University of Reading

Preparing for data archiving

What to do before you deposit your data in a data repository

Contents

Introduction

This is a guide to preparing for the deposit of a dataset in a data repository for long-term preservation and sharing.

Datasets can be valuable research outputs, and you should put as much care into preparing a dataset as you would any other research output.

There are two principal reasons for sharing research data:

- so that others can verify or replicate the results reported in published research findings;
- so that data can be re-used by others for different purposes, which might include, *inter alia*, the investigation of new research questions, evidence-based policy-making, the development of products and services, and teaching and learning uses.

We advise you keep in mind the FAIR Data Principles when preparing for the preservation and sharing of a research dataset: data should **Findable**, **Accessible**, **Interoperable** and **Re-usable**. The data repository is the primary vehicle by means of which data can be made findable and accessible; how you prepare, format, organise, document and license your dataset will have a significant bearing on its interoperability and re-usability. This guide will help you to maximise the 'FAIRness' of your data.

You should invest time in the work of preparation before you deposit your dataset in a repository. A deposit in a data repository can be delayed and in some cases rejected if, for example, you have not correctly identified intellectual property rights in a dataset and obtained relevant permissions, or established an ethical basis for sharing of data collected from participants, or anonymised a dataset where this is required.

Most repositories also have collection policies and minimum requirements that deposits must meet in order to be accepted, so it is important you understand these requirements and are confident you will meet them before you try to deposit your dataset.

This guide is applicable to anyone preparing to deposit data in any data repository. It is illustrated with reference to the University's Research Data Archive and our requirements for depositors. We provide a summary of these requirements in our 8-point Data Deposit Checklist for depositors in the Archive.

Contact

The Research Data Service can be contacted at researchdata@reading.ac.uk / 0118 378 6161.

If you are planning to deposit your data in the University's Research Data Archive, you can book a pre-deposit consultation with us, or get in touch with any questions you have.

1. Define the dataset

Before you deposit your dataset, you will need to define it. It is important to identify the contents of your dataset, as these will also determine what preparation is necessary. A

systematic value assessment of your data can help you to make an informed decision about what to preserve and share. We provide a set of appraisal criteria in our Data Selection and Appraisal Checklist.

These are some considerations to bear in mind:

- University policy requires the preservation and sharing of primary data collected or created in the course of research 'that substantiate published research findings'. Data sharing expectations do not apply to secondary data, i.e. data belonging to other parties that are used in the research. It is not your responsibility to preserve and provide access to these data, and you may not have permission to do these things. Where third-party materials are integral to a dataset, it may be possible to incorporate these into your deposit, subject to permission and in accordance with any stipulated licensing terms;
- Data exist in different manifestations in the course of project: raw data as initially captured; data that have been cleaned to remove noise; data that have been converted from one file format to another, to facilitate processing, or to enhance long-term preservation, accessibility and interoperability when preparing for archiving; derived data products that may have been created during analysis or to facilitate re-use by others, such as tables of mean values or other calculated variables and data visualisations. Not all of these manifestations necessarily need to be included the dataset that is archived. For guidance on deciding what data to deposit, consult the Data selection web page;
- You may need to have an idea of the volume of data you wish to deposit if it is likely to be substantial, as some repositories have limits or may be more or less suited to handling larger and more complex datasets (see the next section);
- A dataset will also consist of documentation and metadata. Documentation might include a readme or user guide, a data dictionary or codebook, copies of data collection instruments, such as questionnaires; experimental protocols; and information sheets and sample consent forms;
- A dataset may include code that has been written to process or analyse data, e.g. Python code written to clean raw data or merge data from separate sources; R scripts written to execute statistical analysis and data visualisations;
- Research software source code may need to be preserved. Source code can be deposited as a standalone item in the Research Data Archive and we offer a number of Open Source licence options. The popular code hosting platform GitHub also has feature that allows a snapshot of a code repository to be archived to the Zenodo digital repository so that it can be preserved and assigned a DOI for citation purposes.

2. Identify the repository and check its requirements

You will need to identify the repository that you intend to use, and check its collection policy or guidance on depositing data to ascertain that your deposit will be eligible.

We provide guidance on choosing a data repository. Some key considerations are highlighted here.

Types of data repository

There are three main categories of repository:

- **Subject and data type**: NERC data centres, the UK Data Service ReShare repository and the Archaeology Data Service are broad subject data repositories. Examples of data type repositories are the databases of the European Bioinformatics Institute (different kinds of genetic data), the Cambridge Structural Database (crystal structures), and OpenNeuro (neuroimaging data). As a general rule, a suitable repository in this category should be your first choice: it will be a community resource and will provide a high level of curation to facilitate interoperability and re-use. In some cases, your choice may be dictated by a funder's requirements: for example, NERC expects its researchers to archive data with the relevant NERC data centre, unless another repository is more suitable;
- **Institutional**: many research-intensive universities now host their own data repositories. There will not always be a suitable external repository for your subject or data type (this kind of repository is mostly found in some areas of the sciences); where this is the case, an institutional service is a good next choice. It will accept any type of data, and will usually provide a good level of data curation. Data submitted to the Research Data Archive will undergo quality and risk management checks. We will enhance descriptive metadata, and advise on organisation, formatting and documentation of data.
- **General-purpose data sharing services**: examples include Zenodo and figshare. These are public free services that can be used to share any type of content, including datasets. They fulfil the basic data repository functions, but they are essentially self-publishing services, and do not provide quality checks or risk management.

Subject, data type and institutional data repositories will have collection policies and eligibility criteria that must be met for a deposit to be accepted. Some examples of collections policies are: CEDA Archive Acquisition and Collections Policy; Archaeology Data Service Collections Policy; European Nucleotide Archive Content description; and the Research Data Archive Collection Policy.

Content and metadata requirements

Subject and data type repositories may have content and metadata requirements, and require or recommend submission of data in specific formats. For example:

- The UK Data Service documents data to the international Data Documentation Initiative (DDI) standard for social science data, which is used to capture structured information about the study, data files and variables, and to assign keywords from the Humanities and Social Science Electronic Thesaurus (HASSET). It also provides a list of recommended and acceptable file formats;

- The CEDA Archive provides guidance on common file formats for the data types it accepts;
- OpenNeuro requires files to be named and organised in conformity with the Brain Imaging Data Structure (BIDS) standard;
- The European Nucleotide Archive requires the submission of sequence data in standard CRAM, BAM or FASTQ formats and deprecates various formats specific to the manufacturers of sequencing instruments.

Dataset volume constraints

Repositories may be more or less suited to handling large data deposits, and some may place limitations on the volume of data that can be deposited. For example:

- The CEDA Archive is built to handle TB-scale datasets that are often generated in the weather and climate modelling fields, and it provides alternative file upload methods suitable for larger datasets;
- The Archaeology Data Services charges for the deposit of datasets, based on the number and size of files, and provides a costing calculator;
- The University's Research Data Archive will accept deposits up to 20 GB in size free of charge; larger deposits may be made at a cost, and are not encouraged, as the Archive is not a large dataset service;
- The Zenodo general-purpose data sharing service will accept deposits up to 50 GB in size through its standard web-based deposit process; it will also accept larger deposits on an ad hoc basis.

Most repositories do not charge for deposits. Exceptions noted above are the Archaeology Data Service and the Research Data Archive for deposits larger than 20 GB. In many cases data archiving costs can be met out of grant funding, provided they have been included in the budget.

Repositories for higher-risk anonymised and identifiable participant data

In most cases data collected from participants can be anonymised and shared openly. Anonymisation of quantitative experimental and observational data may be fairly straightforward and sharing as open data unproblematic.

But some data, even if they have been anonymised, may be considered to present a higher risk of re-identification and harm, for example through linkage to private databases. Qualitative data, such as interviews, may be more difficult to anonymise, because they may contain indirect identifiers which, taken together, or with other information to which they can be linked, would be sufficient to identify a person. Some types of data, such as genetic and other biomedical data may be inherently identifiable.

Some repositories can manage higher-risk anonymised data and data containing identifiable information under a controlled access procedure. A prospective data user may need to fulfil certain conditions to be granted access to the data under a special licence or data sharing agreement. Archiving of and controlled sharing of personal data

must also be in line with the processing purpose(s) notified to the data subject at recruitment. These are examples of repositories that offer controlled access options:

- The UK Data Service ReShare repository has a 'safeguarded data' option suitable for higher-risk anonymised datasets. Prospective data users must be registered with the UK Data Service and will be required to sign a special licence agreement undertaking to maintain the confidentiality of the information supplied;
- The European Genome-phenome Archive is a service for the preservation and sharing of identifiable genetic, phenotypic, and clinical research data;
- The Research Data Archive provides a restricted dataset option. Restricted datasets will be securely preserved by the University and made accessible only to authorised researchers affiliated to a research organisation, subject to approval by a Data Access Committee (including the PI of the original study or a nominated representative), and under the terms of a Data Access Agreement between the University and the recipient organisation.

3. Check your consents

If data have been collected from living persons, check that you have properly-documented consent for data sharing. It is acceptable to disclose data obtained from human subjects without consent if the data have been anonymised, but it is good practice to inform participants of your intention to do this. It is not acceptable to disclose even anonymised data if in your consent procedure you stated that the data would not be disclosed, or would be destroyed at a given time. Identifiable data can be disclosed under a controlled access procedure as described in the previous section, providing that participants have consented to participate in the study on the understanding that data would be shared in this way. The University provides a sample consent form including statements suitable for open data sharing and sharing of data subject to safeguards.

If you are depositing data collected from participants in the Research Data Archive, **you will be required to submit your information sheet(s) and sample consent form(s) used in participant recruitment with your data files**, so that we can confirm you have a basis for data sharing. These documents will be stored in the dataset as Documentation files. Access to them will be restricted, meaning they will not be available for users download.

Consent is best obtained before data collection, but it may be possible to obtain consent retrospectively. In some cases, for example in qualitative research involving the collection of sensitive information, a process consent model may be appropriate. This might involve, for example, obtaining consent prior to conducting an interview, and later seeking approval of the anonymised transcript prior to archiving.

4. Identify dataset creators

It is important to understand who is a creator of your dataset – as well as who is not – because intellectual property rights and permission to distribute the data will be associated with its creators (see the next section). Datasets may be the work of many

hands, and it is not always easy to clearly distinguish its creators from other people who contributed to the work of the project.

According to the Copyright, Designs and Patents Act 1988 a database is 'a collection of independent works, data or other materials which – (a) are arranged in a systematic or methodical way, and (b) are individually accessible by electronic or other means'. It is 'the selection or arrangement of the contents of the database' that constitutes the creative act which attracts copyright.

Therefore, **creators are those who have had a direct creative role in the selection and arrangement of data in the dataset**. This is not the same as being involved in the design of the research or in the original data collection. In most cases, a project PI or student supervisor will not be a creator of the dataset, unless they had a direct authorial hand in its creation. Technicians, contractors and others involved in the collection of data are not usually creators of a dataset, unless they had creative input into the selection and arrangement of the data points.

Authors as defined under the Copyright, Designs and Patents Act 1988 also have a number of moral rights, including the right to be identified as the author of a work, and the right not to have a work falsely attributed to them as an author. For this reason there is also a legal obligation to identify the creators of a dataset accurately.

If you wish to acknowledge the input of contributors to a dataset, for example those who undertook data collection, you can do so in the dataset documentation while distinguishing their role from that of a creator of the dataset. The Research Data Archive has a separate field on the metadata record where contributors can be named and their role specified.

5. Identify the rights-holders

You must clearly identify rights-holders, because your authorisation to archive the dataset depends on their permission. Remember that by archiving data you are also distributing them, and doing this without the authorisation of the rights-holder will be a breach of copyright law. Establishing or obtaining permission is considered in a separate section below.

When you deposit data in the Research Data Archive you will be asked to confirm that you are:

- acting as an employee of the University and […] the University is allowed to deposit the material; or
- acting privately as the owner of any copyright and associated intellectual property rights in the data; or
- otherwise lawfully entitled to distribute the data on behalf of the rights-owner(s).

Whatever repository you deposit your data into, you should always check that you can satisfy one of these criteria.

Owners of intellectual property rights (IPR) in the data will be associated with the creators of the dataset (see the preceding section).

In general, an employer will own IPR created by its employees: the University is ordinarily the rights-holder in IP created by members of staff. Research contracts generally allow ownership of 'arising IP' (i.e. created under the contract) to reside with the originating institution.

Students registered with the University own the IP they create by default, but this may not be the case if they are funded under a third-party sponsorship agreement (excluding public funders such as Research Councils, which do not assign student IP to other parties), or if they have assigned their IP to the University. Usually the third party is a company, e.g. Syngenta, Waitrose, but it may also be a Government-funded agency, such as the Met Office, or a charity that is not primarily a research funder, such as the Donkey Sanctuary. A sponsorship agreement will include Intellectual Property clauses stating which party has ownership of arising IP. Ownership of IP created by a student at another institution will be subject to that institution's IP policy and any relevant agreements.

If a dataset has multiple creators, it may also have multiple rights-holders, which may include the University, students in their own right, and collaborating and partner organisations. We provide a web page with guidance on intellectual property rights and research data.

You may need to investigate any applicable research contracts or student sponsorship agreements to establish what parties hold rights in a dataset. Students and/or their supervisors should have copies of any contracts relating to their research programmes. If you need to locate a copy of a contract, contact your Contracts Manager. Contact us if you have questions about a research contract.

Where datasets incorporate secondary data, the owners of these data will also have the right to determine whether and if so on what terms their data are distributed by you.

6. Decide your licensing preferences

IP should always be published under a licence, so that ownership of the IP and terms of use are clear to others. In accordance with the University's Research Data Management Policy you are expected to share data under an **open licence** wherever possible. The most widely used open data licences are the Creative Commons Attribution (CC BY) licence, which permits re-use of the data provided proper attribution is made, and the Creative Commons Zero Public Domain Dedication (CC0), a waiver of all rights in the work.

In order to license the data you must be the data owner or authorised to assign a license on behalf of the data owner, so the choice of licence may be subject to the permission of other parties (see the next section). For example: a third-party co-creator with commercial interests may request the application of a non-commercial licence; if the dataset incorporates third-party materials these may be made available on an 'All Rights Reserved' basis.

Data held under a controlled access policy (such as UK Data Service safeguarded data and restricted datasets in the Research Data Archive) will be made available under special licence terms. The Data Access Agreement for restricted datasets deposited in the Research Data Archive allows data to be used, subject to authorisation, in confidence for non-commercial research and learning purposes only. The Agreement will be made between the University and the organisation to which the authorised user is affiliated.

Generally the choice of licence belongs with the rights-holders and the licence will be assigned by the depositor on their behalf. Some repositories may mandate the use of certain licences: for example, the Dryad Digital Repository and OpenNeuro both require datasets deposited with them to made available under a CC0 licence. Other repositories may encourage the use of specific licences: the Archaeology Data Service encourages the use of the CC BY or the Open Government Licence, and strongly discourages the use of licences with non-commercial clauses, but leaves the choice to the depositor's discretion.

As a general rule we recommend you use the Creative Commons Attribution 4.0 licence for open data, and this is the default applied to uploaded files in the Research Data Archive. More restrictive licences should only be used if there is a justification for doing so, for example, to protect commercial or other confidential interests. The Archive offers a pick list of licence options, including all flavours of Creative Commons, the Open Government Licence, a variety of Open Source licences for software code, 'All Rights Reserved' and the option to include a licence specific to the dataset ('Licence included with item').

When you deposit data in the Research Data Archive you will be required to choose a licence option for each file you upload. This means a dataset may consist of data files made available under different licences. This may be necessary for some datasets, for example, where they incorporate third-party materials, or where a restricted dataset includes both restricted and public files. But as a general rule you should aim to licence all dataset content under the same licence to facilitate re-use.

We provide a web page with guidance on licensing data. Guidance on licence options for software can be found in our Guide to publishing research software.

7. Obtain permissions if necessary

You must ensure that you have permission to archive and distribute the dataset from: the creators; the rights-holders; parties with contractual rights regarding publication of research outputs; secondary data owners.

Creators

Authors as defined under the Copyright, Designs and Patents Act 1988 have a number of moral rights, including the right to be identified as the author of a work, and the right not to have a work falsely attributed to them as an author. You must therefore ensure that dataset is archived with the knowledge and permission of its creators.

Rights-holders

Where the employer is a University or publicly-funded research organisation, permission to publish the data can be inferred from their policy position on research data: in the case of universities, this is to promote the sharing of data supporting research outputs as openly as possible, while recognising that in some cases restrictions may need to be placed on data sharing for legal, ethical or commercial reasons. If you are an employee of the University, you have a delegated authority to make the data as open as possible. Other parties, including students, studentship sponsors and commercial research partners, will need to give written consent to publication of the dataset.

Whoever owns the data, you should consider carefully whether they can be made publicly available without compromising intellectual property interests, such as any ongoing or intended patent registration, licensing or other commercial activities. Publication of data may automatically invalidate certain intellectual property rights or otherwise cause detriment to the owners of those rights. Contact us if you have any concerns in this area.

Parties to contracts

Research and studentship agreements 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. The standard notice period is 30 days.

Secondary data owners

If your dataset incorporates IP from existing sources, you may need to seek permission to distribute the material. If material has been obtained from a public resource such as a website or a data repository, check the source for any terms of use or licence information. Government and research data are often made available under open licences permitting redistribution, providing acknowledgement of the source is given. But this should not be assumed – the terms of use must always be checked. If you cannot find any information in the published source, or the data have been obtained from a non-public source, you may need to contact the data owner directly. Permission to distribute secondary data may come with licensing conditions. We provide a web page with information about using secondary data.

Seeking permission

To seek permission, you should write to the party concerned, and request permission in writing. Research contracts and sponsorship agreements will nominate a contact person for each party, to whom notices under the contract can be directed. Look for Notice clauses, which are usually towards the end of the contract. Student sponsorship agreements usually provide details of both a legal officer and a supervisor at the sponsoring party. Notices of intention to publish data can be sent to the sponsoring supervisor by email.

When contacting other parties for permission to archive and distribute data, it is important to identify the data unambiguously, and to be clear how they will be made available, and on what terms they will be licensed. While you should always seek to licence the dataset on the most open terms, other parties may legitimately require more restrictive licensing. For example, a commercial partner may not be willing to distribute a dataset under terms that permit re-use for commercial purposes.

8. Form the dataset

Archiving data is not as straightforward as transferring the files from your active storage location into a data repository. Your data will need to be tidied up, put into order, and documented. When forming the dataset, consider the following:

- Identify all the files that will compose the dataset. These might include: raw data files (in the initial collection format); processed data files (e.g. cleaned data; raw data saved to another format; statistical analyses and visualisations); programming code (e.g. analysis scripts); and documentation.
- Ensure the data are stored in suitable formats for preservation and compliance with your chosen repository's requirements. Guidance is provided on suitable file formats for preservation, including a list of formats recommended for deposit in the Research Data Archive. Although common proprietary formats, such as Microsoft Excel and Adobe PDF, are acceptable, you may wish to convert files to open formats, such as CSV for spreadsheets, and .txt or PDF/A for documentation. It is better to preserve image and multimedia files in lossless formats at their highest resolution where quality and resolution are important, but compressed formats may be more suitable for usability.
- If you intend to upload files in specialist or rarely-used formats, your documentation file should provide information about the file format, the equipment/software used to generate it (with relevant version information), and any software required to render the files.
- Make sure your data files are well-formed and readable. Poorly-presented data are harder to read, more likely to contain errors, and inspire less trust. Check the data for errors. Apply consistent style and formatting, and spellcheck your text. Format computer code legibly and include comments to explain what the code is doing. Ensure relevant information is clearly presented in data files, e.g. variable names and definitions, units of measurement, missing value codes. Present actual values; avoid encoded content, such as formulae and conditional formatting in spreadsheets. The Wellcome Trust provides useful guidance on preparing spreadsheet data.
- Redact the data as necessary. Data collected from research participants may need to be anonymised. There is guidance on anonymisation provided by the UK Data Service. Other kinds of information may also need to be removed or obscured, such as commercially-confidential information, locations of endangered species, etc. Link-coded data, where data records are identified by a unique code which is linked to identifiable participant information held in a separate table, are

in data protection law still personal data. They are pseudonymised, not anonymised. For a dataset to be anonymised, and suitable for sharing as open data, you must remove any means of linking data records to identifiable participants, e.g. by destroying all documented records of the link, or by replacing linked IDs in the dataset with unlinked IDs.

- If the dataset is composed of multiple files, make sure they are organised in a logical fashion. You can use file names to sort a list of files. If you are depositing data in the Research Data Archive, you can preserve a folder structure by saving the folders to an archive file format such as Zip or tar.gz.
- Use appropriate and consistent file names, which are descriptive of the file contents, formatted without spaces or special characters, and not longer than 32 characters. Guidance on file naming is provided on the Organising your data web page.
- Check the sizes of the dataset and individual files and make sure they do not exceed any size limitations specified by your chosen data repository. The Research Data Archive allows the deposit of datasets up to 20 GB free of charge and recommends that individual files be no larger than 4 GB. If you have a large dataset and/or a large number of files, it may be advisable to use an archive format to package/compress the files. (You may need to check that your chosen repository accepts deposits in archive file formats; the Research Data Archive does.) Zip and tar.gz are good choices, as they provide lossless compression.
- You could ask a colleague to review your dataset. A pair of eyes unfamiliar with the data may spot mistakes and things you have overlooked. Remember that the people reading your data will have not have your experience of the research context.

9. Prepare the documentation

Every dataset should have at least a basic readme file. Datasets deposited in the Research Data Archive must include a documentation file. This should include the following:

- citation metadata for the dataset (creators, title, publication year, identifier/URL);
- identification of the rights-holder(s) with a licence statement;
- a brief description of the dataset. This might include summary information about what and how much data were collected, the research context in which they were collected, the purpose for which they were collected, and the instruments and methods used;
- information about the project in which data were collected, with any external funding details;
- a description of the contents of the dataset, e.g. as a file listing;
- key interpretative information, e.g. a listing and definition of variables and units used, such as a codebook or data dictionary;
- details of the methods and instruments used to collect, process and analyse the data, and relevant supporting information, such as analysis scripts;

- references to any secondary data sources used;
- references to related publications. If a publication is in process, as much information as possible should be provided to enable identification of the published item, e.g. authors, provisional title, journal (if known), year and status (in preparation/under review, in press).

For deposits in the University's Research Data Archive, a README template text file is provided, which can be used to record basic documentation. Documentation can be saved in .txt, PDF, Word or another text format as preferred.