2023 and me: An overview of the new NIH Data management and sharing policy

Marla Hertz, PhD | Associate Professor
Research Data Management Librarian
Office of Scholarly Communication
UAB Libraries | The University of Alabama at Birmingham
205.934.0694 | mihertz@uab.edu

@ OSC_UAB
After this webinar you will be able to:

➢ Understand
  ➢ The requirements of the NIH data management and sharing policy
  ➢ Which funding mechanisms are affected
  ➢ How the policy affects grant submission

➢ Access resources, templates and sample data management plans

➢ Develop your own Data Management and Sharing Plan (DMSP)
What is a data management and sharing plan?

- **Data**
  - Output/observation from any systematic investigation

- **Metadata**
  - Structured information about data

- **DMSP**
  - A document that outlines what types of data will be collected during a project and how they will be analyzed, stored, and shared

Image files will be stored in .tif format. Data from images will be extracted and analyzed using XYZ software. At the time of publication, data will be made publicly available on the ABC Data Repository and the accompanying software code will be accessible on GitHub.
What encompasses research data management?

Data Life Cycle

- Collect
- Process
- Analyze
- Preserve
- Share
- Reuse
- Plan

Data Management Sharing Plan
Why is NIH updating its DMS policy?

• To ensure research data management beings in the planning stages
  • Increased rigor, transparency and reproducibility in research

• To promote the sharing of scientific data
  • Accelerate the rate scientific discovery

FAIR resources

- https://gofair.us/
The NIH policy has two major requirements

✓ Policy effective on Jan 25, 2023

Required for all grants that generate scientific data:

1. Submission of a 2-page Data Management and Sharing Plan (DMSP)
2. Documentation of compliance with the approved plan (in annual RPPR)

- Description of the data and metadata, including which data will be shared
- Related tools, software, and code
- Standards for data and metadata
- Preservation and timelines
- Distribution and reuse considerations
- Oversight protocols
Policy scope – does this apply to my grant?

All research funded by NIH that **generates scientific data**, including:

- Extramural grants
- Contracts
- Intramural research projects
- All other NIH funding mechanisms

Policy does not apply to funding that does not generate data
How does the policy affect grant submissions?

• DMSP submitted with the proposal
• Can request funds for data management and sharing costs
• Does not impact scoring
• Reviewed by NIH staff
• The DMSP may be updated as the project evolves
• Non-compliance may impact future funding decisions for the recipient institution
What information does NOT fall under data sharing?

- Laboratory notebooks
- Preliminary analysis/ optimization experiments
- Completed case report forms
- Physical objects or specimens
- Drafts of papers, email correspondence or other communications with colleagues
Which data should be shared?

Policy defines 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.”

✓ Underlying data from published studies
✓ Data from a study even if not directly linked to a publication
✓ Null findings that do not result in a publication
Acceptable reasons to not share

**Legal**
Explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure

**Ethical**
Concerns around privacy or safety of research participants
Pre-existing consent policies or agreements prohibit sharing of participant-derived material

**Technical**
Digitization of datasets is impractical

Strength: Strong to Weak
The 6 elements of a DMSP

1. **Description of the data and metadata**, including which data will be shared
2. **Preservation and timelines**
3. **Distribution and reuse considerations**
4. **Oversight protocols**
5. **Standards for data and metadata**
6. **Related tools, software, and code**

Where and when will data be shared?

Explain any limits to sharing, controlled access to data, and protection measures

State which formal data standard you will follow or indicate there are no standards

Who in the research group is responsible for implementing, updating, and reporting on the plan

UCSF prepared templates:
- **Generic**
- **Clinical trial**
- **Social science**

Searchable catalog of over 150 DMPs
1. Data and metadata description

- What types of data will be collected/used?
  - Which of these will be shared?
  - Raw data or aggregated?
- What metadata will be included to facilitate interpretation?
1. Data and metadata description

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

This project will produce _________ [Data type, e.g., imaging, sequencing, experimental measurements] data generated/obtained from _________ [e.g., instrument, method, survey, experiment, data repository]. 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.

Template Language

Template language modified from Generic Template develop at USCF Libraries
1. Data and metadata description

B. Scientific data that will be preserved and shared, and the rationale for doing so:

Based on _______ [ethical, legal, technical] considerations, the following data produced in the course of the project will be preserved and shared: _____ [list] OR All data produced in the course of the project will be preserved and shared.

C. Metadata, other relevant data, and associated documentation:

To facilitate interpretation of the data, _______ [e.g., metadata, documentation, protocols, data collection instruments] will be shared and associated with the relevant datasets.
2. Related tools, software, and code

• Disclose any specialized tools that are needed to access and reuse the shared data

• Provide the names of specific software tools

• What is the availability of tools (open-source vs fee)

• Comment on the expected lifespan of the tools compared to lifespan of the data
2. Related tools, software, and code

- Sharing code
- GitHub - open source sharing
- Archive a copy of code in Zenodo to ensure long-term accessibility
  - Zenodo is a data repository that specializes in preserving software and code

https://zenodo.org/

https://github.com
3. Data standards

- State which formal data standard you will follow or indicate there are no formal standards
  - Data formats
  - Data dictionaries
  - Common data elements
- Increases interoperability
- Reduces amount of time tidying data

Dataset with inconsistent standards:

<table>
<thead>
<tr>
<th>Subject</th>
<th>weight</th>
<th>height</th>
<th>temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>6</td>
<td>98.6</td>
</tr>
<tr>
<td>2</td>
<td>70.3</td>
<td>6 ft</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>190</td>
<td>167.64</td>
<td>100.2</td>
</tr>
<tr>
<td>4</td>
<td>112</td>
<td>5.2</td>
<td>36.6</td>
</tr>
<tr>
<td>5</td>
<td>140</td>
<td>5.5 feet</td>
<td>37.7</td>
</tr>
</tbody>
</table>

NIH Common data elements repository
Example data dictionary

<table>
<thead>
<tr>
<th>Date</th>
<th>Health Area</th>
<th>Registered</th>
<th>Attended</th>
<th>Retention</th>
<th>FTS Neg</th>
<th>FTS Pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-Jul</td>
<td>Okola</td>
<td>18</td>
<td>16</td>
<td>89</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>28-Jul</td>
<td>Mvoua</td>
<td>19</td>
<td>12</td>
<td>63</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>1-Aug</td>
<td>Ngoya</td>
<td>13</td>
<td>11</td>
<td>85</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>2-Aug</td>
<td>Nlong</td>
<td>9</td>
<td>7</td>
<td>78</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4-Aug</td>
<td>Lobo</td>
<td>7</td>
<td>4</td>
<td>57</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>date</td>
<td>date of follow up visit in 2020</td>
</tr>
<tr>
<td>Health Area</td>
<td>location</td>
<td>region of study</td>
</tr>
<tr>
<td>Registered</td>
<td>int</td>
<td>number of subjects registered for follow up</td>
</tr>
<tr>
<td>Attended</td>
<td>int</td>
<td>number of subjects present at follow up</td>
</tr>
<tr>
<td>Retention</td>
<td>percent</td>
<td>percent of registrants that attended</td>
</tr>
<tr>
<td>FTS Neg</td>
<td>int</td>
<td>number of attendees that were negative on the FTS test</td>
</tr>
<tr>
<td>FTS Pos</td>
<td>int</td>
<td>number of attendees that were positive on the FTS test</td>
</tr>
</tbody>
</table>
4. Data preservation, access, timelines

- Where will data be shared?
  - Name repositories

- How will the data be identified?
  - unique identifiers

- When will data will be shared?
  - How long will data remain public
  - Note – it must be no later than the time of publication or end of performance period whichever comes first

Want to learn more?
- MIT tips on data formats and sharing
- Repository registry re3data.org
- NIH data repositories
“Will share upon request” reality check

DATA-SHARING BEHAVIOUR
Of almost 1,800 manuscripts for which the authors stated they were willing to share their data, more than 90% of corresponding authors either declined or did not respond to requests for data. Only about 7% of authors actually handed over data.

- Manuscripts with statement indicating data available on request: 1,792 manuscripts
- Authors who did not respond or declined requests for data: 1,670
- Authors who provided usable data: 120

©nature

Watson Nature 2022. https://doi.org/10.1038/d41586-022-01692-1
Take data preservation out of your hands and choose the right repository

- Maintains data integrity
- Increases data reuse (and citations thereof)

Subject-specific repositories

Generalist repositories

Repository registry re3data.org
5. Data distribution and reuse

• Describe any considerations that may affect the extent of data sharing
  • Legal
  • Ethical
  • Technical

• Instead of not sharing – consider sharing with access control

• For human subject data describe how you will protect the confidentiality of study participants
6. Oversight of the DMSP

• Who in the research group is responsible for the oversight?
  • Include: Name, Position Title, Host Institution, ORCID, email address
  • Typically the PI, could include Co-I, lab manager or senior researcher

• How will compliance be monitored and managed?
  • Orient new researchers to the plan during onboarding
  • Review and update plan annually
  • Report compliance in annual RPPR
Resources to write a DMSP

- Templates
- How to select a repository
- Allowable costs, budgeting
Template and example plans

NIH prepared template (download word document)

UCSF prepared templates:
- Generic
- Clinical trial
- Social science

Other resources:
- Searchable database of over 150 DMPs
- Example Rubric
- OSF toolkit with checklists for PIs, glossaries and other tools

DMPTool – platform to create a DMSP, includes template and example plans
Freedom to select a data repository - How do I chose?

157,693 proteomics publication on PubMed

http://www.proteomexchange.org/
NIH guidance on selecting a repository

NIH strongly encourages subject-specific, open access Data Sharing Repositories as a first choice.


Datasets up to **2 gigabytes**
- **PubMed Central**
  - Stores publication-related supplemental materials and datasets directly associated with publications.

Datasets up to **50 gigabytes**
- **Generalist Repositories**
  - Datasets associated with publications or otherwise and links to PubMed.
  - Options include DRYAD, OSF, Figshare, and Mendeley.

High priority datasets, **petabyte-scale**
- **Cloud Partners** (STRIDES Program)
  - Store and manage large scale, high priority NIH datasets.
  - Options include AWS, Google Cloud, and Microsoft Azure.
## Generalist repositories

Accept data regardless of data type, format, content or disciplinary focus

### Generalist Repository Comparison Chart

<table>
<thead>
<tr>
<th>Size limit</th>
<th>The Dataverse Project</th>
<th>DRYAD</th>
<th>figshare</th>
<th>Mendeley</th>
<th>OSF</th>
<th>Vivli</th>
<th>Zenodo</th>
</tr>
</thead>
<tbody>
<tr>
<td>File size</td>
<td>2.5GB</td>
<td>300GB/dataset</td>
<td>20GB/file</td>
<td>10GB/dataset</td>
<td>5GB/file</td>
<td>10GB/file</td>
<td>50GB/dataset</td>
</tr>
<tr>
<td>Storage space</td>
<td>1TB/researcher</td>
<td>No limit</td>
<td>No limit</td>
<td>No limit</td>
<td>No limit</td>
<td>No limit</td>
<td>No limit</td>
</tr>
<tr>
<td>Cost to researcher</td>
<td>Free</td>
<td>$120 for author</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
</tr>
</tbody>
</table>

Full Generalist Repository Comparison Chart [https://zenodo.org/record/7189481#.Y3PZuXbMKUk](https://zenodo.org/record/7189481#.Y3PZuXbMKUk)
# Desirable Characteristics for Data Repositories

<table>
<thead>
<tr>
<th><strong>Persistent Unique Identifiers</strong></th>
<th>Assigns datasets to a citable PUID to support data discovery and reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Long-term sustainability</strong></td>
<td>Long-term plan for managing data; builds on stable technical infrastructure &amp; funding; contingency plans for unforeseen events</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation</td>
</tr>
<tr>
<td><strong>Curation &amp; Quality Assurance</strong></td>
<td>Provides expertise to improve the accuracy and integrity of datasets and metadata</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Provides maximally open access, consistent with legal and ethical limits</td>
</tr>
<tr>
<td><strong>Free &amp; Easy</strong></td>
<td>Datasets and metadata accessible free of charge and with broadest possible terms of reuse</td>
</tr>
<tr>
<td><strong>Reuse</strong></td>
<td>Enables tracking of data reuse</td>
</tr>
<tr>
<td><strong>Secure</strong></td>
<td>Documentation of meeting accepted criteria for security to prevent unauthorized access or release of data</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td>Documentation of safeguards in compliance with applicable privacy, risk management &amp; continuous monitoring requirements</td>
</tr>
<tr>
<td><strong>Common Format</strong></td>
<td>Datasets and metadata can be downloaded, accessed, or exported in a standards-compliant format</td>
</tr>
<tr>
<td><strong>Provenance</strong></td>
<td>Maintains a detailed logfile of changes to datasets and metadata to ensure integrity</td>
</tr>
</tbody>
</table>

# Additional Considerations for Human Data

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity to Consent</td>
<td>Restricts dataset access to appropriate uses consistent with original consent</td>
</tr>
<tr>
<td>Restricted Use Compliant</td>
<td>Enforces submitters’ data use restrictions</td>
</tr>
<tr>
<td>Privacy</td>
<td>Documentation &amp; implementation of security techniques for human subjects’ data to protect from inappropriate access</td>
</tr>
<tr>
<td>Plan for Breach</td>
<td>Has security measures that include data breach response plan</td>
</tr>
<tr>
<td>Download Control</td>
<td>Controls and audits access to and download of datasets</td>
</tr>
<tr>
<td>Clear Use Guidance</td>
<td>Provides documentation describing restrictions on dataset access and use</td>
</tr>
<tr>
<td>Retention Guidelines</td>
<td>Provides documentation on guidelines for data retention</td>
</tr>
<tr>
<td>Violations</td>
<td>Has plans for addressing violations of terms-of-use and data mismanagement by the repository</td>
</tr>
<tr>
<td>Request Review</td>
<td>Established data access review or oversight group responsible for reviewing data use requests</td>
</tr>
</tbody>
</table>

Allowable costs

• Data curation and development of documentation
• Local data management while project is ongoing
• Preservation and sharing in repositories

• Important consideration for long term preservation of data – funding must be used during the pay period of the grant

Forecasting costs

- Ballpark estimate 5% of the total budget for data management
- Study report forecasting long term costs
- Identify the current and potential value of the data over time
- Budget for the major cost drivers

How-to video
Task Force formed by VPR office
Raise awareness, oversee implementation, and assess impact of the policy on UAB campus

Jonathan Miller, Associate VP for Research Regulatory Oversight, Office of Research (Task Force Chair)
Peggy Biga, PhD, Associate Professor Department of Biology, Chair of the Faculty Senate Research Committee
Victor Darley-Usmar, PhD, HSOM Senior Associate Dean For Research Compliance and Administration
Marla Hertz, PhD, Associate Professor and Research Data Management Librarian, UAB Libraries
Jereme Logan, Associate Risk and Compliance Officer, Office of Compliance and Risk Assurance
Mike Matthews, Interim AVP for Research Business Operations, Office of Sponsored Programs
Ralph Zottola, PhD, Assistant Vice President, Research Computing
Final Thoughts – Winter is coming

Effective date for the DMS Policy is January 25, 2023

- Failure to include a DMSP with the 6 elements will delay submission, possibly to the next grant cycle
- The plan is a living document and can be changed during just-in-time phase (and during award period)

Be Aware: The DMS Policy is the minimum requirement for all of NIH
- NIH Institutes, Centers, and Offices may provide more specific guidance
- Other policies may apply (ex. Genomic Data Sharing policy)
Where to go for help

• Contact the **Office of Scholarly Communication** to request instruction or consultations on
  • Preparing a DMSP
  • Finding an appropriate data repository
• UAB Libraries **online guide** to the policy and example DMSPs
• Central **NIH website** and **AAMC** resources on data sharing policies
• **OSF resource toolkit**

“Open data’s good for society, but it is an added burden on the researchers. It needs to be supported”

- Dominique Roche

*Open data ‘tougher’ than open access and needs ‘mindset change’* [Times Higher Education 2020](#)

Follow us on Twitter: @ OSC_UAB

Marla Hertz
205.934.0694
mihertz@uab.edu

"Open data's good for society, but it is an added burden on the researchers. It needs to be supported”

- Dominique Roche

*Open data ‘tougher’ than open access and needs ‘mindset change’* [Times Higher Education 2020](#)
Looking for volunteers to develop sample DMSPs

1. Collaborative effort to draft a DMSP for the grant application.
   A. Initial meeting to discuss the project and answer questions about the DMS plan/policy.
   B. OSC drafts plan
   C. Joint revisions
2. The plan will be shared publicly
3. Given the impending deadline, this opportunity works best for projects that are either resubmissions or renewals where the research plan is already well developed

If interested contact mihertz@uab.edu