Washington University in St. Louis
University Libraries

Building a Desirable Repository: Meeting the Call of Researchers, Funders, FAIRness, and the OSTP

Sarah Swanz
Olin Library, Data Services
Washington University in St. Louis

Beyond the Numbers Conference 2023

Goal is to make our data FAIR

Findable – in an indexed repository, with a unique, persistent ID and rich metadata

Accessible – repo uses open, standard protocols so the metadata and data can be accessed

Interoperable – data are in formal, standard, open application languages

Reusable – well documented, explicit provenance, open licenses, follows community standards

Goal is to share our data

- Shared data supports transparency in research results
- Data reuse advances science
- Funders require data sharing

Which funders demand sharing?

- Air Force Research Office
- Army Research Office
- Catalog of Federal Domestic Assistance
- Congressionally Directed Medical Research Program
- Defense Advanced Research Projects Agency
- Department of Energy
- Department of Homeland Security
- Department of Housing and Urban Development
- Environmental Protection Agency
- Federal Acquisition Jump Station
- Health & Human Services (HHS)
- National Aeronautics and Space Administration
- National Endowment for the Arts
- National Endowment for the Humanities
- National Institutes of Health
- National Science Foundation
- Office of Naval Research
- The Federal Register
- U.S. Department of Agriculture
- U.S. Department of Education

And more!

Washington University in St. Louis
University Libraries

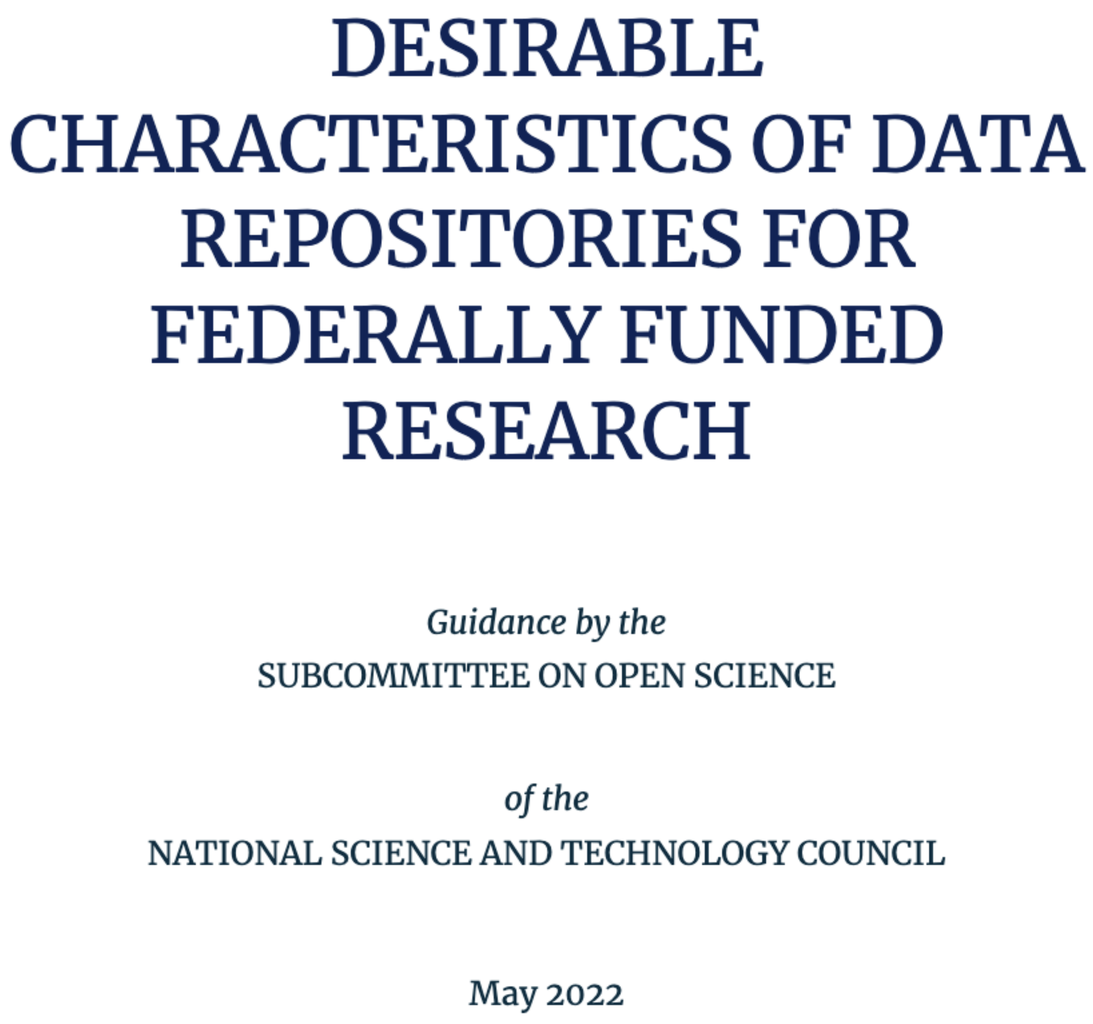
Where can I deposit?

OSTP Memo

From: White House Office of Science and Technology (OSTP)

To: Federal Funding Agencies &

Re: Selecting Repositories



DESIRABLE CHARACTERISTICS OF DATA REPOSITORIES FOR FEDERALLY FUNDED RESEARCH

Guidance by the
SUBCOMMITTEE ON OPEN SCIENCE

of the
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

May 2022

OSTP Memo: Purpose

- Improve consistency of repository selection guidance across Federal agencies
- Promote the FAIR data principles
- Incorporate experiences and comments of agencies, along with public and private sector organizations
- Provide high-level characteristics, rather than an exhaustive set of design criteria, for data repositories
- Allow for implementation flexibility to vary across data repositories
- Remain nimble in the face of evolving technology and data sharing practices

OSTP Memo: Audience

Researchers, Librarians, & Data Managers

Assist researchers in data repository selection under data management and sharing polices

Program Officers, Funders, & Policy Makers

Identify specific repositories designated for use for particular data types & guide development of agency-supported data repositories

Data Repositories & Repository Staff

Inform the characteristics desired by an agency for sharing data resulting from federally-funded research

Updates to Policy Guidance on Increasing Equitable Access to Federally Funded Research Results

To meet core commitments, OSTP is updating policy guidance to promote improved public access to federally funded research results.

EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

Scientific Data

Underlying peer-reviewed scholarly publications resulting from federally funded research should be made freely available and publicly accessible by default at the time of publication

EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES

OFFICE OF SCIENCE AND TECHNOLOGY POLICY

Scientific Data

- *Guidelines for non-peer reviewed publishing also required*
- *Agencies required to provide guidance on repositories [in line with] "Desirable Characteristics of Data Repositories for Federally Funded Research."*

Scientific Data

Public access plans should outline the policies that federal agencies will use to establish researcher responsibilities on how federally funded scientific data will be managed and shared, including:

- *potential legal, privacy, ethical, technical, intellectual property, or security limitations*
- *plans to maximize appropriate sharing of the federally funded scientific data identified such as providing risk-mitigated opportunities for limited data access*
- *specific online digital repository or repositories where the researcher expects to deposit their relevant data, consistent with the federal agency's guidelines*

What is Scientific Data?

Does Not Include:

- laboratory notebooks
- preliminary analyses
- case report forms
- drafts of scientific paper
- plans for future research
- peer-review
- communications with colleagues
- physical objects

Includes:

- the recorded factual material
- of sufficient quality
- which validate and replicate research findings

Tables

Images

Code

Sequences

Simulations

3D

Geospatial

Selecting a right repository

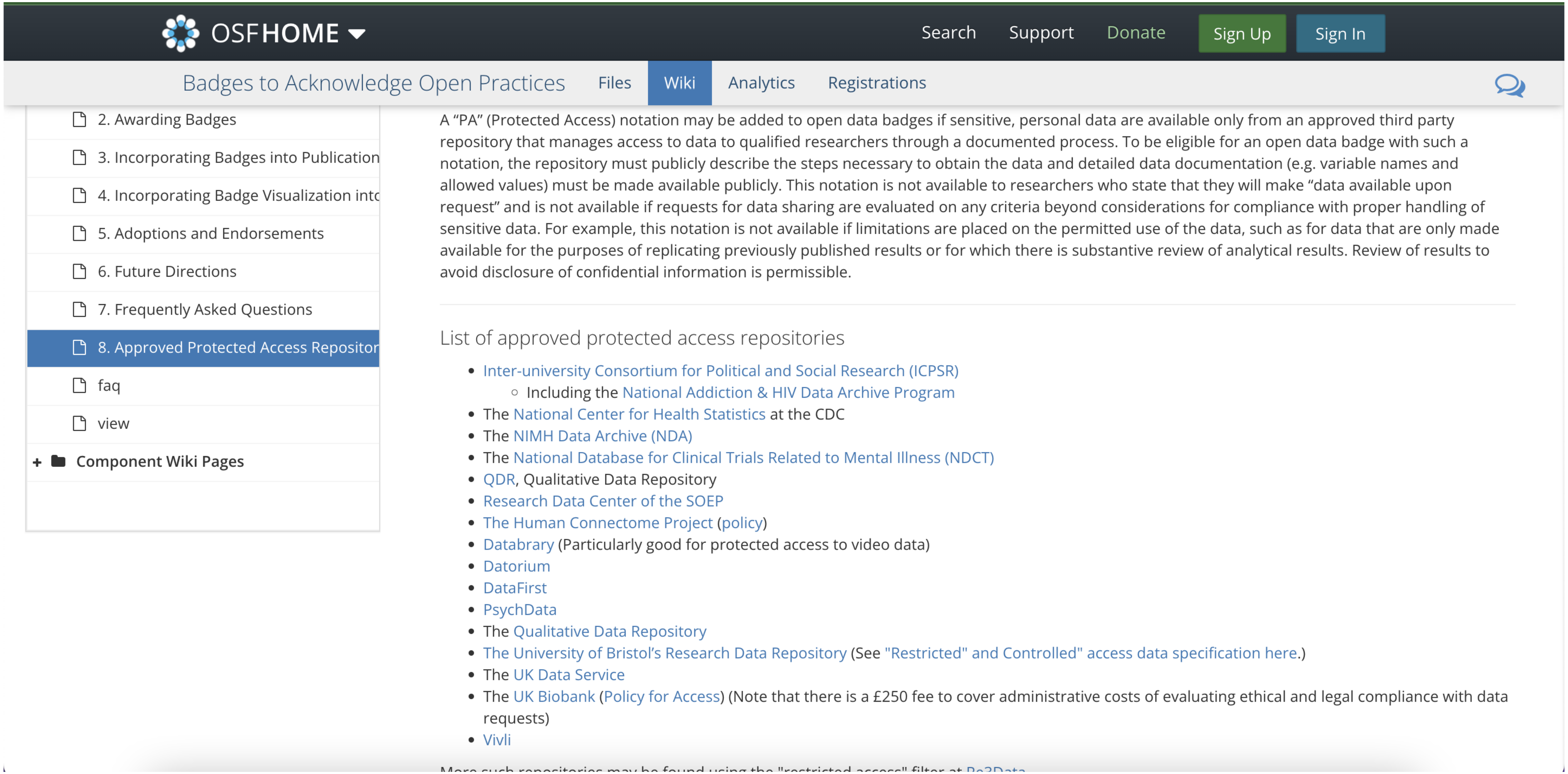
1. If your program indicates a mandatory repository, deposit there
2. If there is a specialized or disciplinary repository, consider depositing there
3. Else, choose an established repository that meets the desirable characteristics guidelines
   - Generalist repository (OSF, Figshare, Dryad, etc.)
   - Institutional repository

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

political sciences

Choosing a restricted repository


OSFHOME
Search
Support
Donate
Sign Up
Sign In
Badges to Acknowledge Open Practices
Files
Wiki
Analytics
Registrations
2. Awarding Badges
3. Incorporating Badges into Publication
4. Incorporating Badge Visualization into
5. Adoptions and Endorsements
6. Future Directions
7. Frequently Asked Questions
8. Approved Protected Access Repositor
faq
view
Component Wiki Pages

A "PA" (Protected Access) notation may be added to open data badges if sensitive, personal data are available only from an approved third party repository that manages access to data to qualified researchers through a documented process. To be eligible for an open data badge with such a notation, the repository must publicly describe the steps necessary to obtain the data and detailed data documentation (e.g. variable names and allowed values) must be made available publicly. This notation is not available to researchers who state that they will make "data available upon request" and is not available if requests for data sharing are evaluated on any criteria beyond considerations for compliance with proper handling of sensitive data. For example, this notation is not available if limitations are placed on the permitted use of the data, such as for data that are only made available for the purposes of replicating previously published results or for which there is substantive review of analytical results. Review of results to avoid disclosure of confidential information is permissible.

List of approved protected access repositories

- Inter-university Consortium for Political and Social Research (ICPSR)
  - Including the National Addiction & HIV Data Archive Program
- The National Center for Health Statistics at the CDC
- The NIMH Data Archive (NDA)
- The National Database for Clinical Trials Related to Mental Illness (NDCT)
- QDR, Qualitative Data Repository
- Research Data Center of the SOEP
- The Human Connectome Project (policy)
- Databrary (Particularly good for protected access to video data)
- Datorium
- DataFirst
- PsychData
- The Qualitative Data Repository
- The University of Bristol's Research Data Repository (See "Restricted" and Controlled" access data specification here.)
- The UK Data Service
- The UK Biobank (Policy for Access) (Note that there is a £250 fee to cover administrative costs of evaluating ethical and legal compliance with data requests)
- Vivli

More such repositories may be found using the "restricted access" filter at Re3Data

Washington University in St. Louis
University Libraries

List of restricted access repositories

What does not meet desirable characteristics of a repository

- A personal website
- A lab website
- A repository you built
- “Available on request”
- Social networking (e.g., academia.edu, researchgate)

Organizational Infrastructure

Free and Easy Access	Provides broad, equitable, and maximally open access to datasets and their metadata free of charge in a timely manner after submission
Clear Use Guidance	Ensures datasets are accompanied by documentation describing terms of dataset access and use
Risk Management	Has documented capabilities for ensuring that administrative, technical, and physical safeguards are employed to comply with applicable confidentiality, risk management, and continuous monitoring requirements for sensitive data
Retention Policy	Provides documentation on policies for data retention
Long-term Organizational Sustainability	Has a plan for long-term management of data, including maintaining integrity, authenticity, and availability of datasets; has contingency plans to ensure data are available and maintained during and after unforeseen events

Digital Object Management

Unique Persistent Identifiers Assigns a dataset a unique persistent identifier (e.g., DOI) to support data discovery, reporting, and research assessment/outputs

Metadata Ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets.

Curation/ Quality Assurance Provides or facilitates expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

Broad and Measured Reuse Ensures datasets are accompanied by metadata that describe terms of reuse and provides the ability to measure attribution, citation, and reuse of data.

Common Format Allows datasets and metadata to be accessed in widely used, preferably non-proprietary, formats consistent with standards used in the relevant disciplines

Provenance Records the origin, chain of custody, version control, and any other modifications to submitted datasets and metadata

Technology

Authentication

Supports authentication of data submitters and facilitates associating submitter PIDs with those assigned to their deposits.

Long-term Technical Sustainability

Has a plan for long-term management of data, building on a stable technical infrastructure and funding plans.

Security and Integrity

Has documented measures in place to meet well established cybersecurity criteria for preventing unauthorized access, modification, or release of data, with levels of security that are appropriate to the sensitivity of data

Additional Considerations for Human Data

Fidelity to Consent	Employs procedures to restrict dataset access and use to those that are consistent with participant consent and changes in consent
Security	Implements appropriate approaches (e.g., tiered access, credentialing of data users, security safeguards against potential breaches) to protect human subjects' data from inappropriate access
Limited Use Compliant	Employs procedures to communicate and enforce data use limitations, such as preventing re-identification or re-distribution to unauthorized users
Download Control	Controls and audits access to and download of datasets
Request Review	Makes use of an established and transparent process for reviewing data access requests
Plan for Breach & Accountability	Has security measures that include a response plan for detected data breaches and procedures for addressing violations of terms-of-use and data mismanagement

Repository Comparisons Across Elements

Organizational Infrastructure

	Institutional (WashU)	General	Domain
Free and Easy Access	yes	varies	varies
Clear Use Guidance	yes	varies	varies
Risk Management	yes	varies	varies
Retention Policy	yes	varies	varies
Long-term Organizational Sustainability	yes	varies	varies

Repository Comparisons Across Elements

Technology

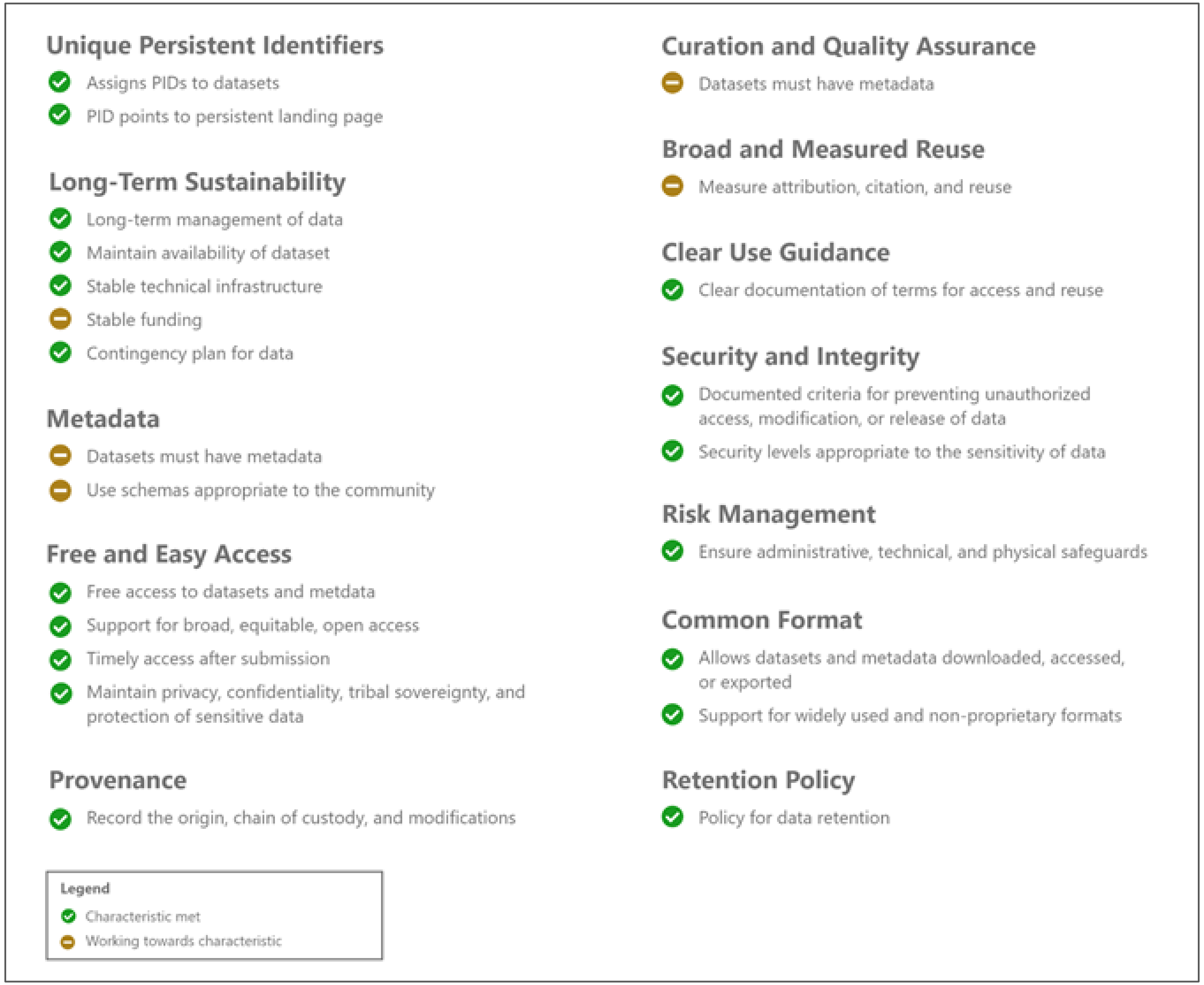
	Institutional	General	Domain
Authentication	yes	yes	yes
Long-term Technical Sustainability	yes	varies	varies
Security and Integrity	yes	varies	varies

Repository Comparisons Across Elements

Digital Object Management

	Institutional (WashU)	General	Domain
Unique Persistent Identifiers	yes	usually	varies
Metadata	yes	yes	yes
Curation / Quality Assurance	yes (free)	usually not	varies
Broad and Measured Reuse	yes	varies	varies
Common Format	yes	varies	varies
Provenance	yes	varies	varies

Case Study: Center for Open Science OSF Platform



Unique Persistent Identifiers

- ✓ Assigns PIDs to datasets
- ✓ PID points to persistent landing page

Long-Term Sustainability

- ✓ Long-term management of data
- ✓ Maintain availability of dataset
- ✓ Stable technical infrastructure
- ⊖ Stable funding
- ✓ Contingency plan for data

Metadata

- ⊖ Datasets must have metadata
- ⊖ Use schemas appropriate to the community

Free and Easy Access

- ✓ Free access to datasets and metdata
- ✓ Support for broad, equitable, open access
- ✓ Timely access after submission
- ✓ Maintain privacy, confidentiality, tribal sovereignty, and protection of sensitive data

Provenance

- ✓ Record the origin, chain of custody, and modifications

Curation and Quality Assurance

- ⊖ Datasets must have metadata

Broad and Measured Reuse

- ⊖ Measure attribution, citation, and reuse

Clear Use Guidance

- ✓ Clear documentation of terms for access and reuse

Security and Integrity

- ✓ Documented criteria for preventing unauthorized access, modification, or release of data
- ✓ Security levels appropriate to the sensitivity of data

Risk Management

- ✓ Ensure administrative, technical, and physical safeguards

Common Format

- ✓ Allows datasets and metadata downloaded, accessed, or exported
- ✓ Support for widely used and non-proprietary formats

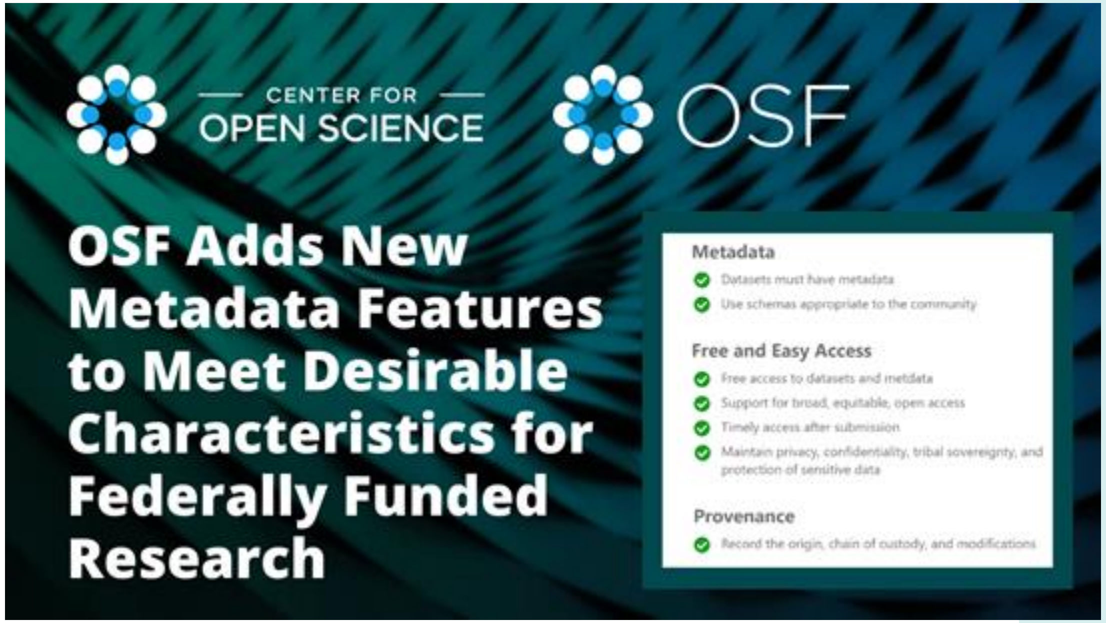
Retention Policy

- ✓ Policy for data retention

Legend
- ✓ Characteristic met
- ⊖ Working towards characteristic

← 2022 Self Assessment

↓ 2023 Announcement



OSF Adds New Metadata Features to Meet Desirable Characteristics for Federally Funded Research

Metadata
- ✓ Datasets must have metadata
- ✓ Use schemas appropriate to the community

Free and Easy Access
- ✓ Free access to datasets and metdata
- ✓ Support for broad, equitable, open access
- ✓ Timely access after submission
- ✓ Maintain privacy, confidentiality, tribal sovereignty, and protection of sensitive data

Provenance
- ✓ Record the origin, chain of custody, and modifications

New WashU Research Data Repository

Planned and developed with Desirable Characteristics in mind

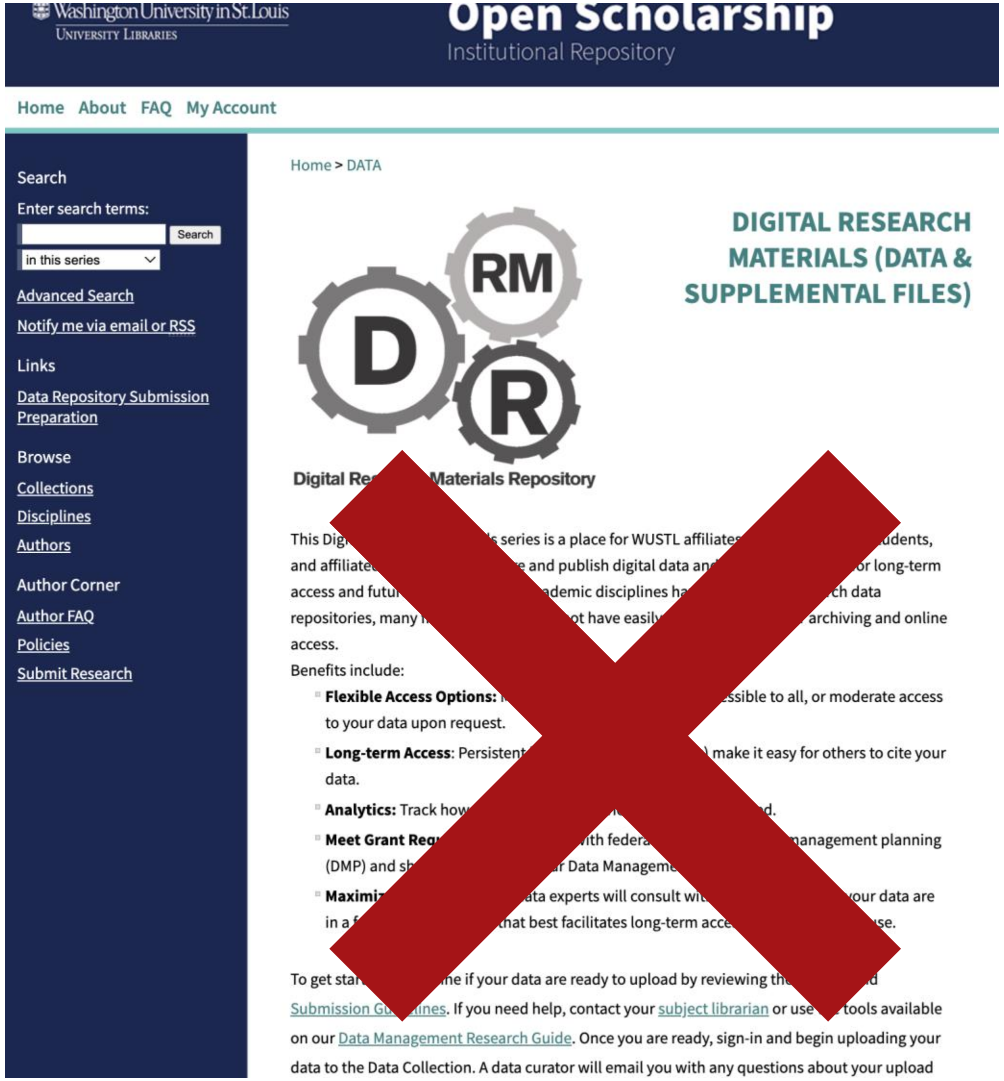
Organizational Infrastructure	Free and Easy Access
	Clear Use Guidelines
	Risk Management
	Retention Policy
	Long-term Organizational Sustainability
Digital Object Management	Unique Persistent Identifiers
	Metadata
	Curation and Quality Assurance
	Broad and Measured Reuse
	Common Format
	Provenance

Technology	Authentication
	Long-term Technical Sustainability
	Security and Integrity
Human Data	*For repositories storing de-identified data from human participants…*
	Fidelity to Consent
	Security
	Limited Use Compliant
	Download Control
	Request Review
	Plan for Breach & Accountability

✓

Washington University in St. Louis
University Libraries

From BePress collection to standalone data repository



Washington University in St.Louis
University Libraries

Open Scholarship
Institutional Repository

DIGITAL RESEARCH MATERIALS (DATA & SUPPLEMENTAL FILES)

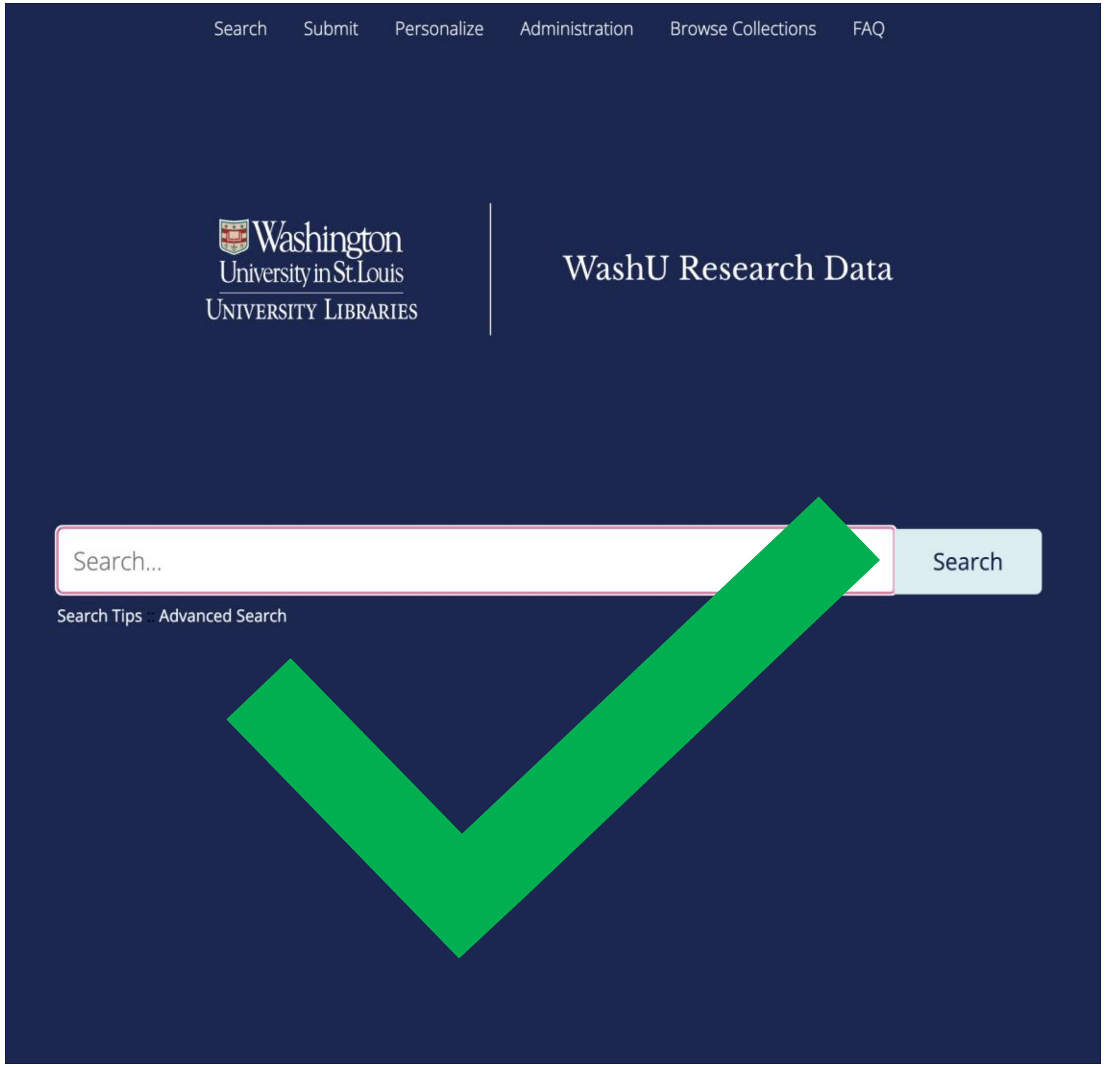
Digital Re Materials Repository

This Dig s series is a place for WUSTL affiliates udents,
and affiliate e and publish digital data an or long-term
access and futu ademic disciplines ha ch data
repositories, many h ot have easily archiving and online
access.

Benefits include:

- **Flexible Access Options:** ssible to all, or moderate access to your data upon request.
- **Long-term Access**: Persistent) make it easy for others to cite your data.
- **Analytics:** Track how d.
- **Meet Grant Req** ith federa anagement planning (DMP) and sh r Data Manageme
- **Maximi** ata experts will consult wi our data are in a f hat best facilitates long-term acce se.

To get sta ne if your data are ready to upload by reviewing the d
Submission Gu ines. If you need help, contact your subject librarian or use tools available
on our Data Management Research Guide. Once you are ready, sign-in and begin uploading your
data to the Data Collection. A data curator will email you with any questions about your upload



Search Submit Personalize Administration Browse Collections FAQ
Washington
University in St.Louis
University Libraries
Search...
Search
Search Tips :: Advanced Search

WashU Research Data

Washington University in St.Louis
University Libraries

https://data.library.wustl.edu/

Built in FAIRness checklist

- □ ORCid complete
- □ DOI assigned
- □ Award information complete
- □ Required DataCite metadata complete
- □ Recommended DataCite metatdata complete
- □ Optional DataCite metadata complete
- □ Curatorial Review complete
- □ Readme "general information" complete
- □ Readme "sharing/access information" complete
- □ Readme "data and file overview" complete
- □ Readme "methodological information" complete
- □ Readme "data specific information" complete
- □ Transformed to common format
- □ Open License Selected
- □ Suggested citation

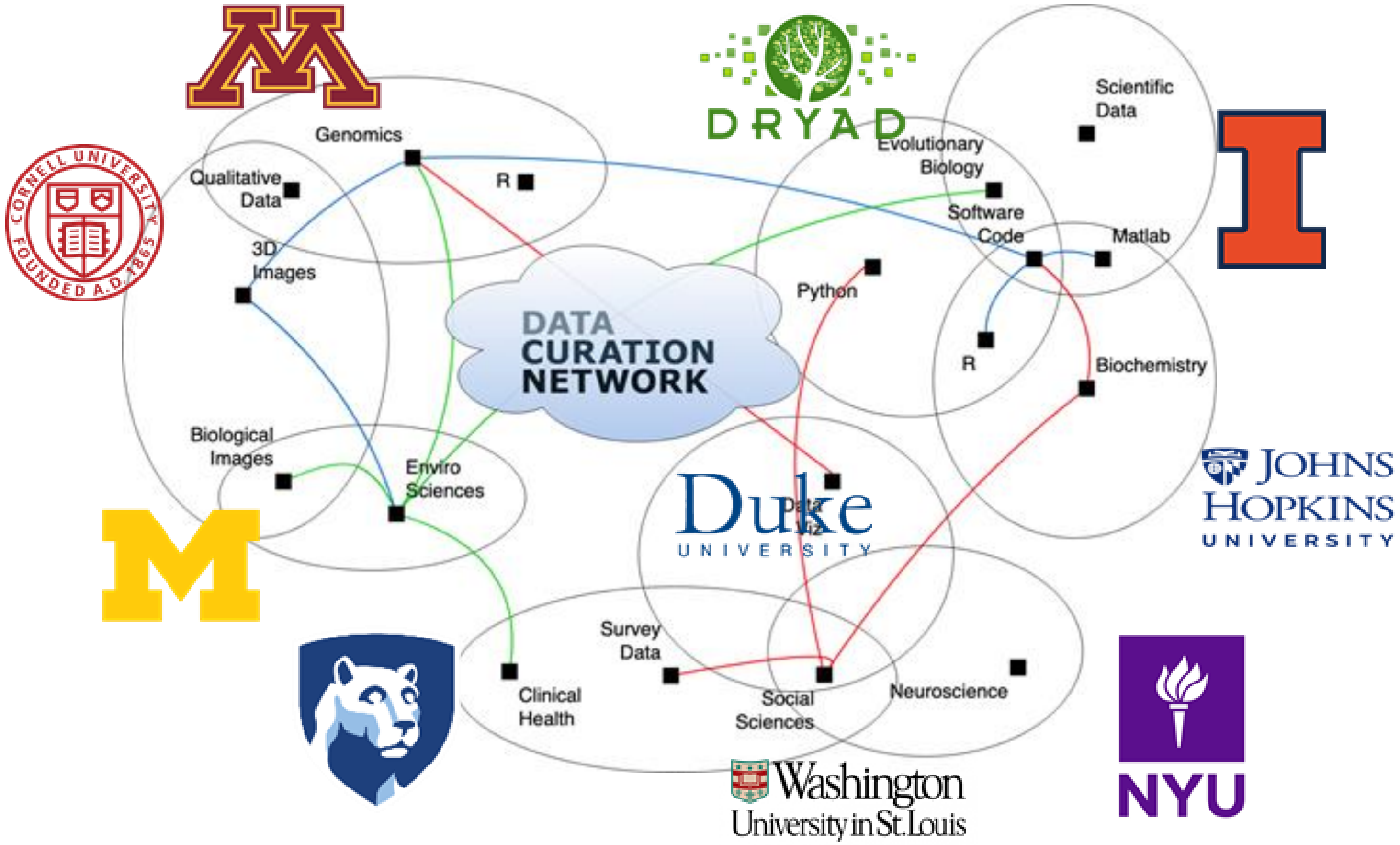
Washington University in St. Louis
University Libraries

Challenge

How to maintain data sharing services and conform to growing requirements?

Data Curation Network

Shared training and expertise



CURATE(D) Steps (checklists)

- **C**heck the files/code
- **U**nderstand the dataset
- **R**equest documentation
- **A**ugment metadata
- **T**ransform file formats
- **E**valuate for FAIRness
- **D**ocument curation log

Washington University in St. Louis
University Libraries

Thank you.

This Photo by Unknown Author is licensed under CC BY-SA-NC

Psst: lunch is next