolutely necessary. How long can I keep my data private?

As a research you enjoy a first-use privilege in respect of the data you collect or generate. It is only fair that you should be entitled to reap the first rewards of the effort you have put in to collect and analyse the data. But this privilege does not last for ever.

As a matter of principle, 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. You can still take steps to preserve your data well ahead of their public release. Most data repositories 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.

Some funders allows funded researchers a specific limited period of time to derive benefit for themselves from research data before it is expected to be made more widely available.

Who else might want to use my data?

Your data may be of interest to:

- Other scientists in the field: they may want to validate or test your results, or use your data as inputs into their own research;
- Policy-makers and those seeking evidence to inform or influence policy, for example directed at mitigating the environmental impacts of anthropogenic activities;
- Industry, for example companies interested in developing new products or services based on your research;
- The general public, who may take a personal interest in or be directly affected by the results of your research.

What is a data repository?

A data repository is an organisational service that exists to preserve and provide access to research data. The organisation managing the repository may be a data centre serving a research domain or community (such as a NERC data centre, or the EC-funded Zenodo repository, which exists to preserve EC-funded research but is freely open to all), a University of other research organisation, or a general-purpose data sharing service (such as figshare or Dryad).

A data repository must meet certain criteria to qualify as a **trustworthy data repository**. It must be a sustainable service that is committed to preserving and maintaining access to data in the long term. A trustworthy data repository will:

- Actively preserve data, e.g. replicating, maintaining bit integrity, migrating to preservation formats;
- Publish metadata to established standards and enable online discovery;
- Assign persistent unique identifiers (e.g. DOIs) to datasets;
- Issue a licence notice for the data, making the terms of use and attribution requirements clear;
- Manages access to data so that they can be used by other people.

You can search for data repositories at https://www.re3data.org/.

Are some kinds of data more valuable than others? Why?

Yes, data value can be considered in various ways.

Data collected from environmental observational are by their nature unique to time and place and cannot be reproduced. If we did not have a data record stretching back centuries (different kinds of weather records, for example) and even millennia (for example, in the information extracted from ice core samples), we would have no understanding of climate change. As we continue to generate observational data, we must continue to ensure their indefinite preservation as part of the historical record.

In comparison, experimental data (including simulation outputs) are the results of reproducible processes, so data can be recreated if lost. Moreover, as science progresses better experiments are devised, instruments are refined, and results may be superseded: for all these reasons experimental data may have a short life, and their value rapidly decline toward zero (although this isn't always the case).

Different stages of data are also more valuable than others. Clearly the raw data are most valuable – if these are lost, there is no recovery other than to repeat the data collection activity (if this is an option). By comparison, processed data, e.g. after cleaning and analysis, should be reproducible if processing methods and instruments are sufficiently well documented, and so may have less absolute value.

But cost and usability are also factors of data value: to re-collect experimental data may be a costly exercise, as also may be to re-process raw data, if the processing is laborious and intensive. Therefore there is often a strong case for retaining data in both raw and processed forms, e.g. you may collect observational data from various sources which are then processed by labour and computation-intensive means into a format suitable for input into a model. There is a clear case for retaining both the raw data and the data in the model-ready format.

Should I preserve all of my data?

You are unlikely to need to preserve all the data you will 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 can be destroyed. 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. 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.

Does my institution have a research data policy?

Most research-intensive universities have research data policies, and most will expect you to manage your data responsibly, and preserve and make accessible data that substantiate your research findings.

If you don't know where to look you should search for 'research data' or research data management' on your University website. If you are still struggling, contact the research support team in your Library or your research office.

I've found some data in a data centre that I want to use. Am I allowed to do this? Are there any restrictions on what I can do?

Data centres exist to preserve and provide access to data, so it's a fair bet that you will be allowed to use the data. But use may be subject to certain licence conditions or terms of use. In the metadata record for the dataset in the data centre catalogue you may find an 'Access to data', 'Licence' or 'Terms of use' section, with information about permitted uses and any requirements you must observe (e.g. often a preferred citation is specified).

In many cases data made available through data centres will be issued under an **open licence**, such as an Open Government Licence or a Creative Commons licence. These are standard licences for distribution of scholarly works, which enable materials to be used in a relatively unrestricted way: often the material can be freely copied, modified and distributed with no other requirement than that the original data creator(s)/source be attributed.

Some variations of open licences introduce further restrictions, for example that data may be used for non-commercial purposes only, or that derived data must be published under the same licence terms.

But not all data are made available under standard open licences, in which case terms of use should be carefully studied prior to undertaking research, so that you know what you can and cannot do with the data and any data you derive from them. Remember that all datasets are the intellectual property of a data owner, and can only be use in accordance with the permission of the rights-holder. If you cannot find a licence or terms of use statement, you should ask the data provider.

I've written some code to process and analyse my data. What should I do with it when I've finished my research?

You should preserve it and make it accessible to others, as you would do with your data. The code is an integral part of your research process, as much as might be a laboratory instrument and the method by which it was used. If someone wished to reproduce your research methods to verify your results, they would need the code, as well as information about its implementation, so you should also think about how you will document the code and make it accessible.

The code should be clearly formatted and commented, and be accompanied by any relevant documentation, so that someone else could make sense of it and run it.

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 in the data repository.

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. An appropriate option here would be to host the code in a public repository, for example using the GitHub platform. Such a service can be used to publish software releases, develop code under version control, track bugs, and provide user support.

If you will be releasing substantial software that has potential for ongoing development and adoption by multiple users, you must take care to ensure the source code is suitably licensed, both to protect your own interests and to ensure others are aware of the terms under which they can use, modify and redistribute the code. You should consider releasing the source code under an Open Source licence, such as MIT, Apache or GNU GPL. For information about Open Source licences, visit https://choosealicense.com/ and https://opensource.org/.