

Gordon and Betty Moore Foundation (GBMF)
Data Sharing Policy and Implementation Guidance¹

January 4, 2005

CONTENTS

DATA SHARING POLICY.....	2
GOALS OF DATA SHARING.....	2
APPLICABILITY OF GBMF DATA SHARING POLICY.....	2
DATA SHARING PLAN DEVELOPMENT FOR PROPOSED GRANTS WITH GBMF.....	3
WHAT TO INCLUDE IN THE GRANT PROPOSAL.....	3
DATA SHARING STRATEGY AND IMPLEMENTATION PLAN.....	3
BUDGET JUSTIFICATION	4
FUNDS FOR DATA SHARING.....	4
GRANT PROPOSAL REVIEW CONSIDERATIONS.....	5
IMPLEMENTATION OF DATA SHARING PLANS.....	5
IMPLEMENTATION	5
TIMELINESS OF DATA SHARING.....	5
PROPRIETARY DATA.....	5
METHODS FOR DATA SHARING	6
<i>Data Sharing Agreements (and examples)</i>	6
DATA DOCUMENTATION	7
DATA ACKNOWLEDGEMENT.....	7
ACQUISITION AND CUSTOMIZATION OF DATA MANAGEMENT SYSTEMS	7
DEFINITIONS.....	9

¹ This document is based on the “NIH Data Sharing Policy and Implementation Guidance” of March 5, 2003. The following policy applies to data sharing activities under GBMF program initiatives, areas of investigation, and stand-alone grants.

DATA SHARING POLICY

It is the policy of the Gordon and Betty Moore Foundation (“Foundation” or “GBMF”) that data produced with Foundation grants and support will be freely shared and made widely available for charitable purposes, thereby enabling the frictionless-flow of data within and between fields. Data will be shared consistent with applicable laws and with full and proper attribution to the data provider.

GOALS OF DATA SHARING

Data sharing and data documentation allow scientists to expedite the translation of research results into knowledge, products, procedures and measurable outcomes related to the Foundation’s grants and initiatives.

There are many reasons to share data. Sharing data facilitates open scientific inquiry, encourages diversity of analysis and opinion; promotes new research; makes possible the testing of new or alternative hypotheses and methods of analysis; supports studies on data collection methods and measurement; facilitates the education of new researchers; enables the exploration of topics not envisioned by the initial investigators; permits the creation of new datasets for new applications when data from multiple sources are combined; and enables open inquiry about the impact of project activities.

In short, *data sharing maximizes the value of each project’s outputs*. In GBMF’s case, open access to data of known origin and quality enables program officers to conduct analysis necessary to optimize their initiative *portfolios*, (i.e., the summation and quantification of outputs from multiple complementary projects). To facilitate data sharing, investigators or organizations who are requested by the Foundation to submit a grant proposal will be expected to include a description of their data-sharing strategy and implementation plan for promoting the visibility and sharing and ensuring the maintenance of final research data for research purposes—including a release schedule—or state why data sharing is not possible. In addition, the implementation plan should include how the investigator will monitor and report on the progress of its data-sharing implementation plan, and, when necessary, modification to data-sharing strategy.

APPLICABILITY OF GBMF DATA SHARING POLICY

The GBMF Data Sharing Policy applies to:

- the sharing of research data for not-for-profit research and charitable purposes
- final research data, and the basic research, laboratory studies, field surveys, and other types of research supported by GBMF. (Note: It is especially important to share unique data that cannot be readily replicated.)
- projects that transform or link pre-existing datasets (as opposed to producing new data) (Note: If there are limitations associated with a data sharing agreement of the original data that preclude subsequent sharing, then the grantee should explain this in the grant proposal.)

GBMF also understands that data distribution regulations vary by country. Investigators collecting data in non-US countries should familiarize themselves with the policies and laws governing data sharing in the country(s) in which they plan to work and address specific limitations in their data-sharing strategy.

DATA SHARING PLAN DEVELOPMENT FOR PROPOSED GRANTS WITH GBMF

Potential grantees are required to discuss their proposed data-sharing strategy and implementation plan with GBMF Program Staff in advance (at least 4 weeks) of submitting a grant proposal to the Foundation.

Final research data are recorded factual materials commonly accepted in the scientific community and require documentation and validation. *So this data includes not only summary statistics or tables, but also all the data on which those statistics and tables are based.* For most studies, GBMF expects final research data will be a computerized dataset. For some, but not all, scientific areas, the final dataset might include both raw and derived data. In all cases data must be accompanied by documentation (e.g., metadata) describing the data and its format (see the Data Documentation section below for further information).

Given the variety of projects that GBMF supports, neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data is stipulated. What works for one field or one study may not work for others.

However, grantees must plan for data sharing, and be aware of the current state of data sharing activities and data-management best practices within their disciplines and fields. This means the grantee is expected to participate in and contribute to the creation of environments that support data sharing by seeking out relevant data sharing activities and networking with others within their discipline. These activities address areas such as:

- relevant online data repositories (e.g., Genbank) and data federations
- procedures for data documentation
- data formatting and data exchange standards
- software (or online data services) that conform to data format and exchange standards
- procedures for quantifying the demand (use) for the data (i.e., number and rate of users and records per data repository and/or data federation per year)
- procedures for the data owner, or provider of the datasets, to receive feedback about the quality of the data

Familiarity with these subjects will allow the grantee to better estimate the costs and benefits associated with their data sharing strategy and implementation plan.

WHAT TO INCLUDE IN THE GRANT PROPOSAL

Investigators should include a detailed description of how final data, and (if applicable) raw data, will be promoted, shared and maintained, or explain why data sharing and maintenance are limited or not possible.

Data Sharing Strategy and Implementation Plan

The precise content of the data sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Potential grantees that are planning to share data should describe and/or identify:

- (1) choice of data access or data distribution mechanism, e.g., name(s) of data federations
- (2) the expected schedule for data sharing
- (3) description of the potential data users
- (4) a baseline of the number of data users
- (5) a forecast of the number of data users for a minimum of 5 years

- (6) description of the potential data contributors including those beyond the primary project team
- (7) a baseline of the number of data contributors
- (8) a forecast of the number of data contributors for a minimum of 5 years
- (9) the form of the final dataset(s)
- (10) the data format standards used
- (11) the long-term data archive/service used to maintain and disseminate the data
- (12) the data documentation to be provided
- (13) whether any additional analytic tools will be needed or provided for the data users
- (14) whether a data sharing agreement will be required and, if so, a brief description of such an agreement (including the criteria for deciding who can receive the data and whether any conditions will be placed on data use)
- (15) the method of data sharing (Investigators who will share data under their own individual processes or mechanisms will describe how they will share the data and address the possible use of a data sharing agreement. The method of data sharing should also provide a plan that incorporates the use of open data-format standards, and standards-compliant applications and online data services.)
- (16) description of how the data sharing agreement(s) relate to the GBMF data sharing guidelines (e.g., complement or conflicts)
- (17) description of the procedure to monitor and report on the success or challenges related to your data sharing strategy and/or implementation
- (18) description of the procedure to track the uptake of the datasets (e.g., metrics for measuring data demand)
- (19) method for promoting the datasets across scientific and other consuming communities (e.g., government, NGO, businesses, etc.)
- (20) description of the support level the data owner or data provider will provide
- (21) description of how the quality of the datasets will be maintained over time
- (22) description of how the data will be used by target users (e.g., provide description(s) of data-use scenarios)

If the potential grantee believes that sharing some or all of the data to be developed with the Foundation's grant funds is detrimental to or otherwise incompatible with the project goals, the investigator will provide a detailed explanation supporting this position in the grant proposal.

Budget Justification

Grantees may request funds for data sharing in their proposal by addressing the relevant financial issues in the budget and budget justification sections. Some investigators have more experience than others in estimating costs associated with preparing the dataset and associated documentation and providing support to data users. As investigators gain experience with the process, their ability to estimate costs will improve but they should decide early how they will record, report, disseminate, and implement on, lessons learned. Investigators working with data archives or online data repositories and data federations can gain insights into the data-sharing process and related data preparation and costs. Investigators concerned about paying for data sharing costs at the end of their grant should make prior arrangements with appropriate data-archive service providers. Investigators facing considerable delays in the preparation of the final dataset for sharing should consult with GBMF Program Staff about how to manage the situation.

Funds for Data Sharing

GBMF recognizes that it takes time and money to prepare data for sharing. Grantees can request funds for data sharing and maintenance in their grant proposal. Investigators who incorporate data sharing in the initial design of their project may more readily and economically establish

adequate procedures for sharing with appropriate documentation and at the same time establish the mechanisms that enable proper attribution (i.e., due credit).

Grant Proposal Review Considerations

GBMF program staff will factor the proposed data sharing strategy and implementation plan into its assessment of a proposed grant. In general, this plan will be weighted as heavily as other components when evaluating a grant proposal.

IMPLEMENTATION OF DATA SHARING PLANS

Implementation

When the Grantee's Principal Investigator (PI) and the other authorized officials of the institution sign the Foundation's Grant Award Letter Agreement (GALA), they are assuring the organization will comply with this Data Sharing Policy. Additionally, when an approved grant (and associated proposal) includes a data sharing plan, *GBMF expects that the grantee will implement the final plan approved by the Foundation.*

Grantees should include any progress made with their data sharing plans in the periodic grant progress report(s) required in the GALA. Additional requirements may be articulated in the GALA depending on the subject matter and scope of the grant. In the final progress report, if not sooner, the grantee should note what steps have been taken with respect to executing the data sharing plan. In the case of noncompliance (depending on its severity and duration) GBMF can take various actions to protect its interests. For example, GBMF may make data sharing an explicit term and condition of subsequent payments.

Grantees should note that, under the GBMF Data Sharing Policy, they are required to preserve and maintain the data for three years following closeout of a grant or contract agreement, unless the GALA specifies a different time period. GBMF recognizes that grantee institutions may have additional policies and procedures regarding the custody, distribution, and required retention period for data produced under research awards. Nothing in this policy is intended to negatively impact or interfere with such policies of the grantee institution.

Timeliness of Data Sharing

Since the value of data often depends on their timeliness, data sharing should occur in a timely fashion. Timeliness is influenced by the nature of the data collected, but GBMF expects data to be released and shared no later than the acceptance for publication of the main findings from the final dataset. Data from small projects can be analyzed and submitted for publication relatively quickly. If data from larger projects are collected over several discrete time periods or phases, it is reasonable to expect that the data be released in phases as they become available or as main findings from a research phase are published. GBMF recognizes that the investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort. GBMF supports the privilege of initial investigators to benefit from first and continuing use of their data, but not to its prolonged exclusive use. The Grantee will propose a plan for the release of data from larger projects, and the final plan for release will be established by the grantee and the Foundation during the course of the grant.

Proprietary Data

Issues related to proprietary data can also arise when co-funding is provided by other donors or the for-profit sector (e.g., the pharmaceutical or biotechnology industries) with corresponding constraints on public disclosure. GBMF recognizes the need to protect patentable and other proprietary data. Any restrictions on data sharing due to co-funding arrangements should be discussed in the data sharing plan section of a proposal and will be considered by program staff. While GBMF understands that an organization's desire to exercise its intellectual property rights

may justify a need to delay disclosure of research findings, the release of data can be delayed up to 60 days, unless other arrangements are made with GBMF.

Methods for Data Sharing

There are many ways to share data, including:

- under the processes or mechanisms of the PI
- through a third party data archive
- as part of a data enclave
- through mixed-mode sharing

The method for sharing chosen will likely depend on several factors, including the state of the data, the size and complexity of the dataset, and the volume of requests anticipated. Early in the life cycle of the data investigators sharing under their own auspices may simply email the data, mail a CD with the data to the requestor, or post the data on their institutional or personal website. Although not a condition for data access, some investigators sharing under their own auspices may form, or participate in, collaborations with other investigators in order to pursue research of mutual interest. Others may simply share the data by transferring them to a data archive facility to distribute more widely to interested users, to maintain associated documentation, and to meet reporting requirements. Third party data archives are particularly attractive for investigators concerned about: a large volume of requests, the burden of adhering to common and open data standards, vetting frivolous or inappropriate requests, or providing technical assistance for users seeking help with analyses.

While data sharing under the auspices of the PI may be adequate early in the life cycle of the data, the data sharing plan should address the latter stages as well, after grant funding has been exhausted or as the project's objectives evolve over time. This will ensure ongoing preservation with like datasets, access of the data and ease the processes of aggregating data from disparate sources. Again, third party data archives/data service providers may be best suited for this purpose.

Datasets that cannot be distributed to the general public due to security considerations or third party licensing or use agreements that prohibit redistribution can be accessed through a *data enclave*. A data enclave provides a controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

Investigators may also wish to develop a "mixed mode" for data sharing that allows for more than one version of the dataset and provides different levels of access depending on the version. For example, a dataset could be made available for general use, with access to more sensitive aspects of the data restricted through the use of a data enclave.

Data Sharing Agreements (and examples)

GBMF supports the use of data sharing agreements whenever possible. Investigators who will share data under their own processes or mechanisms should consider using a *data sharing agreement* in order to make the data freely available or to impose appropriate limitations on users. Such an agreement usually indicates the criteria for data access and conditions for research use, and may incorporate privacy and confidentiality standards, as needed, to ensure data security at the recipient site and prohibit manipulation of data. Some examples of data sharing and usage agreements are:

Creative Commons

<http://creativecommons.org/licenses/by/2.0/>

<http://creativecommons.org>

Russian Longitudinal Monitoring Survey

<http://www.cpc.unc.edu/projects/rlms/data/datauseagreement.html>

GBIF Interim Data Sharing Agreement

<http://www.gbif.org/DataProviders/Agreements/DSA>

GBIF Interim Data Use Agreement

<http://www.gbif.org/DataProviders/Agreements/DUA/>

Data Documentation

Documentation is a key component of data sharing. Regardless of the mechanism used to share data, each dataset will require documentation. (Within the information science field data documentation is referred to as *metadata* or *codebooks*). Proper documentation and adherence to data formatting standards is necessary to ensure that others can use the dataset and to prevent misuse, misinterpretation, and confusion. Documentation provides information about the methodology and procedures used to collect the data, details about codes, definitions of variables, variable field locations, frequencies, and the like. The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.

Data Acknowledgement

It is customary for scientific authors to acknowledge the source of data upon which their findings or manuscript is based. Many investigators include this information in the methods or reference sections of their manuscripts. Journals generally include an acknowledgement section, in which the authors can recognize people who helped them gain access to the data. Authors using shared data should check the policies of the journal to which they plan to submit to determine where such acknowledgements should be placed in the manuscript. For example, most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database (e.g., GenBank) before publication.

Acquisition and Customization of Data Management Systems

New projects tend to spawn new systems for handling data. GBMF strongly discourages the development of new systems that are incompatible with existing systems in the field, unless the grantee presents a compelling reason to the contrary. While GBMF understands the desire to customize systems to serve a particular project, few grantees possess the specialized resources to build data collection and management systems that are scalable, durable over time, are usable outside of their project, and that maximize the potential for data sharing. Moreover, third party funding in areas related to GBMF programs has produced a number of data management systems and open data standards. The use of these systems must be considered, and any expenditure of grant funds for the development of new systems must be approved by GBMF in advance.

GBMF not only encourages the grantee to use pre-existing systems and support open data standards, but to also seek partners such as third-party data archives and specialized systems integrator, and organizations that have made developing scalable data management systems their primary business, or explain in detail why this is not possible.

If a pre-existing data management system is being considered for deployment to a broad group of users (for public access) or for access by users within a specific network, the grantee should explain how the system was designed with this broader use in mind as well as their ongoing deployment and support plan for the system.

If a grant is being sought to develop a large system or database that will serve as an important resource for the community, detailed justification is required to demonstrate why existing data

archives and services are inadequate or have significant barriers to their successful execution. In this case the grantee should not only have specialization in the specific area of science or business operations the system intends to serve but should also have specialization (or a close partnership with a specialist) in open data management systems development, integration, deployment, and support.

DEFINITIONS

Data – The dataset and its associated documentation. (See also the Data Documentation section above and the definitions of Final Research Data and Raw Data.)

Data Archive – A place where machine-readable data are acquired, manipulated, documented, and distributed to others for further analysis and consumption.

Data Enclave – A controlled, secure environment in which eligible researchers can perform analyses using restricted data resources.

Data Federation – A group of data providers working under a common charter to serve data under uniform rules that govern data access and data use.

Data Sharing Agreement – A contract between data provider and data consumers that defines terms of data access and use. These terms should be consistently applied.

Final Research Data – Final research data are recorded factual materials commonly accepted in the scientific community and require documentation and validation. This data includes not only summary statistics or tables, but also all the data on which those statistics and tables are based. For the purposes of this policy, final research data do not include laboratory notebooks; partial datasets; preliminary analyses; drafts of scientific papers; plans for future research; peer review reports; communications with colleagues; or physical objects such as gels or non-vouchered laboratory specimens.

Metadata – The information that describes the data source and the time, place, and conditions under which the data were created. Metadata informs the consumer of who, when, what, where, why, and how data were generated. Meta data allows the data to be traced to a known origin and know quality.

Non-vouchered Laboratory Specimen – Specimens that have no context of time of harvest, location of origin, and/or where the research is not concerned with a species concept.

PI – Principal Investigator

Portfolio – The summation and quantification of outputs from multiple, complementary projects.

Raw Data – Field observations, contents of project-related data study repositories, survey results, results of laboratory studies and preliminary analysis.

Restricted Data - Datasets that cannot be distributed to the general public due to confidentiality concerns, third party licensing or use agreements, national security considerations, or other issues.

Standards-Compliant Applications – Any application that embeds data handling functions (e.g., data collection, management, transfer, integration, publication, etc.) and operates on data in a manner that complies with data format and data syntax specifications produced and maintained by open, standards bodies.

Timeliness – In general, GBMF considers the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset. The Grantee will notify GBMF of any delays to the release of data beyond the publication date or other scheduled dates for the release of data developed with GBMF funds.

Unique Data – Data that cannot be readily replicated. Examples of studies producing unique data include: large surveys that are too expensive to replicate; studies of unique populations, such as native or un-contacted peoples; studies conducted at unique times, such as after a natural disaster; studies conducted over long time scales, and studies of rare phenomena.