The American Association of Immunologists (AAI)

Guide to Developing Your NIH Data Management and Sharing (DMS) Plan

For detailed information, visit the NIH DMS website:

sharing.nih.gov

Have questions? Consult your program officer or email **sharing@nih.gov**

Important Points to Keep in Mind

- This Guide is meant to provide ***examples only***
- Please ***do not copy and paste*** any text into your own plans
  - The "Sample Text" sections are designed to provide example language, not for investigators to use as a template
  - Your language should be tailored to your research proposal
- NIH suggests the DMS plan be ***2 pages or less*** in length
- Hyperlinks and URLs are not allowed in NIH application materials
- This Guide may be updated as new information becomes available

Developed by the AAI Data Management and Sharing Working Group

Ferhat Ay, Ph.D. - La Jolla Inst.
Wendy Garrett, M.D., Ph.D. - Harvard Univ.
Shruti Naik, Ph.D. - New York Univ.
James Faeder, Ph.D. - Univ. of Pittsburgh
Alice Long, Ph.D. - Benaroya Research Inst.
Amber Smith, Ph.D. - Univ. of Tennessee Health Sci. Ctr.

Contact: Emily Kansler, Ph.D., AAI Science Policy Analyst • ekansler@aai.org

An Initiative of the AAI Committee on Public Affairs

Element 1: Data Type

A. 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.*

Considerations/Resources:

- "Amount" will depend on the type of data, e.g., estimated number of participants or samples, number of files generated, or actual file sizes.
- See below for the NIH definition of "scientific data."

> ***"Scientific Data:*** *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."*

Sample Text:

This project will involve performing flow cytometry phenotyping of isolated immune cells from an estimated [number of e.g., participants, samples, experimental groups]. This will generate data in the format of [names of expected file format(s), e.g., .fcs files, FlowJo workspaces, quantification as tabular .xls documents] totaling approximately [amount of data, e.g., 12 GB] in size. Immunofluorescence imaging will also be performed on an estimated [number of e.g., participants, samples, experimental groups] to visualize protein localization which will be collected as [file format(s), e.g., TIFF, JPEG], as well as western blot analyses to quantify level of protein expression, which will be collected as [file format(s), e.g., TIFF, JPEG], totaling approximately [amount of data] in size. In addition, FACS isolated immune cells from an estimated [number of e.g., participants, samples, experimental groups] will be processed to undergo [method(s), e.g., genomics, transcriptomics, proteomics] and will be collected as [file type(s), e.g., .fastq files] which will be transformed by standard workflows into formats as is usual practice.

B. 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.*

Considerations/Resources:

- Data of "sufficient quality to validate and replicate research findings" should be shared; NIH does not expect ALL data be shared.
- If the proposal includes genomic data, the **Genomic Data Sharing (GDS) Policy** applies and should be incorporated into this DMS plan. More information about what data should be shared, and expected timeframes for sharing data, can be found at **Genomic Data Submission and Release Expectations**.
- See information for **sharing software or code** developed in a research project.

Sample Text:

For flow cytometry experiments, the raw files [raw file type(s), e.g., .fcs files, FlowJo workspaces], and analyzed data [analyze data type(s), e.g., .xls format] will be preserved and shared. For imaging experiments, the raw and processed images [file type(s), e.g., as TIFF files] and relevant analyses and/or quantifications will be preserved and shared. The raw western blot images and associated quantifications, where applicable, will be preserved and shared [file type(s), e.g., as JPEG files]. The [method(s), e.g., genomics, transcriptomics, proteomics] processed (after cleaning and quality control) data in the format of [file type(s), e.g., BAM files] will be shared as well as subsequent analyses [file type(s), e.g., DEG heatmaps]. These data are sufficient to validate and replicate research findings proposed in this application.

Alternate Sample Text:

All scientific data as described in Element 1A (raw, processed, and derived) generated of sufficient quality will be shared to promote reuse and reproducibility.

C. 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.*

Considerations/Resources:

- Include all metadata necessary to make the shared data *interpretable* and *reusable* to another researcher examining the data.
- Data repositories often have requirements and/or templates for metadata. ***Plan ahead*** for what metadata are required, and in what format, for the data that will be shared. Also consider standardizing protocols, naming conventions, etc. Establishing this early on will reduce administrative burden when the time comes to deposit and share data.

Sample Text:

For each technique in this proposal, the metadata maintained and shared will include sample collection and processing protocols, antibody lists with concentration and validation information, and key descriptive information about each sample needed to understand and/or make use of this data [e.g., sample names, collection date, etc.], as well as any other metadata as required by [chosen data repository].

Element 2: Related Tools, Software, and 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.

Considerations/Resources:

- This section requires the *disclosure* of relevant tools/software/code but does not require that investigators provide that resource (e.g., some software/code may be available through the investigator upon request, some require a subscription, etc.).
- Consider important factors like software versions, necessary packages, and extensions.
- For code, consider deviations from "standard" and be sure to include such steps to ensure reproducibility.

Sample Text:

To access flow cytometry data users will require analysis software [software name(s), e.g., FlowJo version X, information about access can be found at https://www.flowjo.com/]. Immunofluorescence images can be viewed with [software name(s), e.g., CaseViewer, Fiji/ImageJ, Adobe Illustrator] and analysis software [software name(s), e.g., Zeiss ZEN software]. Western blots can be viewed and analyzed with [software name(s), e.g., Fiji/ImageJ]. To access and replicate [data type name(s), e.g., RNA sequencing data] users will require the following data processing software [software name(s), e.g., R, Python, Cell Ranger], analysis software [software name(s), e.g., Bioconductor, Seurat], and packages [e.g., R scran, R package GSVA]. Any code written to analyze data will be deposited on [repository name, e.g., GitHub]. To access shared statistical analyses users will need [software name(s), e.g., Graphpad Prism].

Element 3: 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.

Considerations/Resources:

- Standards include clinical data standards, standard data formats, data dictionaries, definitions, etc., and are usually approaches to data collection, format, and organization that are agreed upon within a scientific discipline or community to achieve data interoperability.
- Examples include **Sequence Run Archive** (SRA) data standards for dbGaP or **NCI cancer data standards**.
- Consult the sample plans listed in **Writing a Data Management and Sharing Plan** for other examples of standards.
- Consider standards required by repositories, journals, or NIH Funding Opportunity Announcements (FOAs).

Sample Text:

For all genomic data submitted to [repository name, e.g., GEO, dbGaP], [name of standard(s), e.g., SRA data standards] will be followed. For all other scientific data and associated metadata, no common data standards exist/have been widely adopted; however, we will follow the commonly utilized formats and adhere to best known practices in the field.

Element 4: Data Preservation, Access, and Associated Timelines

A. 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.*

Considerations/Resources:

- Some NIH FOAs and journals designate specific data repositories.
- Determine where the data are best deposited, in generalist repositories or discipline-specific repositories. Data can be divided into multiple repositories. See this **list of NIH-supported data repositories** and NIH advice on **Selecting a Data Repository**.
- Note that all NIH-supported genomic human research data must be registered in dbGaP, even if it will be submitted and stored elsewhere. See other guidance on **Where to Submit Genomic Data**.

Sample Text:

All scientific data will be deposited in [repository name, e.g., FigShare, ImmPort] except for the [subset type(s), as needed e.g., genomic] data that will be deposited in [repository name, e.g., NCBI GEO, bdGaP].

B. 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.*

Considerations/Resources:

- Examples of persistent unique identifiers (PIDs) include digital object identifier (DOI) for published data and Open Researcher and Contributor ID (**ORCID**) for researchers.

Sample Text:

All data repositories proposed to be used meet the desirable characteristics set by NIH for data repositories, including the use of persistent unique identifiers [identifier name, e.g., DOI or accession numbers] and keyword search functionality. Related data across repositories will be linked by including the PIDs in the published metadata for each repository. All PIDs will be referenced in any publication.

C. 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.*

Considerations/Resources:

- Under the DMS Policy, "publication" refers to peer-reviewed journal articles. Preprints *do not* count as publications.
- NIH Institutes and Centers may have their own requirements (e.g., **NIAID Data Sharing Guidelines** and **NCI Clinical Trials Access Policy**, see full list **here**).
- NIH has outlined when human and non-human genetic data are expected to be shared in **Data Submission and Release Expectations**.
- 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."

Sample Text:

All non-genomic scientific data will be made available to other users no later than time of an associated publication or end of the performance period, whichever comes first. Data will be made available for at least [timeframe, e.g. 5 years]. For any genomic data generated, data will be submitted in accordance with NIH Data Submission and Release Expectations [e.g., BAM files from human genetic data will be submitted within 3 months after being generated and will be shared no later than 6 months after data submission].

Element 5: Access, Distribution, or Reuse Considerations

A. 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.*

Considerations/Resources:

- NIH describes the following as justifiable factors for limiting sharing of data:
  - "informed consent will not permit or will limit the scope or extent of sharing and future research use
  - existing consent (e.g., for previously collected biospecimens) prohibits sharing or limits the scope or extent of sharing and future research use
  - privacy or safety of research participants would be compromised or place them at greater risk of re-identification or suffering harm, and protective measures such as de-identification and Certificates of Confidentiality would be insufficient
  - explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure
  - restrictions imposed by existing or anticipated agreements (e.g., with third party funders, with partners, with repositories, 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)
  - datasets cannot practically be digitized with reasonable efforts"

Sample Text:

There are no factors affecting access, distribution, or reuse of any of the proposed non-human data. For human data, Institutional Review Board paperwork and informed consent documents will include language describing plans for broad data sharing, with assurances that personal identifying information will be removed.

B. 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).*

Considerations/Resources:

- See information under "Element 5" in **Writing a Data Management and Sharing Plan** for information on allowable limitations for sharing human genomic data.
- Refer to **NOT-OD-22-213**, section "Points to Consider for Choosing Whether to Designate Scientific Data for Controlled Access."
- See information about **Institutional Certifications**, which are required for submission of large-scale human genomic data to NIH-supported repositories under the NIH GDS Policy.

Sample Text:

No non-human scientific data will be under controlled access. Data from human participants who have consented to broad data sharing, that have been sufficiently de-identified with very little risk for reidentification, and that are not considered sensitive, will be openly shared.

Alternate Sample Text:

Due to [justifiable factor(s), e.g., possibility of re-identification, sensitive data, risks to participant privacy, etc.] the human [data type(s) e.g., genomic] data submitted in [repository name, e.g., dbGaP] will be under controlled access.

C. 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).*

Considerations/Resources:

- Consult NIH guidance for **developing informed consent language**.
- See NOT-OD-22-213: **Protecting Privacy when Sharing Human Research Participant Data**.
- See NOT-OD-22-064: **Responsible Management and Sharing of American Indian/Alaska Native Participant Data**.
- Consult **NIAID Genomic Data Sharing Plan Examples**.

Sample Text:

Data from human participants will be sufficiently de-identified [name method if possible] and will be covered by a Certificate of Confidentiality as a term and condition of this award, if granted. [If applicable add: any data that cannot be sufficiently de-identified will be managed under controlled access.]

Element 6: 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).

Considerations/Resources:

- Who can accept the additional administrative responsibility?
- In practicality, how will this be performed?

Sample Text:

Data management, storage, sharing, and overall compliance with the DMS plan will be monitored on a [time frame, e.g., monthly] basis by [title(s) and role(s), e.g., PI, lab manager, technician, designated data stewards] through [planned process, e.g., meetings, written reports, shared documentation, etc.]. The DMS plan will be updated as necessary by [title and role].

All text within blue bubbles comes directly from the NIH DMS Plan Template