



National Institutes  
of Health



# Writing an NIH Data Management and Sharing Plan



Chen Chiu and Betsy Gunia

**Website:** [dataservices.library.jhu.edu](https://dataservices.library.jhu.edu)

**Email:** [dataservices@jhu.edu](mailto:dataservices@jhu.edu)

**Johns Hopkins Research Data Repository:** [archive.data.jhu.edu](https://archive.data.jhu.edu)



Data Services

JHU DATA SERVICES

# HELPING YOU NAVIGATE DATA

WE HELP FACULTY, RESEARCHERS AND STUDENTS



FIND



USE



MANAGE



VISUALIZE



SHARE

FIND OUT  
MORE

**GO TO** [dataservices.library.jhu.edu](https://dataservices.library.jhu.edu)

**EMAIL** [dataservices@jhu.edu](mailto:dataservices@jhu.edu)

**SHARE AT** [archive.data.jhu.edu](https://archive.data.jhu.edu)



JOHNS HOPKINS  
LIBRARIES

Data Services



## • Today's Topics

- Overview of NIH Data Management and Sharing Plans (DMSPs)
- Elements of DMSPs
- Introduction to the DMPTool
- Comments on things we are seeing in draft Plan we are reviewing

(We are NOT talking about the DMS budget or budget justification, but we will provide some resources for budgeting)

You will get a copy of these slides, including live URLs.

# • NIH Data Management and Sharing Policy: Effect On Jan. 25

- [Policy](#) Goals:
  - Advance rigorous and reproducible research
  - Promote public trust in research
- Requirement to write a Data Management and Sharing Plan
  - Written as part of your proposal **for grants due on/after Jan. 25, 2023**
  - Expectation to maximize data sharing with caveats
  - Expectation that data are of sufficient quality to validate and replicate research finding

Data Services can help you write your plan through **consultation** and various **workshops**



## • Implementation Details

- Applies to research that results in/generates scientific data  
([see specific activity codes](#))
- NIH Program staff will assess DMS Plans
- Must comply with the version submitted in a proposal but can update during your regular reports
- If you are subject to both genomic data sharing policy and DMSP, you will submit one Plan!
  - Details here: [Implementation Changes for Genomic Data Sharing Plans](#)

# NIH Data Sharing Site

DATA MANAGEMENT AND SHARING POLICY

GENOMIC DATA SHARING POLICY

OTHER SHARING POLICIES

ACCESSING DATA

ABOUT

## Data Management and Sharing Policy

NIH has a longstanding commitment to making the results of NIH-funded research available. Responsible data management and sharing has many benefits, including accelerating the pace of biomedical research, enabling validation of research results, and providing accessibility to high-value datasets.

About the Data Management and Sharing Policy →



### Planning and Budgeting for Data Management & Sharing

Find out what NIH expects in a Data Management & Sharing plan and what costs are allowed in a request.



### Data Management

Proper data management is crucial for maintaining scientific rigor and research integrity. Learn about best practices for scientific data management.



### Sharing Scientific Data

Under the NIH Data Management & Sharing Policy, investigators are empowered to choose the most appropriate methods for sharing scientific data. Learn more about methods for data sharing and selecting data repositories.

## NIH Resources and Webinars

- [Supplemental guidance](#) on topics like finding a repository, protecting privacy when sharing human research participants data, etc.
- [Data Sharing and Reuse Seminar Webinar Series recordings](#)
- [Data Management and Sharing Policy Webinar Series recordings](#)
- [National Network of the National Library of Medicine \(NNLM\) classes](#)

<https://sharing.nih.gov/> (main page)

<https://sharing.nih.gov/faq> (FAQs)

## • Elements of an NIH Data Management and Sharing Plan

+ Data Type

+ Related Tools, Software and/or Code

+ Standards

+ Data Preservation, Access, and Associated Timelines

+ Access, Distribution, or Reuse Considerations

+ Oversight of Data Management and Sharing

## Example Plan:

### *Using natural language processing to determine predictors of healthy diet and physical activity behavior change in ovarian cancer survivors*

**Paraphrased Abstract:** This study examines whether artificial intelligence can use speech and audio from health coaching calls to predict who is most likely to enact healthy lifestyle behaviors of diet and physical activity.


This project uses previously collected data, recordings of conversations from the Lifestyle Intervention for Ovarian Cancer Enhanced Survival (LIVES) Study, to:

- 1) Develop a machine learning model to identify patterns in the interactions between coaches and their participants that signal a likelihood of optimal behavior change of healthy lifestyle and then
- 2) Decompose the machine learning model in terms of “intervenable factors” to see how well they predict a healthy lifestyle behavior change.





## • Data Type



- + Data Type

- + Related Tools, Software and/or Code

- + Standards

- + Data Preservation, Access, and Associated Timelines

- + Access, Distribution, or Reuse Considerations

- + Oversight of Data Management and Sharing



## • Data Type

NIH asks you to

“Briefly describe the scientific data to be managed, preserved, and shared”

- summary of the types and estimated amount of scientific data
- which scientific data from the project will be preserved and shared
- list of the metadata, other relevant data, and any associated documentation

By "final research data", we mean recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Final research data **do not include** laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

# NIH Guideline For Data Type

A general summary of the types and estimated amount of scientific data to be generated and/or used in the research. Descriptions may indicate

- the data modality (e.g., imaging, genomic, mobile, survey)
- level of aggregation (e.g., individual, aggregated, summarized)
- the degree of data processing that has occurred (i.e., how raw or processed the data will be)

# Sample Text For Data Type

Ref.	Description	Number of Files	Scale	Data type	Preserved	Shared
A)	Audio recordings	1048	50 GB	WAV	5 years	No
B)	Patient demographics, clinical reported outcomes, and patient-reported outcomes	1	100 KB	Excel	5 Years	No
C)	300-second snippets obtained from the LiVES telephone calls. Used for personality analysis	400	1 GB	MP3	5 years	No
D)	Numerical responses from ~4 annotator's personality perceptions obtained from (D)	500	200 MB	CSV	5 years	No
E)	Aggregated speaker-turn annotations for inter-annotator agreement analysis	50	50 MB	RTTM	5 years	No
F)	Linguistic and call content annotations of a subset of the LiVES telephone calls	85	450 MB	JSON	5 years	No
G)	machine-learned models for processing predicting patient outcomes	4	12 GB	Binary or .tflite	10 years	Yes, ReDATA
H)	Computer code for the creation of machine-learned models (G)	10	200 MB	Python files	10 years	Yes, ReDATA

# Data Type: Metadata and Documentation

A set of data that describes and gives information about other data



Data dictionaries  
and codebooks



Standard operating procedures



Self-describing  
file formats



README files



Commented code

# NIH Guideline and Sample Text

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.

We will make accessible all the computer code used to generate the machine-learned models (I) and the **necessary documentation** to use the computer code and models. No other metadata will be made accessible.

**See Data Services' self-paced online module for more information on documenting your research data**

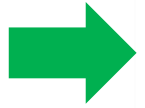
**Documenting Your Research Data**  
learning modules

# Questions to Ask Yourself



- What kind of data will I collect/generate? How much? What file format(s)?
- Will I publicly share any of my data? If not, why?
- Will I use secondary data? Am I allowed to share?
- Do I have code that I can share, such as the code used for analysis?
- What documentation do I need to create so data can be validated and replicated?

## • Related Tools, Software and/or Code



+ Data Type

+ Related Tools, Software and/or Code

+ Standards

+ Data Preservation, Access, and Associated Timelines

+ Access, Distribution, or Reuse Considerations

+ Oversight of Data Management and Sharing



# • Why Share Related Tools, Software and/or Code?

- Why share tools, software and/or code if data are shared already?
  - To reproduce your results
  - To reuse your data
  - To build upon your research
  - To increase the citation of your paper
  - To make your research transparent

```
base64.cc
31 void base64_encode(const uint8_t * data, size_t len, char * dst)
32 {
33     size_t src_idx = 0;
34     size_t dst_idx = 0;
35     for (; (src_idx + 2) < len; src_idx += 3, dst_idx += 4)
36     {
37         uint8_t s0 = data[src_idx];
38         uint8_t s1 = data[src_idx + 1];
39         uint8_t s2 = data[src_idx + 2];
40
41         dst[dst_idx + 0] = charset[(s0 & 0xfc) >> 2];
42         dst[dst_idx + 1] = charset[((s0 & 0x03) << 4) | ((s1 & 0xf0) >> 4)];
43         dst[dst_idx + 2] = charset[((s1 & 0x0f) << 2) | (s2 & 0xc0) >> 6];
44         dst[dst_idx + 3] = charset[(s2 & 0x3f)];
45     }
46
47     if (src_idx < len)
48     {
49         uint8_t s0 = data[src_idx];
50         uint8_t s1 = (src_idx + 1 < len) ? data[src_idx + 1] : 0;
51
52         dst[dst_idx++] = charset[(s0 & 0xfc) >> 2];
53         dst[dst_idx++] = charset[((s0 & 0x03) << 4) | ((s1 & 0xf0) >> 4)];
54         if (src_idx + 1 < len)
55             dst[dst_idx++] = charset[((s1 & 0x0f) << 2)];
56     }
}
```

3 selection regions Spaces: 4



## • Examples: Sharing Related Tools, Software and/or Code

We are going to show you two paragraphs from two published papers. Both described tools for analyzing data in the paragraph. You will read both of them and decide which one of them is more likely for other to reproduce their results.

You will answer the poll question after reading these paragraphs.

## • Example 1

*A custom MATLAB program was used to analyze audio data, and five parameters were applied to characterize the call design of an FM sonar vocalization. These five parameters are duration (ms), bandwidth (kHz), start and end frequencies of the FM sweep (kHz) and sweep rate (kHzms<sup>-1</sup>), all taken from the fundamental. Sweep rate is calculated by dividing bandwidth by duration and describes the slope of the FM call. Data analysis of video recordings was accomplished by digitizing the position of each bat and microphone and reconstructing the 3-D flight path via another custom MATLAB program.*

*~ Journal of Experimental Biology 2009 212: 1392-1404; doi: 10.1242/jeb.027045*

## • Example 2

*We built multiple hierarchical models to examine how increased temperature and other factors influenced shrub seedling recruitment, growth and mortality, as well as tussock grass gap dynamics. For each model, we used Bayesian inference and fitted models in 3.3.2 (R Core Team, 2016) using package rstan 2.14.1 (Stan Development Team, 2016). Detailed information about experimental design and analysis is provided in Supporting Information. Data and source code are available at: [https://github.com/jscamac/Alpine Shrub Experiment](https://github.com/jscamac/Alpine_Shrub_Experiment). To aid in the reproducibility of this work, our code was written using a remake framework (FitzJohn, 2015). This allows others to readily reproduce our entire workflow from data processing, through to producing a pdf of this manuscript by calling `remake::make`. To safeguard against cross-platform issues and future software changes, we have embedded this framework within a Docker image ([https://hub.docker.com/r/jscamac/alpine shrub experiment](https://hub.docker.com/r/jscamac/alpine_shrub_experiment)).*



## • Related Tools, Software and/or Code

Why do you think you can or cannot reproduce these two examples?

- Related tools, software and code to analyze data are not available in Example 1
- Example 2 provides related tools, software and code to analyze their data

## Proprietary Tools/Software

When possible, choose open source tools over proprietary ones or proprietary software but can export files to non-proprietary formats



## • Related Tools, Software and/or Code Examples

- Python scripts for data cleaning
- R scripts to conduct statistical analysis
- Jupyter Notebook documenting data analysis process
- Software downloaded from CDC website for data visualization (provide a link to download the software)
- Proprietary software used for data analysis (cannot be shared but can provide a link to purchase)
- REDCap survey template
- R packages/Python libraries used in data analysis
- MATLAB scripts for data processing
- SQL query to get data from EPIC

# NIH Guideline and Sample Text

- 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.

The machine-learned models will be distributed for their use with the **Python (v.3.7+)** programming language and will require one of the following free Python machine learning libraries (final decision pending): **PyTorch (v.1.10.2)** or **TensorFlow (v.2.8.0)**.

- 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.

**Python (<https://www.python.org/>)\***, **PyTorch (<https://pytorch.org/>)**, and **TensorFlow (<https://www.tensorflow.org/>)** are all freely accessible for most modern computers and at the moment there are no plans to discontinue their support. More information is available at their respective websites.

\*Do not include URLs or hyperlinks in your NIH proposals or DMSPs



# Questions to Ask Yourself



- What tools/software do I use to process, analyze and visualize my data?
  - Are they open source?
  - Can I share my scripts?
  - Can I convert my file formats to non-proprietary file formats?
- When sharing tools/software, provide the following information:
  - License
  - Source website/GitHub page
  - Version
  - Computational environment

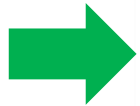
## Standards



+ Data Type



+ Related Tools, Software and/or Code



+ Standards

+ Data Preservation, Access, and Associated Timelines

+ Access, Distribution, or Reuse Considerations

+ Oversight of Data Management and Sharing

## • What are Standards?

### What are standards?

- (Merriam-Webster): something established by authority, custom, or general consent as a model or example
- (Merriam-Webster): something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality



It's 25 degree outside.  
What a beautiful day!

Twenty-five degree!?  
It's freezing!

What's the problem here?

No **standard** unit for measuring temperature



## • What are Data Standards?

What are data standards?

- [NCATS](#): Data standard is an agreed upon set of rules that allow information to be shared and processed in a uniform and consistent manner.
- National Library of Medicine (NLM): [Health Information Technology and Health Data Standards](#)

Standardized ways to document and describe data to ensure interoperability of data across different systems and users.

## Standards Example: COVID-19 Case Classification

Value	Code Name	Code	Code System	Code Description	Actions
Lab-confirmed case	Lab-confirmed case	C36292:C25458	NCI Thesaurus	The outcome of a laboratory test.:Having been established or verified. C36292:C25458	
Point-of-care Confirmed	Point-of-care Confirmed	C154479:C25458	NCI Thesaurus	Services designed to be administered at the patient's bedside or other patient location.:Having been established or verified. C154479:C25458	
Probable Case Based on Clinical and Epidemiologic Evidence	Probable Case Based on Clinical and Epidemiologic Evidence	C178498	NCI Thesaurus	The designation of a case as probable based on clinical criteria and epidemiologic evidence, without confirmatory laboratory results. C178498	
Probable Case Based on Laboratory Results and Clinical or	Probable Case Based on Laboratory Results and Clinical or	C178499	NCI Thesaurus	The designation of a case as probable based on presumptive laboratory evidence and either clinical criteria or epidemiologic evidence. C178499	

<https://cde.nlm.nih.gov/deView?tinyId=QnDNUPpXA7>

## Where to Find Data Standards?



<https://cde.nlm.nih.gov/home>



<https://rd-alliance.github.io/metadata-directory/standards/>



<https://fairsharing.org/search?fairsharingRegistry=Standard>



<https://www.dcc.ac.uk/guidance/standards/metadata/list>

### OMOP on PMAP

Bringing global data standardization to Johns Hopkins.

PMAP is the platform. OMOP is the data.

OMOP on PMAP makes it easier to analyze outcomes at a large scale.

<https://pm.jh.edu/how-it-works/omop/>

Look for data standards on your **funder's** or discipline-specific **data repository's** website

# NIH Guideline and Sample Text

- 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).

## Data and metadata formats:

**Microsoft EXCEL format** is used for patient demographics, clinical reported outcomes, and patient-reported outcomes. We prefer this because these files are accessed often for consultation and EXCEL offers a more flexible interface than other table formats such as CSV.

**WAV format** is used for storing LiVES telephone audio files. MP3 format is used for storing audio snippets (A). We use this format to save space and accessing time.

**JSON format** is used for phone interview metadata, speaker-turn annotations, and linguistic and call content annotations (C) and (E).

# NIH Guideline and Sample Text

- 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).

**Other data:** Data dictionaries for the original LIVES data are stored online in the UA REDCap instance and can be exported as CSV files if necessary. Any data dictionaries resulting from the current study are stored as CSV files in UA Box Health. Data identifiers and unique identifiers such as participant IDs are stored in encrypted Microsoft EXCEL files in the HIPAA-compliant cloud storage UA Box Health. Variable names and definitions for patient demographics, clinical reported outcomes, and patient-reported outcomes are stored in a Microsoft EXCEL file in UA Box Health.



# Questions to Ask Yourself



- Are there any standards in my field?
  - Common vocabularies?
  - Standardized methods to collect data?
  - Common file formats or schemas?
- Do I have a data dictionary/codebook for my data?
- What information can I provide so others can easily understand and reanalyze my data?

# • Data Preservation, Access, and Associated Timelines



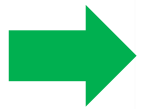
+ Data Type



+ Related Tools, Software and/or Code



+ Standards



+ Data Preservation, Access, and Associated Timelines

+ Access, Distribution, or Reuse Considerations

+ Oversight of Data Management and Sharing



## • Data Preservation, Access, and Associated Timelines

- The name of the repository(ies) where scientific data and metadata arising from the project will be archived
- How the scientific data will be findable and identifiable
- When the scientific data will be made available to others and for how long

# FAIR Principles

## Findable



- Descriptive keywords
- Persistent Identifier (DOI)

## Accessible



- Easy to retrieve by machines and humans
- Data in a repository



## Interoperable

- Open formats
- Consistent vocabulary



## Re-Usable




- Clear reuse licenses
- Good documentation



- FAIR Principles

Depositing Your Data Into a Repository  
Improves the FAIRness of Your Data

## Types of Data Repositories

	Discipline-Specific	Generalist
Open Access		
Controlled-Access		

## Select a Data Repository

*Scenario 1:*

Your ICO or FOA requires a **particular repository**



*Scenario 2:*

**Discipline-specific repository** exists for your research field

*Scenario 3:*

No discipline-specific repository exists or your research is **cross-disciplinary, general repository**



Johns Hopkins Research Data Repository



NIH Guidance: [Selecting a Repository for Data Resulting from NIH-Supported Research](#)  
JHU Guidance: <https://browse.welch.jhmi.edu/nih-data-management/repository-selection>

## • How to Find a Data Repository



- A list of NIH-supported data repositories
- Most accept data from NIH-funded projects (and others), with some exceptions

[https://www.nlm.nih.gov/NIHbmic/nih\\_data\\_sharing\\_repositories.html](https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html)



- A registry of research data repositories
- Search for a repository appropriate to host your data

<https://www.re3data.org/>



Data Services



<https://archive.data.jhu.edu>

- Available for JHU Researchers (SOM, SPH, KSAS, WSE, SOE)
- Free for data smaller than 1 TB



*Still need help finding?*  
Just ask us at JHU Data Services!

[dataservices@jhu.edu](mailto:dataservices@jhu.edu)

[dataservices.library.jhu.edu/](https://dataservices.library.jhu.edu/)



# Johns Hopkins Research Data Repository for OPEN Data



## Johns Hopkins Research Data Repository

JH Research Data Repository <https://archive.data.jhu.edu> available for JHU Researchers.

Contact [dataservices@jhu.edu](mailto:dataservices@jhu.edu) to deposit data.

More information about JH Research Data Repository and FAQs: <http://dataservices.library.jhu.edu/archiving/>

Search this dataverse...



Advanced Search

**Dataverses (48)**

**Datasets (168)**

**Files (3,599)**

### Dataverse Category

Research Project (22)

Research Group (9)

Laboratory (2)

Researcher (2)

Organization or Institution (1)

### Publication Year

2023 (1)

1 to 10 of 216 Results

Sort

Data associated with the publication: Optical properties of organic hazes in water-rich exoplanet atmospheres: Implications for observations with JWST



Jan 6, 2023

He, Chao; Radke, Michael; Moran, Sarah E.; Hörst, Sarah M.; Lewis, Nikole K.; Moses, Julianne I.; Marley, Mark S.; Kempton, Eliza M.-R.; Morley, Caroline V.; Valenti, Jeff A.; Vuitton, Véronique, 2023, "Data associated with the publication: Optical properties of organic hazes in water-rich exoplanet atmospheres: Implications for observations with JWST", <https://doi.org/10.7281/T1/NEACHP>, Johns Hopkins Research Data Repository, V1, UNF:6:8GgIE9In2P+hRwyD8zjoeQ== [fileUNF]

We report the optical properties of two haze analogous to those produced in temperate water-rich exoplanet atmospheres. Their optical constants (the real refractive indices,  $n$ , and the extinction coefficients,  $k$ ) are derived from 0.4 to 28.6  $\mu\text{m}$ , covering optical wavelengths acces...

# NIH Guideline and Sample Text

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

The computer code, machine-learned models, and documentation will be made publicly available through the **University of Arizona Research Data Repository (ReDATA )** and through the **HuggingFace website** (widely used within the machine-learning community) under an Apache 2.0 License. Although there are NRG-oncology approved repositories, most of them deal with specimens and images and none with machine-learned models of language interactions and so none of them were a good fit for this project.

## • Persistent Unique Identifier

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

- A persistent identifier is a long-lasting reference to a document, file, web page, or other object
- DOI (digital object identifier)

**doi:10.7281/T1/QQ770B**



- Journals often mint one for your article, but usually ask for a DOI for your data
- Many repositories, such as the [Johns Hopkins Research Data Repository](#), will mint one for your data

# NIH Guideline and Sample Text

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

The machine-learned models and associated metadata will be findable on ReDATA through a **Digital Object Identifier (DOI)** as well as a search engine at <https://arizona.figshare.com/>. It will also be findable through the search engine at <https://huggingface.co/models>



## • Associated Timelines: NIH and JHU Policy

- NIH data management and sharing policy states “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.”
- JHU data retention policy states “Research Data should be stored ... for a minimum period of 5 years after the date of any publication” (New policy will require 7 years of data retention)

# NIH Guideline and Sample Text

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?

The machine-learned models and code will be available from December 2022 and for at least 10 years (as guaranteed by ReDATA). We cannot guarantee availability on the other two repositories.

# Questions to Ask Yourself



- Will I publicly share any of my data? If not, why not?
- When will I share my data with others?
- Where should I share and preserve my data (e.g., repository name)?
- What policies around data preservation am I beholden to, even if I am unable to publicly share my data?



## • Access, Distribution, or Reuse Considerations



+ Data Type



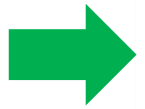
+ Related Tools, Software and/or Code



+ Standards



+ Data Preservation, Access, and Associated Timelines



+ Access, Distribution, or Reuse Considerations

+ Oversight of Data Management and Sharing





## • Access, Distribution, or Reuse Considerations

- Factors affecting subsequent access, distribution, or reuse of scientific data
- Whether access to scientific data will be controlled
- Protections for privacy, rights, and confidentiality of human research participants



## • Describe Limitations to Data Sharing

*Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data.*

This includes:

- Informed consent
- Privacy/confidentiality protections
- Data or code licenses

See [Frequently Asked Questions](#) for examples of justifiable reasons for limiting sharing of data.

## Informed Consent

Regardless of where you plan to share your data, your participants need to understand

- **how** others can access it (e.g., open, DUA),
- **level of deidentification** and
- the **relative risk of reidentification**.

- JHU SOM IRB
  - [Boilerplate](#) for data sharing
  - [Organization Policy](#) on Informed Consent Process and Documentation
- JHU BSPH IRB Consent form [templates](#)
- NIH's Informed Consent template ([link](#))



*Work with IRB to make sure your consent language makes sense given the repository that you are sharing your data in before you collect data.*

## Human Participants Consideration For Data Sharing

- How to decide where to share human participants data?
  - Share fully [de-identified data sets](#) via an **open access data repository**
    - May need Data Trust approval
    - Consent form\* which allows for data sharing
  - Share limited data sets ([LDS](#)) via a **controlled-access data repository**
    - IRB approval
    - Data Use Agreement (DUA): JHU Research Administration can help with this

\* Meyer, M. N. (2018). Practical Tips for Ethical Data Sharing. *Advances in Methods and Practices in Psychological Science*, 1(1), 131–144.  
<https://doi.org/10.1177/2515245917747656>



## • Data and Code License

For ***secondary data***, be aware of any existing license restrictions and mention those in your Plan if applicable

For ***primary data***, you may be able to assign a particular license to your work.

Conditions include:

- Requirement of data/code citation
- Possibility of commercial use
- Ability for others to modify your data/code

As a general rule, you will want to have the most permissive license as possible.



## • Data and Code License

- License for data
  - Creative Commons: <https://creativecommons.org/about/cclicenses/>
  - Creative Commons License Chooser: <https://creativecommons.org/choose/#>
- License for software
  - Types of software licenses: <https://choosealicense.com/licenses/>
  - Choose a software license: <https://choosealicense.com/>

Your repository most likely will either let you choose a particular license or automatically assign one to your data

## Whether Access to Scientific Data will be Controlled

**Public Sharing:** completely open with no restrictions in terms of finding and downloading data.



Johns Hopkins Research Data Repository

**Controlled Sharing:** Different repositories have differing levels of control. Examples of controls include a requirement that people sign a DUA to get access or requirement that a human review the data requestors request.



[Data Sharing Tiers for Broad Sharing of Clinically Derived Data](#) from Data Trust

## • Protections for privacy, rights, and confidentiality

Outline the steps you will take for protecting the privacy, rights, and confidentiality of prospective participants (i.e., through de-identification, [Certificates of Confidentiality](#)).

If you are using a controlled-access repository, discuss the procedures and policies of that repository. For example, a repository may:

- Require all data be deidentified to the Safe Harbour level
- Only share data with researchers once they have been vetted or signed a DUA
- Use HIPPA-compliant storage system

Supplemental Information to the NIH Policy for Data Management and Sharing:  
[Protecting Privacy When Sharing Human Research Participant Data](#)



# NIH Guideline and Sample Text

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

In accordance with the terms of our IRB-approved research plan, we will only share the **resulting ~10 machine-learned models and associated computer code (G-I)**. We'll ensure that we create models where no HIPAA or other text information is recoverable. The models will be made publicly available under the **Apache 2.0 license** with no restrictions on academic or commercial re-use.

# Questions to Ask Yourself



- How do I want others to use and cite my data/software? Which license fits my requirements best?
- Are there Protected Health Information (PHI), Personally identifiable information (PII) or sensitive information in my data?
- Can I fully remove them?
- If not, what restriction I can apply to my data in order to protect the privacy of my research participants?

## Oversight of Data Management and Sharing



+ Data Type



+ Related Tools, Software and/or Code



+ Standards



+ Data Preservation, Access, and Associated Timelines



+ Access, Distribution, or Reuse Considerations



+ Oversight of Data Management and Sharing

# NIH Guideline and Sample Text

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

**The PIs** for this project, Tracy Crane and Steven Bethard, will ensure that the data management plan is followed by auditing the project personnel on a monthly basis and monitoring the project through an online project management tool (Trello).

Sarah Jane Wright is one of the **data liaisons** between the LIVES project and the current project. She is in charge of the patient outcome data, REDCap, questionnaire data, and patient personal records and identifiers. Sarah ensures that sensitive data is accessed on a case-by-case basis in a secure way through REDCap.

# Questions to Ask Yourself



- Who will be in charge of the oversight of this DMSP?
  - Usually this will be the PI and co-PI(s)
- Who will be in charge of different tasks in this DMSP?
  - List each staff member and the task they are in charge of
    - De-identification
    - Day-to-day management
    - Uploading of data to chosen repository
    - Documentation

## • Allowable Costs for Data Management and Sharing

- Curating data and developing supporting documentation
- Local data management considerations
- Preserving and sharing data through established repositories
- De-identifying data


NIH Guidance for [Budgeting for Data Management & Sharing](#)

JHU Guidance for [Estimating Data Management and Sharing Costs](#)

# Sign in DMPTool: <https://dmptool.org>

The screenshot shows the DMPTool website homepage. At the top, the navigation menu includes "Funder Requirements", "Public DMPs" (highlighted with a red box), and "Help". The main banner features the text "Create Data Management Plans that meet requirements and promote your research". On the right, there is a "Sign in / Sign up" form with an "Email address" field and a "Continue" button. Below the banner, three statistics are displayed: "72,852 Users", "327 Participating Institutions", and "71,007 Plans". At the bottom right, there is a "Latest News from DMPTool" section with a "View all news" link and social media icons for Twitter and RSS.

dmptool.org

 **DMPTool**

Build your Data Management Plan

Funder Requirements **Public DMPs** Help

Language ▼

Create Data Management Plans that meet requirements and promote your research

Sign in / Sign up

Email address \*

Continue

Problems signing in? Contact us.

Latest News from DMPTool

Things to know about the updated DMPTool website

View all news

72,852 Users

327 Participating Institutions

71,007 Plans

# Create a New Plan

## Create a new plan

Before you get started, we need some information about your research project to set you up with the best DMP template for your needs.

### \* What research project are you planning?

NIH DMSP workshop demo

mock project for testing, practice, or educational purposes

### \* Select the primary research organization

Research organization

Johns Hopkins University (jhu.edu)

- or -  No research organization associated with this plan or my research organization is not listed

### \* Select the primary funding organization

Funder

National Institutes of Health (nih.gov)

- or -  No funder associated with this plan or my funder is not listed

### Which DMP template would you like to use?

NIH-GEN DMSP (Forthcoming 2023)

We found multiple DMP templates corresponding to your funder.

Create plan

Cancel



# Write Your Plan

## NIH DMSP workshop demo

Project Details

Collaborators

Write Plan

Research outputs

Request feedback

Download

Finalize / Publish

This plan is based on the "NIH-GEN DMSP (Forthcoming 2023) " template provided by National Institutes of Health (nih.gov) - (ver: 3, pub: 2022-03-14).

[expand all](#) | [collapse all](#)

0/12

+ Data Type (0 / 3)

+

+ Related Tools, Software and/or Code (0 / 2)

+

+ Standards (0 / 1)

+

+ Data Preservation, Access, and Associated Timelines (0 / 3)

+

+ Access, Distribution, or Reuse Considerations (0 / 2)

+

+ Oversight of Data Management and Sharing (0 / 1)

+

Briefly describe the scientific data to be managed, preserved, and shared.

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)

**B** *I* [List Icon] [List Icon] [Link Icon] [Table Icon]

Box to type your text

Save

Example answer

DMPTool Fill-In-The-Blank and Sample Answers

**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**

Comments & Guidance

	Guidance	Comments
NIH	JHU	DMPTool

**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 using [this 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.

# Send Your Plan for Feedback

## NIH DMSP workshop demo

Project Details

Collaborators

Write Plan

Research outputs

Request feedback

Download

Finalize / Publish

### Request expert feedback

Click below to give data management staff at Johns Hopkins University (jhu.edu), the Plan Owner's org, access to read and comment on your plan.

Your draft data management plan (DMP) has been sent to JHU Data Services. One of our consultants will provide feedback on your DMP within 2 business days.

Thank you,

JHU Data Services

(<https://dataservices.library.jhu.edu/>)

[dataservices@jhu.edu](mailto:dataservices@jhu.edu)

You can continue to edit and download the plan in the interim.

Request feedback

# Download Your Plan

## NIH DMSP workshop demo

Project Details

Collaborators

Write Plan

Research outputs

Request feedback

Download

Finalize / Publish

### Format

pdf

### Download settings

Optional plan components

- project details coversheet
- section headings
- question text (will only display your answers)
- unanswered questions

### PDF formatting

#### Font

Face

"Times New Roman", Times, Serif

Size (pt)

11

#### Margin (mm)

Top

25

Bottom

25

Left

25

Right

25

Download Plan

# Where to Find Sample Plans?

- [Sample Plans](#) from NIH (e.g., clinical, secondary data, genomic, survey)
- [DMPTool's](#) ongoing collection of [publicly shared data management plans](#) that can be filtered by funder, institution and subject
- DMP examples from [University of Arizona](#)
- [Example DMS Plans on GitHub](#)



## • Observations After Three Weeks of Reviewing Plans

- Using the sample language is fine, but make sure that it makes sense in your Plan
- Remember that NIH wants you to share enough data and documentation so that your research can be reproduced and validated
  - If you do not want to share some parts of your data, a justification for not sharing should be provided ([What are justifiable reasons for limiting sharing of data?](#))
- Give yourself time to write your Plan and plan enough time for Data Services to review your Plan
- One Plan per grant application, doesn't matter how many data products you are going to generate nor individual projects under that application.

# How to Start?

- Use the DMPTool\* and write your Plan, element by element
- Use the "Request Feedback" function to notify us that you want a review
- We will provide feedback within two business days
- Upload your Plan in pdf to the new "Other Plan(s)" field in FORMS-H

\* You don't have to use the DMPTool; just email your draft to us at [dataservices@jhu.edu](mailto:dataservices@jhu.edu)



JOHNS HOPKINS  
LIBRARIES

Data Services

# JHU Resources for NIH Policy

- [Overview of policy and JHU resources/contacts](#) (JHURA)
- [NIH DMSP support](#) and [flyer](#) (Data Services)
- [Guide for NIH Data Management and Sharing Policy](#) (Welch and Data Services)



JOHNS HOPKINS  
LIBRARIES

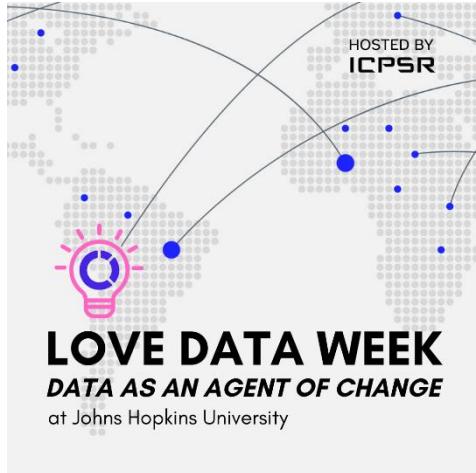
---

Data Services



# Sign up for other Data Services workshops!

<https://jhu.libcal.com/calendar/workshops>



Topics include:

- Love Data Week
- ArcGIS and mapping
- Data cleaning and visualization with R or Python
- De-identifying human subject data
- Data cleaning with Open Refine

Register at: <https://jhu.libcal.com/calendar/workshops>

## Contact JHU Data Services

### GO TO

[dataservices.library.jhu.edu](https://dataservices.library.jhu.edu)

### EMAIL

[dataservices@jhu.edu](mailto:dataservices@jhu.edu)

### SHARE DATA AT

[archive.data.jhu.edu](https://archive.data.jhu.edu)

## Helping you



FIND



USE



MANAGE



VISUALIZE



SHARE

# DATA

SURVEY

<https://www.surveymonkey.com/r/NIH-DMSP>



JOHNS HOPKINS  
LIBRARIES

Data Services