Data Management

Brett Milash brett.milash@utah.edu
Anita Orendt anita.orendt@utah.edu
Center for High Performance Computing

9 November 2022

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*..."

- Data Management

activities and practices that support long term preservation, access and use of data

Why Manage Data?

- Prevent data loss
- Efficiency -- better organization saves time
- Standardize practices
- Promotes reproducible research
- Ease of data sharing – increased visibility of your work
- Required to meet institutional requirements
- Documentation for Intellectual Property (IP) 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

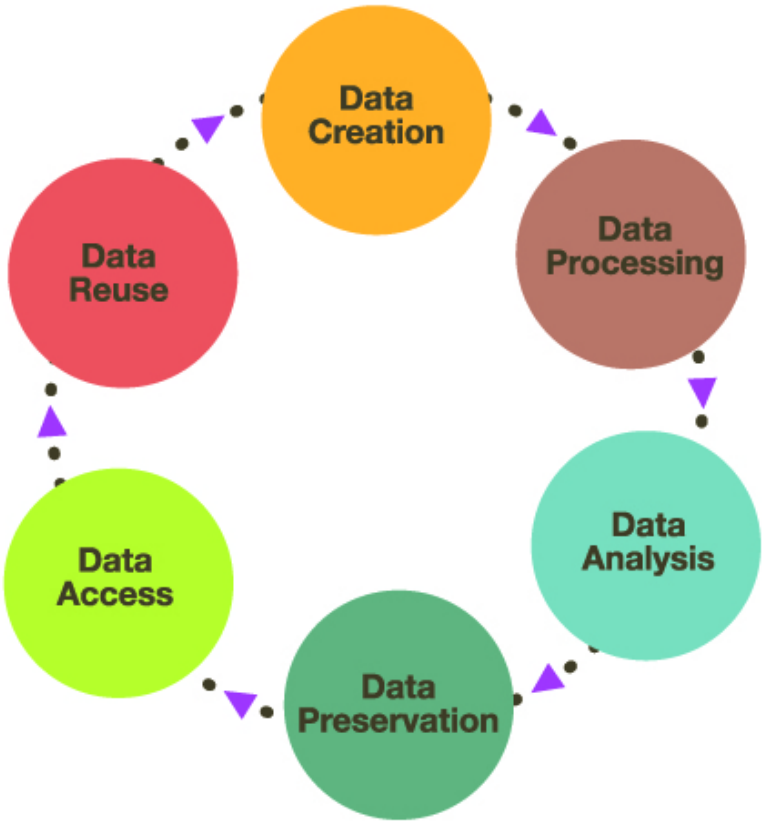

Figure from https://blogs.ntu.edu.sg/lib-datamanagement/data-lifecycle/

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

- https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml
- Think long term; public formats preferred over proprietary

Type of Data	Preferred Format
Tables w/ min metadata	comma separated values file (.csv), tab-delimited file (.tab)
Tables w/ ext metadata	SPSS portable format (.por), eXtensible Mark-up Language (.xml)
Text based data	Rich Text Format (.rtf), Plain Text, ASCII (.txt), eXtensible Mark-up Language (.xml), PDF
Images	TIFF (.tif); also acceptable are JPEG (.jpg), PNG (.png), Adobe Portable Document Format (PDF/A, PDF), (.pdf)
Video	MPEG4 (.mp4); also acceptable motion JPEG 2000 (.jp2)
Audio	Free Lossless Audio Codec (.flac), MPEG audio layer III (.mp3)

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
  - http://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 name allows for identification
- Consider length!
- No special characters, no 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 (http://semver.org)
- Dates are good (yyyymmdd, yyyy-mm-dd best)

Security Concerns

- Safeguard data
  - Multiple copies on separate storage devices
- Safeguard data integrity
  - Use MD5 checksums to detect data corruption during transfer

```
$ md5sum filename > filename.md5
$ cat filename.md5
cb2a149d76a082ea66b62e8e17949d11  filename
```

- Restrict access as appropriate
- Consider the security of system used to store data

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* and *https://regulations.utah.edu/it/rules/Rule4-004C.php*** – deals with the data classifications and encryption requirements/recommendations.

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

Version Control

- A number of options, git is most common
- Make use of git repositories:
  - Gitlab at CHPC: https://gitlab.chpc.utah.edu
  - Github: https://www.github.com
- CHPC Presentation – Friday, Nov 11, 2022
  - 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

Storage Options at CHPC

- Group space – Linux file system on redundant disk array (RAID)
  - Storage: $150/TB for ~7 years
  - Retrieval: free
- Archive storage –object storage similar to Amazon S3
  - Storage: $150/TB for 7 years
  - Retrieval: free
- Group space and archive storage options in both regular environment and protected environment (for restrictive data, PHI)

Backup Strategies at CHPC

- CHPC moved from tape to disk based backup (to CHPC object storage) a few years ago
- 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
  - 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

Other Storage Options Available (1)

- The Hive: https://hive.utah.edu/
  - Public access to data created by University faculty, students, staff
  - Limited to 500 Gb per project
  - 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: http://campusguides.lib.utah.edu/data_storage

Other Storage Options Available (2)

- Google Drive: https://gcloud.utah.edu/
  - Storage: was free, unlimited until July 2022 – now 25 GB for students, 150 GB faculty/staff
  - 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 restricted 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

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 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

Data Repositories (2)

- Institutional repositories
  - https://hive.utah.edu/
- Create your own – can use CHPC VM Farm for hosting
  - Web pages
  - Databases

Data Management Plans

- https://campusguides.lib.utah.edu/c.php?g=160412&p=1051780
- DMPTool – http://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 January 25, 2023
- 2 pages or less – draft format found at https://grants.nih.gov/sites/default/files/DMS-Plan-blank-format-page.pdf
- More information – see https://sharing.nih.gov/data-management-and-sharing-policy
- Examples
  - https://www.nimh.nih.gov/funding/managing-your-grant/nimh-data-sharing-for-applicants-and-awardees#4
  - https://guides.lib.umich.edu/datamanagement/planning
  - https://data.library.arizona.edu/data-management/nih-data-management-sharing-policy-2023

New NIH DMS Plans

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

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)

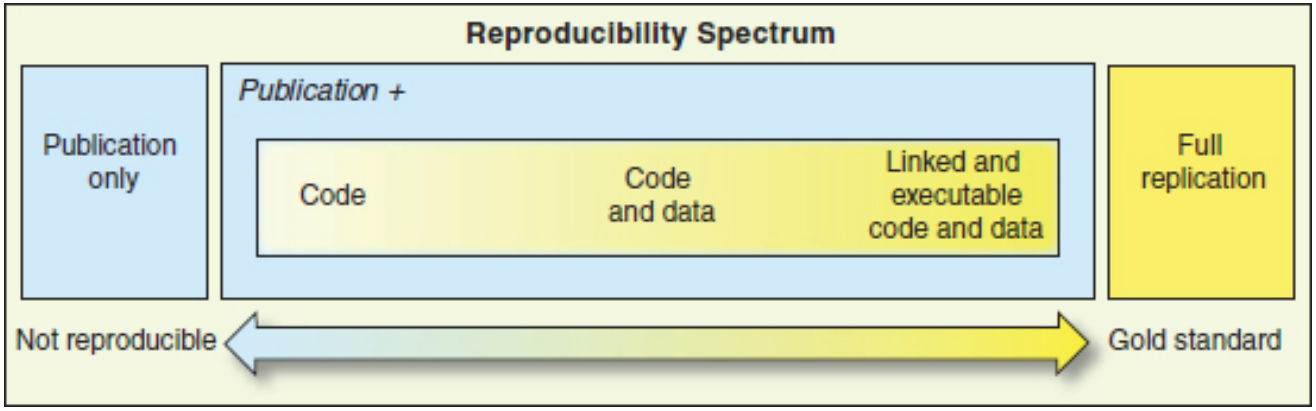

Image from https://www.nap.edu/catalog/21915/statistical-challenges-in-assessing-and-fostering-the-reproducibility-of-scientific-results

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

Building Your Own Containers

- Build a Singularity container at CHPC in Singularity
  - 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
  - Example:
    - Github repo: https://github.com/bmilash/containers/tree/master/scipy-notebook
    - Docker hub: https://hub.docker.com/repository/docker/bmilash/scipy-notebook
    - Retrieve the container with "singularity pull docker://bmilash/scipy-notebook"
- CHPC course: Introduction to Containers
  - 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

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: http://campusguides.lib.utah.edu/identifiers
  - For faculty, graduate students, postdocs, and research associates
- Many publications given DOIs, as are data sets in The Hive

Research Data Mgt and Data Science Resources at the U

- **Data Exploration and Learning for Precision Health Intelligence: The DELPHI data science initiative** - https://uofuhealth.utah.edu/delphi-data-science-initiative
- **One Utah Data Science Hub** - https://research.utah.edu/utah-data-science.php
- **Data Science & Ethics of Technology Initiative (DATASET)** - https://research.utah.edu/datascience.php
- **Utah Center for Data Science** - http://datascience.utah.edu/index.html
- **U Libraries Research Guides on Data Management** - https://campusguides.lib.utah.edu/researchdata
- **REd (Research Education Classes)** – https://education.research.utah.edu

LabArchives

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

- General purpose electronic notebook, licensed by the University
- Cloud based solution -- https://mynotebook.labarchives.com
- Ties in to 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

- CHPC website
  - www.chpc.utah.edu
    - Getting started guide, cluster usage guides, software manual pages, CHPC policies
- Service Now Issue/Incident Tracking System
  - Email: helpdesk@chpc.utah.edu
- Help Desk: 405 INSCC, 581-6440 (9-5 M-F)
- We use chpc-hpc-users@lists.utah.edu for sending messages to users