Preparing for data archiving

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

Contents

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 repos itory

There are three main categories of repository:

- **Subject and data type**: NERC data centres, the UK D

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-

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 D G D W D E D V H L V a collection of µ 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 µthe 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.

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

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 you U F K R V H Q U H S R V L W R U\ 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 shoul7 Tmm7 0 0 1 456.82 367golucu6(e)]TJ10 g0 G230.57 (r)13(e)-3(

vou document

- 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