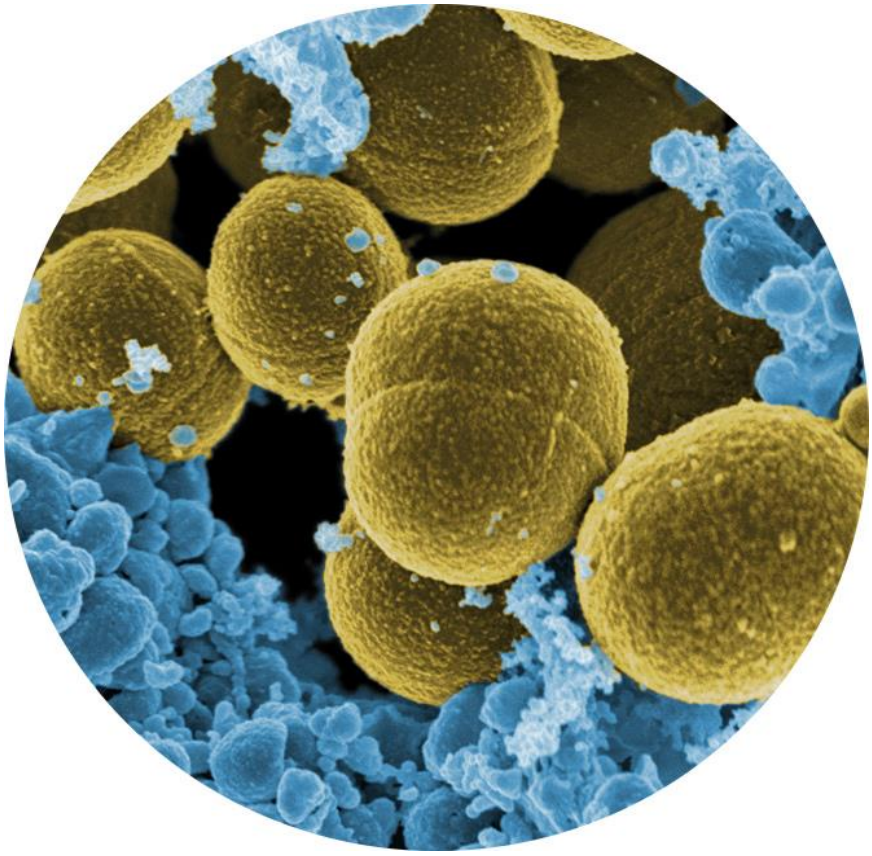
Data Management Planning

NIH funding applicants

Version 1.4 June 2021



University of Bristol

Research Data Service

Image: Staphylococcus aureus bacteria escape NIAID/RML [Public domain]

via Wikimedia Commons

University of BRISTOL

SUMMARY

The NIH expects data underpinning NIH-funded studies to be made "as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data".[1] The following points are intended as a reminder of NIH requirements:

- All final research data generated by studies wholly or partially funded by NIH should be made as accessible as possible whilst respecting participant privacy.
- Applicants requesting $500,000 or more in direct costs in any project year must include a data sharing plan in their application.
- Applications for studies generating large-scale genome data or model organisms must include a data sharing plan, regardless of the level of support sought.
- The data management plan should consist of a single brief paragraph describing how research data will be shared, or why this will not be possible. Further details, for example, costs or risks of data sharing to human subjects, should be included in the relevant sections of the application.
- NIH will cover the costs of data sharing activities if these are included in the grant application
- Certain types of data from NIH-funded research must be deposited in specific repositories; for example, data from genome-wide association studies must be deposited in the Database of Genotypes and Phenotypes.

Introduction

The US National Institutes of Health (NIH), in common with UK and European research councils, "expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers".[2] Effective data sharing can:

- Open up new lines of scientific enquiry
- Encourage new, cross-disciplinary research
- Support studies on new research methodologies
- Provide relevant real-world examples for teaching
- Enable the creation of new, combined datasets from multiple sources.

Data from NIH-supported studies must be preserved for a minimum of 3 years after the grant ends, and should be shared as soon as the main findings have been accepted for publication, or at the time of publication in the case of genomic data.

The NIH policy applies to both studies involving human participants and laboratory research not involving human subjects and is particularly important for studies generating unique data that cannot be easily replicated. Applicants seeking $500,000 or more in direct costs in any year of the proposed project period must submit a data management plan (DMP, described below) with their grant application; some funding calls may require this of all applications regardless of amount. In addition, any applications for support for the development of model organisms, or for the generation of large-scale

[1] NIH Data Sharing Policy and Implementation Guidance, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

[2] Final NIH statement on sharing research data 2003, https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

genome data, must include a data sharing plan, and must also follow the NIH Genomic Data Sharing Policy.[3] However, the general data sharing policy applies to all research wholly or partially funded by NIH even where an upfront DMP is not required.

Your DMP will not be assessed by grant application reviewers, and is not factored into the scientific merit or priority scores. NIH program staff will assess the appropriateness and adequacy of your DMP and any concerns raised by program staff must be resolved before the grant is awarded. You may include funding for data sharing activities in your grant application; if these costs are not included in the grant they may be subsequently recouped from requestors as long as they are reasonable and reflect actual costs.

Format of DMP

Your DMP should take the format of a single brief paragraph describing how your final research data will be shared, or why this will not be possible. This should be included in the 'Research Plan' section of your application and will not count towards the page limit. You may also include a brief description of the expected time frame for making data available, the method by which this will occur, the format of the final dataset and any access restrictions (and how these will be administered).

If you are requesting funds to cover data sharing, you should justify these in the budget sections of your application. Similarly, if data sharing is important to the significance of your project, for example, you are planning to create a database which will provide an important resource for the scientific community, you should describe this in the 'Relevance' section of your application.

If your research involves human subjects, any potential risks to research participants from data sharing should be discussed in the 'Protection of Human Subjects' section. Risks are explained further in the section *Barriers to data sharing*, below.

Defining 'final research data' and 'large-scale genome data'

The NIH data sharing policy defines 'final research data' as "factual material commonly accepted in the scientific community as necessary to document and support research findings".[4] This specifically excludes lab notebooks, drafts, preliminary analyses, partial datasets, and physical objects such as lab specimens. However, such resources may be of use to the wider scientific community, and the NIH encourages grant recipients to "develop patent, license, and material sharing policies with this goal in mind".[5]

Both human and non-human large-scale genome datasets are covered by the NIH data sharing policy and include "[genome-wide association studies] (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data".[6]

[3] NIH Genomic Data Sharing Policy, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

[4] NIH Data Sharing Policy and Implementation Guidance, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

[5] Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources, http://grants.nih.gov/grants/intell-property_64FR72090.pdf

[6] NIH Grants Policy, 8.2, Availability of Research Results: Publications, Intellectual Property Rights, and Sharing Research Resources, section 8.2.3.3, https://grants.nih.gov/grants/policy/nihgps/HTML5/section_8/8.2.3_sharing_research_resources.htm#Genomic

Acceptable methods for data sharing

The NIH policy is not prescriptive about data sharing methods; for many studies, any of the following will be acceptable:

- by direct request to the PI;
- via a data archive;
- via a secure data enclave (such as the UK Data Archive Secure Lab[7]);
- any combination of the above.

However, given the likely time frames and scale and complexity of data generated by studies, the first option is strongly discouraged; personal contact details are likely to change over time, and there is a considerable administrative burden in coordinating data access agreements for datasets requiring restricted access. A data archive will manage this for you and will provide you with a DOI or other unique identifier to your dataset that can be cited in any publications.

If your study generates certain types of data, you must share these via subject-specific data repositories. Non-human genomic data may be deposited in any accepted subject-specific repository, but human genomic data must be deposited in the appropriate NIH-designated data repository (i.e. one that is supported directly or collaboratively by the NIH). More specifically, human GWAS data must be deposited in the Database of Genotypes and Phenotypes, dbGaP.[8]

Barriers to data sharing

If your study involves human participants, any shared data should respect their rights and privacy and should not include any information that would permit the identification of individual participants. Similarly, if there are derived variables in your dataset that could lead to the identification of individuals these should be aggregated or otherwise altered to prevent identification. If this is not possible without undermining the utility of the dataset, you should consider applying controls to your dataset. For example, governing access using a data access agreement, which can include criteria for access and use, privacy and confidentiality standards, and prohibitions on the identification of participants, or use a data enclave to restrict access to a known, secure environment.

In all cases, you should consider data sharing when drawing up consent forms and ensure that this is reflected in participant information sheets.

Example DMPs

Sample data management plans are available from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex - see '*examples of data-sharing plans'* section.

[7] UK Data Service Secure Lab, https://www.ukdataservice.ac.uk/use-data/secure-lab

[8] dbGaP, http://www.ncbi.nlm.nih.gov/gap