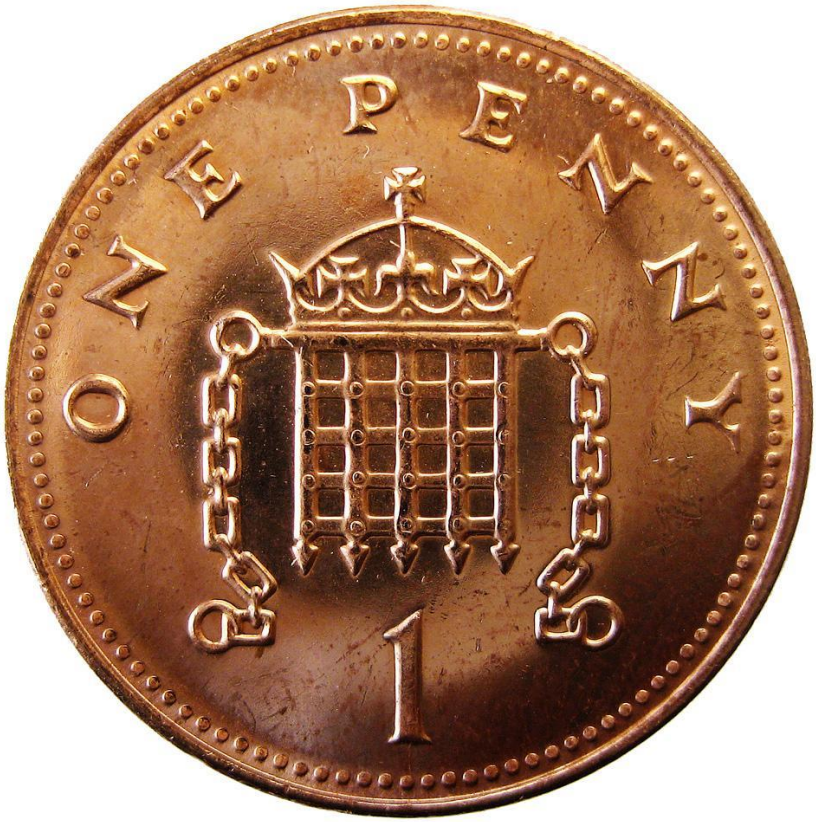
Data Management Planning

Anticipating the costs of research data management

Version 1.6 June 2021



University of Bristol

Research Data Service

Image: Adapted from One Penny, Adrian Rowbotham, Flickr, CC BY-NC 2.0

University of BRISTOL

INTRODUCTION

Many funding bodies now require that award recipients manage their research data, storing and preserving it in the long term and sharing some if not all of that data once the research is completed. Academic publishers too, are increasingly calling for scientific claims to be underpinned by publicly accessible data which can be checked by anyone.

Successful data management always has a financial cost, even if this is only covering a small fraction of a single researcher's time, spent organising files within folders. Often though, when added together, the time spent performing several different aspects of data management (e.g. transcription, data anonymisation and converting between different file types) is more substantial, and so the costs are more significant.

Some of these costs are an integral part of 'doing the research'; others will be incurred only because, for example, a funder requires research data to be made publicly available.

In terms of covering costs, some actions and resources will have no direct cost to a research project (data storage under 5TB, for example, is provided for free to University of Bristol Principal Investigators (PIs)) while other actions (such as carrying out data quality control) will have a direct cost and should be included in research funding applications. This guide is intended to prompt researchers to consider research data management as an important, and potentially costly, research activity, and to enable them to prepare funding applications accordingly.

THE ROLE OF THE DATA MANAGER

Before assessing potential costs for individual data management activities, Principal Investigators planning large-scale projects are asked to consider whether or not employing a Data Manager is appropriate. A career profile for a Data Manager is available from the DaMSSI project.[1] For large (or complex) research activities, employing a Data Manager has several advantages:

- A Data Manager can reduce project costs by ensuring data is sorted and processed as it is generated. This is often cheaper and more effective than leaving data processing until the end of a project;
- A Data Manger can ensure data benchmarks and standards are being met, helping to harmonise the efforts of individuals responsible for generating data and allowing procedural errors to be spotted early on;
- A Data Manager can collect metadata and document methodology while a project is underway. This reduces the risk of important information going unrecorded and being lost;
- Most major funders now take research data very seriously. Dedicating even a fractional role to research data management helps demonstrate

[1] "Data management in perspective: the career profile of data managers", http://www.dcc.ac.uk/sites/default/files/documents/IDCC11/data%20manager%20final.pdf

to a funder that the applicant also appreciates the importance.

For more modest research activities, research data management duties will be assigned to other members of the project team (such as a Research Assistant) or handled directly by the Principal Investigator.

THE RESEARCH LIFECYCLE

A research dataset is typically:

1. Created
2. Used (by the individual or team responsible for its creation)
3. Curated (prior to publication)
4. Published
5. Preserved
6. Re-used (by parties not involved in the creation of the dataset)

Each of these phases has associated costs and even where no 'new' data is to be generated (for example when a freely available, public dataset will be used), researchers are encouraged to consider the potential costs of the other phases. Each phase is discussed below.

Creating new data

The creation of new data is invariably the most expensive phase in this lifecycle, so it isn't difficult to see why research funders are keen that this step is avoided wherever possible; hence their promotion of data re-use.

For many researchers, the creation of data is a familiar activity and the time and resources involved are well understood. Activities resulting in new data (for example time spent analysing samples in the lab or conducting recorded interviews) often form a major part of a Research Assistant's routine duties.

The non-staff costs of creating new data are usually covered directly by a research funder. For instance, some of the University's scientific facilities make a charge to researchers, which should be included within funding applications and so passed on to research funders.

Activity	**Anticipated cost**
Gaining consent for data sharing (for research involving human participants)	Low cost if carried out before new data is created
Data Description (e.g. data in spreadsheet are clearly marked with value and variable labels)	Low cost if carried out as part of data creation
Data Cleaning (e.g. ensuring only relevant data is present or only controlled terms are used)	Low cost if carried out as part of data creation
Documentation (e.g. of methodology, analysis and quality control procedures)	Low cost if carried out as part of data creation
Digitisation	Low cost if simple and small scale (e.g. scanning of a few dozen paper documents).

	Moderate to high cost if complex (e.g. large scale or accurate text mark-up is required)
Organisation of data (e.g. versioning, file naming & folder structure)	Low cost if well planned and then carried out as data is created
Anonymisation	Low cost if well planned and carried out as data is created

Making use of data

Money spent here often supports research efforts immensely. If a robust and fit-for-purpose dataset is created only minimal modification will be required later on, when the same data is shared. Applying logical structures and quality control measures to the data will ensure it sufficiently supports published research claims.

Yet processes such as data standardisation, shifting between different file formats, undertaking quality control procedures and ensuring data is appropriately stored during a research project can have significant costs, some of which are direct costs to the project.

A wide range of software applications capable of carrying out data processing tasks is provided at no direct cost by IT Services.[2] Other, more specialist software can either be purchased or leased, both of which will have associated direct costs.

Grant funding occasionally covers the creation of new, project-software, for example a mobile device app. Research IT[3] can advise on costs involved and are able to undertake some software development work.

High Performance Computing resources are free at the point of use, but costs can also be included in funding applications if guaranteed access is required, or a large request for resources is involved.[4]

The Research Data Storage Facility provides each lead researcher with 5TB of storage without charge, but further data storage can be purchased at a cost of £750 per TB.[5]

Activities in this phase can include:

Activity	**Anticipated cost**
Formatting data (e.g. converting files between different formats)	Low cost if target format is directly equivalent to original format. Can be moderate cost if manual checking is needed (e.g. changing

2 https://uob.sharepoint.com/sites/itservices/SitePages/software.aspx

3 https://uob.sharepoint.com/sites/itservices/SitePages/teaching-learning-and-research.aspx

4 https://www.bristol.ac.uk/acrc/high-performance-computing/

5 https://www.bristol.ac.uk/acrc/research-data-storage-facility/

	between database formats)
Transcription	Moderate cost, depending on quantity. Assume 4-8 hours of transcription per recorded hour
Data storage & security	No cost if under 5TB and RDSF is used

Curating data prior to publication

If data has been carefully planned, created and processed up to this point, only minor modifications will be required in order to publish it, and costs will be correspondingly low. But, if a dataset containing personal information has not been anonymised before the close of a project, weeks of staff time may be required to carry out this activity. It is strongly recommended that these situations are anticipated and so avoided, as money spent at this stage has minimal benefit for the research project which is all but complete. Ideally, this stage would consist of simply processing any recently created data and creating subsets of pre-prepared data in order to underpin specific research claims.

Activities in this phase can include:

Activity	**Anticipated cost**
Data Description (e.g. data in spreadsheet are marked with value and variable labels)	Can be high cost if not carried out as part of data creation
Data cleaning (e.g. ensuring only relevant data is present or only controlled terms are used)	Can be moderate cost if not carried out as part of data creation
Documentation (e.g. of methodology, analysis and quality control procedures)	Can be moderate to high cost if not carried out as part of data creation
Organisation of data (e.g. versioning, file naming & folder structure)	Can be moderate cost if not carried out as part of data creation
Anonymisation	Can be high cost if not carried out as part of data creation
Gaining consent for data sharing (for research involving human participants)	Can be very high cost (or impossible) if not carried out before data is created

The publication and preservation of data to support reuse

Many research funders require data to be made available for a number for years, after a project has ended. However, they are typically unwilling to directly fund ongoing data preservation.

In the vast majority of cases, it is expected that researchers will resolve this issue by making use of one or more research data repository services. The costs of ongoing publication and preservation then become the responsibility of that service. Subject-based, national and institutional data repositories exist[6].

Researchers should be aware that some data repository services charge the data depositor (this covers the cost of adding new data to the repository). Deposit should be made before a project ends and this charge should be included within your funding application.

You can deposit up to 100GB data per project for free in the University's research data repository, data.bris[7]. Costs for other repositories will vary.

Activities in this phase can include:

Activity	Anticipated cost
Data sharing	Low or no cost is using a data repository. Otherwise a significant ongoing cost (usually extending beyond project lifespan)
Repository or discipline specific metadata (e.g. to INSPIRE or DDI standards)	Low cost, often done at dataset level, as part of deposit process

[6] For a searchable list of data repositories, see Re3data.org: https://www.re3data.org/

SUMMARY

Consider whether your project would benefit from the assistance of a Data Manager

- Establish whether any additional costs will be involved in the creation of your data which are not already covered elsewhere in your funding application
- Try to organise and document your data as you go along to avoid the need for any staffing costs associated with cleaning data at the end of the project
- Identify whether you will require any additional software, computing or storage solutions, and speak to the relevant University departments so quotes can be included in your costing
- Always consider the potential costs associated with sharing and preserving your data and use a repository or data centre where possible. Ensure you include deposit charges in your funding application

Please contact the Research Data Service for more information: data-bris@bristol.ac.uk

[7] https://data.bris.ac.uk/data/