ASU Knowledge Enterprise Library Partnership

{

Session 1
Data Management Planning
Before Supercomputing

Session 2
A collaborative approach in
research data sharing at ASU

RMACC 2023
Rocky Mountain Advanced Computing Consortium
SkySong, Arizona State University, Scottsdale, AZ
May 17, 2023

Presented to The 13th annual RMACC HPC Symposium

Data Management Planning Before Supercomputing

Kathryn Claypool, MPH
https://orcid.org/0000-0003-0557-8330

Research Data Management Office
Research Technology Office
**Knowledge Enterprise**

Session 1:
Data management planning is an integral step in the research data life cycle. Large amounts of data and lengthy code accompanying supercomputing runs are no exception. Planning before analysis will benefit research and the researcher by providing a clear strategy for collecting, storing, analyzing, and sharing the data at the end of the research cycle. Supercomputing can require significant storage beyond scratch space, but researchers typically need to be informed of what tools are appropriate and available. Framed within the planning phase of the life cycle, this presentation presents ASU's Storage Selector as a quick and easy tool to find the most appropriate storage resources provided by the university to help researchers choose a proper storage and management solution for their research data at the right time in their project. We will also explore the DMP Tool, developed by the California Digital Library, which provides a resource-rich platform for writing data management plans, including institutional-specific guidance, feedback request, and public plans that can be used as guides.

Open Access

Benefits

- Researchers – Easier to find and use literature, relevant datasets, code blocks
- Institutions – Evens the playing fields for smaller institutions, bringing more competition
- Business – increase employment opportunities, better workforce
- Funders – higher return on investment (ROI) when research data are shared and leveraged. Less duplication of effort.

Create

Explore

Re-use

Share

Preserve
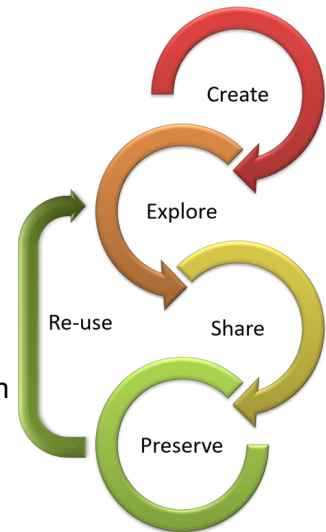
Our work and partnership are driven by our efforts to promote Open Access to scholarly output. We focus on the benefits to researchers over the need to comply with funder mandates.

The 2023 NIH Data Management and Sharing Policy

Previously, the NIH only required grants with $500,000 per year or more in direct costs to provide a brief explanation of how and when data resulting from the grant would be shared.

The 2023 NIH policy is entirely new. Beginning **January 25, 2023**, **ALL** grant applications or renewals that generate Scientific Data must include a **robust and detailed plan** for managing and sharing data during the entire funded period. This includes information on data storage, access policies/procedures, preservation, metadata standards, distribution approaches, and more. You must provide this information in a **data management and sharing plan (DMSP)**. The DMSP is similar to what other funders call a data management plan (DMP).

The **DMSP** will be assessed by NIH Program Staff (though peer reviewers will be able to comment on the proposed data management budget). The Institute, Center, or Office (ICO)-approved plan becomes a Term and Condition of the Notice of Award.

The 2023 NIH Data Management and Sharing Policy has spiked interests and requests towards open data sharing.
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

This screenshot is from the ASU Library guide on the new policy
https://libguides.asu.edu/NIH-2023-DMS/policy

Data Should Be **FAIR**

Findable	**A**ccessible

Persistent Identifiers (PIDs)
Rich metadata
Indexed data repositories
PIDs in metadata

Standard communications protocol
Open, free protocol
Authentication, where necessary
Metadata is alw available

Interoperable	**R**eusable

Vocabularies
Vocabularies are FAIR
Linked metadata

Metadata have multiple attributes
Usage license
Provenance
Community standards

NIH encourages data management and sharing practices to be consistent with the FAIR data principles and reflective of practices within specific research communities.

Data Management Planning

Data management planning includes:

- Data collection plans
- Secure storage during data collection and analysis
- Sharing with collaborators during research
- Sharing datasets for reproducibility and transparency after or with publication

Sharing

Storing

Analyzing
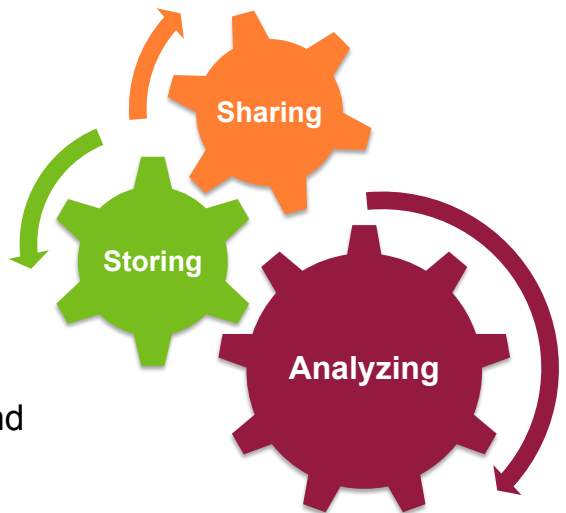
Data management planning is about thinking of basic activities and practices that researchers and teams will agree upon that include actions that happen during the active research process when they will be using research computing and high performance computing resources and actions that they will then follow up on as they prepare to publish their data for reproducibility and transparency after their project is complete.

Research Data Storage Selector

ASU provides a selection tool to give researchers an initial place to find out what options the university provides for storing research data. The web page is adapted from the open source Data Storage Finder developed by Cornell University and available on Github: https://github.com/CU-CommunityApps/CD-finder

Sections

Data Management and Sharing Plan

Data Types

Related Tools, Software, and Code

Standards for data and metadata

Data Preservation, Access, and Associated Timelines

Access, Distribution, or Reuse Considerations

Oversight of Data Management and Sharing
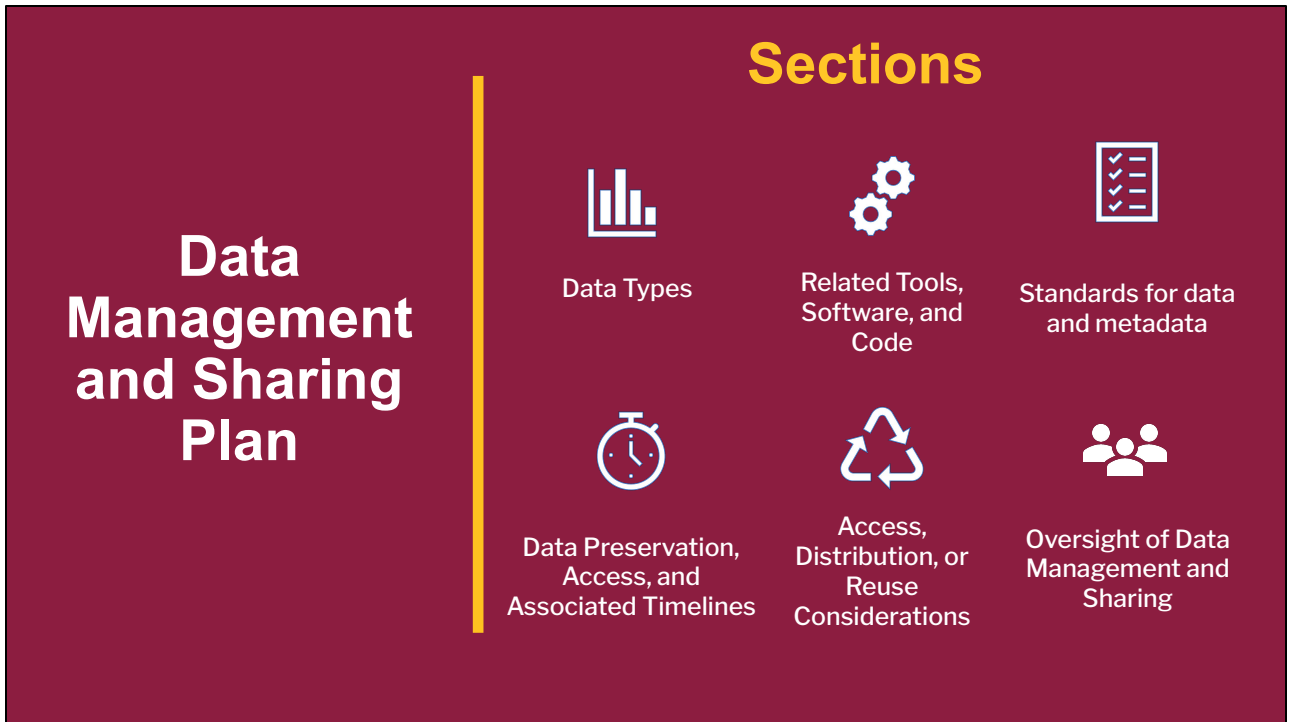
Funder data management and sharing plan requirements appear in different orders and sometimes are broken up or combined but typically all contain six basic elements.

1. Data Types
2. Related Tools, Software, and Code
3. Standards for data and metadata
4. Data Preservation, Access, and Associated Timelines
5. Access, Distribution, or Reuse Considerations
6. Oversight of Data Management and Sharing

The California Digital Library created the DMPTool as a resource to help researchers fill out their data management and sharing plans. While not mandiated ASU strongly recommends using it.
https://dmptool.org

Funder Requirements

Templates for data management plans are based on the specific requirements listed in funder policy documents. The DMPTool maintains these templates, however, researchers should always consult the program officers and policy documents directly for authoritative guidance. Sample plans are provided by a funder or another trusted party.

Template Name ▲	Download	Organization name ⬍	Last Updated ⬍	Funder Links	Create a new plan	Sample Plans (if available)
Alfred P. Sloan Foundation	📄 📄	Alfred P. Sloan Foundation (sloan.org)	10-25-2021	Sloan Grant Proposal Guidelines	➕🗑	
Arctic Data Center: NSF Polar Programs	📄 📄	National Science Foundation (nsf.gov)	10-25-2021	NSF Arctic Data Center DMP Resources NSF Proposal & Award Policies & Procedures Guide (PAPPG)	➕🗑	
BCO-DMO NSF OCE: Biological and Chemical Oceanography	📄 📄	National Science Foundation (nsf.gov)	10-25-2021	NSF OCE Sample and Data Policy NSF GEO DIrectorate Guidance	➕🗑	
DataWorks! Data Management and Sharing Plan Challenge	📄 📄	Federation of American Societies for Experimental Biology (faseb.org)	12-09-2021	Call for Submissions to the DataWorks! Data Management and Sharing Plan Challenge Final NIH Policy for Data Management and Sharing Elements of an NIH Data Management and Sharing Plan Selecting a Repository for Data Resulting from NIH-Supported Research	➕🗑	

The DMPTool provides funder based templates and access to their data management and sharing plan requirements

Users can also find published examples of other plans that are browsable by funding agency, institution, and subject.

User complete a basic form and select the funder to get started and are matched with their institutions local guidance

ASU
Arizona State University

Recycling of Small Packaging

| Project Details | Collaborators | Write Plan | Research outputs | Finalize | Download |

This plan is based on the "Digital Curation Centre" template provided by Digital Curation Centre (dcc.ac.uk) - (ver: 2, pub: 2021-10-25).

expand all | collapse all

13/13 answered

+ Data Collection (2 / 2)   +

+ Documentation and Metadata (1 / 1)   +

+ Ethics and Legal Compliance (2 / 2)   +

+ Storage and Backup (2 / 2)   +

+ Selection and Preservation (2 / 2)   +

+ Data Sharing (2 / 2)   +

+ Responsibilities and Resources (2 / 2)   +

The tool also breaks down the plan based on the select funder template and each section will provide guidance targeted at those components
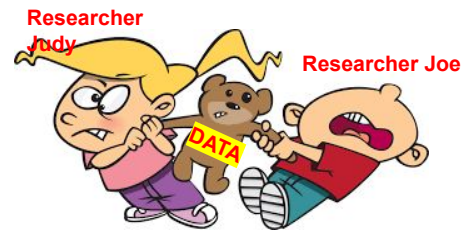
Data Sharing

NIH Policy

- Other Federal funding agencies will be requiring

  data sharing along with publication

Repositories

- Institutional (ASU), discipline-specific (tDAR),

  general (OSF)

- There can be storage limits and costs to evaluate

- Limits for restricted datasets

DUAs

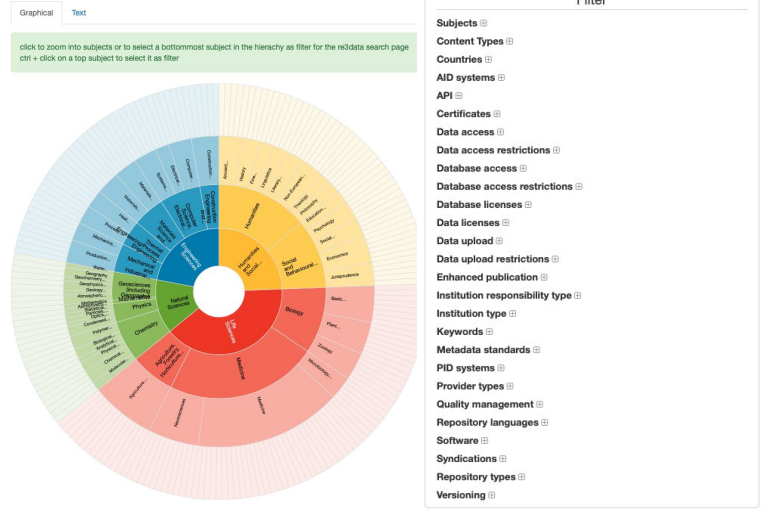**Researcher Judy**

Researcher Joe

DATA

Types of Data Repositories

- Discipline or domain specific (tDAR for Archaeology)

- Institutional (ASU Research Data Repository, KEEP)

- Generalist domain, agnostic (e.g. Zenodo, Dryad, Figshare, etc.)

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Browse by subject

Graphical | Text

click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
ctrl + click on a top subject to select it as filter

Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- AID systems ⊞
- API ⊞
- Certificates ⊞
- Data access ⊞
- Data access restrictions ⊞
- Database access ⊞
- Database access restrictions ⊞
- Database licenses ⊞
- Data licenses ⊞
- Data upload ⊞
- Data upload restrictions ⊞
- Enhanced publication ⊞
- Institution responsibility type ⊞
- Institution type ⊞
- Keywords ⊞
- Metadata standards ⊞
- PID systems ⊞
- Provider types ⊞
- Quality management ⊞
- Repository languages ⊞
- Software ⊞
- Syndications ⊞
- Repository types ⊞
- Versioning ⊞

- NIH recommends that you follow the following guidance when selecting a data repository:
- Utilize the NIH and/or Institute, Center, or Office identified data repository if one exists for your program or data type
- If one doesn't exist, select a data repository that is appropriate for the data generated from the research project and is in accordance with the desired characteristics first considering data repositories that are discipline or data-type specific. If no appropriate discipline or data type specific repository is available, look to institutional or generalist repositories.

A collaborative approach in research data sharing at ASU

ASU Library
Arizona State University

Matthew Harp
orcid.org/0000-0001-6136-851X
Research Data Initiatives Librarian
Open Science and Scholarly Communication
ASU Library

Session 2:
This presentation provides an overview of the ongoing working relationship between the ASU Library Open Science and Scholarly Communication division, Research Data Management Office, and Research Computing. We will explore these teams' interdisciplinary relationships and interdependence as the institution increasingly supports open science practices and initiatives. We will include case studies regarding the decision-making process, data sharing decisions, and opportunities and challenges that arise when transferring research data from a high-performance computing environment to the ASU Research Data Repository. Finally, we will share lessons learned as we intentionally shepherd research data from active project management and storage to final publication and preservation.

Open Science/Open Scholarship definition

"

Open science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity."
- *White House Office of Science and Technology Policy (OSTP), January 11, 2023*

Introduction to Open Scholarship

https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/

Our partnership is informed and drive by commitment to responsibility share the knowledge our university creates with our community. Open Science and Scholarship are core to ASU's identity as our research outputs are ultimately open education and research resources. It compliments Michael Crows vision and his role as a co-chair of Higher Education Leadership Initiative for Open Scholarship (HELIOS).

Interdependent Relationships

Goal: Provide comprehensive support to ASU researchers for everything from data management planning and to open research data sharing

New Repository Tools Launched

KEEP Institutional Repository: scholarship produced by ASU faculty, staff and students

Research Data Repository: research dataset access, discovery and preservation

lib.asu.edu/research

Key University Partnerships

Knowledge Enterprise
- Research Data Management
- Research Computing

University Technology Office

researchdata.asu.edu

The Library has a number of resources and personnel tied directly to Open Science initiatives. The **KEEP Institutional Repository** supports requirements for open sharing of research articles and presentations. The **ASU Research Data Repository** is one of several publication and archiving resources for meeting open requirements for research data.

This works through our partnership with Knowledge Enterprise's Research Data Management Office and Research Computing. This partnership has been a key factor in expanding ASU's data management and sharing capabilities.

The partnership includes education and outreach efforts, formal agreements, and ongoing work towards the development of workflows to streamline the process of making research output more open and accessible.

This portion of our presentation is a quick and high-level look on the publication and archiving phase - the lower half of the cycle - establishing a framework for 'reuse' which many researchers and their students may not be familiar with. Lots of misunderstandings on the difference between active storage and final archiving.. They are not the same thing

What we have found across the spectrum as that there are many folks who are unprepared for sharing their work openly. There are misconceptions on costs and also a lack of preparedness.

3 types of data repositories

Select the most appropriate options

…and code repositories

1. **Disciplinary**

    Data ARchives and Transmission System (DARTS) - space

    tDAR - archeological

    Qualitative Data Repository (QDR) and ICPSR - social science

2. **General**

    Zenodo

    OSF

    Dryad

3. **Institutional (data)**

    ASU

    UA

Discipline-Specific Data Repositories
- [DARTS](#) - multi-disciplinary space science data archive
- [tDAR](#) - archeology data, ASU based but accepts data from outside,
- [QDR](#) - qualitative and multi-method research in the social sciences and related disciplines
- [ICPSR](#) (Institute for Social and Political Research) services for both public-use and restricted-use data,

General repositories are typically do-it-yourself resources and some will provide additional services for a fee which can be allocated for in a proposal
- Zenodo is a general-purpose open repository for research papers, data sets, research software, reports, and any other research related digital artefacts https://zenodo.org/
- The OSF is a similar option but it's more of a collaborative platform and is particularly useful for replication and prepublication workflows https://osf.asu.edu
- Dryad (formerly a biology and ecology repository is now a curated general-purpose repository that provides open access to research data but there are costs associated with Dryad submission. Researchers would need to contact Dryad for details. https://datadryad.org/

Institutional Repositories are typically interdisciplinary but require university affiliation with a respective institution in order to deposit.

- They are typically managed by university libraries and may or may not charge researchers for curation and submission services.
- The [ASU Research Data Repository](#) (Dataverse) or the U of A's [ReData](#) (Figshare) are examples.

An institutional repository is not the default repository

Verify with the repository on any costs associated with archiving and publishing

ASU Research Data Repository

Interdisciplinary research data sharing and archiving

Publication and preservation platform

Indexed as scholarly works

Meets funding agency and institutional retention policy requirements **dataverse.asu.edu**

The ASU Research Data Repository helps ASU affiliated researchers share, store, preserve, cite, explore, and make research data accessible and discoverable.

It is an interdisciplinary repository launched two years ago to serve as both a publication and preservation platform for research datasets. You can find the repository at dataverse dataverse.asu.edu

- The research data repository is a dedicated research data management service platform that serves in the publication and reuse represented in the re-use, share, and exploratory stages of the research lifecycle.
- This repository is intended for public sharing of research data aiding the discoverability of datasets through scholarly indexes and in our general library search so that your works show up along with all the other resources we provide
- Meets funder and institutional requirements

A publication process

Storage
Needs vary

Not a replacement for Google, Dropbox, or other cloud storage solutions

Research data repositories are fixed, selected storage

Curation
Not everything can or should be shared

Sharing research is intentional, informed and requires work

Plan ahead

Open repositories not for restricted or sensitive data

Metadata and Documentation
Metadata vary across the lifecycle and disciplines

README for active research

README metadata for discovery

Reviewed and approved like a manuscript

Storage When repository was first launched in 2020 conversations were focused on storage space but that is misleading

It is not a replacement for Google Drive, Dropbox, or other active data management storage systems

Curation Storage costs are important but not the real work that libraries do in the research space. Similar to shelf space there is work in selection/curation, description/cataloging/metadata, and providing access (who has permission to access and when), there is even use metrics, citation etc.

Metadata What we need (more documentation and organization) but what we encounter are lots of files, little documentation, and data not ready for publishing. There are other components that are just as important like software, code, and methods that may live in other platforms like GitHub, Protocols.io, and the OSF that the repository metadata can record and help build those connections to

This is information begins at the ideation stage such as in a data management and sharing plan and continues to be gathered throughout a project

ASU Library Research Data Repository

(Arizona State University)

dataverse.asu.edu

Metrics | 19,847 Downloads | Contact | Share

Share, find and cite research data produced at Arizona State University.

The Arizona State University (ASU) Research Data Repository provides a platform for ASU-affiliated researchers to share, preserve, cite, and make research data accessible and discoverable. The ASU Research Data Repository provides a permanent digital identifier for research data, which complies with data sharing policies. The repository is powered by the Dataverse open-source application, developed and used by Harvard University. Both the ASU Research Data Repository and the KEEP Institutional Repository are managed by the ASU Library to ensure research produced at Arizona State University is discoverable and accessible to the global community.

Search this dataverse... | Advanced Search

☑ Dataverses (20)
☑ Datasets (57)
☐ Files (8,743)

Dataverse Category
Research Project (7)
Researcher (4)
Research Group (3)
Department (2)
Laboratory (2)

More...

Publication Year
2023 (6)
2022 (24)
2021 (20)

1 to 10 of 77 Results | Sort ▾

2090-2099 Projected Climates and Urban Development Scenarios - Conterminous U.S. (CONUS) Simulation Data
Aug 31, 2022 - United States Regional Climate Change Assessment
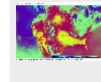
Georgescu, Matei; Brandi, Aldo; Broadbent, Ashley; Krayenhoff, Scott, 2021, "2090-2099 Projected Climates and Urban Development Scenarios - Conterminous U.S. (CONUS) Simulation Data", https://doi.org/10.48349/ASU/3TYXZI, ASU Library Research Data Repository, V3

Simulations representing different climate change, urban development and heat adaptation strategies scenarios for the future Conterminous United States. Refer to CONUS_Sims_README.txt for additional documentation.

Active Schools & Communities Lab at Arizona State University (Arizona State University)
Mar 14, 2023

The Active Schools & Communities Lab works to improve access and opportunities for physical activity, with a focus on school settings. This collection includes tools and datasets generated from affiliated research projects.

The ASU Research Data Repository is an option for ASU-affiliated researchers to share, download, archive, cite, explore, and make research data accessible and discoverable.

Submission of datasets is limited to ASU-affiliated projects and people. The use of datasets and material published in the repository is open to anyone except where otherwise noted due to legal or ethical restrictions.

Visit: https://dataverse.asu.edu

Dataset-Level DOIs

COVID Future Wave 1 Survey Data v1.0.0

Salon, Deborah; Conway, Matthew Wigginton; Capasso da Silva, Denise; Chauhan, Rishabh; Shamshiripour, Ali; Rahimi, Ehsan; Mirtich, Laura; Khoeini, Sara; Mohammadian, Kouros; Derrible, Sybil; Pendyala, Ram, 2021, "COVID Future Wave 1 Survey Data v1.0.0", https://doi.org/10.48349/ASU/QO7BTC, ASU Library Research Data Repository, V1, UNF:6:Z61cxFCm14zzNxiO4fCbfA== [fileUNF]

Cite Dataset ▾   Learn about Data Citation Standards.

...u, Mohammadian, Kouros, Derrible, Sybil, ...

0", https://doi.org/10.48349/ASU/QO7BTC, A ...

...4fCbfA== [fileUNF]

https://www.doi.org/

As part of our service to provide persistent and citable access to research datasets, we provide Digital Object Identifiers (DOI) at the dataset record level for **published** datasets. DOIs are durable, permanent urls for citing your work and getting credit where it is due. The feature is good for timing with an article publishing process where you need to cite your own dataset availability.

These DOI are not activated until a dataset is published but you can create private URLS for sharing datasets amongst colleagues or peer reviewers before publication.

Discovery

Publishing data in the repository supports discoverability of datasets through scholarly indexes and in our general library search so that their datasets show up along with other scholarly resources. It also doesn't have a pay wall like a subscription fee.

This screenshot demonstrates a dataset record from the repository and on the right a choice limit search results to dataset resource types.

File ingest

- [Web Interface](#) (Default)
- Dropbox
- [Command Line](#)
- [Direct Upload](#)
- [Globus](#) (Not yet)

All file types are supported for upload and download in their original format. If you are uploading Excel, CSV, TSV, RData, Stata, or SPSS files, see the guides for tabular support and limitations.

Upload with HTTP via your browser ⌃

Select files or drag and drop into the upload widget. Maximum of 1,000 files per upload. Ingest is limited to the following file sizes based on their format: sav: 0 B.

＋ Select Files to Add

Drag and drop files here.

Select files from Dropbox.

⬙ Upload from Dropbox

Done

Downloading

Web interface, [Download Manager](#), or script

There are multiple ways of getting files into the repository all with their prose and cons. Downloading is typically limited to the web interface but users can take advantage of download managers to monitor the process without having selectively download each file.

- ASU's repository primarily S3 with additional options
- Not just about links but curation and documentation
- Data management doesn't end here

Additional storage options

Data management practices still apply

The repository is a web based cloud system. Large files may require essentially a sneaker network of direct drive shipping

The curated publically available data may reside in a research data repository but for ease of access and security they may still need to utilize local or on-premise resources. This is where are partnership continues. We make informed decisions on resource provisioning and determining where data can be stored and who should have access to it.

Size and data security classifications may require metadata only records or a new feature of remote trusted storage which ASU has yet to fully test

Use case:

Unlimited Google support no longer a thing. Scratch space is temporary. So where do they go next?

This example [Projected Climates and Urban Development Scenarios](https://doi.org/10.48349/ASU/3TYXZI) is a single dataset that accounts for almost the entire 'storage' of our repository.

A major challenge with this project was sheer size of the collection and included Zip files that were much larger than the 3-5 gig web interface limitation.

Artificial Social Intelligence for Successful Teams is a published dataset from a DARPA funded project that study human subject interactions using Minecraft.

They had file organization challenges, documentation needs, and needed to develop an understanding of what files should be shared for the purpose of reproducibility. Even though we were able to work with them to get their files into the repository accessing them is still a challenge for end users. The number of files presents indexing challenges and performance issues when working with the record and like the other dataset means users will need to use a download manager or other option for accessing the entire dataset.

Featured Dataset: https://doi.org/10.48349/ASU/BZUZDE

Note: Large files take time to download and preview

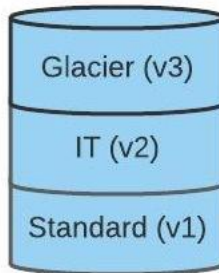# S3 Versioning and Hidden Costs
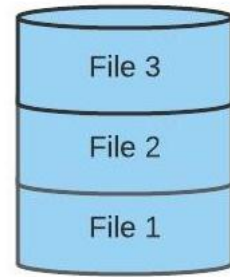
Deleting a file doesn't really delete it	Transitioning creates a new file	Dataverse doesn't know about S3
delete marker / File (V1)	Glacier (v3) / IT (v2) / Standard (v1)	File 3 / File 2 / File 1
Delete File	**Transition File**	**Replace File**
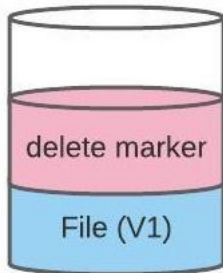
These files are all stored in the cloud. AWS makes "version" and dataverse sees any re-upload as a "new" file .. plus we are copying everything to another AWS account (which has versioning) AND we are copying everything to wasabi

In the previous use cases we were dealing with a lot of files, and many deletes and re-uploads. We found that this was significant in relation to how AWS S3 handles file versions. First, if you delete a file, it doesn't really delete the file. it creates a new version which is a delete marker, but retains the other copy of the file as a previous version.

Also, when you transition a file to a different storage class (which we are doing to save storage cost), it doesn't just change the storage status of the current file, it creates a new file in the second storage class. So, you have multiple versions of the same file.

Finally, Dataverse manages it's own versioning, and is not aware of S3 at all .. so if a user replaces a file, or deletes and re-uploads a file, the second file is considered a NEW FILE in S3, so again you have multiple copies of the same file, which adds up to increased costs that we were not really aware of.

Lessons Learned

Change the conversation from storage to publication

Developing pathways from research computing to publication

Documentation is just as important as the dataset files

As we shepherd research data from active project management and storage to final publication and preservation, proper documentation and vetting will be required. We are emphasizing that this is a publication process that has intentional actions that requires work on both the researcher and those of us in the support side.

There are two considerations when sharing data. First is how users will access datasets files and documentation the other is that anything that goes into our repository gets duplicated.

So we need to be upfront and request information and documentation early. For example, would a layperson know what they are looking at? Ask and ask again for documentation - sometimes it's just about getting on the same page.

We also realized throughout this project the need to make sure that our researchers have a clear organization structure (especially if they have many, many files) before doing any uploading into a repository.

Finally, we realize that other organizations have probably faced similar challenges and we welcome any advice and suggestions on better approaches. We would love to talk to any community members who are facing the same issues.

Questions?

Thank you!

Jon Rawlinson, CC BY 2.0, via Wikimedia Commons

jonrawlinson.com

There is a lot of work ahead including more planning, more outreach, new agreements and responsibilities, and more collaboration. Part of our goal is to develop a proactive communication strategy that gets in front of research teams as early so they can plan for their data sharing commitments that are waiting down the road.