Part II

Best Practices: Publishing, Disseminating, and Maintaining Data

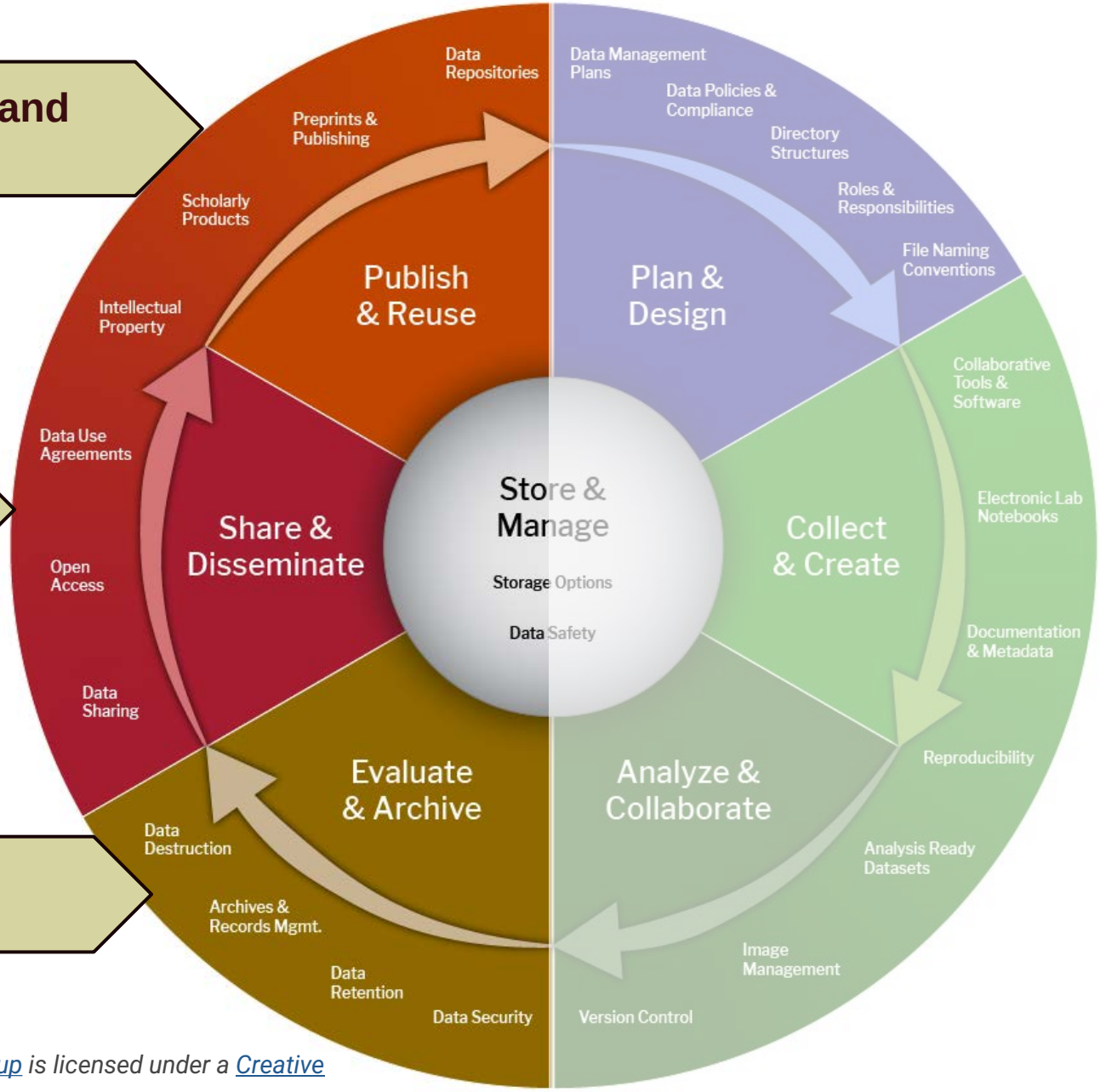
Heide Jackson
heidej@umd.edu

MPRC
Maryland Population Research Center

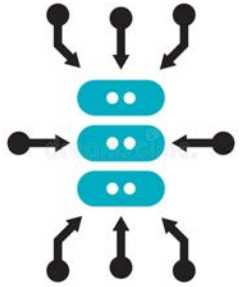
Planning for data use and reuse.

Key terms to know for data sharing.

Transition from internally to externally facing data.

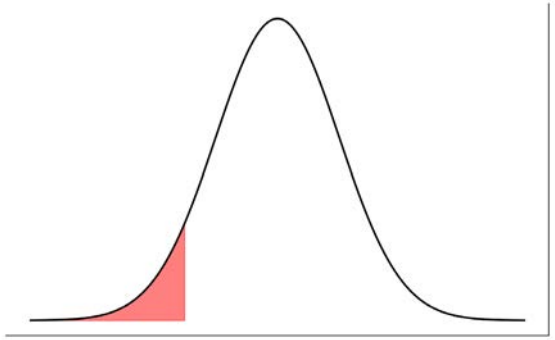


Internal to External: Human Subjects Considerations



Aggregation

Suppression



Top or Bottom Coding

Restricted Data Release

Internal to External: Data Destruction and Retention

- Know the timeline for data destruction and retention.
- Automate data destruction if possible.
- Know how to destroy and document the destruction.
- Many options. No one solution. Once the data is external, it cannot truly be destroyed.

Internal to External: Documentation

- Avoid internal jargon!
- Ensure any cross-tabulations or values reported use the public version of the data.
- Provide examples for suggested use:
  - Consider examples across a wide range of software packages.
  - Choose examples for essential data manipulation and best usage practice.

R

```
data
    %>% group_by(var1 = haven::as_factor(var1))
    %>% summarize(var1 = weighted.mean(var1, PERWT)
```

STATA

```
svyset cluster [pweight=perwt], strata(strata)
svy: mean var1
```

SAS

```
proc sort data = datasetname;by strata cluster;

run;

proc surveymeans data = datasetname;

weight perwt;

strata strata ;

cluster cluster;

var var1;

run;
```

Source: https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

Internal to External: Project Branding

- Consider a project name and identity.
- Identify ways to connect to users.
  - A project website
  - Social media accounts
  - Email list serve

Key Terms for Public Release: The Data Management and Sharing Plan (DMSP)

- The externally facing version of a data management plan.
- Their structure, content and length are determined by the funder.
  - NSF Guidelines
  - NIH Guidelines
- Unlike the DMP, the DMSP is a static document. This will be submitted with a grant application.

DMSP Templates

Example: NIH Data Management and Sharing Plan

OMB No. 0925-0001 and 0925-0002 *(Rev. 07/2022 Approved Through 01/31/2026)*

DATA MANAGEMENT AND SHARING PLAN

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. The Plan is recommended not to exceed two pages. Text in italics should be deleted. There is no "form page" for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format* shown below.

Public reporting burden for this collection of information is estimated to average 2 hours per response, including the time for reviewing instructions, searching existing data sources, gathering, and maintaining the data needed, and completing and reviewing the collection of information. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: NIH, Project Clearance Branch, 6705 Rockledge Drive, MSC 7974, Bethesda, MD 20892-7974, ATTN: PRA (0925-0001 and 0925-0002). Do not return the completed form to this address.

Element 1: Data Type

A. **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,

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.

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.

Example: NSF Data Management and Sharing Plan

DMR TEMPLATE

DATA MANAGEMENT AND SHARING PLAN[i]

Given that most proposed research involves the generation of scientific data, proposals are subject to the NSF Data Sharing Policy and require submission of a Data Management and Sharing Plan. Data should be findable, accessible, interoperable, and reusable (FAIR). Refer to the guidance on https://new.nsf.gov/funding/data-management-plan for developing this plan. The Plan should not exceed two pages. The Plan may be provided in the format shown below. ***Only the sections appropriate for the project need to be completed.*** *A valid Data Management and Sharing Plan may include only the statement that no detailed plan is needed if the statement is accompanied by a clear justification. The Plan is not intended to circumvent the number of pages that are allowed for the Project Description.* ***Text in italic with a yellow background should be deleted before submission.***

Element 1: Data Types

A. **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.

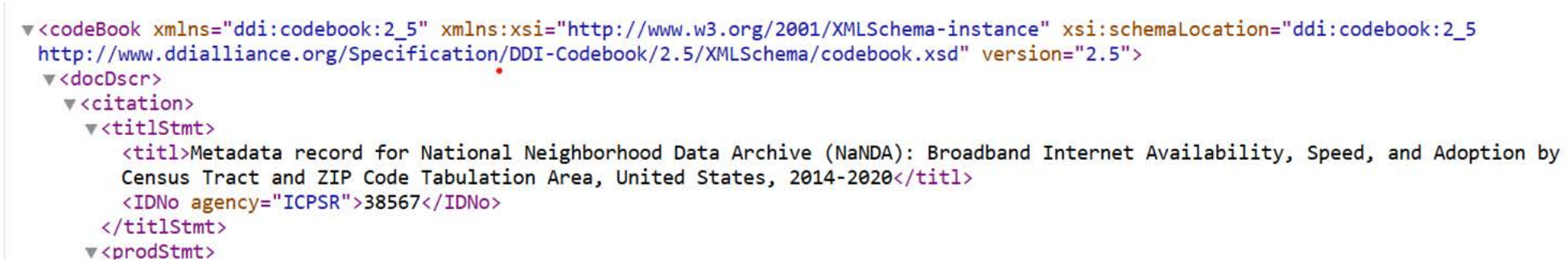
B. **Of the generated scientific data in 1.A, the scientific data that will be preserved and shared, and the associated rationale:**
Describe which types of scientific data from the project will be preserved and shared and briefly provide the rationale for this decision.

C. **Metadata and associated documentation:**
Briefly list the metadata (provides information about the scientific data) and any associated documentation (e.g., study protocols, research methods, and data-collection instruments) that will be made accessible to facilitate interpretation of the preserved and shared scientific data.

Key Terms for Public Release: Metadata

- Metadata: ‘information about our data collections that help others discover, understand, and use them’ (Source: ICPSR)
- Should follow specific guidelines such as the Data Documentation Initiative (DDI) and the Dublin Core.
- Why are metadata important?
  - Metadata help your data be FAIR (Findable, Accessible, Interoperable, and Reusable) by helping digital systems classify and identify your data and its key elements

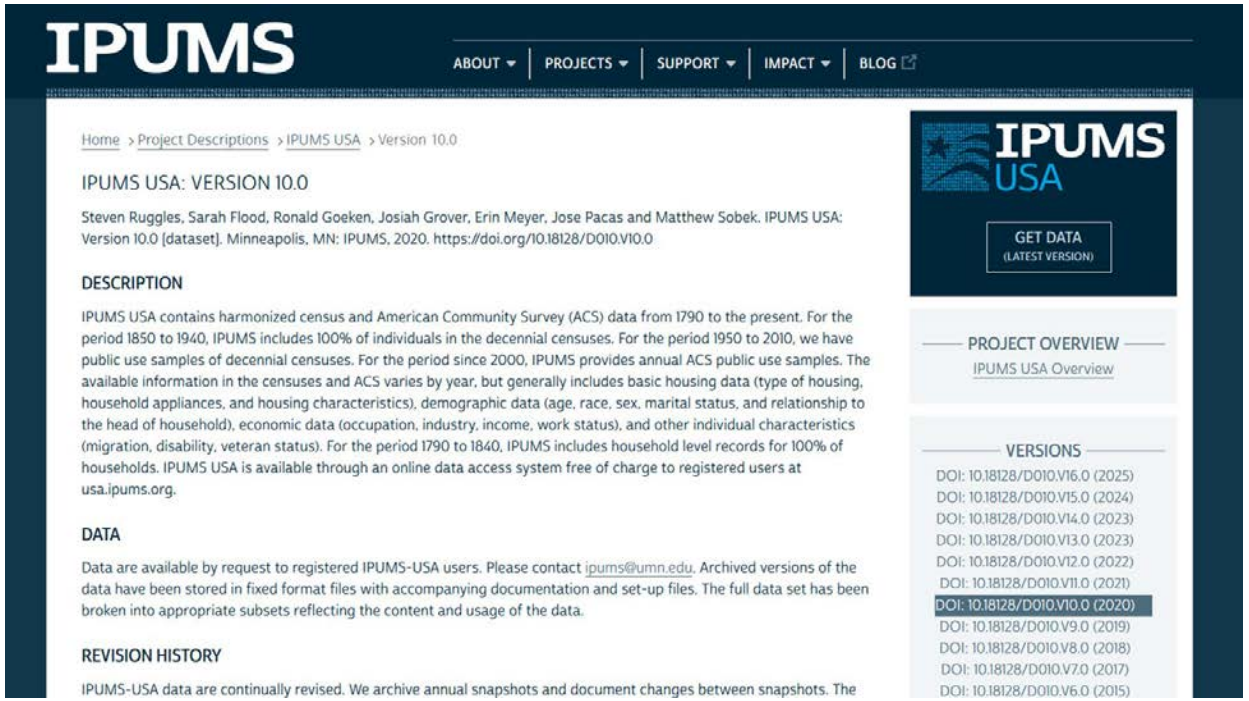
Example of Metadata



```
▼<codeBook xmlns="ddi:codebook:2_5" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="ddi:codebook:2_5
 http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd" version="2.5">
  ▼<docDscr>
    ▼<citation>
      ▼<titlStmt>
          <titl>Metadata record for National Neighborhood Data Archive (NaNDA): Broadband Internet Availability, Speed, and Adoption by
          Census Tract and ZIP Code Tabulation Area, United States, 2014-2020</titl>
          <IDNo agency="ICPSR">38567</IDNo>
        </titlStmt>
      ▼<prodStmt>
```

Key Terms for Public Release: Unique Persistent Identifiers

- For data, digital object identifiers (DOIs) are often used. Regulated by the Digital Object Identifier Foundation.
- DOIs provide a persistent and unique way to identify a resource even as technology changes.
- DOIs also can serve as public facing identifiers and project version control.
- Broadly share the doi not the data!



https://doi.org/10.18128/D010.V10.0

Key Terms for Public Release: Data Repository, Data Deposit, and Data Curation

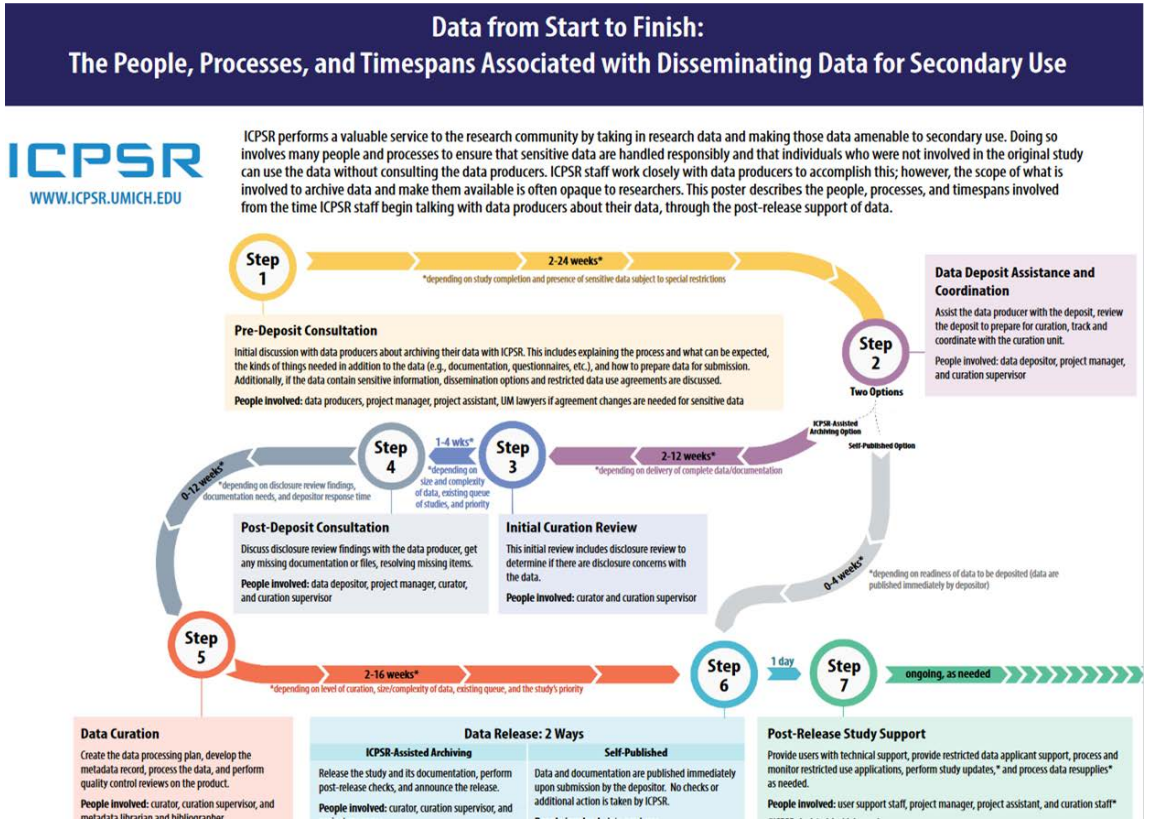


Image Source: ICPSR

- Data Repository: 'a centralized place to hold data, share data publicly, and organize data in a logical manner.' Harvard Dataverse
- Data Deposit: the process for submitting data for dissemination through a data depository.
- Data curation: the processing and maintaining of project data.

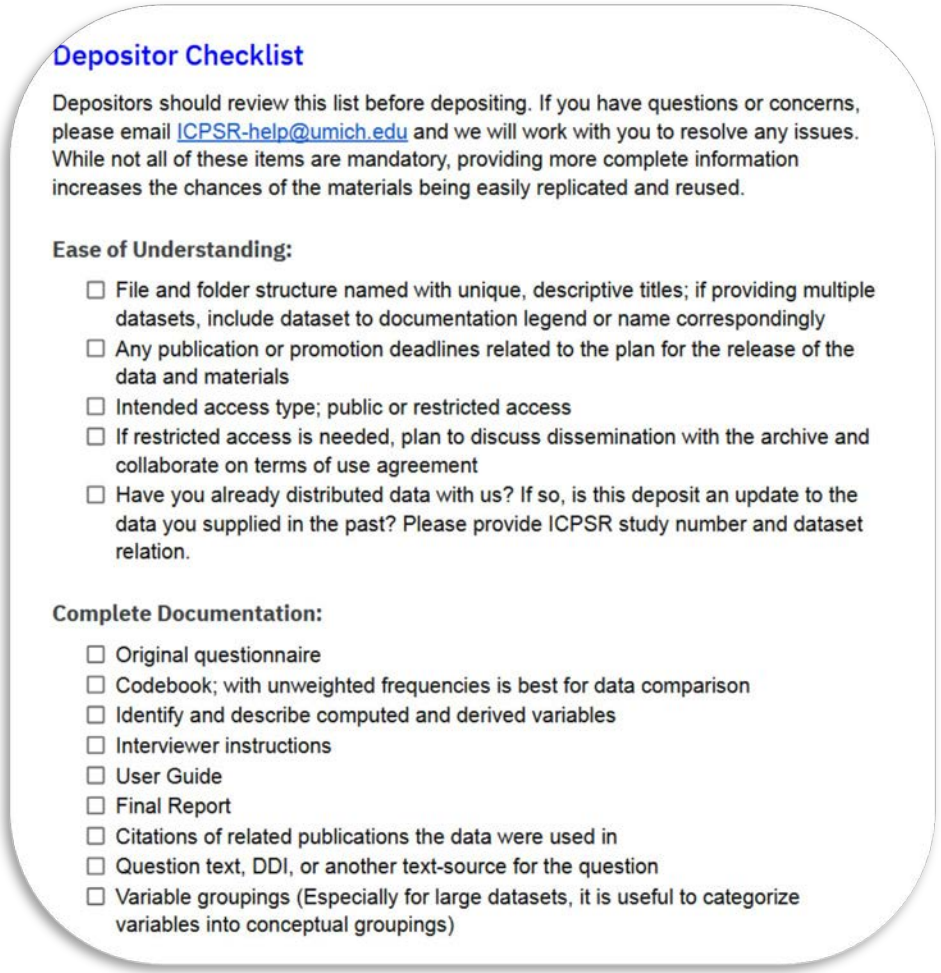
What is Shared with a Repository?

Depends on the repository!

At a minimum:

- Documentation
- Data
- Metadata

ICPSR Depositor Checklist



Depositor Checklist

Depositors should review this list before depositing. If you have questions or concerns, please email ICPSR-help@umich.edu and we will work with you to resolve any issues. While not all of these items are mandatory, providing more complete information increases the chances of the materials being easily replicated and reused.

Ease of Understanding:

- ☐ File and folder structure named with unique, descriptive titles; if providing multiple datasets, include dataset to documentation legend or name correspondingly
- ☐ Any publication or promotion deadlines related to the plan for the release of the data and materials
- ☐ Intended access type; public or restricted access
- ☐ If restricted access is needed, plan to discuss dissemination with the archive and collaborate on terms of use agreement
- ☐ Have you already distributed data with us? If so, is this deposit an update to the data you supplied in the past? Please provide ICPSR study number and dataset relation.

Complete Documentation:

- ☐ Original questionnaire
- ☐ Codebook; with unweighted frequencies is best for data comparison
- ☐ Identify and describe computed and derived variables
- ☐ Interviewer instructions
- ☐ User Guide
- ☐ Final Report
- ☐ Citations of related publications the data were used in
- ☐ Question text, DDI, or another text-source for the question
- ☐ Variable groupings (Especially for large datasets, it is useful to categorize variables into conceptual groupings)

Selecting a Repository: Some Popular Options

ICPSR and OpenICPSR

Harvard Dataverse

Open Science Foundation

Github

PubMed Central

Project or Department Website (Not Recommended!)

Considerations for Selecting a Repository

Funder requirements

Popularity with your target user base

Institutional partnerships for reducing data curation and archival costs

Specialized repository features

Costs

Information to Get from a Funder

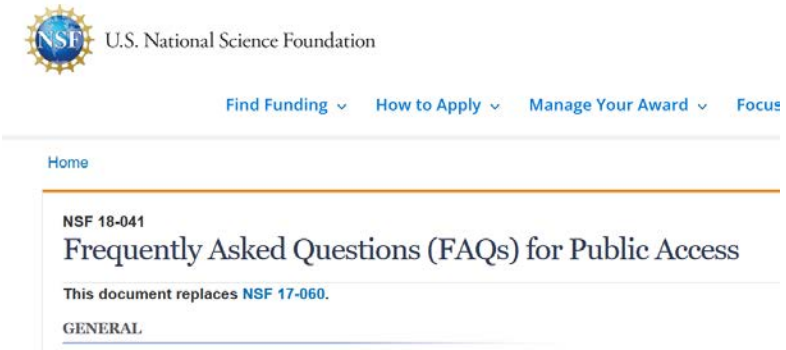
Desirable Characteristics for All Data Repositories

When choosing a repository to manage and share data resulting from Federally fu

- **Unique Persistent Identifiers:** Assigns datasets a citable, unique persistent ic support data discovery, reporting, and research assessment. The identifier po de-accessioned or no longer available.
- **Long-Term Sustainability:** Has a plan for long-term management of data, inc on a stable technical infrastructure and funding plans; and having contingenc unforeseen events.
- **Metadata:** Ensures datasets are accompanied by metadata to enable discove ideally widely used across, the community(ies) the repository serves. Domain generalist repositories.
- **Curation and Quality Assurance:** Provides, or has a mechanism for others to integrity of datasets and metadata.
- **Free and Easy Access:** Provides broad, equitable, and maximally open access submission, consistent with legal and ethical limits required to maintain priva data.
- **Broad and Measured Reuse:** Makes datasets and their metadata available w attribution, citation, and reuse of data (i.e., through assignment of adequate m
- **Clear Use Guidance:** Provides accompanying documentation describing term data use committee).
- **Security and Integrity:** Has documented measures in place to meet generally release of data, with levels of security that are appropriate to the sensitivity c
- **Confidentiality:** Has documented capabilities for ensuring that administrative

NIH Supported Data Repositories

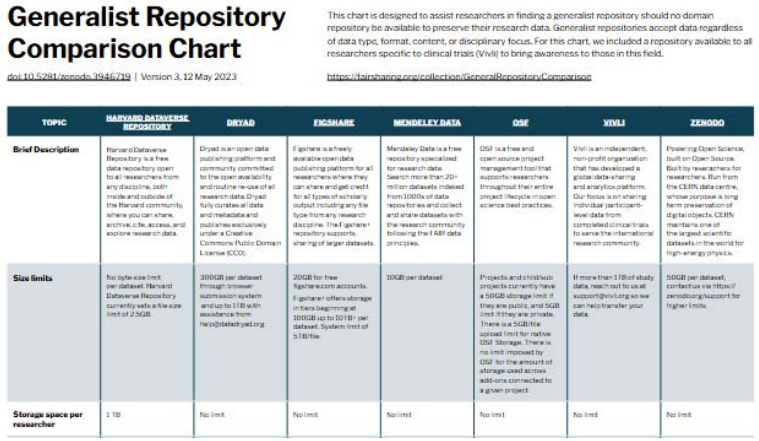
Detailed NSF Information



Questions to Ask a Repository

- How is metadata handled? Who creates the metadata?
- Does the repository allow an embargo period? How about a limited release?
- How does the repository manage persistent identifiers and version control?
- Does the repository do any data de-identification review or data curation?
- Does the repository support restricted data releases? Who approves researcher access?
- What are the costs associated with the deposit? Are there size limitations?
- Will users have any costs to access the data? use and reuse agreements? Are there user terms of use?
- How is the data stored? What is the long term plan for data preservation?
- What is your user support policy?

A Comparison of Generalist Repositories



Generalist Repository Comparison Chart

This chart is designed to assist researchers in finding a generalist repository should no domain repository be available to preserve their research data. Generalist repositories accept data regardless of data type, format, content, or disciplinary focus. For this chart, we included a repository available to all researchers specific to clinical trials (Vivli) to bring awareness to those in this field.

doi:10.5281/zenodo.1946719 | Version 3, 12 May 2023

https://fairsharing.org/collection/GeneralRepositoryComparison

TOPIC	HARVARD DATAVERSE REPOSITORY	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse Repository is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.	Dryad is an open data publishing platform and community committed to the open availability and routine re-use of all research data. Dryad fully curates all data and metadata and publishes exclusively under a Creative Commons Public Domain License (CC0).	Figshare is a freely available open data publishing platform for all researchers where they can share and get credit for all types of scholarly output including any file type from any research discipline. The Figshare+ repository supports sharing of larger datasets.	Mendeley Data is a free repository specialized for research data. Search more than 20+ million datasets indexed from 1000s of data repositories and collect and share datasets with the research community following the FAIR data principles.	OSF is a free and open source project management tool that supports researchers throughout their entire project lifecycle in open science best practices.	Vivli is an independent, non-profit organization that has developed a global data sharing and analytics platform. Our focus is on sharing individual participant-level data from completed clinical trials to serve the international research community.	Powering Open Science, built on Open Source. Built by researchers for researchers. Run from the CERN data centre, whose purpose is long term preservation of digital objects. CERN maintains one of the largest scientific datasets in the world for high-energy physics.
Size limits	No byte size limit per dataset. Harvard Dataverse Repository currently sets a file size limit of 2.5GB.	300GB per dataset through browser submission system and up to 1TB with assistance from help@datadryad.org.	20GB for free figshare.com accounts. Figshare+ offers storage in tiers beginning at 100GB up to 10TB+ per dataset. System limit of 5 TB/file.	10GB per dataset	Projects and child/sub projects currently have a 5GB storage limit if they are public, and 5GB limit if they are private. There is a 5GB/file upload limit for native OSF Storage. There is no limit imposed by OSF for the amount of storage used across add-ons connected to a given project.	If more than 1 TB of study data, reach out to us at support@vivli.org so we can help transfer your data.	50GB per dataset, contact us via https://zenodo.org/support for higher limits
Storage space per researcher	1 TB	No limit	No limit	No limit	No limit	No limit	No limit

Questions? Discussion

Heide Jackson
Associate Research Professor
Maryland Population Research Center
heidej@umd.edu

MPRC
Maryland Population Research Center