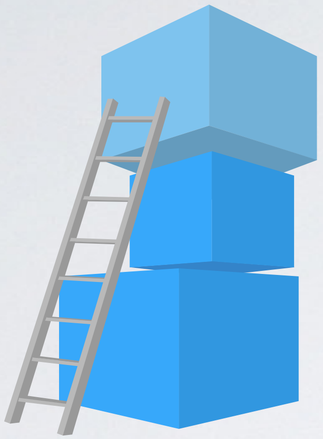
Getting Credit for Sharing Your Data (Part II): Sharing Your Data

Alaina Pearce
azp271@psu.edu

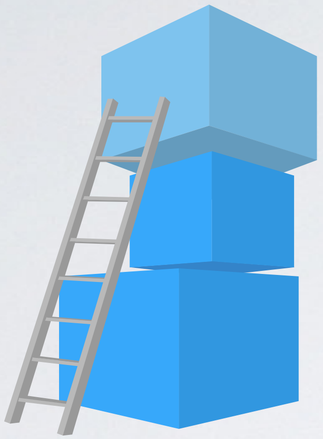
Why Share Your Data?



Advance Rigorous & Reproducible Research

- Enable validation of research results
- Make high-value datasets accessible
- Accelerate future research directions
- Increase opportunities for citation and collaboration

Why Share Your Data?



Advance Rigorous & Reproducible Research

- Enable validation of research results
- Make high-value datasets accessible
- Accelerate future research directions
- Increase opportunities for citation and collaboration

Promote Public Trust

- Foster transparency and accountability
- Demonstrate stewardship over taxpayer funds
- Maximize research participants' contributions
- Support appropriate protections of research participants' data

Why Share Your Data?

Advance Rigorous & Reproducible Research

- Enable validation of research results
- Demonstrate stewardship over taxpayer funds
- Maximize research participants' contributions
- Support appropriate protections of research participants' data

You (May Be) Required To by Funder

(In terms & conditions of your award)

NIH Data Management & Sharing Policy

Effective Date: January 25, 2024

Requirements:

- Submission of a Data Management & Sharing Plan (DMSP) for ALL NIH-funded research that generates scientific data
- Expects researchers to 'maximize appropriate data sharing' in established repositories
  - Sharing should occur no later than *publication* or end of award
  - Requires researchers to justify any sharing exceptions (i.e., ethical, legal, technical factors
- Allows research to request funding for personnel costs or other fees related to data management and share actives, but the money must be spent during the grant's award period

NIH Data Management & Sharing Policy

Effective Date: January 25, 2024

Requirements:

- Submi esearch that g
- Expec ries
  - Sha
  - Rec al factors
- Allow ata management and share actives, but the money must be spent during the grant's award period

Approved DMSPs Become a Term and Condition of Your Award and Non-Compliance Could Impact Future Funding Decisions

NIH Data Management & Sharing Policy

What is *scientific* data?

In Scope

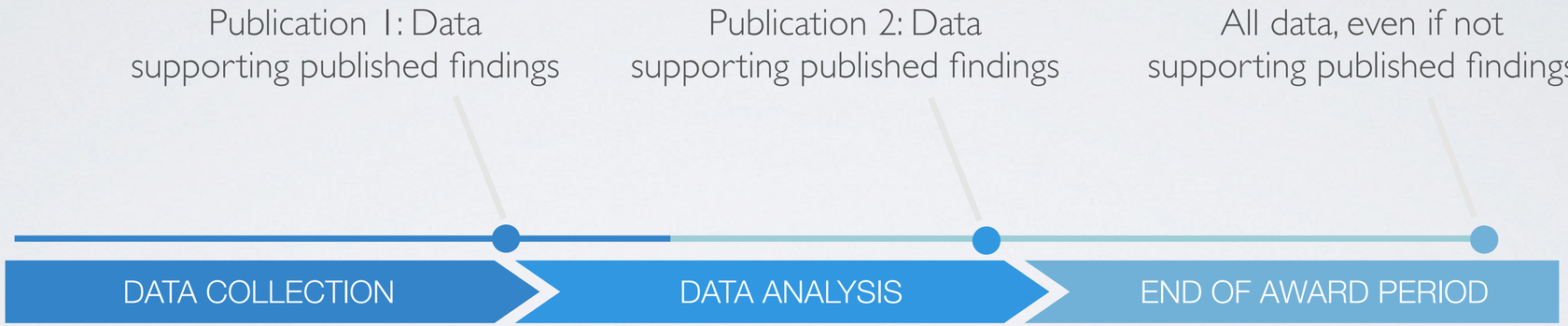
"recorded factual material…of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications"

Out of Scope

Lab notebooks, preliminary analyses, case report forms, physical objets, drafts of scientific papers, plans for future research, peer reviews

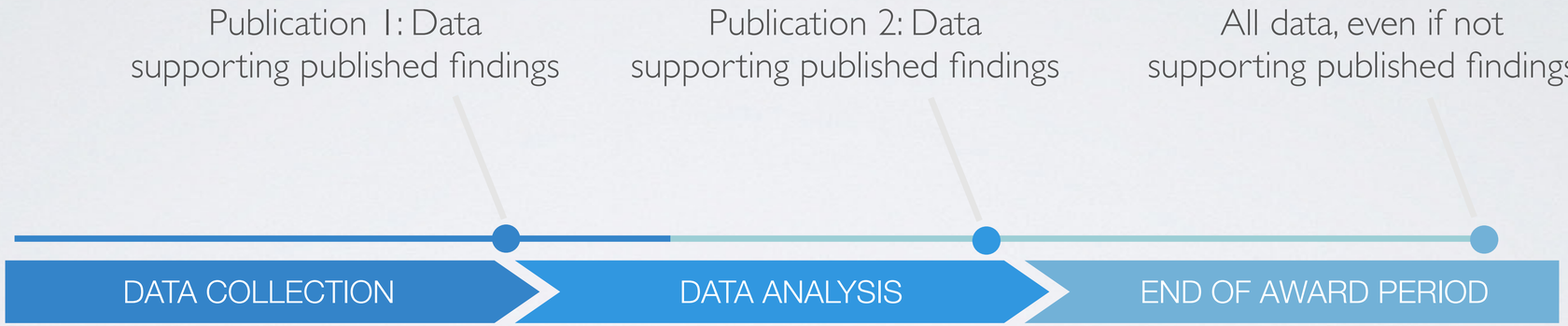
NIH Data Management & Sharing Policy

When to Share:



NIH Data Management & Sharing Policy

When to Share:



When Limit to Sharing - Justifiable ethical, legal, and technical factors:

- Informed consent will not permit or limits scope of sharing or use
- Privacy or safety of research participants would be compromised and available protections insufficient
- Explicit federal, state, local, or Tribal law, regulation, or policy prohibits disclosure
- Restrictions imposed by existing or anticipated agreements with other parties

NIH vs Draft NSF DMSP

	NIH	NSF (Draft)
What must be shared?	Scientific data: "Recorded factual material…of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications"	•Primary data •Samples •Physical collections •Scripts/code •Curriculum material •Other supporting materials created or gathered during award period
Timeline	•At publication •End of Award (even if not supporting published finding)	At publication
Page Limit	2 pages	Must use NSF webform (no page limit)
Required Elements	•Data Types •Related tools, software/code •Standards •Data preservation, access, & timelines •Access, distribution, & reuse considerations •Oversight	•Data and Research Product Categories •Access Policies and Limitations •Data Standards and Metadata •Data Provenance •Public Archiving •Retention Time •Accountability
Repository Requirements	Encourages use of established repositories	Must provide a persistent identifier (e.g., DOI)

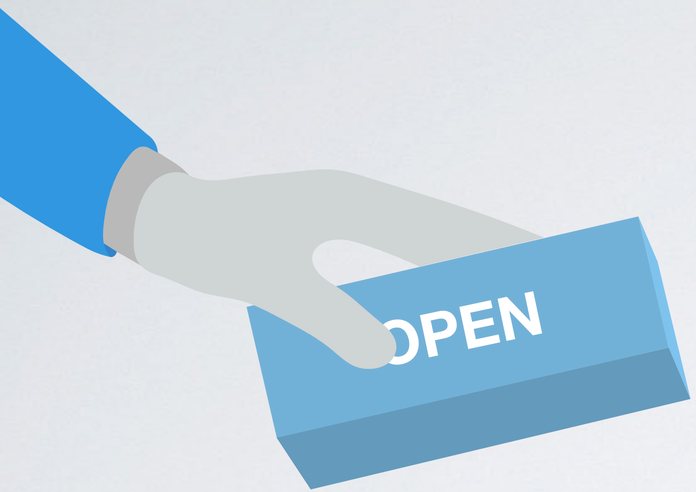
What Data Do I Share?

All data must be managed - Not ALL data can be shared

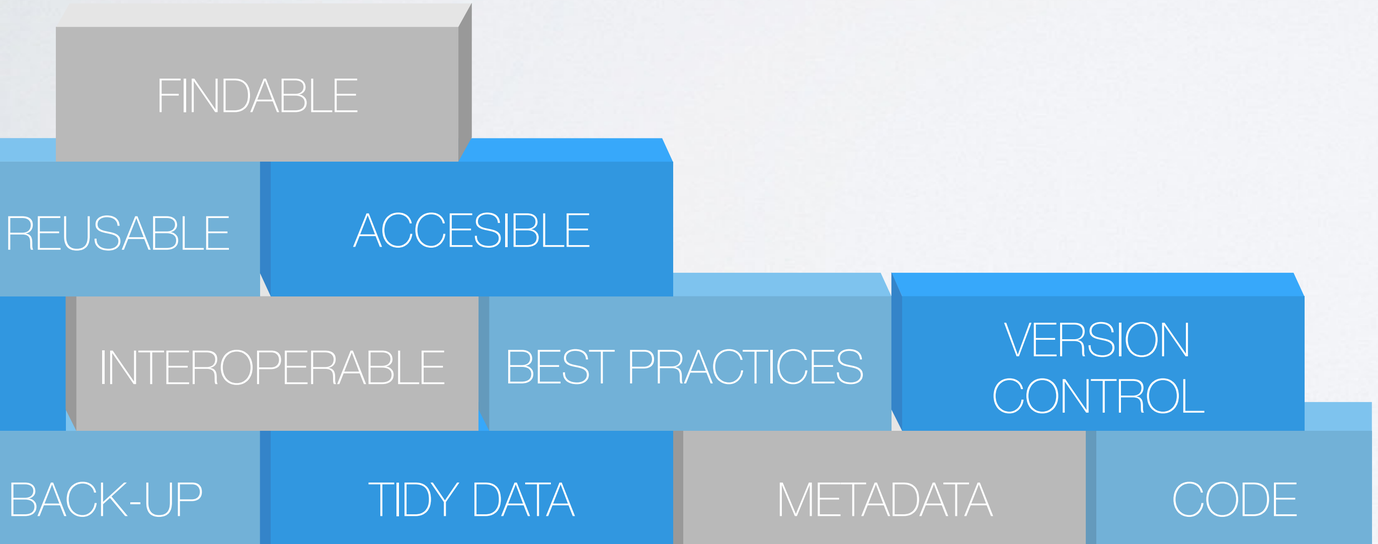
- "Raw" data need to support published findings or contribute to new studies
- Accompanying code/materials to support data

- Copyrighted data
- Data contains personal identifiable information, health information, or other sensitive information about individuals, communities, or locations
- You intend to make a patent application and must avoid prior disclosure

How To Prepare Your Data



- *Good Enough* Data Practices
- **FAIR Data**
- Repositories
- Metadata
- Institutional Resources



FAIR Data

Data is documented with rich metadata, has a unique and persistent identifier (e.g., DOI), and is in a resource that can be searched or indexed.

Metadata and data is both human and machine readable and is stored in a trusted repository.

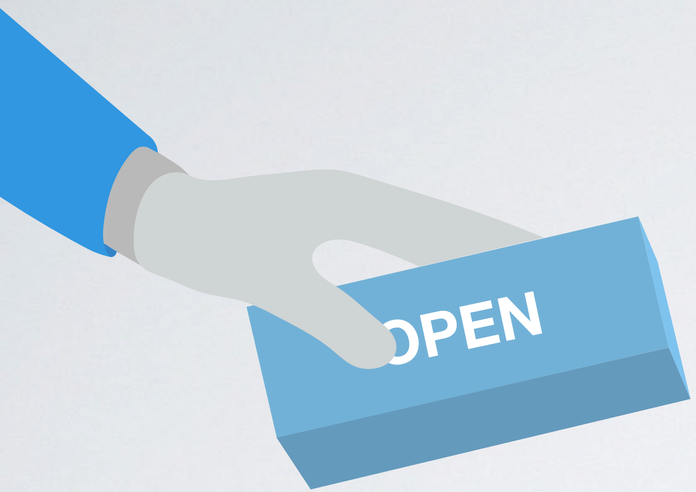
Data is in open formats with a common structure and metadata uses accepted, disciplinary terminology (e.g., controlled vocabulary or ontologies)

Data has clear licenses and provenance and align with community/disciplinary standards

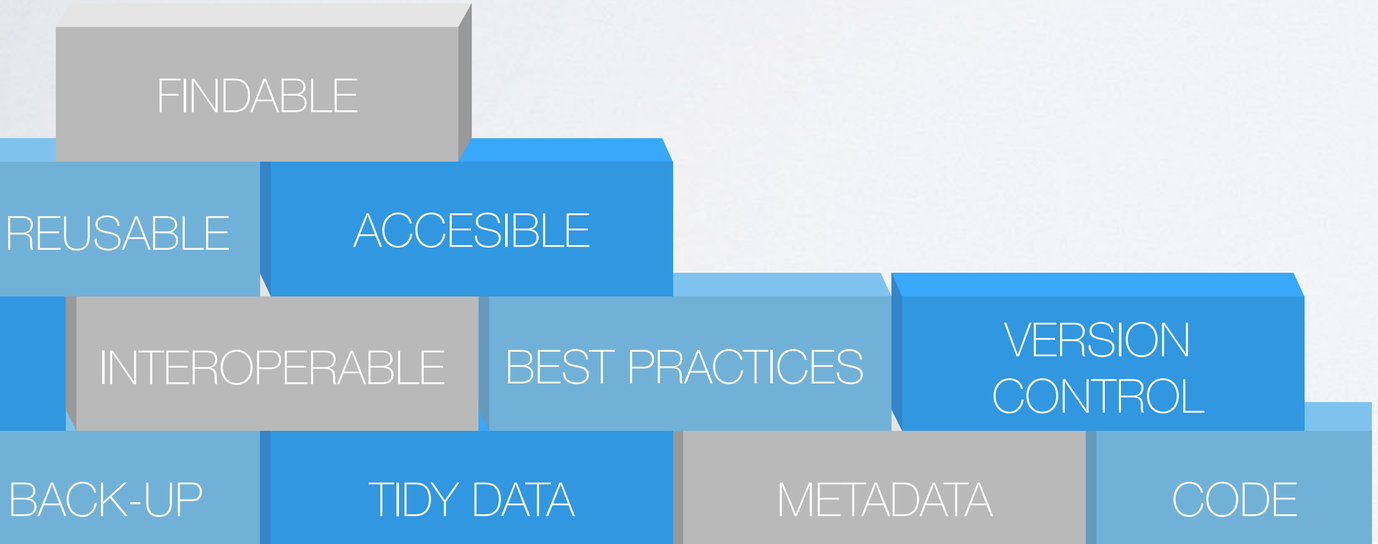
FAIR Data

	Researcher	Repository
Findable	• Files are unambiguously named • All files are fully described with rich metadata (e.g., README) • All data are fully described with rich metadata (e.g., data dictionary, data manual)	• Provides a persistent identifier (e.g., DOI) • (Meta)data are discoverable through an open search protocol
Accessible	• Identify appropriate access controls, if necessary • Clearly document how to find and access data, including authentication/access protocols	• (Meta)data are retrievable via a persistent identifier with clear download instructions • Provides reliable storage with long-term sustainability • (Meta)data are machine actionable
Interoperable	• Use standard and open file formats • (Meta)data follow disciplinary notation and terminology (e.g., controlled vocabularies and/or ontologies) • If available, follows (meta)data standards (e.g., BIDS, DDI, Pysch-DS)	• Provides a standard interface and protocol for data deposit and retrieval • Metadata is structured and can be parsed by algorithms
Reusable	• Choose open license • Detailed documentation about the data, its provenance, and relevant data collection and/or processing steps • Well documented context so user can understand the data's purpose and limitations	• Versions are tracked • Has license and/or clear terms of use • Mechanism for documenting origin, history, and who to cite

How To Prepare Your Data

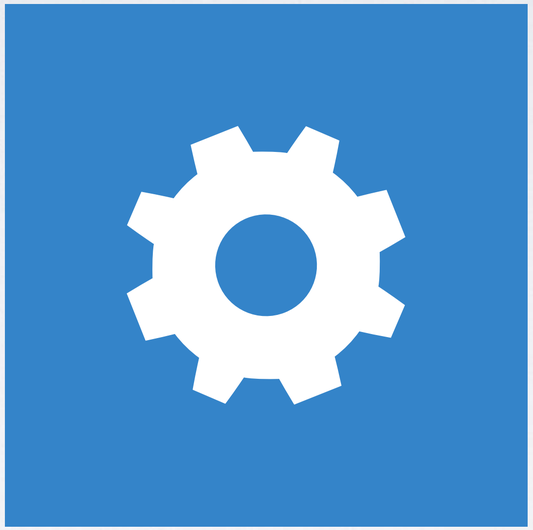


- *Good Enough* Data Practices
- FAIR Data
- **Repositories**
- Metadata
- Institutional Resources



Repositories

Curation
(from metadata enhancement to individualized help)



Technical Support:
FAIR principles and data preservation

Access:
Persistent identifiers and licenses/terms of use

Repositories

Domain Specific

Qualitative Data Repository

OpenNEURO

NCBI

ICPSR

dbGaP GENOTYPES and PHENOTYPES

Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

Databrary

NIMH Data Archive

NDA | ABCD | CCF | NIAAA$_{DA}$ | OAI

Institutional

ScholarSphere | PennState University Libraries

datacommons@psu

data. energy.environment.society.technology

Generalist

DRYAD

zenodo

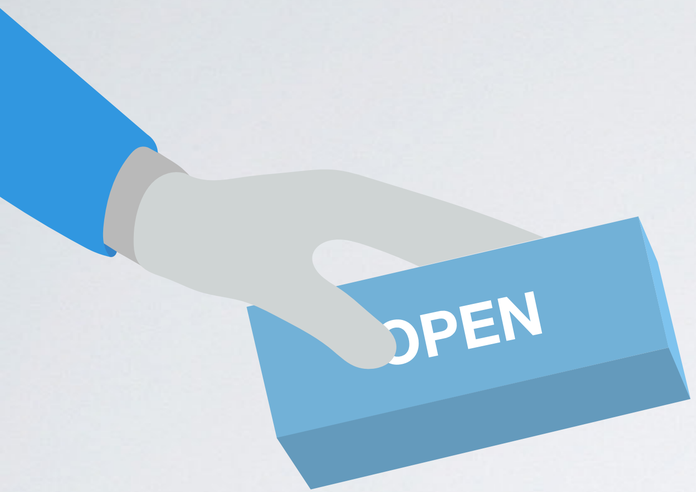
figshare

OSF

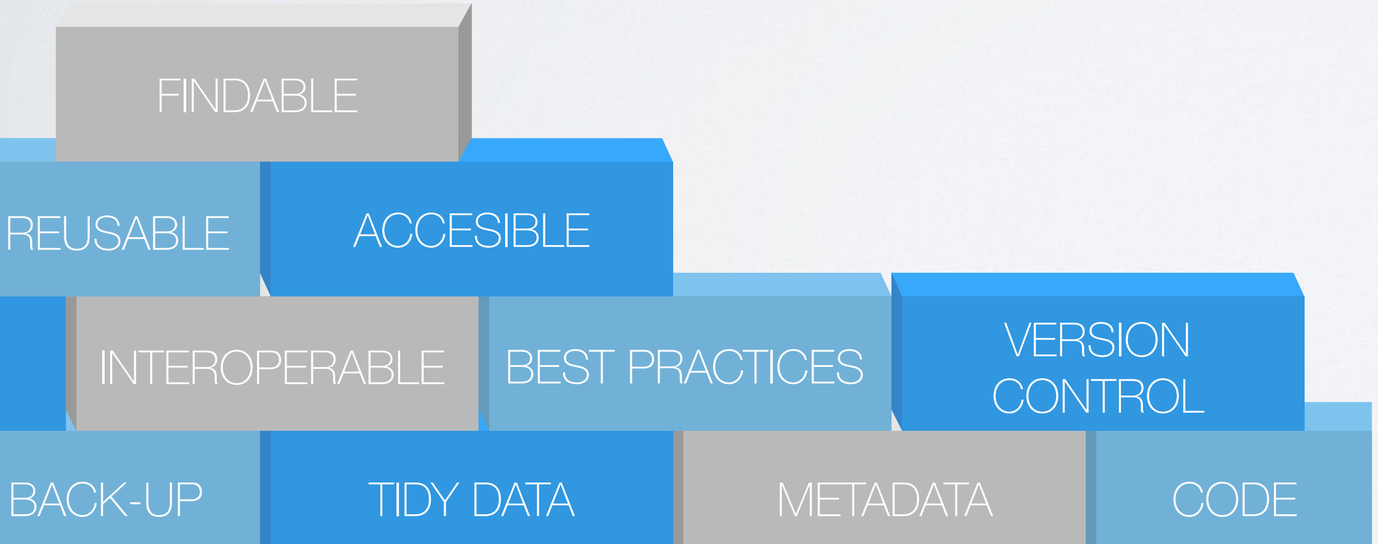
Repositories

	Public-Public (Open Access)	Public-Private (Metadata Open Access)	Private-Private (Closed Access)
Metadata is fully discoverable	✓	✓	✗
Data are accessible	✓	✓	✗
Data can be downloaded immediately	✓	✗	✗
Data access is mediated	✗	✓	✗
Good for level 1 data	✓	✗	✗
Good for identifiable/highly sensitive data	✗	✓	✓

How To Prepare Your Data



- *Good Enough* Data Practices
- FAIR Data
- Repositories
- **Metadata**
- Institutional Resources



Metadata

Highly structured data laid out in fields, often with controlled vocabularies - the who, what, when, where, and why of data

Metadata

Highly structured data laid out in fields, often with controlled vocabularies - the who, what, when, where, and why of data

Data Dictionary

- Variable names and definition
- Data units and formats
- Min and max values
- Coded values and their meanings
- Null/NA representations
- Known issues with the data
- Relationships with other variables
- How it was measured

Metadata

Highly structured data laid out in fields, often with controlled vocabularies - the who, what, when, where, and why of data

Data Dictionary

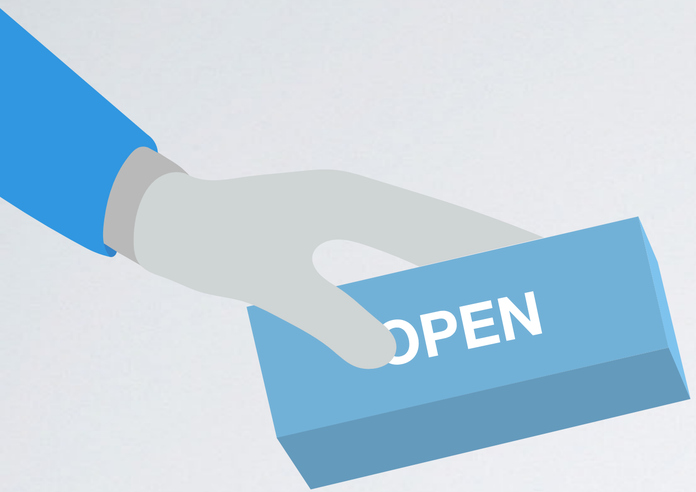
- Variable names and definition
- Data units and formats
- Min and max values
- Coded values and their meanings
- Null/NA representations
- Known issues with the data
- Relationships with other variables
- How it was measured

README.txt (or .md)

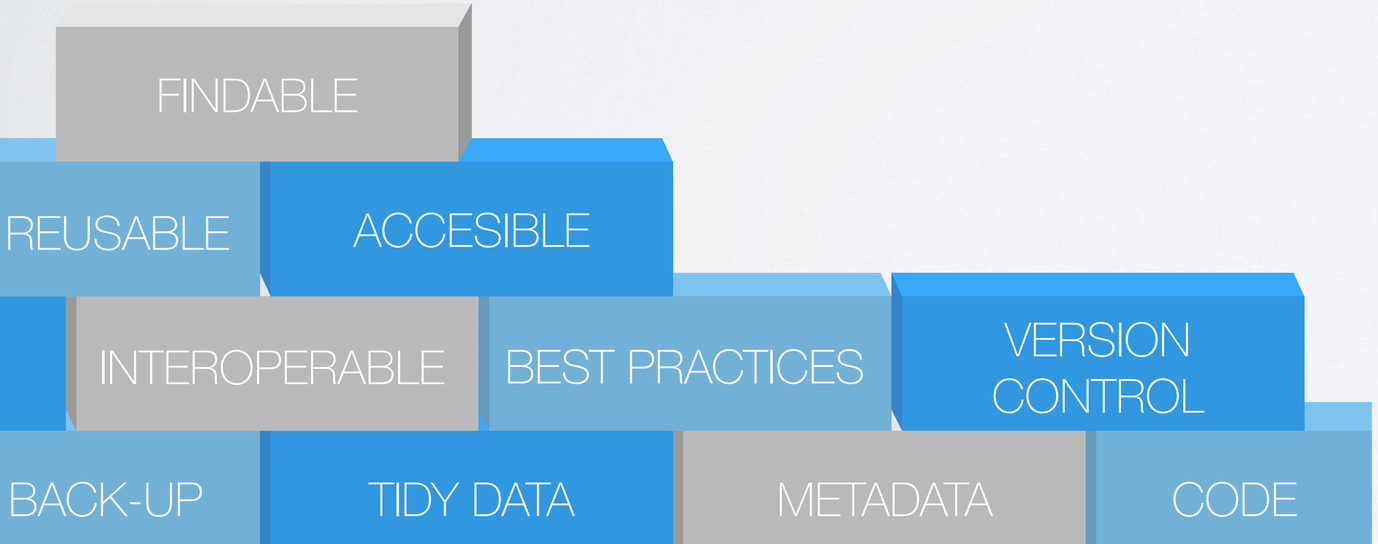
- Information about the files it accompanies
  - Often describes the directory structure
- Licensing information
- Project description
- Funding information
- How to use the data/files being shared
- Author name and contact information
- Any disclosures

README Template Worksheet

How To Prepare Your Data



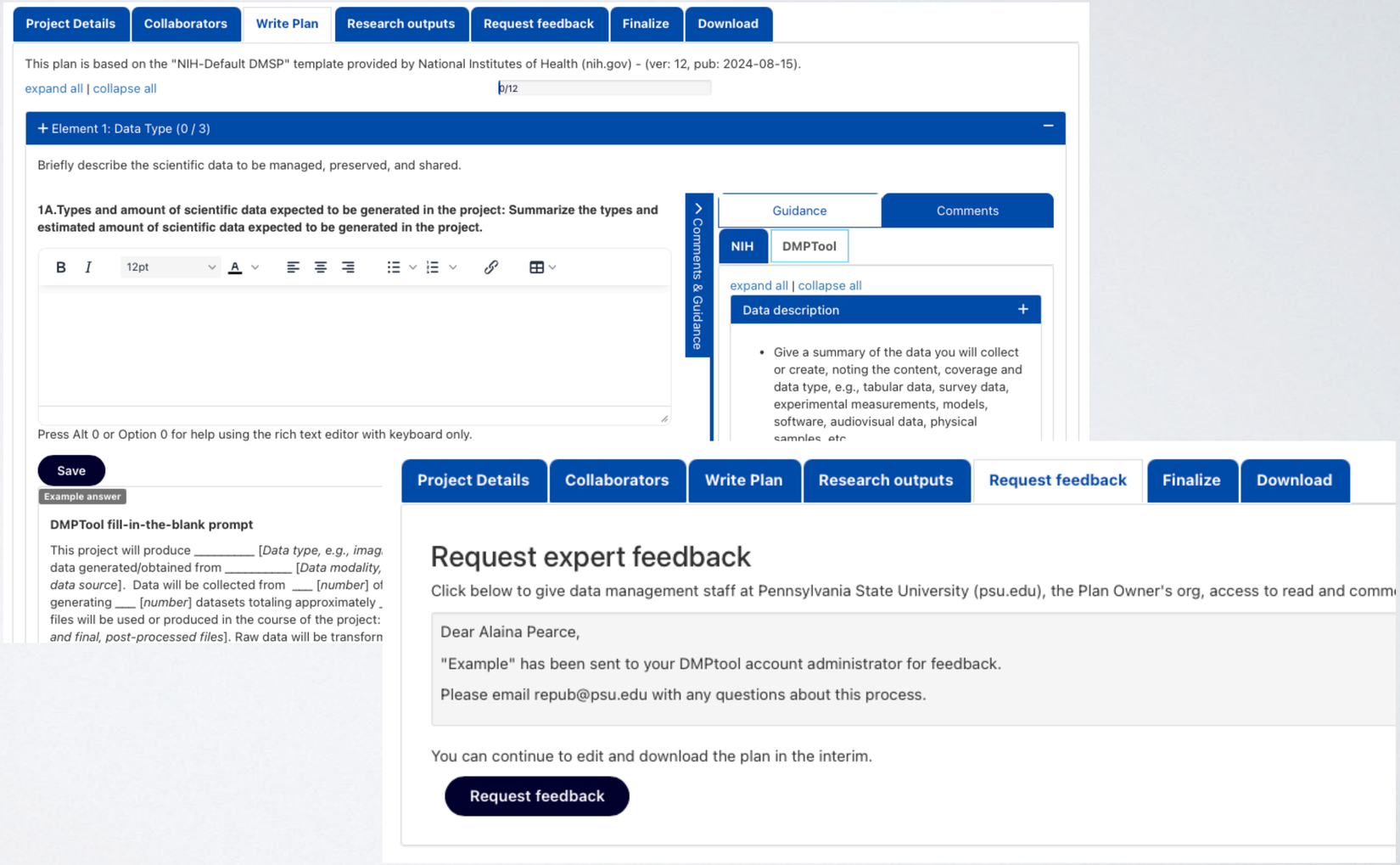
- *Good Enough* Data Practices
- FAIR Data
- Repositories
- Metadata
- **Institutional Resources**



Institutional Tools & Resources

DMP Tool

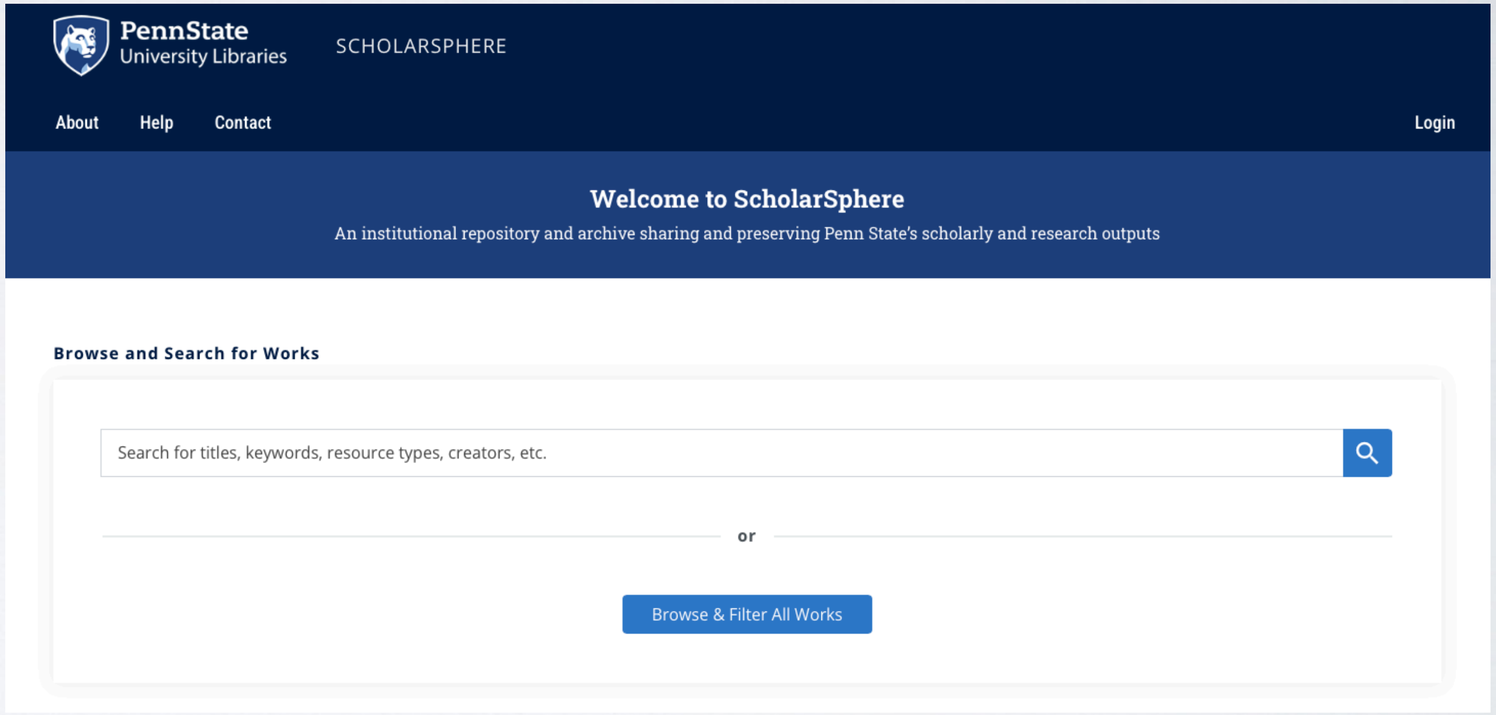
www.dmptool.org



Institutional Tools & Resources

ScholarSphere

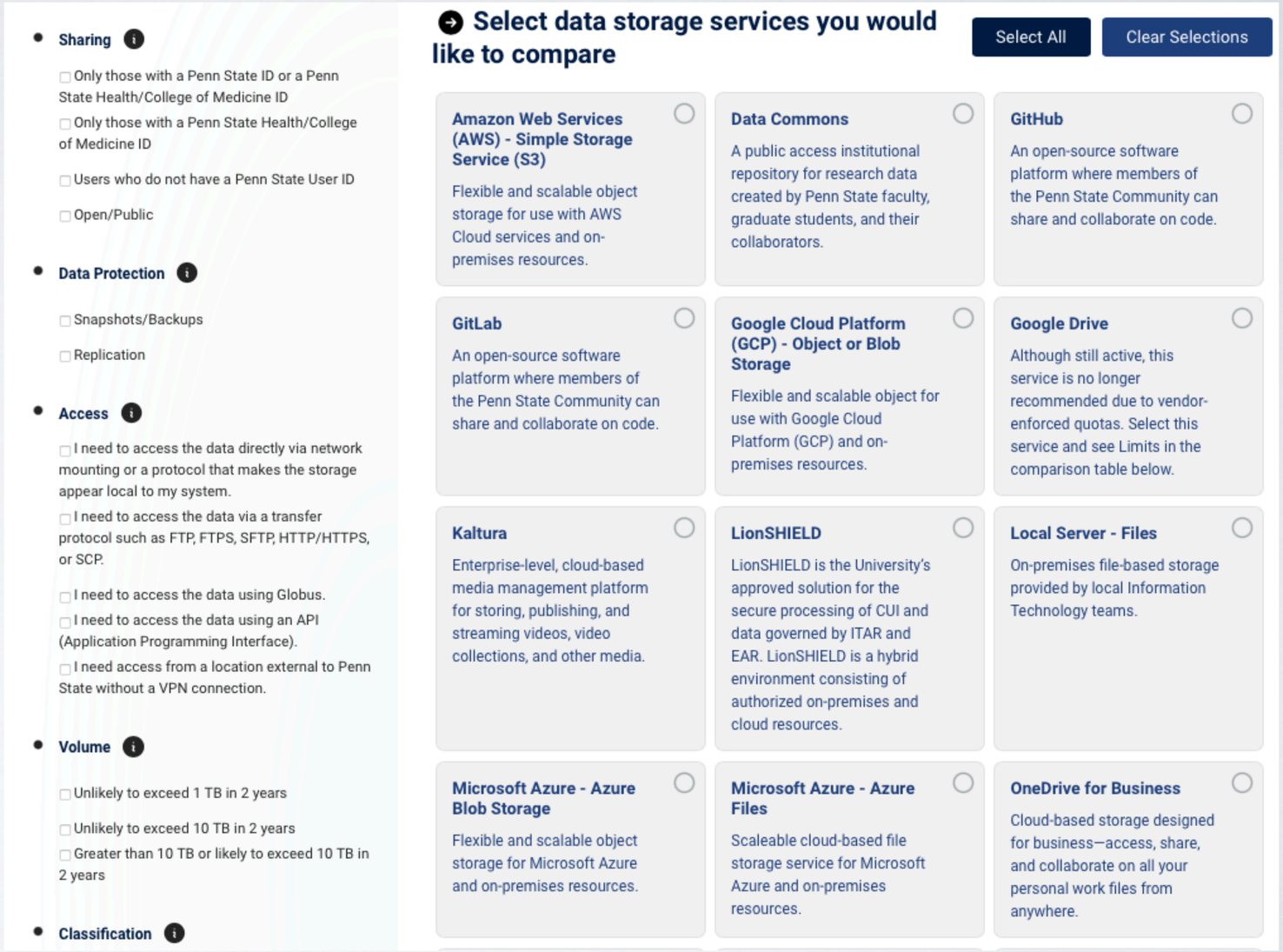
scholarsphere.psu.edu



Institutional Tools & Resources

Penn State Data Storage Finder

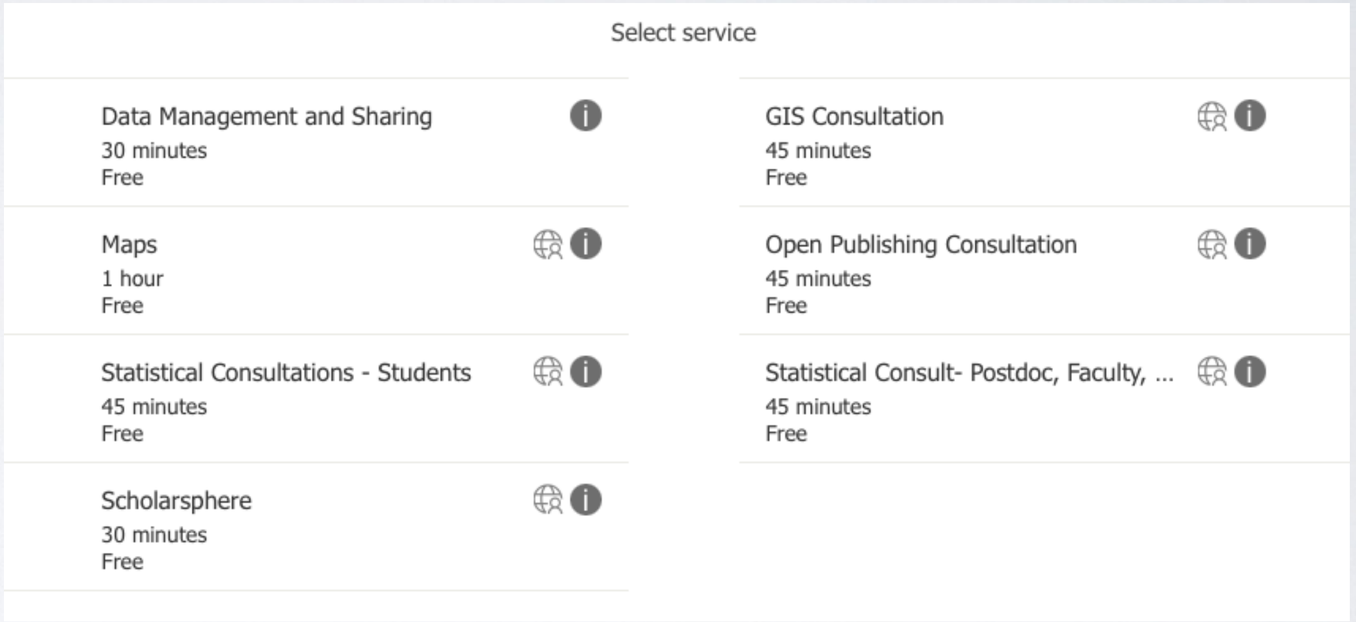
datastoragefinder.psu.edu



Institutional Tools & Resources

Data Learning Center

libraries.psu.edu/about/departments/units/data-learning-center



NIH Data Management & Sharing Policy

Data Management and Sharing Plan Elements

Data Types

Related Tools, Software, and/or Code

Metadata Standards

Data Preservation, Access, and Associated Timelines

Access, Distribution, or Reuse Considerations

Oversight of Data Management and Sharing