Anticipating the costs of research data management

Research Data Management Service

University of Glasgow.

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 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 5GB, for example, is provided for free in Enlighten: Research Data to University of Glasgow researchers) 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 employing a Data Manager may be appropriate. A career profile for a Data Manager is available from the DaMSSI project. 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.

University of Glasgow
Research Information Management Service
December 2021

- 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 to a funder that the applicant also appreciates the importance.

For more modest research activities, research data management duties will be assigned to members of the project team 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. 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 researcher'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 collected or 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 collection or creation
Data cleaning (e.g. ensuring only relevant data are present or only controlled terms are used)	Low cost if carried out as part of data collection or creation
Documentation (e.g. methodology, analysis and quality control procedures)	Low cost if carried out as part of data collection or 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 if accurate text mark-up is required).
Organisation of data (e.g., versioning, file naming and folder structure)	Low cost if well planned and then carried out as part of data collection or creation
Anonymisation	Low cost if well planned and then carried out as part of data collection or creation

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 when the same data is shared. Applying logical structures and quality control measures to the data will ensure it sufficiently supports published research claims.

Processes such as data standardisation, converting 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. Other, more specialist software can either be purchased or leased, both of which will have associated direct costs.

High Performance Computing resources are also available through IT services for researchers who need additional computational resource.

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 between database formats).
Transcription	Moderate costs, depending on quantity. Assume 4-8 hours of transcription per recorded hour.
Data storage and security	No cost for most routine active data storage on University systems (OneDrive, OwnCloud and networked filestores). If you anticipate that your data are going to be particularly large, it may be worth speaking to your local IT contact to check that provision can be made. No cost for repository storage if Enlighten: Research Data is used and dataset us under 5Gb in size. For datasets over this size, costs can be viewed here.

Commented [MD1]: Link to costs on webpages

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 collection or creation
Data cleaning (e.g. ensuring only relevant data are present or only controlled terms are used)	Can be moderate cost if not carried out as part of data collection or 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 collection or creation.
Organisation of data (e.g. versioning, file naming and folder structure)	Can be moderate cost if not carried out as part of data collection or creation.
Anonymisation	Can be high cost if not carried out as part of data collection or 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 collected or created.

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 generally 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.

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 5GB data per project for free in the University's research data repository, Enlighten: Research Data. Costs for other repositories will vary.

Activities in this phase can include:

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

Summary

Firstly, 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 teams 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 Management Team for more information: research-datamanagement@glasgow.ac.uk

Acknowledgement

This guide has been modified from the University of Bristol 'Anticipating the costs of research data management' guide. Version 1.5 (October 2020) which was shared under a CC-BY licence.

University of Glasgow
Research Data Management Team
December 2021