Data Management

Mike Ramshaw

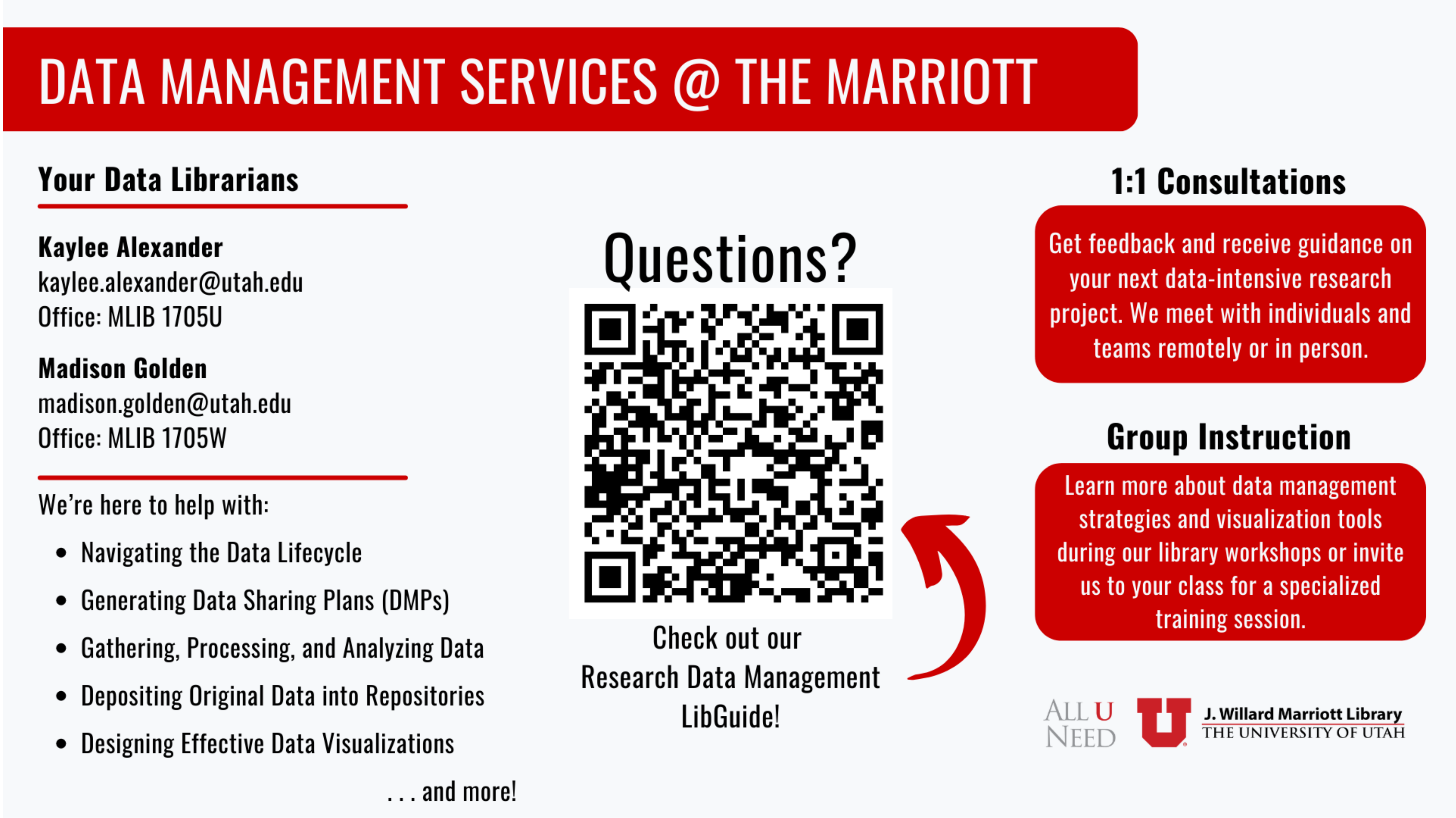
michael.ramshaw@utah.edu

Special Guests from the Marriott Library:

Kaylee Alexander & Madison Golden

kaylee.alexander@utah.edu madison.golden@utah.edu

Spring 2026



DATA MANAGEMENT SERVICES @ THE MARRIOTT

Your Data Librarians

Kaylee Alexander
kaylee.alexander@utah.edu
Office: MLIB 1705U

Madison Golden
madison.golden@utah.edu
Office: MLIB 1705W

We're here to help with:

- Navigating the Data Lifecycle
- Generating Data Sharing Plans (DMPs)
- Gathering, Processing, and Analyzing Data
- Depositing Original Data into Repositories
- Designing Effective Data Visualizations

. . . and more!

Questions?

Check out our Research Data Management LibGuide!

1:1 Consultations

Get feedback and receive guidance on your next data-intensive research project. We meet with individuals and teams remotely or in person.

Group Instruction

Learn more about data management strategies and visualization tools during our library workshops or invite us to your class for a specialized training session.

ALL U NEED
J. Willard Marriott Library
THE UNIVERSITY OF UTAH

What is data and data management?

The Office of Management and Budget (OMB) defines research data as:
“...*the recorded factual material commonly accepted in the scientific community as necessary to validate research findings*...”

The National Library of Medicine (NNLM) defines data management as:
"Research data management is the organization, documentation, storage, and preservation of the data resulting from the research process..."

Why Manage Data?

- Prevent data loss
- Efficiency -- better organization saves time
- Standardize practices
- Promotes reproducible research
- Ease of data sharing – increase the visibility of your work
- Required to meet institutional requirements
- Documentation for Intellectual Property (IP) concerns and other regulatory concerns
- Required by funding agencies

Goal of data management is to ensure data are well-managed in the present, and prepared for preservation in the future

Data Lifecycle

- Planning
  - What information, format, amount
- Documenting
  - Metadata, vocabulary
- Organizing
  - Version control & where stored
- Storing
- Access
  - Who, how
- Preservation
  - Where, software, media

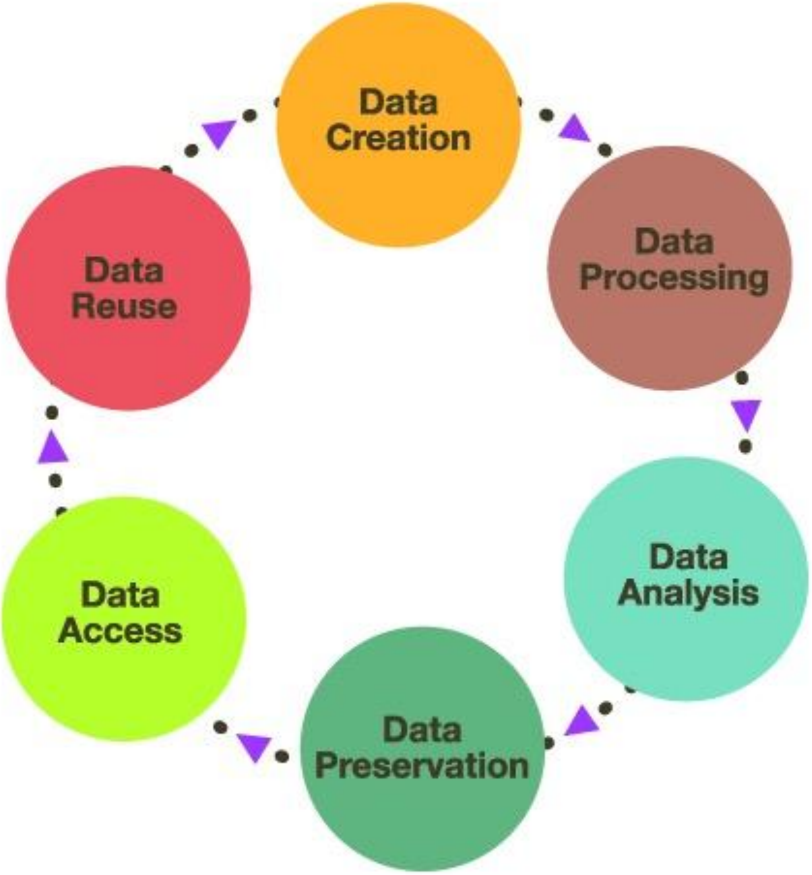


Figure from https://libguides.ntu.edu.sg/rdm/researchdatalifecycle/

Data Management Essentials

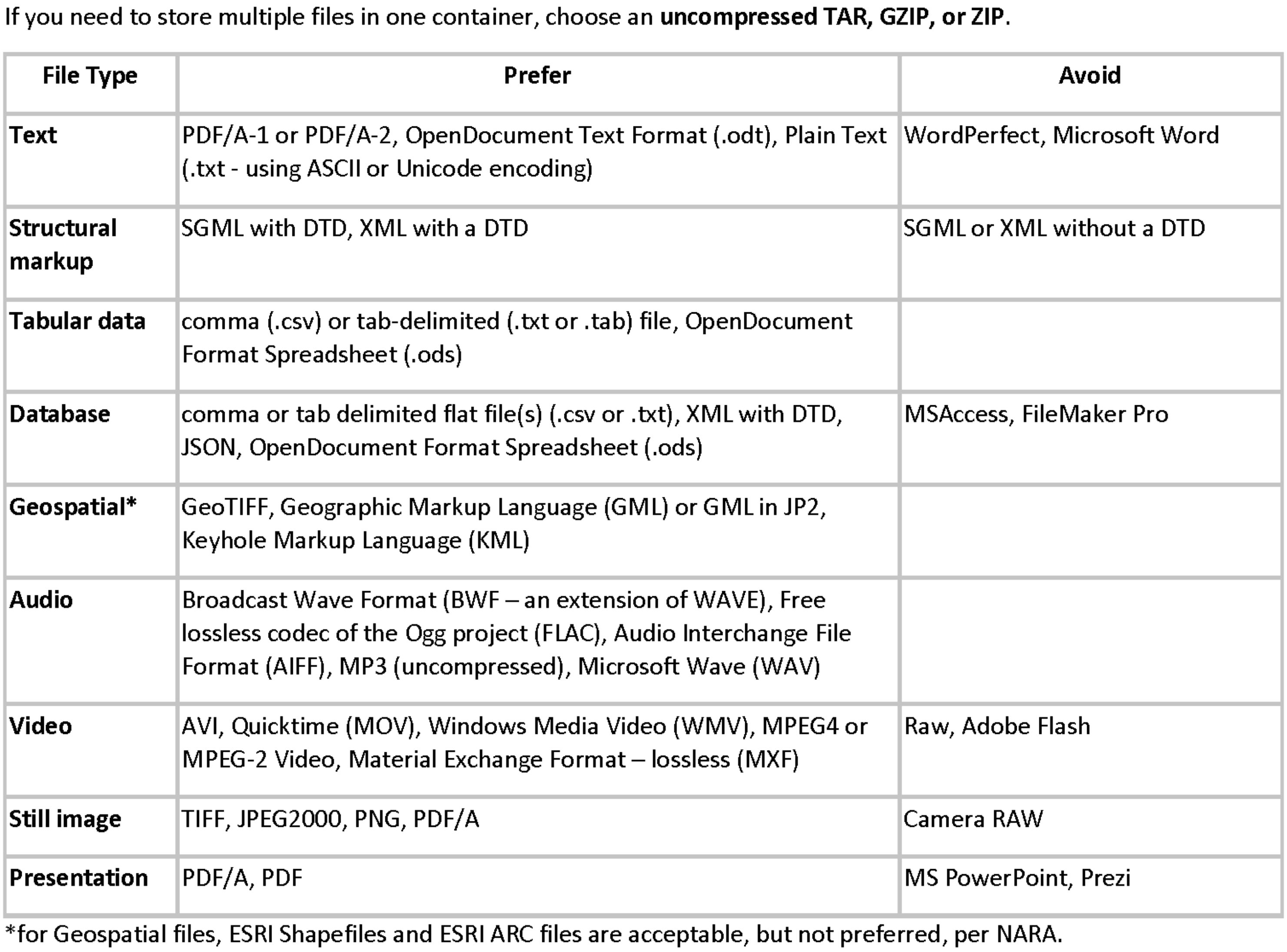
- Keep in **sustainable formats**
- Include **metadata**
- Organize
- Store and back them up
- Keep them secure

Have a plan in place before you start data collection!

Good reference for best practices:

https://guides.library.stanford.edu/data-best-practices

Sustainable formats



If you need to store multiple files in one container, choose an **uncompressed TAR, GZIP, or ZIP**.

File Type	Prefer	Avoid
Text	PDF/A-1 or PDF/A-2, OpenDocument Text Format (.odt), Plain Text (.txt - using ASCII or Unicode encoding)	WordPerfect, Microsoft Word
Structural markup	SGML with DTD, XML with a DTD	SGML or XML without a DTD
Tabular data	comma (.csv) or tab-delimited (.txt or .tab) file, OpenDocument Format Spreadsheet (.ods)	
Database	comma or tab delimited flat file(s) (.csv or .txt), XML with DTD, JSON, OpenDocument Format Spreadsheet (.ods)	MSAccess, FileMaker Pro
Geospatial*	GeoTIFF, Geographic Markup Language (GML) or GML in JP2, Keyhole Markup Language (KML)	
Audio	Broadcast Wave Format (BWF – an extension of WAVE), Free lossless codec of the Ogg project (FLAC), Audio Interchange File Format (AIFF), MP3 (uncompressed), Microsoft Wave (WAV)	
Video	AVI, Quicktime (MOV), Windows Media Video (WMV), MPEG4 or MPEG-2 Video, Material Exchange Format – lossless (MXF)	Raw, Adobe Flash
Still image	TIFF, JPEG2000, PNG, PDF/A	Camera RAW
Presentation	PDF/A, PDF	MS PowerPoint, Prezi

*for Geospatial files, ESRI Shapefiles and ESRI ARC files are acceptable, but not preferred, per NARA.

Metadata

- **Structured** information about data
  - a shorthand representation of the data
- Enhances data discoverability and reuse
  - Allows you to easily find and reuse your own data
  - Enables you to discover, evaluate, and reuse the data of others
  - Helps others discover, reproduce, reuse, and cite your data
- Metadata standards by discipline
  - – https://www.dcc.ac.uk/resources/metadata-standards
  - If no standards – be consistent and document system

Organization

- Identify and keep track of what data you have & where it is
- Define what you need to keep
- Organize by folders
- Have a README text file documenting structure details
- Subfolders with consistent naming convention

What’s in a Filename?

- Be **consistent and descriptive** such that file names allow for identification
- Consider length!
- Avoid special characters and spaces
  - Use dashes, “camel case”
  - CapitalizingFirstLetterOfEachWord
- If numbering for version control – use leading 0’s for scalability & ordering
- Consider semantic versioning: *major.minor.patch* version numbers (https://semver.org)
- Dates are good (yyyymmdd, yyyy-mm-dd is best)
- For some CHPC-specific tips, we’ll switch to Mike…

Version Control

- Several software options, *git* is most common
- Make use of git repositories:
  - Gitlab at CHPC: https://gitlab.chpc.utah.edu
  - Github: https://www.github.com
- CHPC course: Using Git for Version Control - March 31
  - https://www.chpc.utah.edu/presentations/IntroGit.php
- Other CHPC documentation
  - https://www.chpc.utah.edu/presentations/GitCheatsheet.pdf
  - https://www.chpc.utah.edu/documentation/software/git-scm.php
  - https://youtu.be/nvC6QkWTjr8 Git presentation link
- Not for datasets!

Integrity and Security

- Safeguard data integrity e.g. https://stacks.cdc.gov/view/cdc/153156
  - – Use SHA256 checksums to detect data corruption during transfer

```
$ sha256sum filename
Cb2a149d76a082ea66b62e8e17949d11...  filename
```

- Safeguard data
  - – Multiple copies on separate storage devices

Restricted vs Sensitive Data

Restricted Data	Sensitive Data
• Personally Identifiable Information (PII) • Protected Health Information (PHI) • Payment Card Industry (PCI) • Financial information • Donor information	• Intellectual Property • Employee information • Student information • Current litigation materials • Contracts • Physical building and utilities detail documentation

- ***https://regulations.utah.edu/it/4-004.php*** covers data classifications and encryption requirements, recommendations.

- CHPC Protected Environment, Box, and Office 365 Cloud all satisfy this storage requirement for sensitive and restricted data.

What does that mean in practice?

- CHPC has 3 different environments (GE, PE, Citadel)
- GE (General Environment) – weather data, open source data
- PE (Protected Environment) – PHI/PII, health data
- Citadel (brand new, case by case) – New NIST standards
- If you’re not sure which is right for you, please contact us! helpdesk@chpc.utah.edu

Reproducible Research

- The practice of distributing all data, software source code and tools required to reproduce results
- Key Components – Automation, version control, keep track of software used (including version) & architecture of system used, saving the right content (raw data, input files)

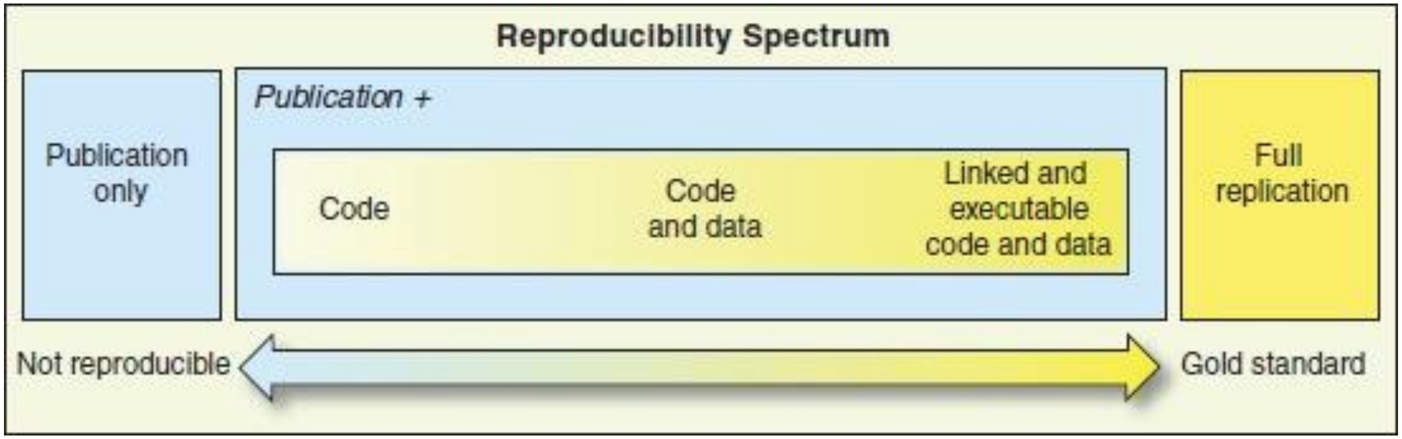


Image from https://www.nap.edu/catalog/21915/statistical-challenges-in-assessing-and-fostering-the-reproducibility-of-scientific-results

Preserving Software Environments: Containers

Ways of communicating your analysis software setup:

- Good: document all software versions and options
- Better: put a script in your git repository that performs the analysis
- Best: create a container with all the software and the environment in which it runs

Containers:

- Hold software files, configuration files, scripts, even data files
- Provide complete environment in which software can run
- Can be run interactively, to apply your analysis to a different data set
- NOTE: you can not run Windows container in Linux

Building Your Own Containers

- Build a container at CHPC with Apptainer
  - https://www.chpc.utah.edu/documentation/software/singularity.php
- Build a container from your github repository:
  - Create repository on https://hub.docker.com and link to your github repository
  - Add a Dockerfile to your github repo – Docker hub will build the container
- Reproducibility – use someone else's

  e.g. **apptainer pull redis.sif docker://redis:8.2.1**

- CHPC course: Introduction to Containers – March 19
  - https://www.chpc.utah.edu/presentations/Containers.php

Reproducible Research: CloudLab

- www.cloudlab.us
- profiles can also be published, giving other researchers the exact same environment—hardware and software—on which to repeat experiments and compare results.
- Enables researchers to repeat or build upon each others' work

- University of Utah is 1 of the 3 sites!

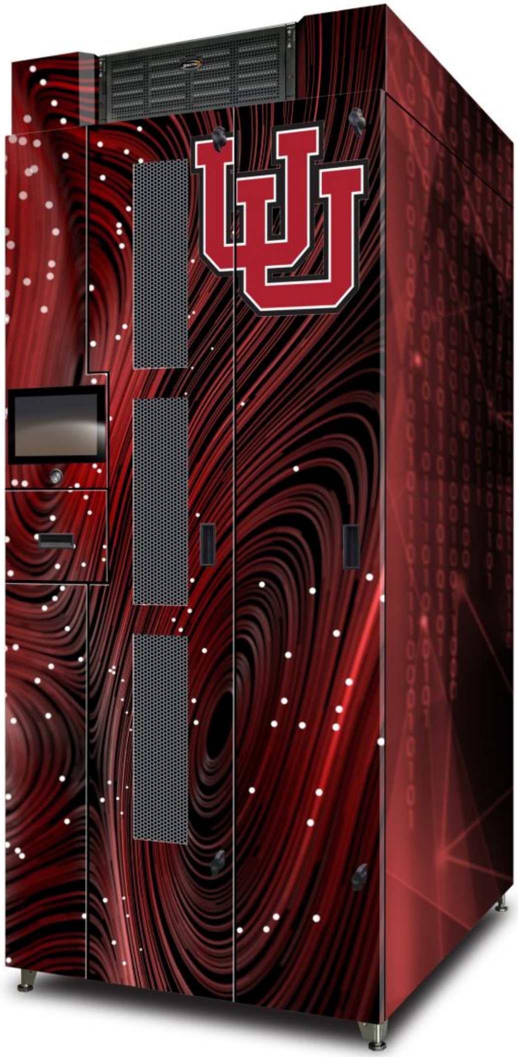
Storage Options at CHPC

- Home directory – 50GB default but can buy more
- Group space – Linux file system on redundant disk array (RAID)
  - Cost: $150/TB for ~7 years (no backup) or $450/TB for ~7 years (backed up)
  - Retrieval: free
  - This is a single copy, not backed up
- Archive storage –object storage similar to Amazon S3
  - Cost: $150/TB for 7 years
  - Retrieval: free
- Group space and archive storage options in both regular environment and protected environment (for restricted data, PHI)
- Scratch space – TEMPORARY 60 day limit
- https://www.chpc.utah.edu/resources/storage_services.php

Backup Strategies at CHPC

- CHPC will continue to provide backup of purchased home directory spaces in general environment and all PE home directories
- New general environment group spaces and PE project spaces backup options – specify at purchase time
  - CHPC backup to in-house object storage
    - Requires purchase of sufficient amount of object storage space (2x if all needs to be backed up)
  - Owner driven backup to
    - in-house object storage
    - Box
    - Other storage external to CHPC
- CHPC provides tools for Owner drive backup: globus, rclone, fpsync
- All of these at DDC but Disaster Recovery options now available

State Archive Prototype



- Funded by NSF CC* Award (free of charge for lifetime of grant)
- Infrastructure
  - 2.8PiB Ceph-based object storage (DDC - SLC)
  - 7.2PiB Spectralogic BlackPearl Tape Capacity (TDC – St. George)
- Workflow
  - Users copy important datasets to primary disk-based archive
  - Datasets automatically synced to TDC
    - Two copies to tape
    - Globus Timers
- Connect and share with HIVE, CHPC S3, National CI (ACCESS)
- PID/DOI tagging
- Initial Institutions – University of Utah and Utah State
- Open data only
- 50 TB limit per group

Other Storage Options Available (1)

- The Hive: https://hive.utah.edu/
  - Public access to data created by University faculty, students, staff
  - Limited to 200 GB per project, with individual files limited to 50 GB each
  - Automatically assigned a DOI
- Box: https://box.utah.edu/
  - 1 TB limit total, 50 GB file size limit
  - OK for sensitive, restricted data
- Office 365 Cloud: https://o365cloud.utah.edu
  - 1 TB limit total, 2 GB file size limit
  - OK for sensitive, restricted data

See: https://campusguides.lib.utah.edu/data_storage

Other Storage Options Available (2)

- Google Drive: https://gcloud.utah.edu/
  - Storage: 25 GB for students, 150 GB faculty/staff (as of January 2026)
  - Public data only! Nothing sensitive, restricted, no IP, PII, PHI, etc
  - Retrieval: free, but…
    - Upload limited to 750 GB/day, and no more than 2 files/minute
    - Download limited to 10 TB/day
  - Backup to Google Drive using rclone: https://www.chpc.utah.edu/documentation/software/rclone.php
  - For PE explore google cloud government
    - https://cloud.google.com/solutions/government/
    - Not part of the free storage via the University agreement
- Amazon S3 Glacier https://aws.amazon.com/glacier/
  - Storage: $0.0036/GB/mo ($310/TB/7 years)
  - Retrieval: $0.01/GB

Data Repositories

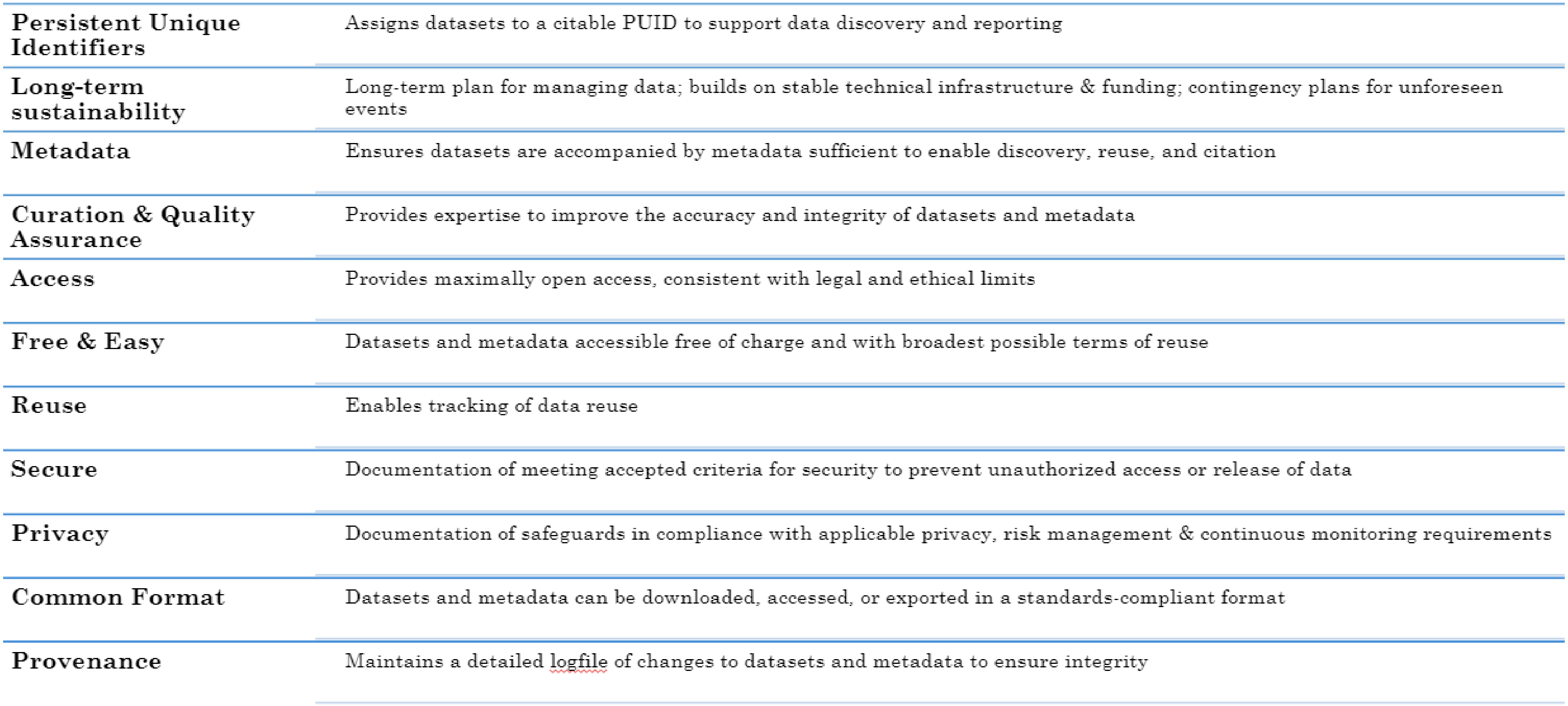
- Subject based repositories index
  - https://www.re3data.org/
- General purpose repositories, including:
  - https://figshare.com/
  - https://datadryad.org/
  - https://dataverse.org/
  - https://data.mendeley.com
  - https://www.cos.io/products/osf
  - https://vivli.org
  - https://zenodo.org (EU / CERN)

Data Repositories (2)

- Institutional repositories
  - https://hive.utah.edu/ General environment only!
- Create your own – can use CHPC VM Farm for hosting
  - Web pages
  - Databases
  - If small, can be free, on community servers
    - https://www.chpc.utah.edu/resources/hosting.php
  - If larger, may need own VM
    - https://www.chpc.utah.edu/resources/virtualmachines.php

What makes a good repository? Librarians will know!

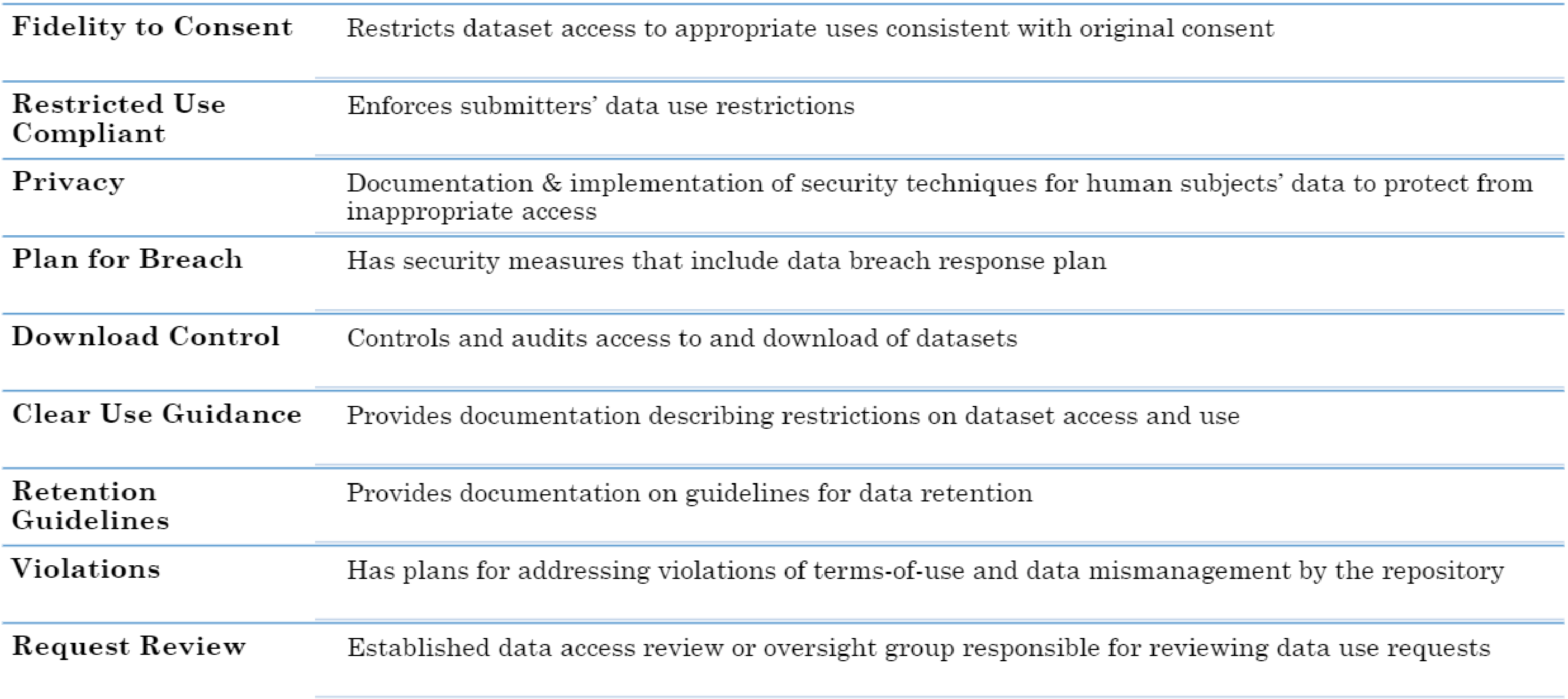
Desirable Characteristics for Data Repositories



Persistent Unique Identifiers	Assigns datasets to a citable PUID to support data discovery and reporting
Long-term sustainability	Long-term plan for managing data; builds on stable technical infrastructure & funding; contingency plans for unforeseen events
Metadata	Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation
Curation & Quality Assurance	Provides expertise to improve the accuracy and integrity of datasets and metadata
Access	Provides maximally open access, consistent with legal and ethical limits
Free & Easy	Datasets and metadata accessible free of charge and with broadest possible terms of reuse
Reuse	Enables tracking of data reuse
Secure	Documentation of meeting accepted criteria for security to prevent unauthorized access or release of data
Privacy	Documentation of safeguards in compliance with applicable privacy, risk management & continuous monitoring requirements
Common Format	Datasets and metadata can be downloaded, accessed, or exported in a standards-compliant format
Provenance	Maintains a detailed logfile of changes to datasets and metadata to ensure integrity

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

Additional Considerations for Human Data



Fidelity to Consent	Restricts dataset access to appropriate uses consistent with original consent
Restricted Use Compliant	Enforces submitters' data use restrictions
Privacy	Documentation & implementation of security techniques for human subjects' data to protect from inappropriate access
Plan for Breach	Has security measures that include data breach response plan
Download Control	Controls and audits access to and download of datasets
Clear Use Guidance	Provides documentation describing restrictions on dataset access and use
Retention Guidelines	Provides documentation on guidelines for data retention
Violations	Has plans for addressing violations of terms-of-use and data mismanagement by the repository
Request Review	Established data access review or oversight group responsible for reviewing data use requests

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

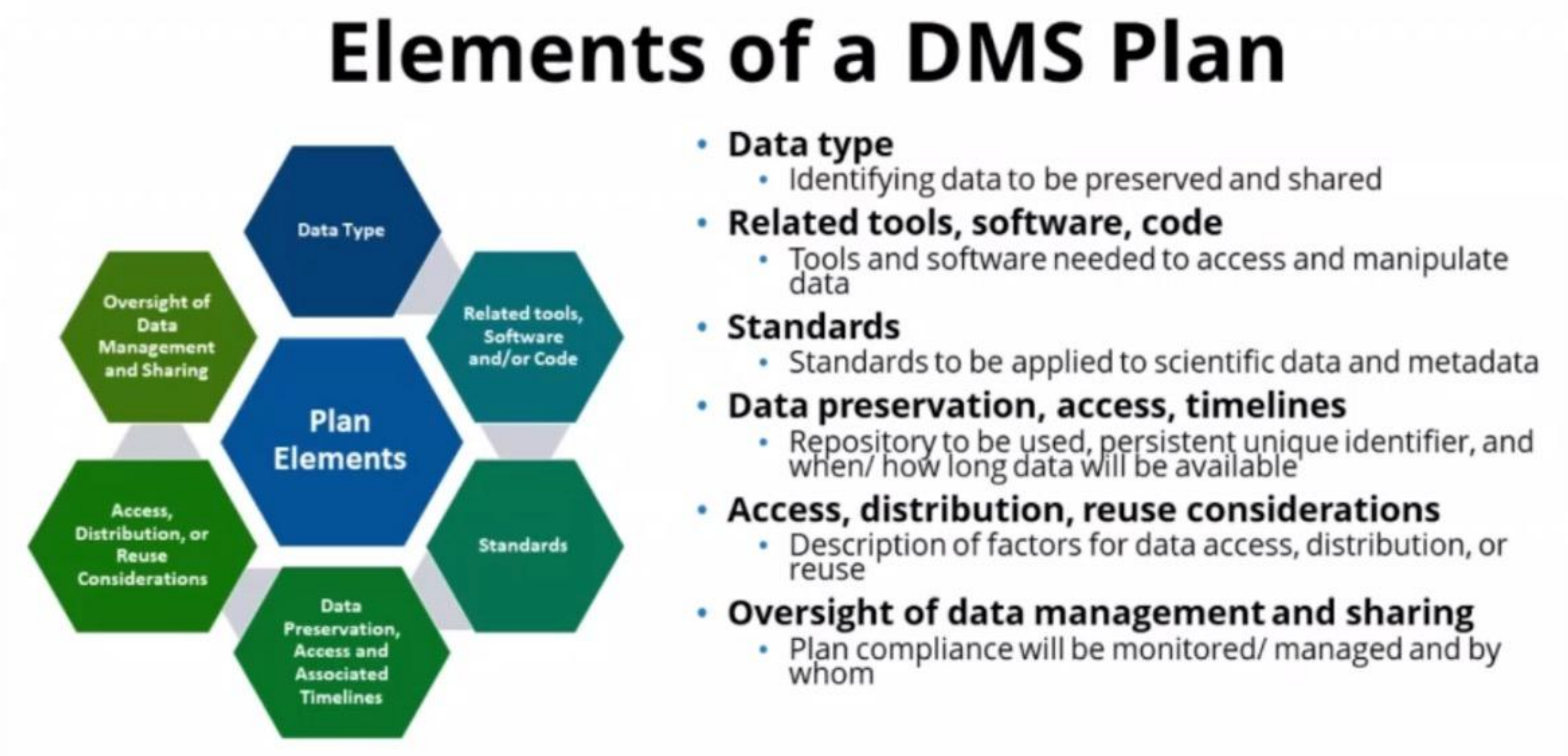
Data Management Plans

- https://campusguides.lib.utah.edu/c.php?g=160412&p=1051780
- DMPTool – https://dmptool.org – sign in with institutional credentials
  - Have templates for different funding agencies
- Other Help
  - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525
- Plan includes (varies by funding agency):
  - Types of data including file formats
  - Data description, including metadata schemas
  - Data storage
  - Data sharing, including confidentiality and privacy restrictions
  - Data archiving and responsibility
  - Data management costs

New NIH DMS Plans Requirements

- Required in Proposals as January 25, 2023
- 2 pages or less – draft format found at https://grants.nih.gov/grants/forms/data management-and-sharing-plan-format-page
- More information – see https://sharing.nih.gov/data-management-and-sharing-policy
- Examples
  - https://www.nimh.nih.gov/funding/managing-your-grant/nimh-data-management-and-sharing-for-applicants-and-awardees
  - https://guides.lib.umich.edu/datamanagement/planning
  - https://data.library.arizona.edu/data-management/nih-data-management-sharing-policy-2023

New NIH DMS Plans


Data Type
Related tools, Software and/or Code
Oversight of Data Management and Sharing
Plan Elements
Access, Distribution, or Reuse Considerations
Standards
Data Preservation, Access and Associated Timelines

Elements of a DMS Plan

- **Data type**
  - Identifying data to be preserved and shared
- **Related tools, software, code**
  - Tools and software needed to access and manipulate data
- **Standards**
  - Standards to be applied to scientific data and metadata
- **Data preservation, access, timelines**
  - Repository to be used, persistent unique identifier, and when/ how long data will be available
- **Access, distribution, reuse considerations**
  - Description of factors for data access, distribution, or reuse
- **Oversight of data management and sharing**
  - Plan compliance will be monitored/ managed and by whom

Research Data Management and Data Science Resources at the U

- **Data Exploration and Learning for Precision Health Intelligence (DELPHI) data science initiative** - https://uofuhealth.utah.edu/delphi-data-science-initiative
- **One Utah Data Science Hub** - https://www.research.utah.edu/one-utah-data-science-hub/
- **Utah Center for Data Science -** https://datascience.utah.edu
- **U Libraries Research Guides on Data Management** https://campusguides.lib.utah.edu/researchdata
- **REd (Research Education Classes)** – https://education.research.utah.edu

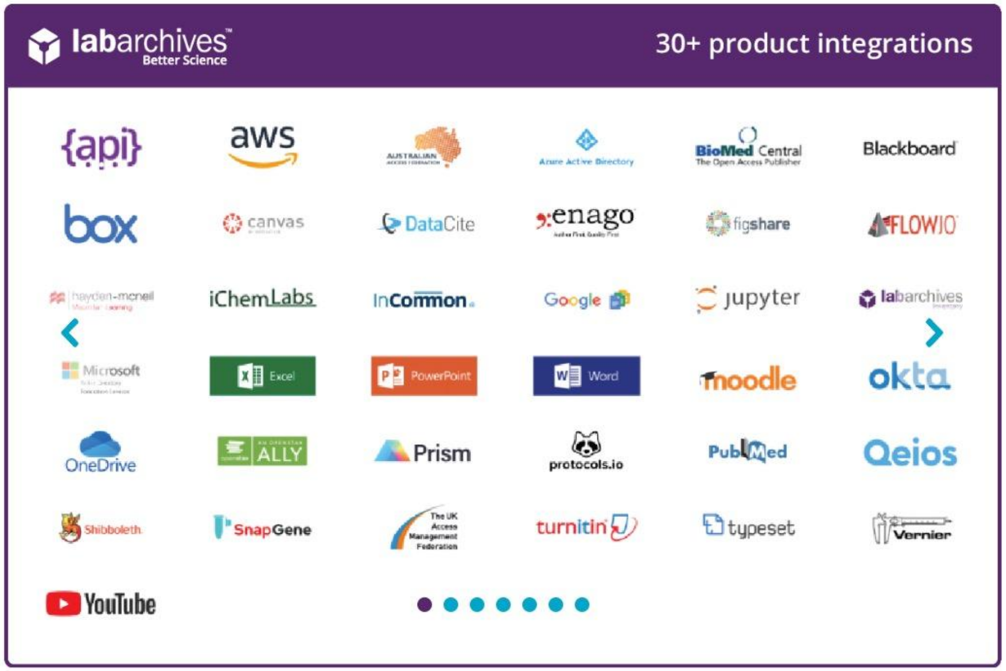
Referencing Data: DOI’s

- What’s a DOI: Digital Object Identifier
  - – Persistent identifier, forwards request to current location
  - – Useful for citation purposes, when dataset location could move
  - – For example: https://doi.org/10.1109/5.771073
- How do I get one: https://campusguides.lib.utah.edu/identifiers
  - – For faculty, graduate students, postdocs, and research associates
- Many publications given DOIs, as are data sets in The Hive

LabArchives

https://campusguides.lib.utah.edu/labarchives

- General purpose electronic notebook, licensed by the University
- Cloud based solution - https://mynotebook.labarchives.com
- Ties into the University of Utah Box
- Integration with many other tools, including REDCap
- Allows for shared notebooks – great for collaborations
- Working towards compliance with security requirements



Getting Help

- Online documentation: www.chpc.utah.edu
- Onboarding, Policies, and more
- CHPC Email: helpdesk@chpc.utah.edu
- Research Data Drop in Session Thursdays Noon – 1
- Research Librarians: mlib-data@lists.utah.edu