Data Mangement 101

Best Practices

Course Site:
http://lib.ucsd.edu/dm101

Reid Otsuji
Data Curation Specialist Librarian
rotsuji@ucsd.edu

Course Overview

Part 1:

A. Why is data management important?

B. Best practices to Consider

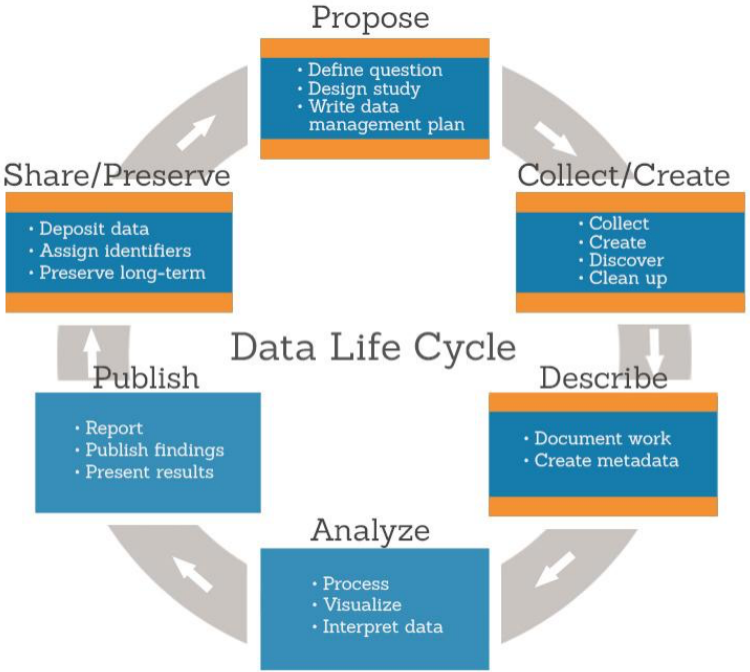
Part 2:

Hands-on with the Unix Shell

Why is Data Management Important?

Part 1a:
Why Manage Data?

Why is Data Management Important?



Managing data in the Data Life Cycle:

- Data management planning
- Documenting project/file details
- Choosing file formats
- File organization & naming conventions
- Access control & security
- Backup & Storage
- Sharing and Preservation

Why is Data Management Important?

A good data management strategy:

- Establish best practices for your data management
- Plan to share well-documented data
- Well prepared data saves time
- Create a concise data management plan for your grant proposal
- Reduces cost of creating, protecting and storing data

Ensures your data will be available to future generations to enable reproducible research

Why is Data Management Important?

Benefits of good data management:

- Promotes successful data collection.
- Ease of using and sharing data.
- Helps to increase research impact and visibility.
- Standardize data management practices and policies in your research lab.

Saves you time, effort and resources during the research project.

Best Practices for Data Management

- Organization
- TEIR Protocol v2
- Documentation and Description
- Metadata
- Data Clean-up
- Basic Storage
- Backup
- Preservation

Best Practices for Organization

File and Folder organization
Choose a consistent filing system that will make sense to you or someone else five years from now.

Choose a logical directory hierarchy. For example: TIER Documentation Protocol.

Assign descriptive file names. E.g. DOLInterview_DoeJane_20061207

//Project001/SiteB/SiteB_2010_rawdata.txt

Is better than . . .

//Project001/SiteB/2010/rawdata.txt

Organization: TIER Documentation Protocol

Developed by Haverford College –

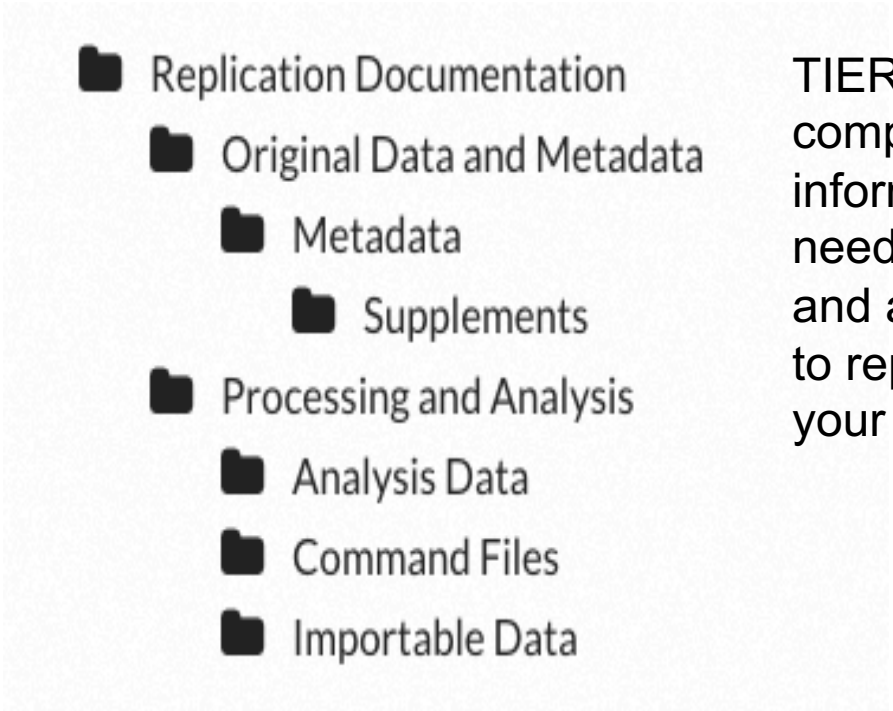
The **Teaching Integrity in Empirical Research** or TIER protocol, is a protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper.

All documentation, do-files, scripts, raw data, metadata, that are presented in a paper are organized in a specific file structure.

This file structure keeps your data organized and offers easy replication of results reported in a paper.

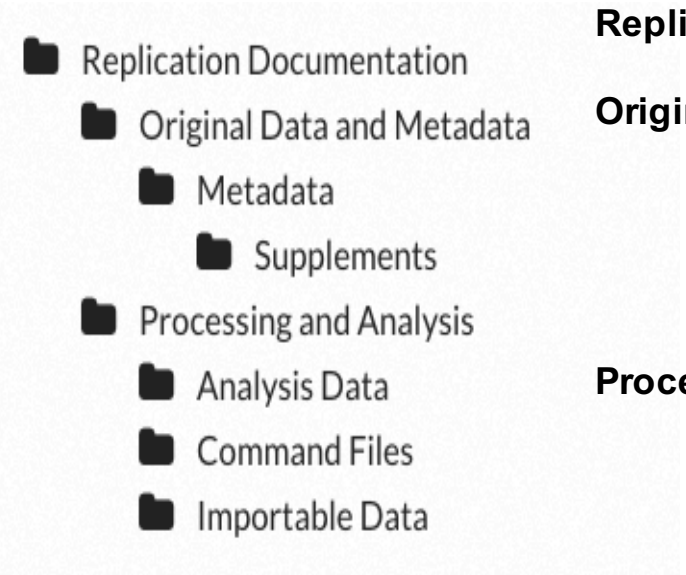
https://www.haverford.edu/project-tier/protocol-v2

Create and use the folder hierarchy



TIER file structure contains all the data, computer programs, and explanatory information an independent researcher would need to be able to replicate the data processing and analysis you conducted for the project and to reproduce exactly all the results reported in your paper.

What to put in the folders



Replication documentation folder -readme file, copy of final paper

Original Data and Metadata folder:

original data folder – all original data

metadata folder – metadata guide – info about your original data

supplements folder – user guides or codebooks

Processing and Analysis Folder:

Importable data folder – file version of import data: .csv, .dta

Command files folder – do-files or scripts used for data processing and analysis to reproduce results

Analysis data folder – analysis data files, data appendix

Documentation & Description

- Describe the method used to create derived data products.
- Consider creating templates for data collection.
- At the file level: Take consistent notes on file changes, name changes, dates of changes, etc.
- Include critical information, such as date or location, in the data table, not just as metadata embedded in the file name.

Metadata

Metadata is data about your data.

Creating metadata, i.e., information about your data's contents, structure, and permissions, makes it possible for others to find and use your data properly.

Without good metadata, you might not be able to reuse your own data five years from now!

Data Clean-up

OpenRefine (http://openrefine.org/), for making sure records and variables are consistently coded, filling in known blanks, replacing text selectively, transforming data, and more.

Basic Storage

- Computers and shared servers can be good places for **temporary** storage of your working files.
- Store copies of data in open, **stable formats** (e.g., ascii, .txt, .csv, .pdf) for long term accessibility. . .
- Use flash drives **only for file transfer.**
- Cloud storage can be a convenient way to store and share temporary working files.
- For long-term storage, data should be put into well-managed **preservation system.**

Backup

- Rule of 3: Keep 2 copies onsite, 1 offsite.

  LOCKSS concept – Lots Of Copies Keeps Stuff Safe

- Backup regularly and frequently - automate the process if possible.

Preservation

- Preservation is the act of making sure your data are secure and accessible for future generations.
- Long-term preservation is not merely storage or backing up of your data.
- Identify data with long-term value. Preserve the raw data and any intermediate/derived/time consuming products that are expensive to reproduce or can be directly used for analysis.
- Preserve any scripted code and data that was used to clean and transform the raw data.
- Example:

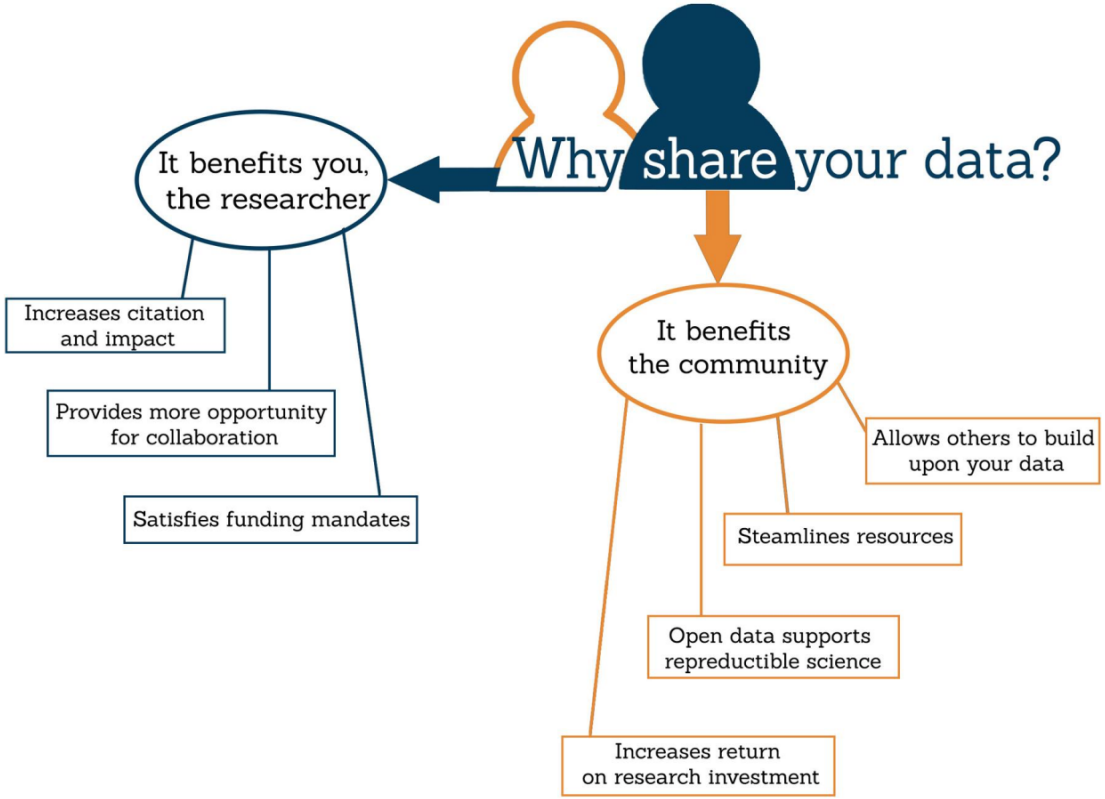
Save tabular data in a delimited text format.

Save data in uncompressed and unencrypted formats, where possible.

Benefits of Sharing Your Data

Data sharing allows for reproducibility, transparency, and data re-use in research.

Sharing is easier if data are managed well from the start of a project.



Hands-On with the Unix Shell

Course Site:
http://lib.ucsd.edu/dm101

Etherpad:
https://public.etherpad-mozilla.org/p/gps-dm101

Socrative Room Name: UCSDGPSDM101