2020

Introduction to Research Data Management

Ian Harmon
West Virginia University, ian.harmon@mail.wvu.edu

Introduction to Research Data Management

Ian Harmon, Scholarly Communications Librarian

What counts as “data”?

- Not limited to numbers on spreadsheet
- Any and all information collected and used in research and scholarship across all disciplines
- Videos, images, digitized texts, artifacts, geospatial coordinates, drafts of manuscripts, notes, code, algorithms, etc.
- All researchers have data to manage

What is Research Data Management?

Research Data Management refers to the processes involved in organizing, storing, and sharing the data that is generated through the course of a research project.

The goal of data management is to have datasets that can be easily understood, consulted, and used in the future.

Why manage data?

- Spend less time managing data and more time on research
- Data will be easier to find, use, and analyze
- Easier for collaborators to to understand and use your data
- Makes data reusable by other researchers
- Can help ensure you get credit for the datasets you produce
- Comply with funder requirements

Components of Research Data Management

- File naming and organization
- Data Documentation
- Storage and Backup
- Archiving and Sharing
- Data Management Plans (DMP)

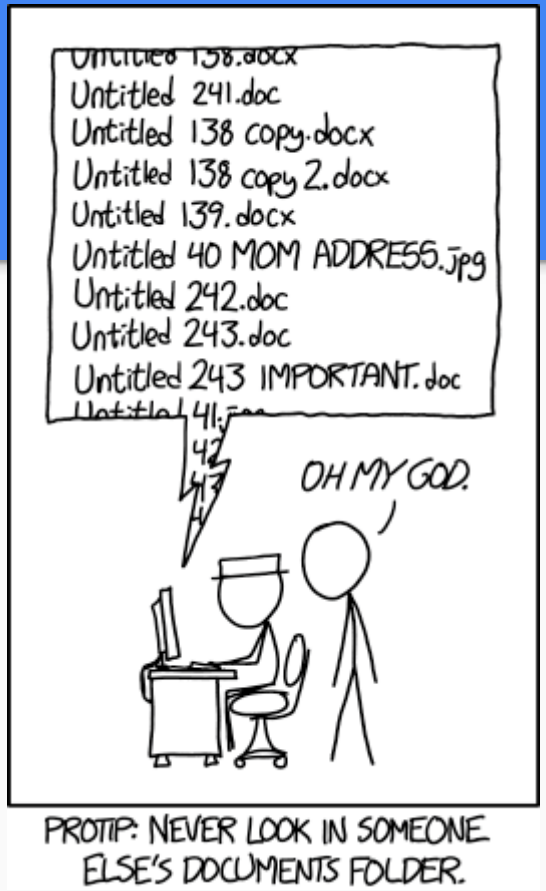
File names and organization

File naming conventions

Descriptive but short (around 25 characters or less)

Built up of elements associated with the project, e.g.:

- File creator's name/initials
- Project lead's name/initials
- Project name
- Date created
- Version number
- Location



Source: https://xkcd.com/1459/

File naming conventions

Guidelines for file names:

- Files should be named in a consistent manner
- File names should be short, but informative
- Avoid the use of special characters and spaces (use - or _ instead)
- Use consistent formatting for dates (YYYYMMDD is recommended)
- Include a version number with leading zeros (e.g. V02).
- Avoid generic file names (e.g. MyData.csv)

File naming conventions

Examples:

YYYYMMDD-ProjectName-Version

- 20191223-DataManagementGuide-V01.pdf

ProjectName-Creator-Version

- DissChapter2-Harmon-V03.txt

Organizing Your File Structure

- Use folders hierarchically to keep your materials organized
- Be consistent about where you store various types of files
- Identify important contextual information
  - How do you think when you look for a file?
  - Avoid overlapping categories
- Don’t let folders get too big
- Don’t led folder structure get too deep

Organizing Your File Structure

Thesis/Dissertation Chapter

- Drafts
  - Chapter 1
    - Version 1
    - Version 2
- Raw Data
- Processed Data
- Figures
  - Chapter 1

Project

- Experiment 1
  - Data
  - Observations
  - Analysis
- Experiment 2
  - Data
  - Observations
  - Analysis

File organization

- Make a system
- Document your system
- Follow your system

Data Documentation

Data Documentation

- Your organizational system
  - What are you file naming conventions?
  - What is your folder hierarchy?
- Your workflow
  - How did you move from raw data to finished product? What would someone else need to know to go through the same process?
- Your data
  - What do the fields in your spreadsheet mean? What units did you use? What do your abbreviations or acronyms mean?

Data Documentation

- Where to access (do you have a permanent url?)
- Known problems, limitations, etc.
- Ethical/privacy restrictions
- Licensing (how can others use your data, e.g. Creative Commons licenses)
- Recommended citation

Creating a README file

- Descriptions of each file (including its format if not obvious)
- Names, contact info for PIs, co-investigators, etc.
- Dates of data collection, dates files created
- Location information (if relevant)
- Key words
- Names and definitions
- Units of measurement
- Licensing and access information, recommended citation.

Creating a README file

Example of a dataset with README file: https://doi.org/10.18130/V3/NEPGOL

Backup and Storage

Backup and Storage: 3-2-1 Rule

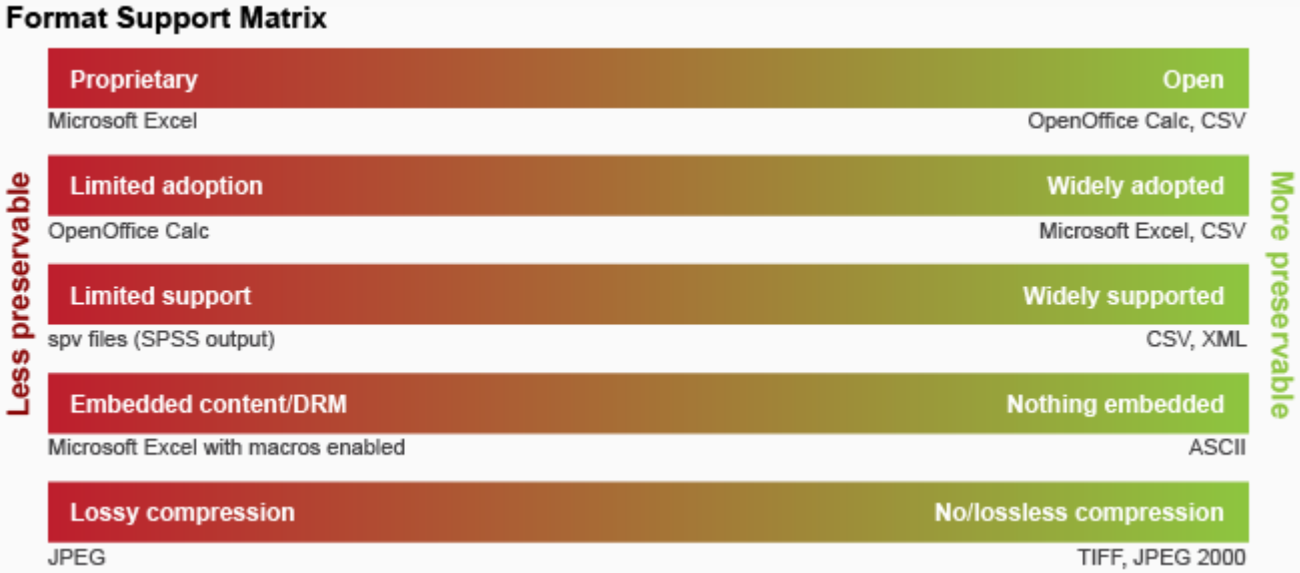
- Keep 3 copies of all important files
- Store your files on at least 2 different media types (e.g. laptop and cloud storage)
- Store at least 1 copy offsite - (e.g. cloud storage, an external drive kept at home)

*Establish a regular time to run backups.

Backup and Storage: File Formats

- Non-proprietary, open
- Commonly used
- Unencrypted
- Uncompressed

Backup and Storage: File Formats



Source: https://guides.library.illinois.edu/introdata/preservation

Backup and Storage: File Formats

Proprietary Format	**Recommended Alternative**
MS Word (.doc, .docx)	Plain text (.txt), PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Excel (.xls, .xlsx)	Comma Separated Values (.csv)
Photoshop (.psd)	TIFF (.tif, .tiff)

Library of Congress Recommended Formats:
https://www.loc.gov/preservation/resources/rfs/

Archiving and Sharing

Why share data?

- Data is scholarship
- Provides a more complete picture of your work
- Promotes new research in your field
- Encourages collaboration
- Supports transparency and reproducibility
- Allows others to cite, increasing your impact

How to share data?

- Deposit in a disciplinary repository (e.g. Humanities Commons, ICPSR)
  - Registry of Research Data Repositories
- Deposit in the WVU Research Repository
- Submit your data with a journal article as a supplementary file

Data Management Plans

Funding Agency Requirements

Many funding agencies have begun requiring the inclusion of data management plans in all funding proposals.

- SPARC Data Sharing Requirements by Federal Agency

Creating a Data Management Plan

- DMPTool.org
- WVU Office of Sponsored Programs
- Eberly College Research Office

Additional Resources

- Introduction to Research Data Management
- Digital Curation Centre
- Research Data Management Toolkit

My contact information

ian.harmon@mail.wvu.edu

304.293.0329