Resource created by the Working Group on NIH DMSP Guidance

Data Terms related to the NIH DMS Plan and Policy

Review this guide for definitions and descriptions of terms related to the sections (aka Elements) of the NIH Data Management and Sharing Plan.

Data Management and Sharing Plan: A plan describing the data management, preservation, and sharing of scientific data and accompanying metadata. See the Elements required for the NIH DMS Plan 2023 and the Format Guide for supplying this information.

ELEMENT 1: Data Type

A. Types and Amount of Scientific Data

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

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.

Type

Types can include experimental data (e.g. results from laboratory experiments or clinical trials), observational data (e.g. from instruments, sequencing, survey data), simulations (from test models) and/or derived from existing datasets)

See: Ten simple rules for maximizing the recommendations of the NIH data management and sharing plan (Rule 1); Types of Data.

Amount

Depends on the data and how it is typically measured. It may define number of participants, number of files, or could reflect size of storage space anticipated.

Examples

- Participants: 256-channel EEG data and fMRI images from ~50 research participants
- File Space: Between 10-15 GB across all files

B. Scientific Data that will be preserved and shared

Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

Data Preservation

Data preservation consists of a series of managed activities necessary to ensure continued stability and access to data for as long as necessary. For data to be preserved, at minimum, it must be stored in a secure location, stored across multiple locations (e.g. 'Rule of Three'), and saved in open file formats that will likely have the greatest utility in the future. Part of the preservation process can include depositing data in an institutional, discipline-specific, or generalist data repository, all which allow for publication and preservation. From NCDS Data Glossary.

Examples

The new NIH Data Management and Sharing Policy requires data be preserved and shared, so a medical researcher submits their COVID data to the National COVID Cohort Collaborative (N3C), as listed in the Open Domain-Specific Data Sharing Repository.

Sharing

Data sharing refers to the practice of making data available to other research stakeholders, including other investigators, research subjects, and the broader public. Data sharing can happen indirectly through publications and other scholarship, but under the NIH DMS Plan, sharing is encouraged through an established repository (see Element 4).

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.

Metadata

Metadata can be defined as Information about a data set that is structured (often but not always in machine-readable format) for purposes of search and retrieval. Metadata elements may include basic information (e.g. title, author, date created, etc.) and/or specific elements inherent to datasets (e.g., spatial coverage, time periods).

Documentation

Information needed for the data to be understood, interpreted, and used. Documentation can describe the research project as well as the resulting data. Dataset documentation for tabular (e.g. spreadsheet data) can include variable names and descriptions, explanation of codes and classification schemes used, etc. See: Ten simple rules… (Rule 2)

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.

Tools

Specialized tools such as software (e.g. Excel, Python, SPSS) or equipment that would be needed for another researcher to be able to access or manipulate shared data for replication or reuse. Consider important attributes such as the version of an application, as well as any needed packages or extensions needed to review and analyze the data, and how this information can be found.

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.

Interoperability

Data interoperability refers to the ways in which data is formatted to allow diverse datasets to be merged or aggregated in meaningful ways. Well documented data records inform users about the utility of data sets for their research. Data interoperability relies on metadata and data documentation. Without proper documentation researchers would not know which datasets and variables are comparable. Data interoperability is frequently accomplished through the use of data standards, which are community-agreed-upon approaches to the collection and organization of data.

Examples

Options for collecting age of subject data: recording a specific age (e.g. age: 7), dates of birth to experiment (e.g. subject age: Jan 1, 2015 to Jun 30, 2022), age in months (e.g. subject age: 30 months)

Metadata Standard

Common approaches to the collection and organization of data that are agreed upon by a community to support data interoperability. Standards may apply to the project as a whole or to the individual files and variables within a dataset.

Examples

- Sequence Run Archive data standard for dbGaP
- NIMH Data Archive Data Dictionary

Metadata Schema

A machine readable structured framework or plan to document the basic information about data sets (e.g. the title of the dataset, project title, key terms, authors). Repositories often require schemas to meet their standardized search parameters, like a subject index or geographic filter.

Examples

- General: Dublin Core (DC)
- Biology: Darwin Core
- Ecology: Ecological Metadata Language (EML)
- Social Sciences: Data Documentation Initiative (DDI)

Common Data Standards

Standards that are adopted by a wide range of members of the discipline, and which also apply to data and/or metadata. See also this definition for data standards. Note that there is no specific guidance for how commonly-adopted a standard must be to count as a *common* data standard.

Content Standards

Agreed-upon shared guidance on how items in datasets are collected and represented. Content standards may include definitions of conditions, units of measurement, terminology allowable *within* each variable or data element, and any other guidance about the content of data points when collected, stored, transformed, or loaded to another system.

Encoding Standards

Agreed-upon shared guidance on the technological process by which data and metadata files are made into a computer-readable format. Common encoding standards include html, pdf, csv, docx, and more. Less common encoding standards requiring specific software may need to be addressed in Element 2.

Data Elements

A basic unit of information that has a unique meaning and subcategories (data items) of distinct value. Examples of data elements include gender, race, and geographic location. NIST

In the context of standards, a DMS Plan might discuss data elements by stating that the data elements will be drawn from the NIH Common Data Elements or another shared codebook defining the content and metadata for variables.

Example

Blood pressure, as a two-variable element: systolic sub-element and a diastolic sub-element.

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.

Repository

Established web-based platforms designed for data deposit in order to preserve and share data, often publicly, though some may support mediated requests. Common types of data repositories include discipline-based (e.g. National COVID Cohort Collaborative (N3C)), generalist repositories that house data regardless of type, format, content, or subject matter (e.g. Figshare, Zenodo), and institutional repositories, typically maintained by researchers' institutions for housing and sharing scholarship and data. NIH encourages researchers to select a data repository that is most appropriate for their data type and discipline.

B. How scientific data will be findable and identifiable

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

Unique Persistent Identifier

Also referred to as persistent identifier (PID), a citable identifier to support data discovery, reporting, and research assessment. A well-known persistent identifier (PID) is a Digital Object Identifier (DOI) used to locate specific digital objects, such as journal articles. See DMPTool Persistent Identifiers.

Indexing Tools

In the context of the DMSP, refers to discoverability of shared datasets, e.g. through an established repository or by ensuring that any non-repository location is findable by search engines that index sites in order to facilitate discovery of datasets.

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.

ELEMENT 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse

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.

Reuse

Data Reuse, is the analysis of existing data for related or new research. Access to data sets and reuse of data can help determine the validity and reproducibility of scientific hypotheses. See NCDS Data Glossary.

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).

Controlled

Controlled access refers to a data sharing model that requires a request for access to all or part of the dataset(s). Some data, e.g. about human participants, remains sensitive even after other measures (de-identification, etc.) have been taken, either because it cannot be sufficiently anonymized or because doing so would make the data less useful. In those cases, planning for controlled access may be necessary.

NIH strongly encourages investigators to plan for how data management and sharing will be addressed in the informed consent process. Investigators should communicate with prospective participants about how their scientific data are expected to be used and shared.

There are a number of approaches to controlled access including: Data Access Request (DAR), Data Use Agreement (DUA), Data Access Condition (DAC), or Data Use Limitation (DUL)

Examples

- Data Access Request (DAR) for NIH-Designated Data Repositories (e.g. dbGaP)
- Data Use Limitation guidance for Genomic Data Sharing Policy
- Accessing Restricted Data at ICPSR
- Access Controls for the Qualitative Data Repository (QDR) (good list of different levels of 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).

Data Privacy

Research data, particularly when containing information about human subjects such as protected health information, may have special concerns around maintaining the privacy of the participants and require de-identification and other steps to maintain data security. The Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) are the most commonly used federally mandated privacy policies. See Data Privacy.

De-identification

A method to protect research participants' privacy. Strategies for data de-identification are outlined in the *Supplemental Information to the NIH Policy for Data Management and Sharing: Protecting Privacy When Sharing Human Research Participant Data*. Recommendations include relying on the standards for identifiability outlined in the Common Rule (participant identity cannot "readily be ascertained") and in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (using Expert Determination or Safe Harbor methods).

Data Confidentiality

Certificates of Confidentiality (CoCs) protect the privacy of research subjects by prohibiting disclosure of identifiable, sensitive research information to anyone not connected to the research except when the subject consents or in a few other specific situations.

Informed Consent

The process by which a volunteer confirms his or her willingness to participate in a research study, such as a clinical trial, after having been informed of all aspects of the study that are relevant to the volunteer's decision to participate. Informed consent is documented by means of a written, signed, and dated informed consent form. Regarding sharing of data from such studies, see Informed Consent for Secondary Research with Data and Biospecimens: Points to Consider and Sample Language for Future Use and/or Sharing

Human Research Participants

According to 45 CFR 46, a human subject is "a living individual about whom an investigator (whether professional or student) conducting research:

- Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
- Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens."

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).

Data Stewardship

The process of Data Stewardship involves data management processes, including effective curation, control, and use of data assets and can include creating and managing metadata, applying standards, managing data quality and integrity, and additional data governance activities related to data curation.

Additional Glossaries

For other research data management related terms, see these helpful glossaries:

Data Glossary – NNLM National Center for Data Services (NCDS)

Research Data Management Terminology – CODATA (Committee on Data of the International Science Council (ISC)

Glossary – Toolkit for Patient-Focused Therapy Development – National Center for Advancing Translational Sciences (NCATS)