# Fred Hutch Data Science Lab DMSP Guidance for Researchers

[Fred Hutch Data Science Lab Guidance for Researchers 1](#_Toc120802466)

[Introduction 1](#_Toc120802467)

[Element 1: Data Type 1](#_Toc120802468)

[A. Types and amount of scientific data expected to be generated in the project 1](#_Toc120802469)

[B. Scientific data that will be preserved and shared, and the rationale for doing so 3](#_Toc120802470)

[C. Metadata, other relevant data, and associated documentation 3](#_Toc120802471)

[Element 2: Related Tools, Software and/or Code 4](#_Toc120802472)

[Element 3: Standards 4](#_Toc120802473)

[Element 4: Data Preservation, Access, and Associated Timelines 4](#_Toc120802474)

[A. Repository where scientific data and metadata will be archived 5](#_Toc120802475)

[B. How scientific data will be findable and identifiable 6](#_Toc120802476)

[C. When and how long the scientific data will be made available 6](#_Toc120802477)

[Element 5: Access, Distribution, or Reuse Considerations 6](#_Toc120802478)

[A. Factors affecting subsequent access, distribution, or reuse of scientific data 6](#_Toc120802479)

[B. Whether access to scientific data will be controlled 7](#_Toc120802480)

[C. Protections for privacy, rights, and confidentiality of human research participants 7](#_Toc120802481)

[Element 6: Oversight of Data Management and Sharing 7](#_Toc120802482)

## Introduction

This guidance was adapted from NIH DMSP Guidance Working Group, <https://osf.io/uadxr/> with contributions from the Fred Hutch community.

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on [sharing.nih.gov.](https://sharing.nih.gov/) The Plan is recommended not to exceed two pages. There is no “form page” for the Data Management and Sharing Plan.

## Element 1: Data Type

### A. Types and amount of scientific data expected to be generated in the project

Summarize the types (for example, 256-channel EEG data and fMRI images) and amount (for example, 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.

*Fred Hutch Data Science Lab practical guidance:*

Depending on the data type, you’ll need to describe the data in slightly different ways. Here are some examples.

* Genomics:

Genomic data come in a variety of shapes and sizes depending on the actual data type. Be specific about which type of genomic data you’ll be generating or using. File sizes for each “sample” for a general genomics assay and run vary considerably depending on assay properties and experimental design (such as depth of coverage targeted). Per-samples sizes listed here are averages we’ve seen at the FHCC Genomics Shared Resource:

|  |  |  |  |
| --- | --- | --- | --- |
| **Assay** | **File Type** | **Per sample Size** | **Notes** |
| Bulk RNA-seq | Paired Fastq | 2-4G | Depends on library prep & goals |
| RNA Exome | Paired Fastq | 3G |  |
| Whole Exome | Paired, Fastq | 3G | HS platform dependent |
| CRISPR | Single Fastq | ≥500M | sgRNA library dependent |
| CUT&RUN | Paired Fastq | ≥500M | Ab dependent |
| CUT&Tag | Paired Fastq | ≥500M | Ab dependent |
| ChIP-seq | Fastq | 0.5-5G | Ab dependent |
| ATAC-seq | Fastq | 3-5G |  |
| 10x scRNA-seq | Paired Fastq | 10G | Target cell number dependent |
| 10x Multiome | Paired Fastq | ≥20G | Target nuclei number dependent |
| 10x Visium | Paired Fastq | ≥5G | Spots under tissue dependent |
| Small Genome | Paired Fastq | ≥2G | Genome size dependent |
| PacBio Amplicon | CCS BAM | 0.5-20G | Amplicon size & target depth dependent |
| PacBio Small Genome | CCS BAM | highly variable | Genome size dependent |

* Cellular imaging
* Clinical datasets
* Proteomics data
* Flow cytometry data
* Clinical Trials data

### 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. 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.

*Fred Hutch Data Science Lab practical guidance:*

Things to consider for this section include:

* What versions of what data are most relevant to share in order for someone else to validate the results of your study?
* What is a useful file type or stage of processing where enough has been done to the data in order for them to be shareable in such a way that with sufficient access and metadata, another group could leverage the data for another study or to replicate yours?
* Perhaps raw data isn’t what is appropriate to share due to legal or ethical implications of the entire dataset, but analyzed, interpreted or summarized aspects of those data ARE acceptable to share and can be done so in compliance with any applicable restrictions.
  + Examples of this include:
    - not sharing normal exome bam files generated with the express intention of being used as a reference for tumor exome bam file analyses (and the patient did not consent to sharing of their non-tumor related data).
    - Individual level data may not be ethical to share due to privacy issues but aggregate data can be shared and still be in compliance.

### 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.

* 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
* **For data subject to the GDS Policy:**
  + Data types expected to be shared under the GDS Policy should be described in this element. Note that the GDS Policy expects certain types of data to be shared that may not be covered by the DMS Policy’s definition of “scientific data”. For more information on the data types to be shared under the GDS Policy, consult [Data Submission and Release Expectations](https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/data-submission-and-release-expectations).

*Fred Hutch Data Science Lab practical guidance:*

Keep in mind what types of metadata (data about the data themselves) might be needed for another group to be able to leverage the data you share.

## Element 2: Related Tools, Software and/or Code

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

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.

*Fred Hutch Data Science Lab practical guidance:*

* GitHub is an excellent resource to store, share and manage versioning of research software and code based assets that are involved in your data analysis.
* For non-open source software and code, recording specific instructions about how the software was obtained can help others access those software resources in the future.

## 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.

*Fred Hutch Data Science Lab practical guidance:*

Often you can learn about what standards exist for your data from the repository (if there is one) where you intend to deposit the data for sharing. When no clear standard exists for a data type, it is useful to think about the FAIR guiding principles for data being Findable, Accessible, Interoperable and Reusable (<https://www.nature.com/articles/sdata201618>). Sometimes knowing how you work with your own data type, the FAIR principles will help you identify what makes the most sense to share and in what format or what standard.

## Element 4: Data Preservation, Access, and Associated Timelines

Describe 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](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/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.
  + Note that 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.
* **For data subject to the GDS Policy:**
  + For human genomic data:
    - § Investigators are expected to submit data to a repository acceptable under the Genomic Data Sharing Policy. See [Where to Submit Genomic Data](https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/where-to-submit-genomic-data).
    - § Human genomic data is expected to be shared according to NIH’s [Data Submission and Release Expectations](https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/data-submission-and-release-expectations), but no later than the end of the performance period, whichever comes first.
  + For Non-human genomic data:
    - § Investigators may submit data to any widely used repository.
    - § Non-human genomic data is expected to be shared as soon as possible, but no later than the time of an associated publication, or end of the performance period, whichever is first.

### 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](https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository)).

*Fred Hutch Data Science Lab practical guidance:*

Some things to consider about where you will share your data include:

* What kind of data is it?
* What existing repositories exist for that type?
  + What metadata will be required for you to submit/deposit data there?
* If there is not a known repository, how will you design a sharing strategy?
  + What metadata do you think someone would need in order to leverage the data you share?

Here are some specific examples:

* Genomics data

Primary repositories for raw sequence data will be the Gene Expression Omnibus for data that can be made publicly available and dbGaP for data that require access controls. Both repositories are backed by the Sequence Read Archive (SRA) for storage of raw sequence data, typically in Fastq format although uBAM can also be submitted. The SRA is managed and supported by the NCBI, and there is no current charge for data hosting (although you’d figure this would have to change at some point).

In addition to raw data, both GEO and dbGaP can accept derived results (e.g. gene-barcode matrices, CRISPR sgRNA counts, etc.), but the supporting unmanipulated raw data is absolutely required.

The Fred Hutch Bioinformatics group can and has assisted with submission to both of these resources, typically on a fee-for-service basis. We can assist with assembly of metadata and with data transfer for GEO/SRA. For dbGaP the investigator is solely responsible for assembly and submission of metadata; this triggers generation of a metadata sheet and protected upload area from the SRA, which Bioinformatics can then assist with.

GEO, dbGaP, and SRA all provide stable IDs to various levels (Project accession, SRA read accession, sequencing platform, etc.). Primary references would be to a GEO series accession (e.g. GSE198265), dbGaP study accession (e.g. phs001805.v1.p1 to pick a random yet compelling example), or Sequence Read Archive run accession (e.g. SRR18284544).

### 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.

*Fred Hutch Data Science Lab practical guidance:*

### 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.

*Fred Hutch Data Science Lab practical guidance:*

## Element 5: Access, Distribution, or Reuse Considerations

### Factors affecting subsequent access, distribution, or reuse of scientific data

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See [Frequently Asked Questions](https://sharing.nih.gov/faqs%23/data-management-and-sharing-policy.htm) for examples of justifiable reasons for limiting sharing of data.

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 ICO will assess whether an applicant’s DMS plan appropriately considers and describes these factors. For more examples, see [Frequently Asked Questions](https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm) for examples of justifiable reasons for limiting sharing of data.

**Expectations for human genomic data subject to the GDS Policy:**

* Informed Consent Expectations:
  + § For research involving the generation of large-scale human genomic data from cell lines or clinical specimens that were created or collected **AFTER the effective date of the GDS Policy (January 25, 2015):**
  + § NIH expects that informed consent for future research use and broad data sharing will have been obtained. This expectation applies to de-identified cell lines or clinical specimens regardless of whether the data meet technical and/or legal definitions of de-identified (i.e. the research does not meet the definition of “human subjects research” under the Common Rule).
  + § For research involving the generation of large-scale human genomic data from cell lines or clinical specimens that were created or collected **BEFORE the effective date of the GDS Policy:**
  + § There may or may not have been consent for research use and broad data sharing. NIH will accept data derived from de-identified cell lines or clinical specimens lacking consent for research use that were created or collected before the effective date of this Policy.
* Institutional Certifications and Data Sharing Limitation Expectations:
  + § DMS Plans should address limitations on sharing by anticipating sharing according to the criteria of the [Institutional Certification](https://sharing.nih.gov/genomic-data-sharing-policy/institutional-certifications/about-institutional-certifications).
  + § In cases where it is anticipated that Institutional Certification criteria cannot be met (i.e., data cannot be shared as expected by the GDS Policy), investigators should state the institutional Certification criteria in their DMS Plan, explaining why the element cannot be met, and indicating what data, if any, can be shared and how to enable sharing to the maximal extent possible (for example, sharing data in a summary format). In some instances, the funding NIH ICO may need to determine whether to grant an exception to the data submission expectation under the GDS Policy.
* Genomic Summary Results:
  + § Investigators conducting research subject to the GDS Policy should indicate in their DMS Plan if a study should be designated as “sensitive” for the purposes of access to Genomic Summary Results (GSR), as described in [NOT-OD-19-023](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html).

*Fred Hutch Data Science Lab practical guidance:*

### 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).

*Fred Hutch Data Science Lab practical guidance:*

For sensitive data or those that originate from humans that pose a potential risk to the research participant, it is often the case that a review of the conditions under which the data can be shared based on patient or participant consent forms, or other legal or regulatory oversight is performed at your institution. These restrictions then can be placed on the data so they can be shared, but only under specific limitations that ensure the rights and privacy of those participants can be respected.

### 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).

*Fred Hutch Data Science Lab practical guidance:*

## 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).

*Fred Hutch Data Science Lab practical guidance:*

Remember that DMS Plans can and should be updated during multiple stages of a grant lifecycle, though this requires NIH approval. Designating who will be primarily responsible for updating the plan itself, ensuring compliance with it, and reporting on the work and processes involved during the annual RPPR process is important here. At the Fred Hutch we currently do not have plans to review all data management and sharing plans submitted by Fred Hutch staff, nor to ensure compliance. Currently the primary mechanism for oversight and compliance with these plans rests with the team executing the research that generates the datasets. It is important for you to describe a plan for how you will identify when data needs to be shared, define who will be responsible for work of processing and sharing of the data, and how often or at what specific milestones in the project you will assess your compliance with the plan you outlined.