National Institutes of Health (nih.gov): NIH-GEN DMSP (Forthcoming 2023)

Data Type

A general summary of the types and estimated amount of scientific data to be generated and/or used in the research. Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

Guidance:

NIH Guidance
The final DMS Policy defines Scientific Data as: “The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens.”

Even those scientific data not used to support a publication are considered scientific data and within the final DMS Policy’s scope. We understand that a lack of publication does not necessarily mean that the findings are null or negative; however, indicating that scientific data are defined independent of publication is sufficient to cover data underlying null or negative findings.

Additional Guidance
Research projects vary widely in the types of data produced. In this section, you will describe the categories, amounts, and degree of processing of your data.

Example Answer:

This project will produce _____ [Data type, e.g., imaging, sequencing, experimental measurements] data generated/obtained from _____ [e.g., instrument, method, survey, experiment, data repository]. Data will be collected from ___ [number] of research participants/specimens/experiments, generating ___ [number] datasets totaling approximately ___ [amount of data] in size. The following data files will be used or produced in the course of the project: _____ [list input data files, intermediate files, and final, post-processed files]. Raw data will be transformed by ____ [analysis, method] and the subsequent processed dataset used for statistical analysis. To protect research participant identities, _____ [e.g., individual, aggregated, summarized] data will be made available for sharing.

A description of which scientific data from the project will be preserved and shared.

Guidance:

NIH does not anticipate that researchers will preserve and share all scientific data generated in a study. Researchers should decide which scientific data to preserve and share based on ethical, legal, and technical factors that may affect the extent to which scientific data are preserved and shared. Provide the rationale for these decisions.

Example Answer:

Based on _____ [ethical, legal, technical] considerations, the following data produced in the course of the project will be preserved and shared: ____ [list] **OR** All data produced in the course of the project will be preserved and shared.

A brief listing of the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

Example Answer:

To facilitate interpretation of the data, _____ [e.g., metadata, documentation, protocols, data collection instruments] will be shared and associated with the relevant datasets.

Related Tools, Software and/or Code

An indication of whether specialized tools are needed to access or manipulate shared scientific data to support replication or reuse, and name(s) of the needed tool(s) and software.

Guidance:

Additional Guidance

The file formats in which data are saved in a digital format can be divided into two general categories.

- Proprietary - The specification of the data encoding format is not released or restricted in some way. Proprietary formats can only be easily opened and manipulated by particular software tools.
- Open - The specification of the data encoding format which can be used and implemented by anyone. Open formats can often be easily opened and manipulated by a large number of software tools.

Example Answer:

If no specialized tools are needed to access or manipulate the data:

_____ [Data type - Imaging data, survey data, etc] data will be made available in _____ [csv, txt, dicom, etc] format and will not require the use of specialized tools to be accessed or manipulated.

If specialized tools are needed to access or manipulate the data:

_____ [Data type] data will be made available in _____ format, which requires the use of specialized tools, such as _____ [include list of tools] to be accessed and manipulated.

If applicable, specify how needed tools can be accessed, (e.g., open source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.

Example Answer:

- The ____ tool, which can be used to ____ is available free of charge through ____ [source name]
- The ____ tool, which can be used to ____ is available for a fee of ____ through ____ [source name].
- Custom tools to ____ will be/have been developed by the research team.
    - Requests for these tools should be directed to ____ [include details of members of the research team].
    - These tools will be shared openly via ____.

Standards

An indication of what standards will be applied to the scientific data and associated metadata (i.e., data formats, data dictionaries, data identifiers, definitions, unique identifiers, and other data documentation).

Guidance:

NIH Guidance

While many scientific fields have developed and adopted common data standards, others have not. In such cases, the Plan may indicate that no consensus data standards exist for the scientific data and metadata to be generated, preserved, and shared.

Additional Guidance

A *standard* specifies how exactly data and related materials should be stored, organized, and described. In the context of research data, the term typically refers to the use of specific and well-defined formats, schemas, vocabularies, and ontologies in the description and organization of data. However, for researchers within a community where more formal standards have not been well established, it can also be interpreted more broadly to refer to the adoption of the same (or similar) data management-related activities or strategies by different researchers and across different projects.

It is possible that your work will employ multiple formal standards or a mix of formal standards and other data management strategies. You should be as specific as possible when describing the standards used for each type of data included in your proposal.

Example Answer:

To facilitate their efficient use, all of our data and materials will be structured and described using the following standards:

If there are formal data standards for some/all of the data:

Whenever possible, we will use _____ [common data elements, standardized survey instruments, etc] to structure and organize our data.

Our ____ data will be structured and described using the ____ standard, which has been widely adopted in the ____ community. [Add additional information about this standard, if applicable - e.g.

implementation in data repositories, utility in combining/reusing datasets]

If there are not formal standards:

Formal standards for ____ data have not yet been widely adopted. However, our data and other materials will be structured and described according to best practices.

Data will be stored in common and open formats, such as ____ for our ____ data. Information needed to make use of this data [e.g. the meaning of variable names, codes, information about missing data, other metadata etc] will be recorded in ____ [data dictionaries/codebooks] that will be accessible to the research team and will subsequently be shared alongside final datasets.

Information about our research process, including the details of our analysis pipeline will be maintained contemporaneously, using ____ [lab notebooks, protocols, etc]. This information will be accessible to all members of the research team and will be shared alongside our data.

Data Preservation, Access, and Associated Timelines

The name of the repository(ies) where scientific data and metadata arising from the project will be archived.

Guidance:

NIH Guidance

NIH has provided additional information to assist in selecting suitable repositories for scientific data resulting from funded research: NOT-OD-21-016.

Selecting a Data Repository

1. For some programs and types of data, NIH and/or Institute, Center, Office (ICO) policy(ies) and Funding Opportunity Announcements (FOAs) identify particular data repositories (or sets of repositories) to be used to preserve and share data. For data generated from research subject to such policies or funded under such FOAs, researchers should use the designated data repository(ies).

1. For data generated from research for which no data repository is specified by NIH or the NIH ICO (as described above), researchers are encouraged to select a data repository that is appropriate for the data generated from the research project and is in accordance with the desired characteristics, taking into consideration the following guidance:

1. Primary consideration should be given to data repositories that are discipline or data-type specific to support effective data discovery and reuse. NIH makes a list of such data repositories available (see Open Domain-Specific Data Sharing Repositories).
2. If no appropriate discipline or data-type specific repository is available, researchers should consider a variety of other potentially suitable data sharing options:
    1. 1. Small datasets (up to 2 GB in size) may be included as supplementary material to accompany articles submitted to PubMed Central (see PMC policies).
        2. Data repositories, including generalist repositories (see Generalist Repositories) or institutional repositories, that make data available to the larger research community, institutions, or the broader public.
        3. Large datasets may benefit from cloud-based data repositories for data access, preservation, and sharing.

Example Answer:

All dataset(s) that can be shared will be deposited in _____ [Add appropriate NIH-supported data repositories]OR _____ [Add appropriate subject or disease repositories]

Sample Language for Dryad Data Repository

Dataset(s) resulting from this research will be shared via the generalist repository Dryad, which provides metadata, persistent identifiers (i.e., DOIs), and long-term access. Dryad is the institutional data repository supported by the University of California and all data is shared under a CC0 waiver, which makes the dataset(s) publicly available. Data will be made available as soon as possible or at the time of associated publication. Dryad datasets are backed up to Merritt, the UC's CoreTrustSeal-certified digital repository, for long-term storage and accessibility. Procedures in place to ensure dataset preservation include storage of data files in multiple geographic locations, regular audits for fixity and authenticity, and succession plans in the event of repository closure.

How the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

Example Answer:

The _____ [Insert repository name] provides metadata, persistent identifiers (i.e., insert whether DOI, handles, other), and long-term access. This repository is supported by _____[Insert funder/organization] and dataset(s) are available under a _____ [Insert license information] **OR** through a request process _____ [Insert information about request process].

When the scientific data will be made available to other users (i.e., the larger research community, institutions, and/or the broader public) and for how long.

Guidance:

NIH Guidance

NIH encourages scientific data be shared as soon as possible, and no later than time of an associated publication or end of the performance period, whichever comes first. Researchers are encouraged to consider relevant requirements and expectations (e.g., data repository policies, award record retention requirements, journal policies) as guidance for the minimum time frame scientific data should be made available. NIH encourages researchers to make scientific data available for as long as they anticipate it being useful for the larger research community, institutions, and/or the broader public. Identify any differences in timelines for different subsets of scientific data to be shared.

Example Answer:

Data will be made available as soon as possible or at the time of associated publication.

Access, Distribution, or Reuse Considerations

Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to:

- Informed consent (e.g., disease-specific limitations, particular communities' concerns).
- Privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures) consistent with applicable federal, Tribal, state, and local laws, regulations, and policies.

Guidance:

Additional Guidance

Certain kinds of data, especially human subjects data, require extra preparation before they can be shared to ensure participant privacy. In this section, you will describe your approach to preparing human subjects data for sharing and note any additional restrictions or policies that will impact access to your data.

Whether access to scientific data derived from humans will be controlled (i.e., made available by a data repository only after approval).

- Any restrictions imposed by federal, Tribal, or state laws, regulations, or policies, or existing or anticipated agreements (e.g., with third party funders, with partners, with Health Insurance Portability and Accountability Act (HIPAA) covered entities that provide Protected Health Information under a data use agreement, through licensing limitations attached to materials needed to conduct the research).
- Any other considerations that may limit the extent of data sharing.

Guidance:

Additional Guidance

Certain kinds of data, especially human subjects data, require extra preparation before they can be shared to ensure participant privacy. In this section, you will describe your approach to preparing human subjects data for sharing and note any additional restrictions or policies that will impact access to your data. If you are working with human subjects you should also describe how you will address data management and sharing in your informed consent process. You will also need to describe your methods for ensuring privacy and confidentiality, including how you will de-identify your data. If you have decided that a controlled access repository (where researchers must apply to access data) is a better fit for your data than an open repository, you should describe the repository's access procedures. Finally, if there are any other laws, policies, or existing agreements that impact your ability to share your data they should be described here.

Example Answer:

For researchers working with human subjects data

In order to ensure participant consent for data sharing, IRB paperwork and informed consent documents will include language describing plans for data management and sharing data,

describing the motivation for sharing, and explaining that personal identifying information will be removed.

To protect participant privacy and confidentiality, shared data will be de-identified using the _____ method. [Describe de-identification method, noting any other applicable laws or policies such as HIPAA].

For researchers selecting controlled access repositories

Given the sensitive nature of the dataset, de-identified human subjects data will be made available in _____ data repository, which restricts access to the data to qualified investigators with an appropriate research question who sign a data use agreement. [Describe data repository access methods and security measures].

Oversight of Data Management and Sharing

Indicate how compliance with the Plan will be monitored and managed, frequency of oversight, and by whom (e.g., titles, roles).

Guidance:

NIH Guidance

This section should address titles and roles overseeing data management and sharing, within the investigator team or as key personnel.

Personnel costs required to perform the types of data management and sharing activities are allowable. Examples of costs may include time and effort for data curation processes; local specialized infrastructure (only those not covered by institutional F&A costs); or fees for preserving and sharing data. Reasonable, allowable costs for management and sharing may be included in NIH budget requests. Funds for these activities must be spent during the performance period, even for scientific data and metadata preserved and shared beyond the award period. See NIH's supplementary guidance on allowable costs for data management and sharing

Additional Guidance

List the roles responsible for data capture, metadata production, data quality, storage and backup, data archiving, and data sharing. Include the name (if available), title, affiliation, and ORCIDs where possible.

If this is a collaborative project across institutions, explain how data management tasks will be addressed across partners.

Identify which individual (or role) will be responsible for implementing, updating, and revising the DMSP.

Explain how the necessary resources (for example personnel time) to prepare the data for sharing/preservation have been budgeted. Consider and justify any resources needed to adhere to the DMP. These may include curating data and developing documentation, infrastructure necessary to provide local management and preservation, and data deposit fees.

Example Answer:

The following individuals [or just the position titles if unknown] will be responsible for data collection, management, storage, retention, and dissemination of project data, including updating and revising the Data Management and Sharing Plan when necessary.

- Name, Position Title, Host Institution, ORCID, email

Sample Language for budgeting requirements

This project includes the following costs associated with data management and sharing.

For data curation and the development of related documentation, the project is requesting $_____. These funds will allow us to prepare data for sharing including de-identification of data, the incorporation of metadata to ensure discoverability and the data transfer process to _____repository for preservation and access. An additional cost of $ _____ is required to cover data deposit fees for _____ repository, which will cover _____ years of hosting.