An Introduction to Managing Research Data

Author	University of Bristol Research Data Service
Date	1 August 2013
Version	3
Notes	
URI	data.bris.ac.uk
IPR	Copyright © 2013 University of Bristol

Within the Research Councils UK (RCUK) statement on Expectations for Societal and Economic Impact[1], there is a clear directive that those who receive funding are expected to "*take responsibility for the curation, management and exploitation of data for future use".*

Introduction

Many funding bodies now require that their funding recipients create plans at the bidding stage for managing their research data, storing and preserving it in the long term and sharing some, or all of that data once the research is completed.

Academic publishers too, and increasingly calling for scientific claims to be underpinned by publically accessible data which can be checked by anyone.

Managing data is an essential area of responsible research conduct. As members of the University, all researchers have individual responsibility for appropriately managing the research data that they create. This document introduces you to the concept of research data and research data management, explains what constitutes research data and how it differs from other types of information. The document is particularly appropriate for postgraduate students and early career researchers who would like to learn more about managing their research data prior to submitting a funding application.

What is research data?

Research data is information that is created, manipulated or cited in the course of funded or unfunded research. Newly created research data is often arranged or formatted in such a way as to make it suitable for communication, interpretation and processing, often by a computer. Advances in technology are transforming research. Although the growth of data has been most apparent in large scale research projects, small projects will also generate important research data.

For the purposes of this document research data shall be regarded as that which is created in a digital form (born digital), converted to a digital form (digitised) or significantly altered within the digital realm, during the course of research activities. Research data does not typically include data generated in the course of personal activities, desktop or mailbox backups, or data produced by non-research activities such as University administration or teaching. The same information may be research data one point in time, but not at another time, depending on whether information is being used for research purposes.

[1] RCUK expectations for societal and economic impact. Research Councils UK (RCUK), www.rcuk.ac.uk/documents/innovation/expectationssei.pdf

For example, a scanned photographic image of an old municipal building in a historical archive is an archived image in an image bank. When used by a researcher to study the history of a city, the photographic image becomes research data.

Research data comes in an endless variety of formats and may include any of the following:

- Documents (text, MS Word), spreadsheets
- Scanned laboratory notebooks, field notebooks, diaries
- Online questionnaires, transcripts or surveys
- Digital audio or video recordings
- Transcribed test responses
- Database contents
- Digital models, algorithms or scripts
- Contents of an application (input, output, log files for analysis software, simulation software, schemas)
- Documented methodologies and workflows
- Records of standard operating procedures and protocols

The following *research records* may also need to be managed during and after the life of a project but are not generally considered to be research data:

- Correspondence (electronic mail and paper correspondence)
- Grant applications
- Ethics applications
- Research progress reports
- Research publications
- Master lists
- Internal social media communications such as blogs, wikis etc

Why manage research data?

Research data management concerns the organisation and curation of data, from its entry into the research cycle through to the dissemination and archiving of valuable results and the checking of those results by third parties. It also includes activities that ensure research data is ‘fit for purpose’. In the course of your research you are likely to create a significant amount of data and this can quickly become disorganised, lost, out-of-date or meaningless. By actively managing your research data in an appropriate way you will ensure that funding and regulatory requirements are met; transparency and accountability are maintained, data remains accurate, reliable and complete; research data keeps its integrity and research results may be replicated; duplication of effort is kept to a minimum; data security is enhanced and the risk of data loss is minimised. Above all good research data management will enhance the long-term value of your research, allow it to be better shared amongst a wider research community, increasing its visibility and impact.

Ethical considerations

All major research funders recognise that some research data cannot be shared as it is particularly sensitive. Most data however, can be shared, at least on part, if certain precautions are taken. Commonly used precautions are data anonymisation, seeking permissions to share data *at the time of data collection* and restricting access via the need for a signed secondary user agreement.

The immense benefits of sharing data within medicine and the biological and social sciences are a powerful argument for doing so. Your plans to share research data should reflect your ethical plan, if you have one.

Copyright and IRP

Research funders also recognise the challenges faced by researchers who wish to commercially exploit the research data they have created. This is broadly encouraged by funders who typically grant a defined period of 'exclusive data access' after a project has ended, during which researchers can exploit any commercial potential the research data may have.

However, it is generally a requirement of funding that, after this period has ended, research data will be made 'freely available' (i.e. without charge).

Research data formats

Some disciplines have well defined guidelines on which file formats to use and how to describe or catalogue a dataset to best support secondary data users. Other disciplines are still developing these. A major barrier to data sharing is the widespread use of non-standard formats or rapidly obsolete commercial formats. Many research funders do not dictate the formats you should use but do ask that you justify your decisions to use certain data formats over others. When selecting a data format to use, your own research needs must come first.

However, if you find need to use a non-standard or rarely used format, you should consider converting your data to a more widely re-usable format once your own data analysis is complete.

The re-usable format you select should be as accessible as possible to as many people as possible. In order to make use of any data a number of digital technologies must be available; these are known as technological 'dependencies'. These may be fairly common technologies such as; a desktop PC, the Windows 7 operating system and Adobe Reader 9 software. Alternatively, the technology required to access data might be rare and hard to acquire or even unique. You should address this challenge as far as it is possible, by minimising the number of unnecessary technological dependencies involved in using your data.

You should also favour 'open' technologies rather than proprietary ones whenever possible. Proprietary technologies are owned by a commercial company or group of companies. Commercial pressures may lead to the withdrawal of a particular piece of hard or software and its replacement with a new version. Open technologies are supported by community of users and so do not have the same commercial vulnerabilities.

If you're unsure which file formats to use the UK Data Archive publish a list of recommended deposit formats[2]. These formats are appropriate for many non-specialised uses.

[2] UK Data Archive File Formats Table, www.data-archive.ac.uk/create-manage/format/formats-table

Describing research data

Without an accurate description of what a dataset actually consists of, why it was created or collected and what secondary users are permitted to do with it, the value of research data is greatly reduced. Wherever a dataset is made available it should be accompanied by a useful description.

Metadata is ‘data about data’ and is information (or ‘cataloguing information’) that enables data users to find and or use a dataset. A description of a dataset is often kept in a separate, dedicated database or spreadsheet. As with data formats, some disciplines offer precise guidelines on how to describe a dataset. While, for other disciplines, no such guidelines yet exist.

If you find you have to take a pragmatic approach to describing data, it may help to imagine a secondary data user attempting to make sense of your data in your absence, after your project has concluded. If presented with only the data itself a secondary user may well be faced with the difficult task of ‘unpicking’ it. How will they make sense of your file and folder naming conventions? What extra information would they need to make the maximum use of your data? How were new datasets derived from raw data?

Ensuring the quality of data

Quality should be considered whenever data is created or altered, for instance at the time of data collection, data entry or digitisation. Funders often ask for information about the procedures you will carry out to ensure data quality is maintained, such as putting time aside to validate data or entering values into pre-prepared databases or templates.

Sharing research data

Most major funding bodies require non-confidential research data to be not only retained but also actively shared at the end of a funded project. This will either be done via deposit of data into an established national ‘data centre’ (supported by a research funding body), via a discipline-specific data repository (usually supported by a number of different organisations) or via an institutionally research data repository (the University of Bristol has the data.bris service).

The option for researchers to provide access directly is also sometimes acceptable to funders. However, data may have to be retained and shared for many years (potentially even in perpetuity) as a condition of funding and during this period the contact details and responsibilities of individual researchers can be expected to change. These factors usually make this option impractical.