

Version 1.0 (2022-10-07): This document will be updated regularly. For the most current version, visit <https://research.iu.edu/policies/nih-data-mgmt-sharing-policy.html>.

How to use this document

This document provides guidance to support researchers in developing a Data Management & Sharing (DMS) Plan, as described in the National Institutes of Health (NIH) [Final Policy on Data Management and Sharing](#) and in [supplemental information about the elements of a DMS Plan](#). It is aligned with the DMS Plan template provided by the NIH and enhanced with additional guidance, samples, and resources for support.

This guidance, along with a DMS Plan template in FireForm (*available in November 2022*), has been developed by the IU Research Data Management Plan Working Group.

The following information is provided:

- *Policy description* provides language from the NIH policy ([NOT-OD-19-013](#)) as well as the supplemental information describing the elements of a DMS Plan ([NOT-OD-19-014](#)).
- *Guidance* offers additional information regarding the scope and level of detail suggested.
- *Sample responses* – where possible, we have developed or adapted sample responses that represent common research scenarios, workflows, and issues.

Instructions

- Review this guidance document for support in meeting the baseline expectations set by the NIH policy.
- Review the Program Announcement/Funding Opportunity Announcement/Request for Applications you intend to apply for to identify any *additional* requirements from Institutes, Centers, or Offices (ICO) related to data management and sharing. For example, some ICO may indicate use of specific repositories.

Key Policy Resources

Final Policy on DMS: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>

Elements of a DMS Plan: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-014.html>

IU support: <https://research.iu.edu/policies/nih-data-mgmt-sharing-policy.html>

Data Type

- Policy description:* Briefly describe the scientific data to be managed, preserved, and shared including:
- Summarize the types (e.g., 256-channel EEG data and fMRI images) and amount (e.g., from 50 research participants) of scientific data to be generated and/or used in the research. Descriptions may include the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing.
 - Describe which scientific data from the project will be preserved and shared. 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. The plan should provide the reasoning for these decisions.
 - 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.

- Guidance:* Consider addressing the following dimensions of your data:
- A high-level description of the **research methods or source used to generate and/or collect data** – e.g., observational, interview, survey, simulation, experimental, etc.
 - The **content of the data** – e.g., numeric, text, image, audio, video, instrument specific, models or algorithms
 - The **file format or structure** – e.g., csv, plain text, xml, jpg, pdf, AIFF or audio interchange file format, html, DICOM, etc. Some file formats are proprietary, while others are open standards. Typically, open standard file formats like csv are more sustainable.
 - The **level of processing that has been applied** – e.g., raw, cleaned, validated, aggregated, or normalized.
 - Whether the data are **stable or dynamic** – e.g., will the data be collected longitudinally, change over time, or be revised and updated.
 - **Amount of data** – estimate the number of files by file type or total storage required (in GB, TB, PB, etc.)
 - The **scope of the data** – what data are necessary to validate and replicate research findings.

After describing your data, clearly identify which data will be preserved and shared. This policy does not require that *all* research data must be shared.

Sample responses: Example for bench science

The project will generate three main sources of laboratory data: microscopy data, observational data, and physical data objects such as Western films. Imaging data from microscopes in GCT, CZI, and OME-TIFF file formats is captured and saved on the instrument control device, then transferred to two locations. While some microscopy image files will be saved to the project storage space in Microsoft Teams, there are too many large files for Teams to be the primary storage location. Monthly, the microscopy files will be transferred from instrument control computers to external hard drives and laptops, then moved to storage on a data server maintained by the departmental IT team, which is managed in alignment with the IU Information Security Program (<https://informationsecurity.iu.edu/program/index.html>).



We will also generate observational data and scans of physical data objects. These are digitally scanned as JPG or TIFF and saved to the instrument control machines, with a backup to the departmental server. Measurements derived from these objects are saved into Excel spreadsheets, which are stored in the project notebook in LabArchives. We sometimes create DNA sequencing data and qPCR data, which are generated by the IUSM Genomics Core (<https://medicine.iu.edu/service-cores/facilities/medical-genomics>). Sequence files are saved in the project storage space in Microsoft Teams, with copies backed up to the departmental server. We estimate that the project will generate approximately 250GB of data. Each Excel spreadsheet will contain a data dictionary to define abbreviations, fields, and valid parameters. Genomic sequence data will be documented in LabArchives pages that describe the source organism, isolate, sequence, phenotype, and project information. Microscopy image files are documented using the REMBI schema (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8606015/>) in a database operating on the IU Research Database Complex (<https://kb.iu.edu/d/amuw>). Data will be discussed at weekly project meetings with corrections recorded in meeting notes stored within the project lab notebook on LabArchives. Gene expression data will be deposited to GEO within 6 months of creation. Selected data will be described, packaged into TAR files, and archived for long-term storage on the Scholarly Data Archive (<https://kb.iu.edu/d/aiyi>). These data will be retained for 10 years after the end of the project period.

Example for human subjects research involving PII, but not PHI

Participants (n = 1000) will provide self-reported medical history information during interviews with study personnel; these data are entered into the REDCap project database. Study personnel collect a family medical history from participants, which is recorded on paper and then scanned (PDF) and uploaded into REDCap. Data pertinent to sample collection are recorded on de-identified data sheets until the information is uploaded to REDCap by study personnel. Data from CT scans are entered into an Excel spreadsheet (50 columns by 1000 rows). These data are aggregated with the REDCap database prior to analysis. All downstream assays of samples in the lab utilize only study participant IDs (de-identified). Elements from the final database will be harmonized and deposited into the NIMH Data Archive.

Data Source, Content, Format	Number of Files	Storage Required
Medical history interviews/pedigrees, PDF	1000	5GB
Medical history interviews/pedigrees, REDCap database	1	<1GB
Clinical data, REDCap database	3	<1GB
CT scan images, DICOM	3000	100GB
CT scan measurements, Excel spreadsheet	1	<1GB

Example(s) for human subjects research involving PHI

Example 1

There are two sources of clinical data: 1) data gathered via structured interview with participants; and 2) data extracted from the electronic medical record, including



demographic characteristics, anthropomorphic characteristics, and results of radiologic tests.

Participants may have genetic testing or other molecular assays performed on a research basis, which utilize the biospecimens that are collected as part of the study. The clinical data are entered into IU REDCap (<https://kb.iu.edu/d/bdhl>). Family data are recorded as pedigrees, scanned electronically, and stored in the Microsoft Secure Storage (<https://kb.iu.edu/d/bgfb>) space created for the project. The results of radiological studies are recorded on an electronic spreadsheet, which is also stored in Microsoft Secure Storage (<https://kb.iu.edu/d/bgfb>).

Example 2

This project involves four data streams:

- 1. Participants will provide self-reported data that will be entered into the REDCap database.*
- 2. Participants will be asked interview questions and interviewer will enter data into the REDCap database.*
- 3. Field data will be recorded on paper (stored in study binders) and entered into an IU system approved for use with critical data including PHI.*
- 4. Information will be collected from the Electronic Medical Record (EMR) and entered into an IU system approved for use with critical data including PHI.*

Specimens are collected specifically for research purposes and include: recording of a history and physical examination, photographs and recording of skin lesions and clinical data, surface cultures, skin biopsies and blood. Data include subject demographics, contact information, HIV serology and pregnancy tests. The history, physical exam, laboratory and daily visit data are recorded on a paper chart, where the linkage to a study number and subject identifiers can be made. All laboratory specimens are coded with the subject number and have no other identifiers. Data concerning each subject (age, gender, ethnicity, trial, date of infection, date of biopsy, days infected, outcome of each infected site, hypertrophic scar formation, specimens that are stored) are stored in an IU system approved for use with critical data including PHI. The database does not contain patient identifiers. For the blood drawing protocol, specimens are coded with a participant number without other identifiers; data concerning each subject (age, gender, ethnicity, date of donation, amount donated) are recorded in REDCap.

Example for aggregating existing data (for secondary use) with newly generated data

There are two data components to this project. The first will use three sources of existing data: [location redacted] County Birth Certificates, outpatient EHR data from [Health System redacted] via the [name redacted] data repository, and geographical data from [name redacted]. The second part of the project will use human-centered design techniques to develop a communication strategy for parents/caregivers regarding their child's weight and obesity risk. Participants will be engaged in focus group activities that can include discussion, collage, card sorting, and other activities such as cognitive interviews. Focus group sessions will be recorded (audio and video) and photographs may also be captured. Data will include participant demographics as well as information shared during the focus group session, which may include participant

provided health information or their child's health information. Data will be stored in REDCap and Qualtrics. Audio files will be securely transmitted to a transcription service that has received university approval for use with HIPAA protected information (i.e., a Business Associate Agreement is in place). The transcription service will upload completed transcripts to a designated space within the project's Microsoft Secure Storage (<https://kb.iu.edu/d/bgfb>) space.

Data Source, Content, Format	Number of Files	Storage Required
Clinical data, REDCap database (200 records x 300 fields)	3	2GB
Photographs	600	10GB
Audio recordings	30	5GB
Video recordings	30	100GB
Transcriptions (.docx files)	30	<1GB

Related Tools, Software and/or Code

Policy description: Indicate 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. If applicable, specify how needed tools can be accessed.

Guidance: Consult the Program Announcement/Funding Opportunity Announcement/Request for Applications, as they may include more specific requirements than the DMS Policy.

Sample responses: We use the Open Microscopy Environment (OME) Files standard (<https://www.openmicroscopy.org/ome-files/>), so we use the Open Microscopy Environment (OME) Bio-Formats tool (<https://www.openmicroscopy.org/bio-formats/>).

We use ChemDraw (<https://perkinelmerinformatics.com/products/research/chemdraw>) to document pathways and reactions.

We use NVivo 13 (QSR, <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>) to code and analyze interview transcripts.

Standards

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

Sample responses: We use the CDISC standards CDASH (<https://www.cdisc.org/standards/foundational/cdash>) and SDTM (<https://www.cdisc.org/standards/foundational/sdtm>) for collecting and reporting clinical data.

This study will generate multiple streams of data, which will be managed according to the following standards:

- *Brain Imaging Data Structure (BIDS; <https://bids-specification.readthedocs.io/en/stable/>) for neuroimaging data*
- *ARRIVE (Animal Research: Reporting of In Vivo Experiments; <https://arriveguidelines.org/arrive-guidelines>) Guidelines for reporting animal study data*
- *NIMH Data Archive (NDA) Data Dictionary (<https://nda.nih.gov/nda/harmonization-standards.html>) for human participant data*

We use the MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments, <https://doi.org/10.1373/clinchem.2008.112797>

We use the guidelines and data quality measures endorsed by the ENCODE Consortium & Data Coordination Center. See <https://www.encodeproject.org/data-standards/>

For fluorescence microscopy data, we use the 3D Microscopy Metadata Standards (3D-MMS; <https://doryworkspace.org/metadata>).

We attempt to stay current with the implementation of the ACT (Accrual to Clinical Trials Network Repository) ontology for metadata organization. Link to github: <https://github.com/dbmi-pitt/ACT-Network/tree/master/ontology/ACTOntologyV4.0>

We use the DICOM standards for medical images. USP<823> requirements in radiopharmaceutical production: https://www.usp.org/sites/default/files/usp/document/our-work/chemical-medicines/water_mark-footnote_usp32-nf27_chapter_823.pdf

International Epidemiology Databases to Evaluate AIDS Data Exchange Standards: <https://iedea.github.io/> and <https://redcap.vanderbilt.edu/plugins/iedea/des/>

Data Preservation, Access, and Associated Timelines

Description: **NOTE:** The NIH encourages scientific data to be shared as soon as possible, and no later than the time of an associated publication or end of the performance period, whichever comes first. NIH also 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.

Give plans and timelines for data preservation and access, including:

- The name of the repository(ies) where scientific data and metadata arising from the project will be archived. See [Selecting a Data Repository](#) for information on selecting an appropriate repository.
- How the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.
- When the scientific data will be made available to other users and for how long. Identify any differences in timelines for different subsets of scientific data to be shared.

<i>Considerations:</i>	Consult the Considerations for Sharing Research Data at https://forms.office.com/r/ssQjhtD9m3
<i>Guidance:</i>	<p>Decisions about sharing research data may be dependent on terms in existing Research Agreements or Data Use Agreements, and/or licenses applicable to data used for secondary analysis. Additionally, for human subjects data that will be collected prospectively, you will need to ensure the Informed Consent Statement describes the planned sharing. Thus, you may need to consult with support units/groups to make appropriate and informed choices about what data can be shared, with whom, how, and when and how sharing should be described to research participants. We strongly recommend discussing with your study team the Considerations for Sharing Research Data prior to submitting the proposal.</p> <p>If existing agreements stipulate controlled access, your choice of repositories may be constrained.</p>

Access, Distribution, or Reuse Considerations

- Description:* Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to:
- Informed consent
 - Privacy and confidentiality protections consistent with applicable federal, Tribal, state, and local laws, regulations, and policies
 - Whether access to scientific data derived from humans will be controlled
 - Any restrictions imposed by federal, Tribal, or state laws, regulations, or policies, or existing or anticipated agreements
 - Any other considerations that may limit the extent of data sharing. Any potential limitations on subsequent data use should be communicated to the individuals or entities (for example, data repository managers) that will preserve and share the scientific data. The NIH IC will assess whether an applicant’s DMS plan appropriately considers and describes these factors.

<i>Considerations</i>	Consult the Considerations for Sharing Research Data document at https://forms.office.com/r/ssQjhtD9m3
<i>Guidance:</i>	<p>If not already addressed in the previous section, describe considerations for access and reuse.</p> <p>Decisions about sharing research data are often dependent on participant consent, terms in existing Research Agreements or Data Use Agreements, and/or licenses applicable to data used for secondary analysis. Thus, you may need to consult with support units/groups to make appropriate and informed choices about what data can be shared, with whom, how, and when. We strongly recommend reviewing the document Considerations for Sharing Research Data prior to submitting your proposal.</p>

Oversight of Data Management and Sharing

Description: Indicate how compliance with the DMS plan will be monitored and managed.

<i>Guidance:</i>	<p>We recommend identifying study personnel who will be responsible for oversight of various types of data activities, including data collection, data entry, data screening and processing, data analysis, data visualization, reporting, record-keeping, metadata creation, prepare datasets for sharing, data deposit and/or dissemination.</p> <p>While the Principal Investigator (PI) is ultimately accountable, the responsibilities of data management and sharing should not be assigned to a single person.</p>
<i>Sample responses:</i>	<p><i>PI: data analysis, data reporting, data management training for study personnel</i></p> <p><i>Research Technicians: data collection, data entry, record-keeping, metadata creation</i></p> <p><i>Post-doc: data screening & processing, data analysis, prepare datasets for sharing, data deposit and/or dissemination, data reporting</i></p> <p><i>Research Scientist: data analysis, data visualization, data management training for study personnel</i></p> <p><i>Research Coordinator: record-keeping, metadata creation, prepare datasets for sharing</i></p>