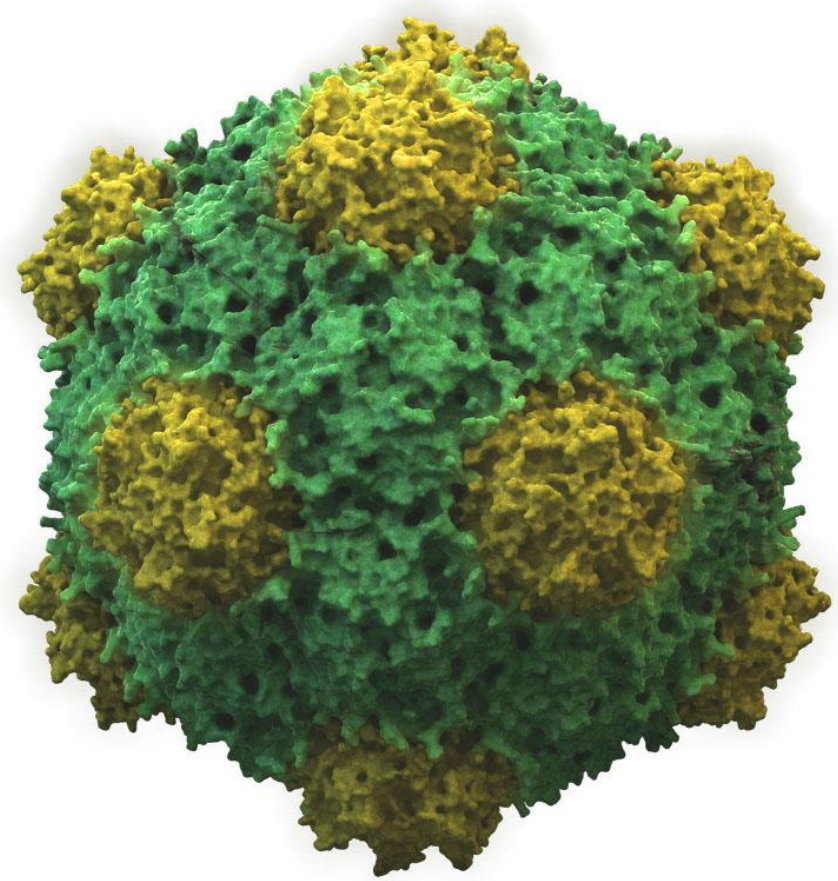
Data Management Planning

BBSRC funding applicants

Version 2.8 June 2021



University of Bristol

Research Data Service

Image: Structure of the icosahedral Cowpea mosaic virus, Wikipedia, CC-BY-SA

University of BRISTOL

SUMMARY

- Data must be released no later than the publication of findings and within three years of project completion.
- Commercial interests should not "unduly delay or prevent sharing" of data funded by BBSRC.
- Data must be available for a minimum of ten years after project end and in a form appropriate for secondary use.
- Sharing via an appropriate, established repository is expected in research areas where such repositories exist. BBSRC currently supports a number of resources that are available for the bioscience community to use [1].
- All applicants for research funding should complete a Data Management Plan as part of their research grant proposal in order to demonstrate a willingness to share data.
- Compliance with the BBSRC Data Sharing Policy is checked by BBSRC via ResearchFish.
- Certain data types are considered by the BBSRC to have an especially high re-use value.
- Funding to support the management and sharing of research data can be requested as part of the full economic cost of a research project.

INTRODUCTION

The Biotechnology and Biological Sciences Research Council (BBSRC) supports the view that publicly-funded research data is a public good, produced in the public interest and should be openly available to the maximum extent possible. The re-use of data can lead to new scientific understanding.

Holders of BBSRC grants are encouraged and expected to follow the BBSRC Policy[2] in order to practice and promote data sharing and create a scientific culture within which data sharing is embedded.

This guide is intended for BBSRC applicants who are required to submit a Data Management Plan along with their application.

Who does the BBSRC Policy apply to?

All applicants for BBSRC research grants must observe the BBSRC data sharing policy. Applicants should complete a Data Management Plan as part of their research grant proposal in order to demonstrate a willingness to share data.

Grantholders are required to record data sharing activities on ResearchFish and compliance with the submitted Data Management Plan will be monitored through this mechanism.

The BBSRC Data Sharing Policy does not currently apply to studentships or fellowships (though it is planned to extend it further). However, applicants to these funding streams are encouraged to consider the policy when making applications.

The BBSRC Data Sharing Policy also applies to institutions in receipt of core strategic grants and researchers funded by such institutions. The

[1] https://bbsrc.ukri.org/research/resources/

[2] https://bbsrc.ukri.org/about/policies-standards/data-sharing-policy/

monitoring of institutional compliance is carried out as part of the Institute Assessment Exercise.

Exceptions to data sharing

BBSRC acknowledges that in exceptional circumstances data sharing may not be possible or may not be desirable. If you believe this applies to your research, you should use your Data Management Plan as an opportunity to explain why you feel this is the case. You may feel the costs of sharing data with anticipated low re-use would make data sharing not worthwhile.

Note that only in rare circumstances will BBSRC accept that ethical considerations preclude the sharing of *all* the research data you generate. It is far more likely that some additional actions (such as anonymisation of data and seeking appropriate permissions) may be required in order to facilitate re-use of the data.

BBSRC recognises the need to protect opportunities for commercialisation of research outputs, but they state that commercial interests should not "unduly delay or prevent sharing" of data funded by BBSRC.

Research data with a high re-use value

Added value can be gained by using certain datasets for purposes other than those for which they were originally designed. Researchers in all areas are encouraged to consider data sharing where there is scientific merit in doing so. However, the following types of research data are considered by the BBSRC to have a high re-use potential:

Data from high volume experiments

Datasets containing hundreds of measurements generated in parallel from a single experimental sample (for example, omics, sequencing etc.).

Low throughput data from Long Time Series (LTS) or cumulative approaches

Standardised measurements collected at regular intervals forming a resource that can be subjected to retrospective analysis. Often this type of data is of particularly high value as it cannot be substituted or replaced.

Models generated using systems approaches

Models created using iterated systems approaches can be as important as the data they generate. They are sharable and re-usable assets with a high re-use potential. Such models should be freely available to any researchers wishing to reproduce experiments.

An appropriate repository in which to deposit models is the BioModels Database.[3] Authors are encouraged to submit models to the BioModels Database before publication of an associated paper (depositors will receive an identifier that they can use in the publication), but models will only be publicly available on the BioModels Database once the paper has been published.

[3] https://www.ebi.ac.uk/biomodels/

Costs of sharing research data

BBSRC recognises that data sharing has time and cost implications. Funding to support the management and sharing of research data (for example, staffing and physical resources such as storage and networking capability) can be requested as part of the full economic cost of a research project.

THE DATA MANAGEMENT PLAN

In order to demonstrate an intention to comply with the BBSRC Data Sharing Policy, applicants are required to complete a Data Management Plan at the time of application. An additional page of the Case for Support is allocated for this purpose. Using this page for any other purpose will result in automatic rejection of the proposal.

Your Data Management Plan will be assessed by peer reviewers separately from the rest of the proposal. However, if an inappropriate Data Management Plan is submitted, an applicant's credibility will suffer. Proposals with exceptional scientific value but a poor Data Management Plan may be offered a conditional award; alternatively, a redrafted Data Management Plan may be requested. In individual cases, BBSRC reserves the right to take a more prescriptive approach to data sharing.

It is suggested that applicants use the following headings in their Data Management Plan:

Data types

An estimate of the volume and type of data that you expect to generate (e.g. experimental measurements, models or images). BBSRC encourages researchers to outline *all* research data and *not only that which directly underpins a research publication.*

Data and metadata standards

In order to maximise re-use value, BBSRC researchers should generate and manage data using widely accepted formats and methodologies. In some disciplines these are well defined; in others, standards are still being developed.

An example of an established standard is the widespread use of models in SBML and CellML formats. The BioModels Database only accepts models in these formats in order to help to verify the reproducibility of results.

Numerous conversion tools exist (such as SBFC), which will transform models in other formats into SBML or CellML models. If you find you do need to use a non-standard technology during the course of your research, consider standardising datasets and models prior to sharing them.

Where no clear disciplinary guidelines exist concerning which formats to use, your own research needs must come first. If you find you do need to use a non-standard format, you should consider converting your data to a more widely re-usable format once your own data analysis is complete. If you're unsure which file formats to use, the UK Data Archive publishes a list of

recommended deposit formats.[4] These formats may also be appropriate for use throughout your research.

A major barrier to data sharing is the widespread use of non-standard, highly specialised file formats. In order to make use of data, a number of digital technologies must be available, which are known as technological 'dependencies'. These may be fairly common technologies such as a desktop PC, the Windows 10 operating system and Adobe Reader DC 15 software, or the technology required to access data might be rare and hard to acquire, or even unique. You should address this problem by minimising the number of technological dependencies involved in using your data as much as possible.

Where dependencies are inevitable you should favour 'open' technologies rather than proprietary ones. Proprietary technologies are owned by a vendor or group of vendors. Commercial pressures may lead to the withdrawal of a particular piece of hardware or software, in favour of a new and possibly incompatible replacement. In contrast, 'open' technologies are supported by a community of users and do not have the same commercial vulnerabilities.

Metadata is 'data about data' or 'cataloguing information' that enables data users to find and/or use a digital output. In your Data Management Plan you should briefly outline plans for documentation, both to meet your own needs (i.e. to ensure that you can find what you want, when you need it) and those of later users.

Where no clear disciplinary metadata standards exist, it may help to imagine a secondary user attempting to make sense of your output in your absence. If presented with only the data, they may be faced with the difficult task of 'unpicking' it. So, for example, how would they make sense of file and folder naming conventions? Has any special software been used in the creation of an output that must also be available in order to use it? How was secondary date derived from primary data?

Relationships

Relationships to any other datasets available for re-use.

Secondary uses

Briefly describe the re-use potential which the dataset or model will have once it is complete.

Data sharing

BBSRC recognises two broad approaches to data sharing: 1) via a third party, such as an established, online public repository or as supplementary information accompanying a journal article, and 2) by the award holder providing data on request. A combination of these two approaches may be appropriate and either of them may be subject to specific access mechanisms (such as a requirement to have a data sharing agreement) in order to protect confidential or otherwise sensitive data.

[4] UK Data Archive File Formats Table, https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

Sharing via an appropriate, established repository is expected in research areas where such repositories exist. BBSRC currently supports a number of resources that are available for the bioscience community to use.[5]

The University of Bristol has its own Research Data Repository which researchers from any discipline may wish to use. This repository can provide ongoing access to research data for extended periods of time and issue unique DOIs for deposited datasets. For smaller datasets, there is no cost. If you are planning to deposit larger datasets with the repository, a cost may be incurred. Contact the data.bris service[6] as early as possible if you believe you'll need to make use of Bristol's data repository.

Note that the responsibility for ensuring data is retained remains with the grant holder even when data is deposited with an external archive for sharing purposes. Therefore BBSRC recommends that a copy is retained locally for security purposes.

Researchers have the option of providing data directly to third parties on request. However, researchers choosing to do so should consider the BBSRC requirement to make data available for a minimum of ten years after the project ends, in a format appropriate for secondary use. Some updating of both data formats and accompanying metadata is to be expected during this period.

BBSRC suggests that any data preparation which is required before sharing (such as standardisation or quality checking) should be done within the lifetime of the funded project, in order to avoid subsequent loss of staff or motivation.

Data sharing timeframe

Some communities have established time frames for releasing data (for example, the Cambridge Structural Database[7], where a twelve-month delay between publishing the first paper on a structure and making co-ordinates public for secondary use is typical, or Metabolomics (MeT-RO)[8] where a six-month delay in publication can be requested). It is the responsibility of the applicant to reference such disciplinary guidelines in the Data Management Plan.

Where no clear community guidelines exist, BBSRC expects data to be released no later than the publication of findings and within three years of project completion.

A delay is permissible to protect intellectual property but should not prevent sharing entirely. Long term or large-scale projects may also choose to release data in waves as it is generated or as findings are published.

Primary data should be securely retained, in an accessible format, for a minimum of ten years after project completion.

[5] https://bbsrc.ukri.org/research/resources/
[6] The University of Bristol's Research Data Service *data.bris*, data.bris.ac.uk
[7] Cambridge Structural Database https://www.ccdc.cam.ac.uk/
[8] MeT-RO http://www.metabolomics.bbsrc.ac.uk/MeT-RO.htm

Proprietary data

Specify any data which will not be freely available for re-use and a brief explanation of why this is the case. For example, a co-funder may have requirements which conflict with those of the BBSRC.

Format/s

List of the formats of final datasets or models (see Data and metadata standards, above).

SAMPLE BBSRC DATA MANAGEMENT PLAN

The following is intended as an illustration of a BBSRC Data Management Plan. It is drawn from a real world BBSRC proposal prepared by the School of Biological Sciences and submitted to the BBSRC. The plan is made public with the kind permission of the applicant, Dr Christos Ioannou.

Full details of experiments and costing were covered in the wider 'Case for Support'. This document is not available. The following paragraph explains the general purpose and nature of the project:

This project is interested in how differences between individuals in groups affects both group-level outcomes (such as decision making) and individual-level outcomes (such as the time available for an animal to feed in a particular place). The project is entirely lab based, with High Definition video being used to record fish shoals from above. From this video, the position of each fish is tracked using video-tracking software, and the resulting data are time series of fish coordinates. These coordinate data provide a high temporal and spatial resolution to answer the hypotheses of interest.

DATA MANAGEMENT PLAN

Data types: In each of the experiments outlined in the Case for Support, there will be generated a time series of coordinate data for fish positions in each trial. This primary data is not of the volume obtained in high-volume experiments (such as sequencing) or long time series data; access will be provided to encourage reuse whether the trials form part of a published article or not. In our analyses, from this primary data group-level properties and/or summary statistics over the time series will be calculated. However, providing the data in the raw format of coordinate data along with metadata will allow for the most diverse range of possible uses.

From previous work, we estimate the data files to be less than 1TB over the course of the study. Still images and video will be kept locally on multiple external hard drives for data analysis and verification, and will be backed-up where possible with the remainder of the 5TB each PI is provided with free of charge by Bristol's Research Data Storage Facility. The University of Bristol Research Data Storage Facility (RDSF) provides secure, long-term storage, exclusively for research data. This £2m investment provides nightly backup of all data, with further resilience provided by three geographically distinct storage locations. A tape library is used for backup purposes and also for long-term offline data storage. The RDSF is managed by Bristol's Advanced Computing Research Centre (ACRC) which has a dedicated steering group and a rigorous data storage policy (https://www.acrc.bris.ac.uk/acrc/RDSF_policy.pdf).

Data formats and metadata standards: To facilitate the most widespread reuse, data will be provided in text (.txt) and comma delimited (.csv) formats. Metadata will accompany these data to explain what hypotheses the data were originally designed to test, how the data were collected and who collected the data, where and when. In cases where data is processed, for example by calculating average turning angle by the fish, a full description of the procedure used will be given.

Secondary uses: Our project is strongly focused on analysis of empirical data. However, a strength of recent work in collective behaviour is the synergy between theoretical and empirical work, thus our data will be provide a rich data set in inspiring modelling beyond the scope of the proposed work, provide model parameters for more realistic models, and provide data to test model predictions against. Analytic tools developed in the future can also be tested using the data provided from this study.

Data sharing: All data generated during the project will be made freely available via the University of Bristol's Research Data Repository. DOIs to these data will be provided (as part of the DataCite programme) and cited in any published articles using this data and any other data generated in the project (to allow data unrelated to any published work to be found). Data deposited into Bristol's Research Data Repository will be maintained for a minimum of 20 years. There are no security, licensing or ethical issues related to the expected data, and all data used in the project will be generated directly as a result of the project, without any pre-existing data being used.

Data sharing timeframe: Any data relevant to a published article will be made available alongside the article when published. All other data will be made available within three-years of the project completion; the PI, CCI, will take responsibility to ensure that this is carried out after the PDRAs have finished.