Research Engagement, Library

DATA PREPARATION GUIDE
University of Reading Research Data Archive

Contents

Introduction

Before you deposit data in the University of Reading Research Data Archive you will need to prepare and organise your material, and make a number of decisions. As you prepare your Dataset, bear in mind the two principle reasons for publishing data:

- So that independent researchers can validate/replicate the results reported in published research findings;
- So that the data can be re-used by other researchers for new research purposes.

For guidance on deciding what data to deposit, consult the Data selection web page.

Before you start to deposit a new Dataset or create a Collection, use this guide to make sure you have everything prepared. The main requirements covered here are summarised our 8-point Data Preparation Checklist.

If you do not submit a valid Dataset or fail to provide all the information required, we will need to contact you to complete the deposit process before the Dataset can be published.

Submissions that are not within the scope of the Archive Collection Policy will be rejected.

To ensure your Dataset is eligible for deposit and that your submission meets the criteria for acceptance, check the Archive Collection Policy and Submission Policy.

If you are creating a Metadata Record for a Dataset or Collection held in an external service or offline, but are not depositing any data files, not all of the requirements stated below will apply.

If you require advice or assistance, contact the Research Data Manager at r.m.darby@reading.ac.uk / researchdata@reading.ac.uk or call 0118 378 6161.

Check your permissions and rights

- Do you or the University have the rights to make the data available?
- Have you obtained permission from any other rights-holders and is it clear what you can and cannot do with the data?
- Can you provide references with appropriate attribution to any source data?
- Can you make the data available without compromising any intellectual property interests, for example, patent registration or commercial exploitation?
- Are you aware of, and are you comfortable with, the rights you are passing on to the Archive/University of Reading?

When you deposit the data into the Archive you will be asked to confirm that you are: acting as an employee of the University and that the University is allowed to deposit the material; or that you are acting privately as the owner of any copyright and associated intellectual property rights in the data; or that you are otherwise lawfully entitled to distribute the data on behalf of the rights-owner(s). If the Archive receives notification that published data violate any third-party rights, or the University has reasonable grounds to believe that it may do so, the relevant data will be immediately removed from public view pending completion of an investigation, in accordance with the Archive Takedown Policy.

If you are an employee of the University, data created by you in the course of your employment belongs to the University, unless any contracts or collaboration agreements stipulate otherwise. As an employee of the University you are entitled to distribute the data created by you or your colleagues in the course of research by depositing them in the Research Data Archive or other suitable data centre or repository, in accordance with University policies and subject at all times to any third party contract terms or rights underpinning the creation, ownership and publication of the data.

If you are a student at the University, by default you own the IP in any data you have collected or generated, but you should make sure that you have not shared or assigned IP elsewhere as part of any agreement with the University or any third party.

If other parties hold rights in the data you will need to obtain permission from them prior to deposit of the data. This permission may qualify how you can dispose of the data and license the data. Where you have used data belonging third parties, you should be prepared to provide appropriate attribution by means of a citation or reference that can be included in the Metadata Record.

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 distribution or commercial activities. Publication of data may automatically invalidate certain intellectual property rights or otherwise cause detriment to the owners of those rights. Please contact the Archive to discuss any concerns in this area.

Lastly, make sure you are aware of the rights to administer the data that you will grant to the Archive through the Deposit Agreement and the Archive's Data, Metadata and Preservation Policies.

Group your files into Datasets

- Are you depositing an individual Dataset?
- Are you depositing multiple Datasets or items?

Consider how best to package the files that constitute your Dataset(s). A Dataset is a set of files or a database constituting a discrete and clearly defined set of data and/or the means of generating data related to a research activity or project, or supporting one or more research publications. Large or multi-phase projects, or research that has given rise to multiple publications, might be more suitably decomposed into multiple Datasets, which can be grouped together as Collections.

When assembling the files for your Dataset, make sure that the Dataset is complete, and includes both data and relevant documentation. Documentation is necessary to record the nature of data collection and processing. Every Dataset must include at least one documentation file.

Prepare your data

- What is the total volume of your Dataset and do you need permission to deposit the data?
- Are your files in the most suitable formats for preservation and sharing?
- Have you provided information about non-standard or specialist file formats?
- Are your files appropriately named and organised?
- How can your files be most efficiently uploaded to the Archive?
- Are any of your files too large for web upload?

Check the total volume of the files you propose to deposit. If it is **greater than 20 GB**, contact researchdata@reading.ac.uk to discuss your deposit. Deposits exceeding 20 GB in volume may be subject to a charge; unauthorised deposits above this limit will be rejected.

Ensure files are in the most appropriate format for long-term preservation and sharing. A list of Recommended File Formats is provided.

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 text documents. It is better to preserve image and multimedia files in lossless formats at their highest resolution, but compressed formats may be more suitable for usability.

The same data may be deposited in different formats. For example, you might want to include an image in your dataset as both a high-resolution TIFF file, and a compressed JPEG, which is easier to view and download. The UK Data Service provides guidance on recommended file formats for preservation and sharing.

If you intend to upload files in specialist or rarely-used formats, a documentation file should be deposited describing the exceptional file format and providing information about any software required to render the files.

Use consistent, meaningful file names wherever possible, so that potential users can understand at a glance what a file is and how it sits in the Dataset as a whole before downloading it. You can also use file naming to sort your files and determine the order in which they display. The UK Data Service provides advice on file naming.

Files can be uploaded to the Archive singly or in zip bundles. If you have a large number of files consider grouping them for upload to the Archive according to their content or file format. Keep a folder structure for the files in your zip bundle.

The Archive currently has a size limit of 10 GB per file for web upload. If you have files larger than 10 GB which you would like to deposit, please contact us for advice on the options available. Note that some browsers have a file size limit of 4 GB on file uploads. If you have trouble uploading files or have large amounts of data contact us to arrange alternative means of deposit.

Prepare your documentation

- Have you provided sufficient information for another user to understand, validate and use your data?
- Have you fully documented how the data were collected and processed?

The Dataset should contain enough information for another user to be able to understand the data, and how they were produced and transformed in the research activity.

Every Dataset must include at least one documentation file. At the minimum you should include a README file that provides a basic Dataset description, e.g. a file listing and some description of the Dataset. A README file template is available for you to use.

You may need to include documentation regarding data collection and processing materials and methods, instrument user guides and technical reports, lab notebook records, survey questionnaires, data dictionaries and codebooks, information sheets, sample consent forms and interview instructions.

Check consent and confidentiality

- Have you sufficiently anonymised your data, or obtained explicit consent from any data subjects whose identity could be revealed from the data?

Data collected from human subjects can be made available if they have been effectively anonymised. It is your responsibility to ensure that your data are made available in compliance with the Data Protection Act and the obligations of ethical research conduct. The UK Data Service provides guidance on data anonymisation.

Personally-identifying data (such as interviews, images and audiovisual recordings) may be deposited in the Archive if you have obtained explicit consent from the subjects for publication of the data. If you do deposit personal data, you should clearly indicate this to the Archive Administrators in the **Comments and suggestions** field in the Metadata Record. You will be required to provide the Archive with evidence of the consent obtained from the research participants. The UK Data Service provides guidance on obtaining consent for data sharing.

You can if necessary make sensitive data available through the Archive on a Restricted Access basis, and allow only authorised users to download the data in compliance with specified terms of use. This option should only be used if you have a legitimate reason for restricting access to the data.

The University provides a standard Licence for Restricted Data that allows data to be used, subject to authorisation, in confidence for non-commercial research and learning purposes only. This Licence can be selected when you upload your data files. You may also specify and/or upload your own licence terms.

Decide your embargo and access control requirements

- Does your Dataset need to be placed under embargo before it can be made available?
- Do you wish to require readers to register before they can download your data files?
- Do you wish to make data files available on a Restricted Access basis?

You can place files in your Dataset under embargo for a specified period, so that the Metadata Record and file names are visible, but the file contents cannot be viewed or downloaded.

An embargo enables you to deposit your data when the research is still fresh on your mind, for example while preparing for publication, but to delay release of the data, e.g. until the research findings have been published or any intellectual property has been exploited.

Most public funders allow researchers a period of exclusive use of their data, but data should be made available within a reasonable time after they have been collected, and no later than publication of your research findings, unless there is good reason for longer-term restrictions on access.

Data files should not be embargoed for longer than 12 months from the date of deposit, unless you have good reason to restrict access for longer. If you set a longer embargo period for any files you must enter a justification for the attention of the Archive Administrator in the **Comments and suggestions** field in the Metadata Record.

You will also need to select the Access settings for each file or zip bundle:

- **Open:** anyone can download the file. This is the default and should be used unless there is a valid reason to apply a different access setting;
- **Registered users:** only staff and research students with a University account can download the file. This option should only be used on a temporary basis to allow colleagues access to a pre-final version of the Dataset; the final version of a Dataset should be either **Open** or **Restricted**;
- **Restricted:** access to the file can only be granted by an Authorising Agent. This option should only be selected when there is a valid reason to do so, for example based on the sensitive nature of data.

Decide how to license your data

- Have you understood how your data will be licensed for use?

You will be required to choose a licence option for each file and zip bundle you upload to the Archive. In order to license the data you must be the data owner or authorised to assign a license on behalf of the data owner.

For open data you should use Creative Commons licences. The University recommends you use Creative Commons Attribution 4.0, which permits re-use of the data provided that proper attribution is made. Other licence options are available, including the Creative Commons Attribution Non-Commercial licence. For more guidance on Creative Commons licences go to the Creative Commons website.

For software code standard Open Source licence options are available, including Apache 2.0, MIT, GNU GPL 3.0 and LGPL 3.0. Guidance on licence options for software can be found in our Guide to publishing research software, and at choosealicense.com.

For Restricted files use the University of Reading Licence for Restricted Data, which allows data to be used, subject to authorisation, in confidence for non-commercial research and learning purposes only. Alternatively, you can upload your own licence.

The terms on which you can license your data may depend on the interests of any third parties who have rights in the data. But you are strongly encouraged to use the recommended standard licences unless there is a justification for using an alternative option.

Have your research project information to hand

You will be required to provide information about your research project, including details of any external funding and references for any related publications or other resources, such as project websites.