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, , 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.

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,

systematic value assessment of your data can help

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

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.

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.

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

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;

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.et documentation

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.

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.

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.

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.

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).