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

Data Type

Types and amount of scientific data expected to be generated in the project:
Summarize the types and estimated amount of scientific data expected to be generated in the project.

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 has specific definitions for what Scientific Data is, and what proposals are considered to be producing scientific data

Per the Policy, "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."

NIH Genomic Data Sharing (GDS) Policy Considerations

Check if your research is subject to NIH GDS (Genomic Data Sharing) policy usingthis criteria and list those data and the levels of processing here.

Individual NIH Institutes and Centers (IC) may have additional expectations or requirements for genomic data sharing as well. Please check the IC-specific genomic data sharing requirements.

Example Answer:

DMPTool fill-in-the-blank prompt

This project will produce _____ [*Data type, e.g., imaging, sequencing, experimental measurements*] data generated/obtained from _____ [*Data modality, e.g., instrument, method, survey, experiment, data source*]. 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.

If working with human subjects, consider adding: Data collection will be performed at clinical sites in the ____ [*location*] area(s) with ____ [*population(s) being studied; i.e., T2 diabetes*].

Sample answer from DMPTool: Basic sciences data

In this proposed project, data will be generated via the following methods: cell culture, light microscopy, confocal microscopy, real-time quantitative polymerase chain reaction (PCR), and stereological counting techniques. This data will be collected from a minimum of 3 independent experiments, with each independent experiment consisting of 3 groups, Wild-type (Rest+/+), heterozygous (Rest+/–), and homozygous (Rest–/–) from both embryonic stem (ES) cells and the corresponding neural stem/progenitor (NS/P) cells. The total size of the data collected is projected to be 300 GB.

We expect to generate the following data file types and formats during this project: Carl Zeiss microscopic image file (.CZI), images (.TIFF), tabular (.CSV), and Affymetrix GeneChip files (.CEL).

Raw data files will be analyzed to generate CSV files containing counts of cell type, total number of stem cells, and to enable statistical analysis.

Scientific data that will be preserved and shared, and the rationale for doing so:
Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

Guidance:

NIH 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.

Additional Guidance from DMPTool

If human subjects data will be collected and only de-identified subsets are to be shared, consider specific de-identification approaches that fit the population and purposes. Guidance on protecting privacy is at NOT-OD-22-213

NIH GDS Policy Considerations

If you are generating genomic data, follow specific sharing requirements (data submission and release expectation) under the NIH GDS policy (five levels of processing and associated expectations for data submission and release).

Example Answer:

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

OR

All data produced in the course of the project will be preserved and shared.

Addendum to DMPTool fill-in-the-blank prompt: Add the following for working with human data:

The final dataset will include _____[*e.g., self-reported demographic and behavioral data from interviews with participants and laboratory data from blood and urine specimens provided*]. We will share de-identified individual-participant level (IPD) data. Appropriate measures such as _____ [*describe specific de-identification practices to be used*] will be used for data de-identification and sharing, and informed consent forms will reflect those plans.

Sample answer from DMPTool: Minimal answer when most data will be shared openly

In this proposed project, the cleaned, item-level spreadsheet data for all variables will be shared openly, along with example quantifications and transformations from initial raw data. Final files used to generate specific analyses to answer the Specific Aims and related results will also be shared. The rationale for sharing only cleaned data is to foster ease of data reuse.

Metadata, other relevant data, and associated documentation: Briefly list 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:

DMPTool fill-in-the-blank prompt basics

To facilitate interpretation of the data, _____[*e.g., data dictionary, metadata, documentation, statistical analysis plans, bench protocols, data collection instruments*] will be created, shared, and associated with the relevant datasets.

DMPTool fill-in-the-blank prompt addendum: Add relevant parts of the following if working with clinical trials

In addition to _____ [*individual participant data (IPD) dataset being shared by restricted access and/or aggregate data being shared openly*], the researcher will share the _____ [*describe any other elements of the final data package not already addressed*]. Documentation and support materials will be compatible with the clinicaltrials.gov Protocol Registration Data Elements.

Sample answer by DMPTool

To facilitate the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets. The README file will include method description, instrument settings, RRIDs of resources such as antibodies, model organisms, cell lines, plasmids, and other tools (e.g., software, databases, services), and Protocol DOIs issued from protocols.io. The data dictionary will define and describe all variables in the dataset.

Guidance:

NIH Guidance

In addition to the documentation examples, consider metadata that will provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, any intermediate or descriptive observational variables).

Related Tools, Software and/or Code

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they

can be accessed.

Guidance:

Additional Guidance from DMPTool

Tool(s) and software should be identified, then plans should specify how the 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). When known, the longevity or period of time for which custom or proprietary tools will be available should be addressed.

In addition, 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:

DMPTool fill-in-the-blank prompt: 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.

DMPTool fill-in-the-blank prompt: If specialized tools (open source or proprietary) are needed to access or manipulate the data:

_____ [*Data type - Imaging data, survey data, etc*.] 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.

These tools will be shared openly via ____.

OR

These tools are fee-based, proprietary software. Alternative access to the data will be provided by [*describe the strategy for other sites to see or work with the data - potential strategies include committing to provide links to file viewers, or exporting files to a nonproprietary format for limited use and reuse*].

Sample answer from DMPtool: Animal studies with computational modeling

The raw data generated via the confocal microscope is in the Carl Zeiss (.czi) file format. Zeiss software or Fiji ImageJ is required to access the raw data. The raw data generated via the Affymetrix Mouse Genome 430 2.0 Array is in the .CEL format. Statistical programs such as MATLAB or R can be used to analyze the raw data present in the CEL file.

Fiji ImageJ is open-source software that can be downloaded freely online. Links to this or other open-source viewers will be included with the documentation for the shared dataset. Matlab is available for purchase from Mathworks. R is a free software environment for statistical computing and graphics. RStudio is a free R development environment that runs on most operating systems. R Scripts produced through the course of the research will be made publicly available on the lab's GitHub repository, and will be provided as Supplementary files for any publications through a Zenodo-GitHub link. Code will be available no later than when a publication has been submitted.

Standards

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist

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 from DMPTool

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:

DMPTool fill-in-the-blank prompt

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.*] along with references to the sources of those standardized names and metadata items will be included wherever applicable.

Addendum DMPTool fill-in-the-blank for 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*]

Addendum DMPTool fill-in-the-blank prompt for if there *are no* 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 which are as follows: [*list appropriate best practices*].

Sample answer from DMPTool

In accordance with FAIR Principles for data, we will use open file formats (e.g. JPEG, MP4, CSV, TXT, PDF, HTML, etc.) and persistent unique identifiers (PIDs) such as RRIDs for resources (e.g., organisms, plasmids, antibodies, cell lines, software tools, and databases) and DOIs for protocols using protocols.io. The bioimaging community has not yet agreed on a single standard data format that is generated by all acquisition systems, but we will use OME-Files for data that will be preserved and shared.

Data Preservation, Access, and Associated Timelines

Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository)

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.

Additional Guidance from DMPTool

See NOT-OD-21-016 and other guidance on selecting a repository for details on repository considerations. In brief, first consideration (option 1) goes to whether the FOA or Institute specifies a repository, in which case that repository must be used. Next priority (Option 2A) goes to approved Open Domain-Specific Data Sharing Repositories). If neither of those considerations fit, consider (Option 2B) other potentially suitable options: PubMed Central attachments, approved generalist repositories, or your organization's institutional repository.

NIH GDS Policy Considerations

If your research is subject to GDS policy, please refer to recommended repositories on the "Where to Submit Genomic Data" page on the NIH sharing site.

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*]

DMPTool fill-in-the-blank prompt

All dataset(s) that can be shared will be deposited in _____ [*Add appropriate NIH-supported data repositories*] OR _____ [*Add appropriate discipline- or data-specific repository, generalist*

repository, or your institutional data repository]

Sample answer from DMPTool: Minimal information, imaging study

Imaging data will be deposited into NCI's Imaging Data Commons. All other data described above in the "data to be shared" section will be deposited into Zenodo.

Sample answer from DMPTool: Minimal information, clinical study

Aggregate clinical trials data from all arms of the study will be available in clinicaltrials.gov, along with related metadata. All other data described above in the "data to be shared" section will be deposited into the National Addiction & HIV Data Archive Program (NAHDAP) repository.

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

Example Answer:

DMPTool fill-in-the-blank prompt

The _____ [*repository name*] provides metadata, persistent identifiers [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

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

Sample answer by DMPTool: Minimal information

[Repository Name] provides searchable study-level metadata for dataset discovery. [Repository] assigns DOIs as persistent identifiers, and has a robust preservation plan to ensure long-term access. Data will be discoverable online through standard web search of the study-level metadata as well as the persistent pointer from the DOI to the dataset.

Sample answer by DMPTool: Vivli clinical trials data repository

Vivli provides access to data and documentation through study-level metadata specific to clinical trials description, long-term preservation and access, and Vivli-issued DOIs. In addition to DOIs, Vivli records are cross-searchable by the clinicaltrials.gov registration ID. Access request processes are described in detail with each data record.

Sample answer by DMPTool: Expanded identifier discovery information

We will use Persistent Unique Identifiers (PIDs) to improve data findability across all dissemination outputs. PIDs used will include ORCID iDs for people, DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. We will also use indexed metadata, such as MeSH terms with a unique URL to make scientific data easily findable. We will keep our ORCID Records up to date with DOIs for our datasets and publications, ROR, and funder IDs to increase findability.

Guidance:

NIH Guidance

Unique Persistent Identifiers: The repository assigns datasets a citable, unique persistent identifier, such as a digital object identifier (DOI) or accession number, to support data discovery, reporting, and research assessment. The identifier points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

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.

Genomic data has further guidance on release expectations and timelines.

Example Answer:

DMPTool fill-in-the-blank prompt

Shared data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of _____[*duration*] years after the end of the funding period.

Sample answer by DMPTool

All scientific data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

Access, Distribution, or Reuse Considerations

Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.

Guidance:

NIH Guidance

Genomic data may have further considerations to address. The NIH now expects a single data sharing plan at time of funding application satisfies both the Genomic Data Sharing (GDS) Policy and the DMS Policy (per NOT-OD-22-198).

Additional DMPTool Guidance

Some data may require extra preparation before they can be shared. This is the section to describe what legal, ethical, or technical issues will require limiting the sharing of your data. Examples may include existing legal limits such as data licenses or use agreements, issues of proprietary IP development, technical limits about the size or structure of the data, or ethical issues for human subjects privacy.

Key issues in justification of human subjects data specifically may be informed consent (e.g., disease-specific limitations, particular communities' concerns) or privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures). Specific steps for human subjects data preparation can be addressed in the protections for privacy subquestion below.

NIH GDS Policy Considerations

How to access genomic data varies depending on which repository you selected. Please refer to the Accessing Genomic Data from NIH Repositories page on the NIH sharing site.

Example Answer:

DMPTool fill-in-the-blank prompt

There are no anticipated factors or limitations that will affect the access, distribution or reuse of the scientific data generated by the proposal.

OR

Due to _____ [*ethical/legal/technical considerations*], access/distribution/reuse of the resulting scientific data will be limited and approved/monitored by _____ [*describe the approach to limiting access/distribution/reuse*].

Sample answer by DMPTool for animal studies

To address safety and security concerns related to capturing and distributing pictures or video of vertebrate research animals, access and distribution of behavioral video files generated in our lab during brain inactivation studies of non-human primates will be limited as described and justified in the IACUC protocol governing the project and in compliance with the "Image Recordings of Research Animals" Standard Operating Procedure at our institution. There are no other factors that will impact access, distribution, and reuse for all other scientific data generated by this study.

Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after

approval).

Guidance:

Additional DMPTool guidance

Check the repository you intend to use to find out more about whether and how the repository supports controlled access.

Example Answer:

Researchers who are not using controlled access repositories can skip this section or state:

Controlled access will not be used. The data that is shared will be shared by unrestricted download.

DMPTool fill-in-the-blank prompts for researchers selecting controlled access repositories

Given the sensitive nature of the dataset, data will be made available in _____ data repository, which restricts access to the data to _____ [describe restriction, e.g. to qualified investigators with an appropriate research question and approved data use agreement (DUA)]. Data can be accessed by _____ [describe data repository access methods and measures].

Sample answer from DMPTool for clinical trials data sharing in the Vivli repository

Data will be available by controlled access only. To access data arising from this project, users must complete the Vivli data request form and sign the Vivli Data Use Agreement (DUA), which limits subsequent use to the terms of the approved request and requires that users maintain data security, and refrain from any attempts to re-identify research participants or engage in any unauthorized uses of the data. To get access to the data, the user must submit a valid scientific question, include a statistical analysis plan, and complete all required fields on the Vivli data request form. Vivli will review the data request for completeness.

Protections for privacy, rights, and confidentiality of human research participants:
If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

Guidance:

Additional DMPTool 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.

Issues to consider:

- 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.

Example Answer:

This subsection applies to studies involving human research participants. Other studies can generally skip question 5.3

DMPTool fill-in-the-blank prompt for 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 of 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 _____ methods [*describe de-identification method, noting any other applicable laws or policies such as HIPAA*].

Oversight of Data Management and Sharing

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

Guidance:

NIH This element refers to oversight by the funded institution, rather than by NIH. The DMS Policy does not create any expectations about who will be responsible for Plan oversight at the institution.

Additional DMPTool Guidance:

Describe how and by whom compliance with this Plan will be managed. If oversight and roles will include the addition of study personnel for oversight of data management and sharing, describe reasonable, allowable personnel costs in the budget justification rather than the DMS Plan.

Example Answer:

DMPTool fill-in-the-blank prompt

Lead PI ____[*name*]___, ORCID: __[*ORCID ID*]___, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI and Co-I team as part of general [campus(es)] stewardship, reporting, and compliance processes.