Introduction to Data Management and FAIR Data

Katie Mika

Data Services Librarian

Harvard Library & Institute for Quantitative Social Science

katherine_mika@harvard.edu

Learning Objectives

1. Apply **FAIR Principles** to share data effectively on Harvard Dataverse
2. Design practical **data management & sharing strategies** to meet funder requirements

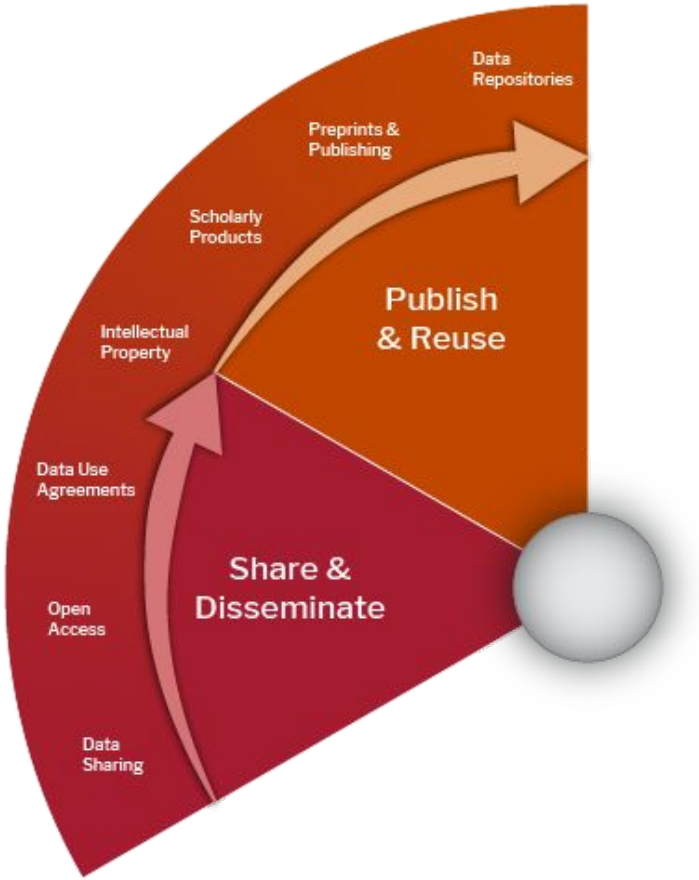
What is RDM?

"The ***active and ongoing*** management of data ***through its lifecycle*** of interest and usefulness to scholarship, science, and education."

The University of Illinois' Graduate School of Library and Information Science

Data Sharing

- Depositing data in a repository
- Choosing data licenses
- Applying metadata to make published data more findable
- Write good documentation so shared data is actually reusable
- Steward shared data over its useful term



FINDABLE

Increases visibility, citations, and impact of research

Supports knowledge discovery and innovation

ACCESSIBLE

Streamlines and maximizes ability to build upon previous research results

Attracts partnerships with researchers and business in allied disciplines

INTEROPERABLE

Supports and promotes inter- and cross-disciplinary data and reuse

REUSABLE

Promotes use and reuse of data allowing resources to be allocated wisely

Improves reproducibility and reliability of research results

The FAIR Guiding Principles (Wilkinson et al. 2016)

Who makes data FAIR?

Repositories:

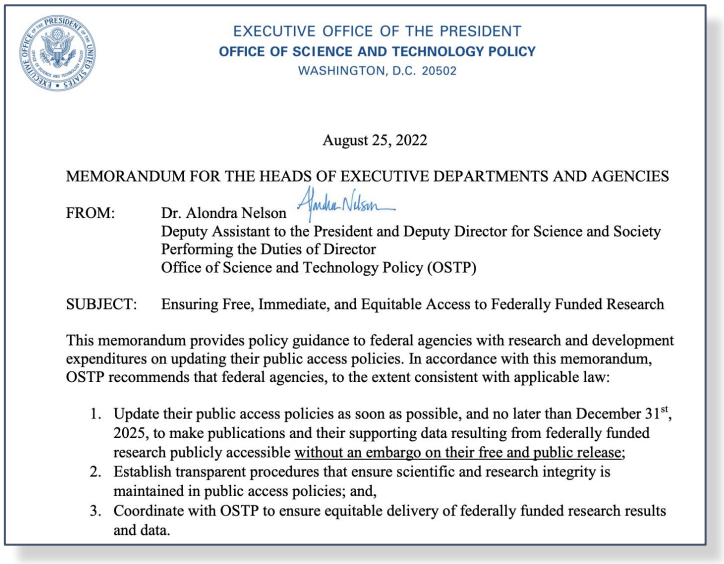
- Assign persistent identifiers
- Structure metadata records according to a disciplinary standard or schema
- Index data as searchable resources
- Retrieve datasets according to an open protocol that supports authentication
- Preserves data files and metadata
- Provenance and versions are tracked

Researchers:

- Structure data clearly and apply good data management practices
- Document data and software
- Richly describe data using standardized metadata fields
- Apply license and/or clear terms of use

Data curation is the process of transforming your collection of data files, notes, code, visualizations, and other vital research materials into an organized digital object that is optimized for reuse

Response to federal mandates



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

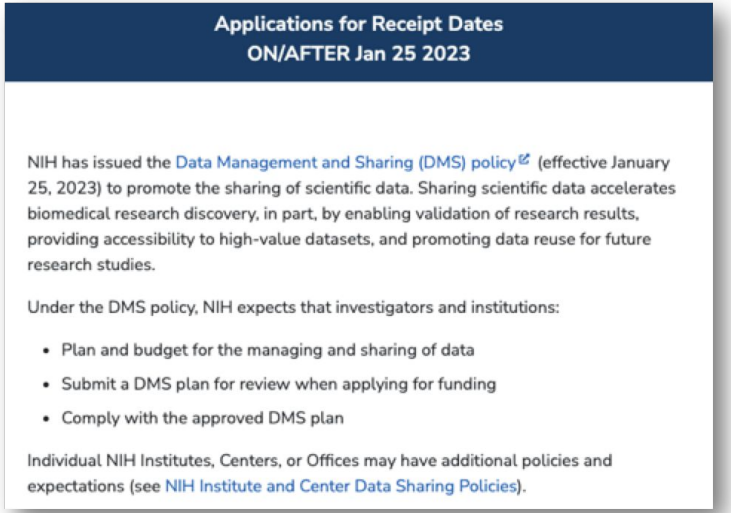
FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

[OSTP Memo: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research]



Applications for Receipt Dates
ON/AFTER Jan 25 2023

NIH has issued the Data Management and Sharing (DMS) policy (effective January 25, 2023) to promote the sharing of scientific data. Sharing scientific data accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies.

Under the DMS policy, NIH expects that investigators and institutions:

- Plan and budget for the managing and sharing of data
- Submit a DMS plan for review when applying for funding
- Comply with the approved DMS plan

Individual NIH Institutes, Centers, or Offices may have additional policies and expectations (see NIH Institute and Center Data Sharing Policies).

[NIH Data Management and Sharing Policy]

NIH Data Repository "Desired Characteristics"

NIH promotes the use of established data repositories because deposit in a quality data repository generally improves the ***FAIRness*** (Findable, Accessible, Interoperable, and Re-usable) of the data.

- **Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research:** https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html
- **Open Domain-Specific Data Sharing Repositories:** https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
- **Generalist Repositories:** https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html

NIH Data Management and Sharing Plan

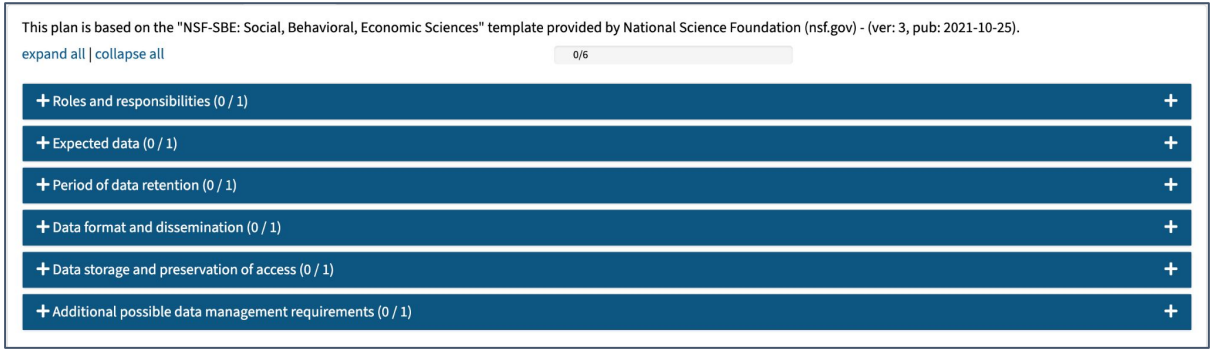
Specific guidance for Harvard Dataverse users: DMP Guidance

Harvard-wide templates available on DMPTool & RDM @ Harvard

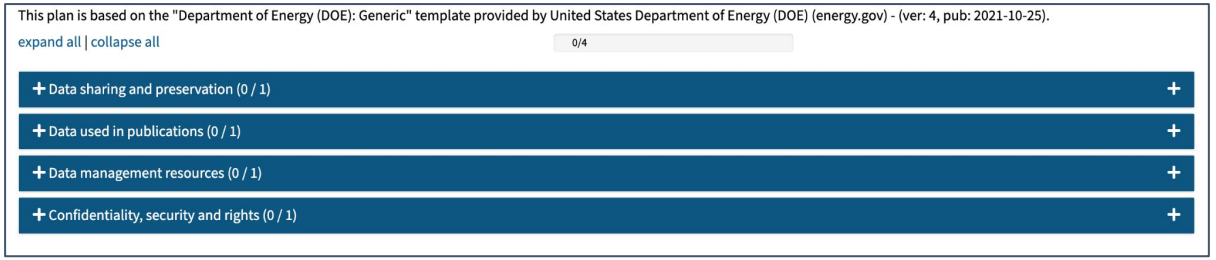
Specific elements relevant for depositing data in Harvard Dataverse

- **Data types:** content overview, overall size and individual file size, security level & licensed data
- **Related tools, software, & code:** open source tools, custom software
- **Standards:** disciplinary, metadata standards (descriptive & technical)
- **Data Preservation, Access, and Associated Timelines:** de-identification, embargos, general records schedule
- **Access, Distribution, or Reuse Considerations:** IP & licenses
- **Oversight of Data Management:** ongoing administration & stewardship

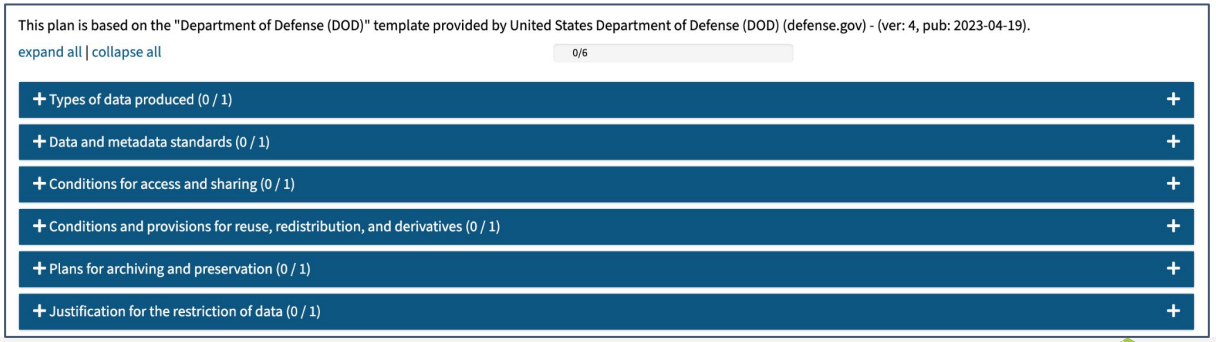
NSF: Social, Behavioral, Economic Services Directorate (NSF-SBE)



Department of Energy (DOE)



Department of Defense (DOD)



Data Types

- Briefly summarize the data types, including file formats, that you create or collect
- Describe what data will be shared and why
- List metadata, other relevant data, and documentation that will accompany shared data

Examples:

Murunga, Nickson; Nyawanda, Bryan; Nyiro, Joyce U.; Otieno, Grieven P.; Kamau, Everlyn; Agoti, Charles N.; Lewa, Clement; Gichuki, Alex; Mutunga, Martin; Otieno, Nancy; Mayieka, Lilian; Ochieng, Melvin; Kikwai, Gilbert; Hunsperger, Elizabeth; Onyango, Clayton; Emukule, Gideon; Bigogo, Godfrey; Verani, Jennifer R.; Chaves, Sandra S.; Nokes, David James; Munywoki, Patrick K., 2022, "Replication Data for: Surveillance of respiratory viruses at health facilities from across Kenya, 2014", https://doi.org/10.7910/DVN/VFCZN4, Harvard Dataverse, V7, UNF:6:ek1fu3CwyMKIiLVkPQPC8Q== [fileUNF]

Tools & Software

- Will specialized tools be needed to access or manipulate shared scientific data?
- What programs or systems will you use to manage and analyze the data?

Examples:

Kozitsin, Ivan, 2022, "Replication Data for: Optimal control in opinion dynamics models: towards a unified framework", https://doi.org/10.7910/DVN/6D6OGG, Harvard Dataverse, V5

Dataverse Uploader GitHub Action: https://github.com/marketplace/actions/dataverse-uploader-action

RSpace: https://documentation.researchspace.com/article/h14qd5tvjj-dataverse-integration

Open Science Framework (OSF): https://help.osf.io/article/208-connect-dataverse-to-a-project

Metadata Standards

1. **Citation Metadata:** any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset;
2. **Domain Specific Metadata:** with specific support currently for Social Science, Life Science, Geospatial, and Astronomy datasets; and
3. **File-level Metadata:** varies depending on the type of data file (for more details see File Handling section below) and include options like file tags, descriptions, and hierarchy preservation.

Examples:

Slavery, Abolition, Emancipation, and Freedom Collection: https://dataverse.harvard.edu/dataverse/SAEF

James, Toby S.; Garnett, Holly Ann, 2023, "Electoral Management Survey, (EMS-2.0)", https://doi.org/10.7910/DVN/Z7XVMC, Harvard Dataverse, V1, UNF:6:C5AOksQAF5Dn6QOp3II7+w== [fileUNF]

Data Preservation & Stewardship

- Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived.
- Consider how the dataset will be managed over time

Examples:

Harvard Dataverse Preservation Policy: https://support.dataverse.harvard.edu/harvard-dataverse-preservation-policy

Mallick, Swapan; Reich, David, 2023, "The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes", https://doi.org/10.7910/DVN/FFIDCW, Harvard Dataverse, V8

Access & Reuse

- Include any restrictions related to informed consent; privacy and confidentiality protections; Tribal, state, and local laws, regulations, and policies; and any other considerations.
- State whether access to the scientific data will be controlled.
- Describe how the privacy, rights, and confidentiality of human research participants will be protected (if applicable).

Examples:

Phillips, Jonathan, 2023, "*Locating what comes to mind in empirically derived representational spaces*", https://doi.org/10.7910/DVN/ISZQG3, *Harvard Dataverse*, V2

International Food Policy Research Institute (IFPRI), 2023, "Assessment of Nutrition-Sensitive Value Chains in the Feed the Future Zone of Influence in Tajikistan", https://doi.org/10.7910/DVN/BMKEMG, Harvard Dataverse, V1, UNF:6:x/dlNUoZ79srF5dZfHr4ug== [fileUNF]

McArthur, Charles C.; King, Stanley H., 1992, "*Harvard Student Study, 1960-1964*", https://doi.org/10.7910/DVN/4POMAK, *Harvard Dataverse*, V3

FAIR Summary

Findable: DOIs, **rich metadata,** indexing, **data citation**

Accessible: Standard & open retrieval protocol, authentication, persistent records, open metadata, **access permissions**, file format agnostic

Interoperable: Standard metadata schemas, **controlled vocabularies**, ***open formats***, APIs & external tools, metadata harvesting

Reusable: **Accurate & comprehensive metadata, clear Terms of Use,** tracked versions, **data stewardship** & preservation, file previews

Join the Dataverse Community!

Harvard Dataverse: dataverse.harvard.edu

The Dataverse Project: dataverse.org

Test out features at demo.dataverse.org

Preservation policy and other governance information: https://support.dataverse.harvard.edu/harvard-dataverse-preservation-policy

Get help: support@dataverse.harvard.edu

Leave an issue about a bug or a feature: https://github.com/IQSS/dataverse/issues

Dataverse-Users Google Group: https://groups.google.com/g/dataverse-community

Bi-Weekly Community Call: https://dataverse.org/community-calls

Thanks!

Katie Mika
katherine_mika@harvard.edu