Fixing Your Files:
A Research Data Management Primer

Aditya Ranganath
Data Librarian
aditya.ranganath@colorado.edu

What we’ll talk about

- What is research data?
- Data Management vs. Data Management Plans
- Data Storage & Access
- Backups & Versioning
- Documentation
- File Names & Structures
- File Formats & Units of Measurement
- Data Management Plans
- Tools

You can’t do everything

- I’m going to talk about a lot of different things
- Doing *anything* is good
- Pick one or two things and concentrate on doing them well

What is Research Data?

- “Recorded factual material commonly accepted in the scientific community as necessary to validate research findings” (US Office of Management and Budget)
    - Primary / Secondary
    - Qualitative / Quantitative
    - Experimental / Observational

Data Management vs. Data Management Plans

(and Data Management and Sharing Plans)

Data Management vs DMPs

- Data management refers to the things researchers do to stay organized as they create, collect, describe, store, and work with research data
- Data Management Plans (DMPs) are a written description of the data management strategy for a particular research project, and how the data will be utilized and stored during and after a project.

Data Management vs DMPs

- Data management is what you ***do.***

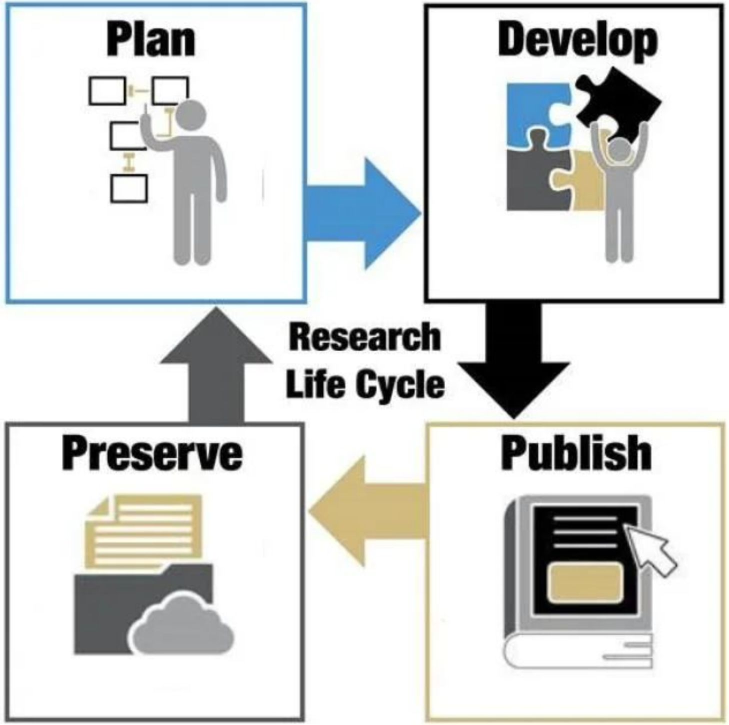
- Data Management Plans are where you write about what you’re going to do. (Usually because a grant requires one.)

Why is data management important?

- Protects data from loss
- Saves you time
- You can find your data when you need it
- Helps new members of research teams understand processes faster
- Facilitates reproducibility
- Improves the quality of published data
- Keep sensitive data secure

xtranorm

Research Lifecycle



What Data Management isn’t

- Data management is ***not*** data sharing
- You can manage your data without sharing it
- You can share your data without managing it
  - (please don’t do this)

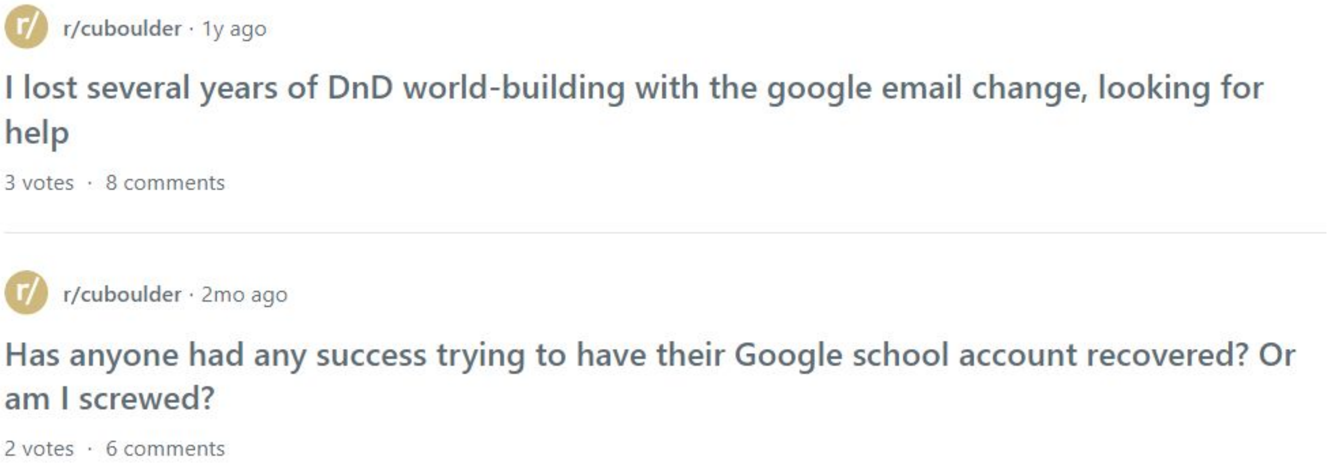
Data Storage & Access

Data Storage & Access

- Where will the data be stored?
  - When the data is being collected
  - When the data is being analyzed
  - After the project is over
- Who can access the data?
  - How can they access it?
  - Is there sensitive or confidential data?
  - What security measures are in place to protect the data?

Things to think about

- Who’s paying for data storage?
- How long will you have access to this storage?
  - What happens when you graduate?



r/cuboulder · 1y ago

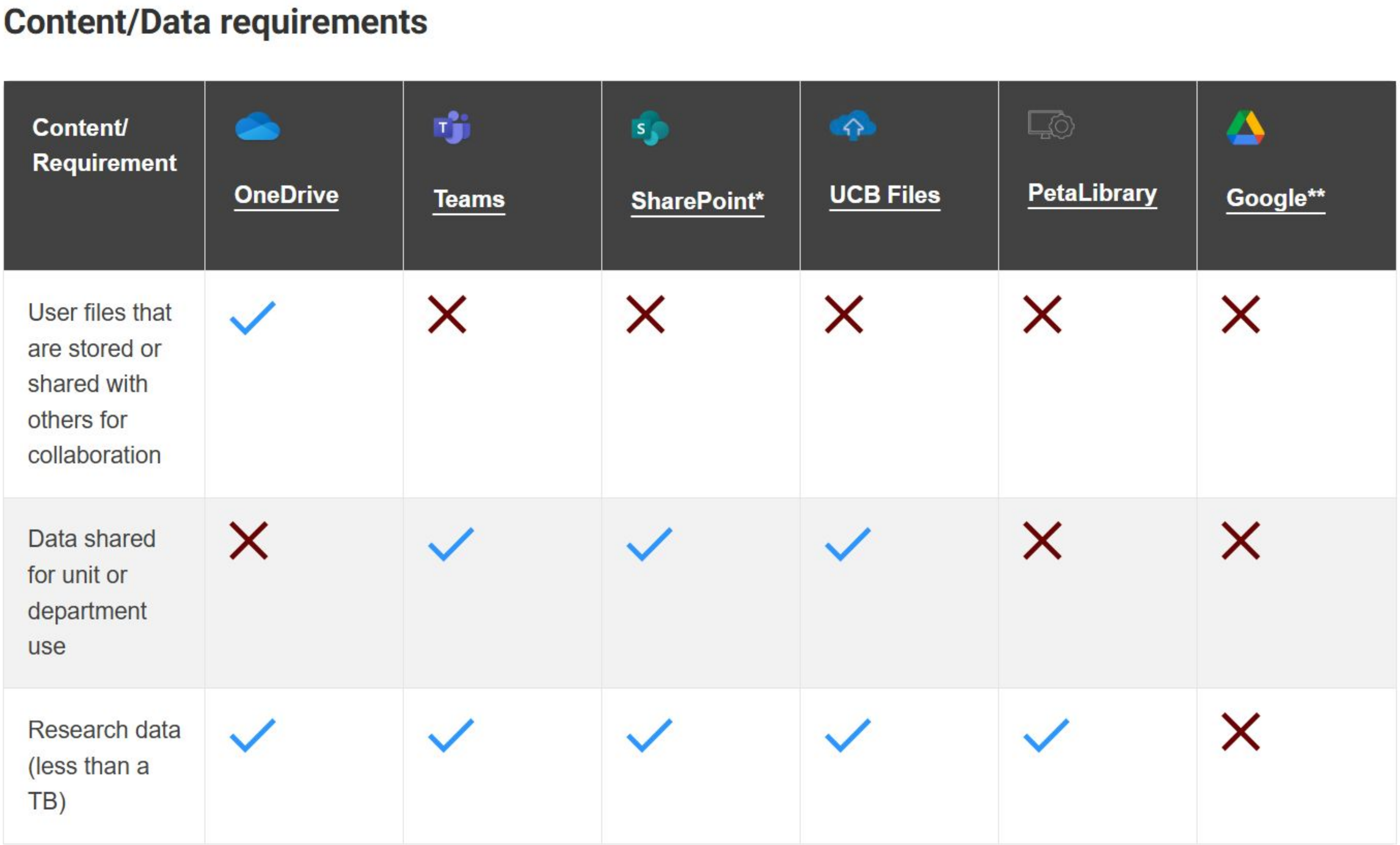
I lost several years of DnD world-building with the google email change, looking for help

3 votes · 8 comments

r/cuboulder · 2mo ago

Has anyone had any success trying to have their Google school account recovered? Or am I screwed?

2 votes · 6 comments



Content/Data requirements

Content/ Requirement	OneDrive	Teams	SharePoint*	UCB Files	PetaLibrary	Google**
User files that are stored or shared with others for collaboration	✓	✗	✗	✗	✗	✗
Data shared for unit or department use	✗	✓	✓	✓	✗	✗
Research data (less than a TB)	✓	✓	✓	✓	✓	✗

https://oit.colorado.edu/services/file-transfer-storage-infrastructure

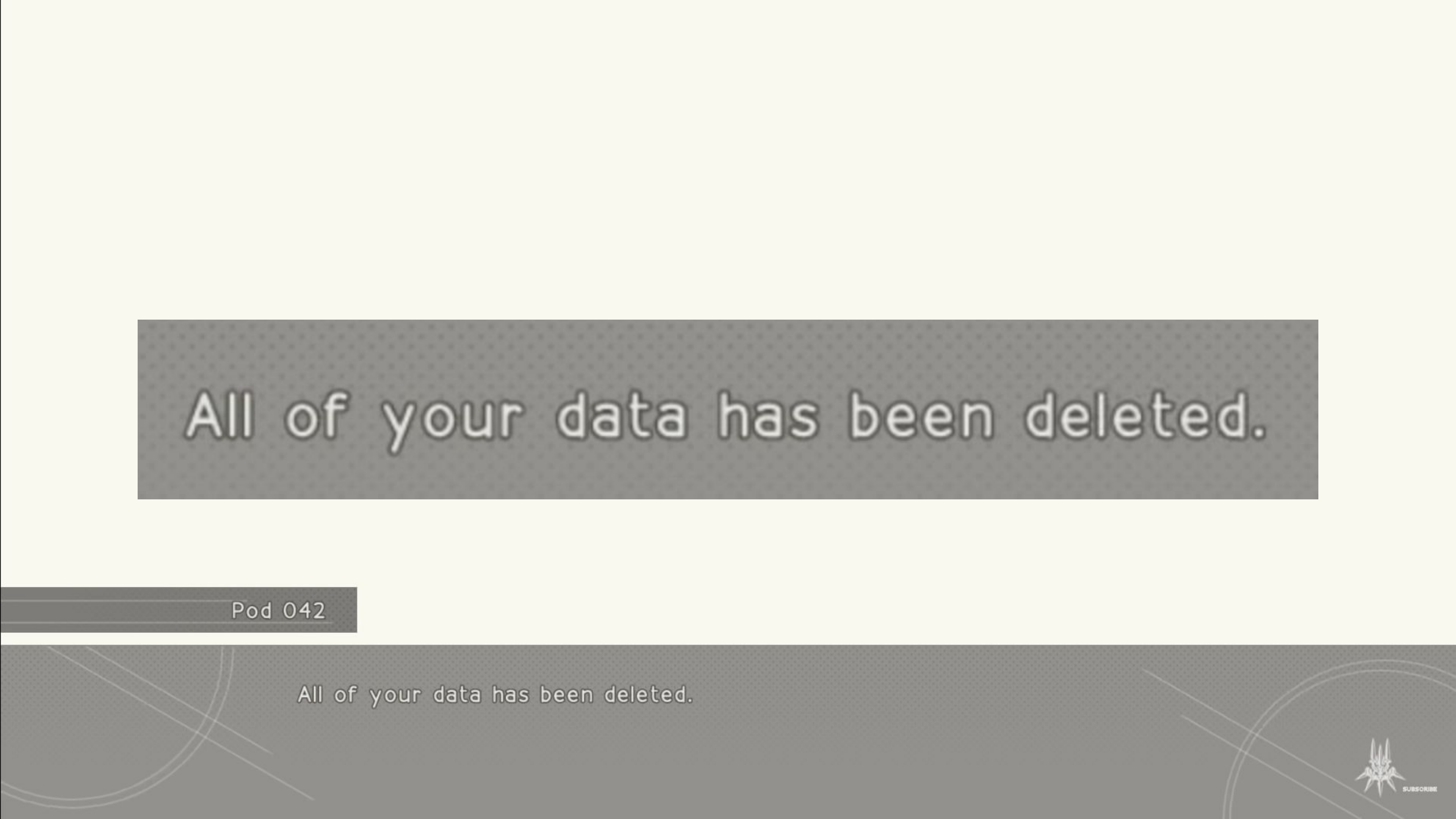
Big Data

- Larger data sets means increased complexity!
- It’s harder to store and provide access to data when you’re working with terabytes or petabytes of data

Backups & Versioning

Why backup your data?

- Technology failure
- Natural disasters
- Theft
- Human error
- Rogue AIs



All of your data has been deleted.

Pod 042

All of your data has been deleted.

3-2-1 Rule

- Three copies
  - One primary and two backups
- Two formats/media
  - e.g. External hard drive & cloud storage
- One off-site
  - Where is your cloud storage located?

3-2-1 Rule

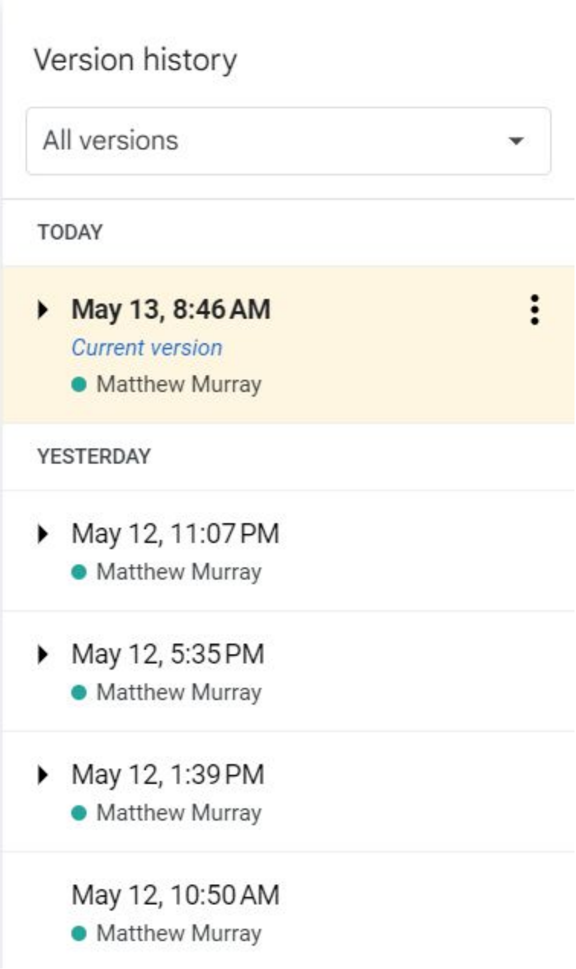
- Not always feasible
    - Big data (terabytes or petabytes of data)
    - Frequently changing data
    - Data collected in places with terrible internet
- Ask yourself:
    - How often do you need to backup your data?
    - If you lost a specific piece of data could your research project continue?

USBs are ***Not*** Data Storage

- You will lose them

Versioning

- Versioning is when you save specific versions of your files
- Some software (Git!) does this automatically
- Can be as basic as having “raw” and “cleaned” versions
- Be consistent in giving version numbers



Version history

All versions

TODAY

May 13, 8:46 AM
Current version
● Matthew Murray

YESTERDAY

May 12, 11:07 PM
● Matthew Murray

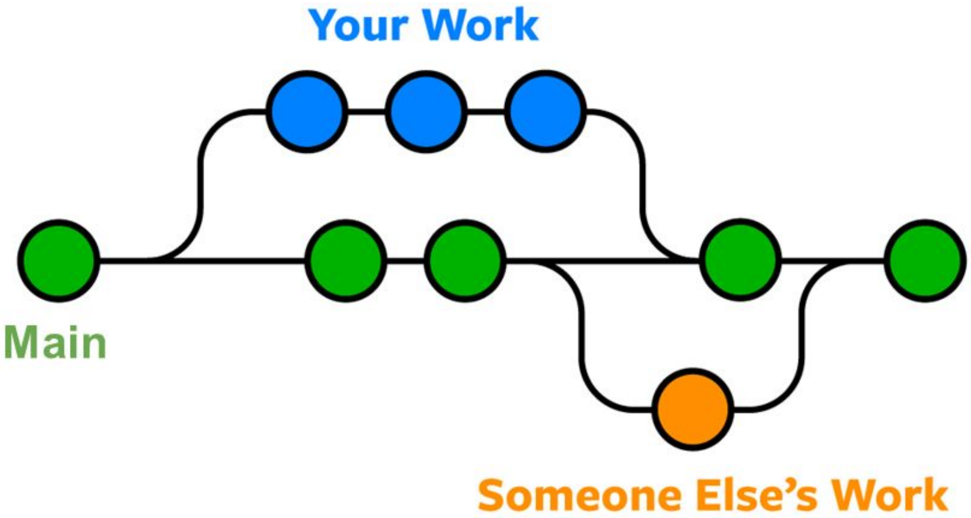
May 12, 5:35 PM
● Matthew Murray

May 12, 1:39 PM
● Matthew Murray

May 12, 10:50 AM
● Matthew Murray

Git/GitHub

- Can be used for anything (not just code!)
- Manage and view of different branches
- Collaboration through merging of branches
- Tracks history



Deleting Files

- Deleting files is also an important part of data management
- I have a monthly calendar reminder to spend some time and delete files I no longer need

Documenting Data

What is Documentation?

- Documentation is capturing your research process (from data collection through analysis)
- Includes the Five Ws (and One H)
    - What was done, Who did it, When it happened, Where it happened, Why it was done, and How it was done

Why is Documentation Important?

- You can keep track of what still needs to be done
- Your future self will thank you for writing down what you did
- Allows others to understand your process and replicate your work

Examples of Documentation

- Data Dictionaries & Codebooks
  - Contain a description of elements in a dataset, including names, definitions, acronyms, and other relevant information.
- README files
  - A plain text file that provides information and instructions about a project, including its purpose, usage instructions, known issues, and contact information for support or collaboration.

Codebooks

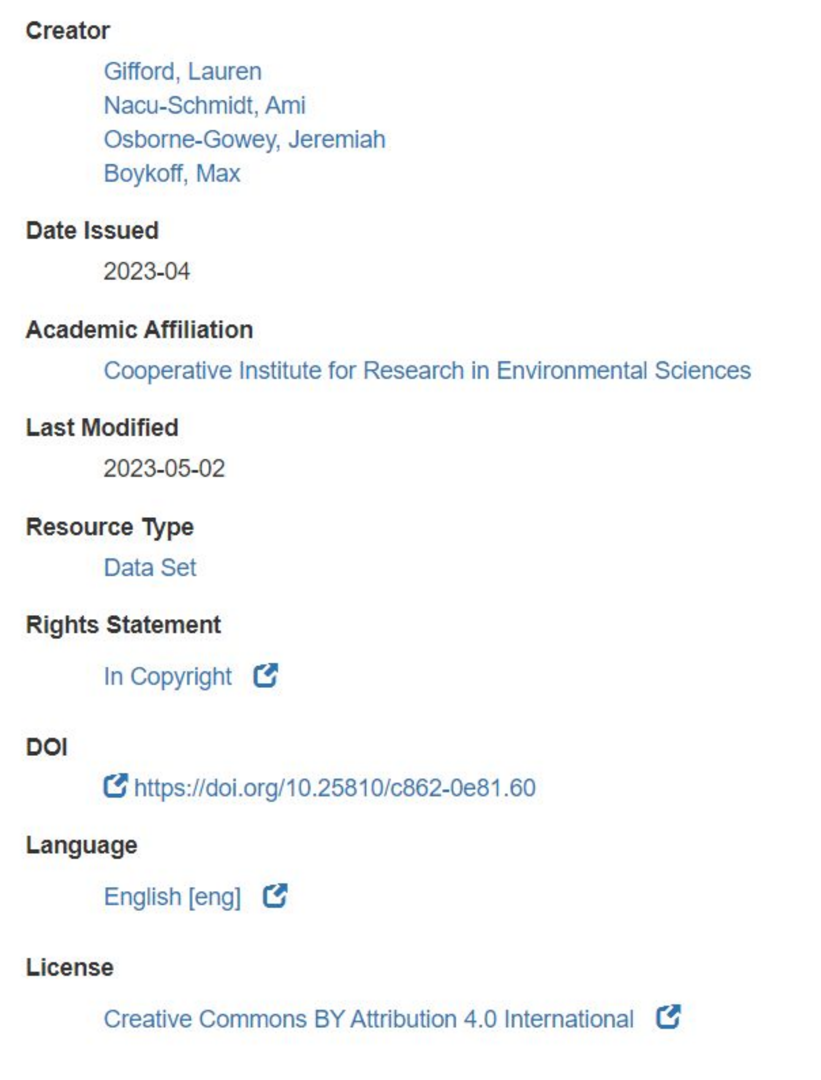
- Define the variables and their units
- Explain the formats for dates, time, geographic coordinates
- Define any coded values and missing values
- Allows others outside of your research group to understand the data

README Files

- Title of Dataset
- Authors
- Contact information
- Date of data collection
- Licenses/restrictions placed on the data
- Links to publications that cite or use the data
- Recommended citation for the data
- Structure and organization of the data files
- List of software (with version numbers) and instruments

Metadata

- Data about data!
- Descriptions that help you find and understand data
- Different fields/disciplines use different metadata standards



Creator

Gifford, Lauren
Nacu-Schmidt, Ami
Osborne-Gowey, Jeremiah
Boykoff, Max

Date Issued

2023-04

Academic Affiliation

Cooperative Institute for Research in Environmental Sciences

Last Modified

2023-05-02

Resource Type

Data Set

Rights Statement

In Copyright

DOI

https://doi.org/10.25810/c862-0e81.60

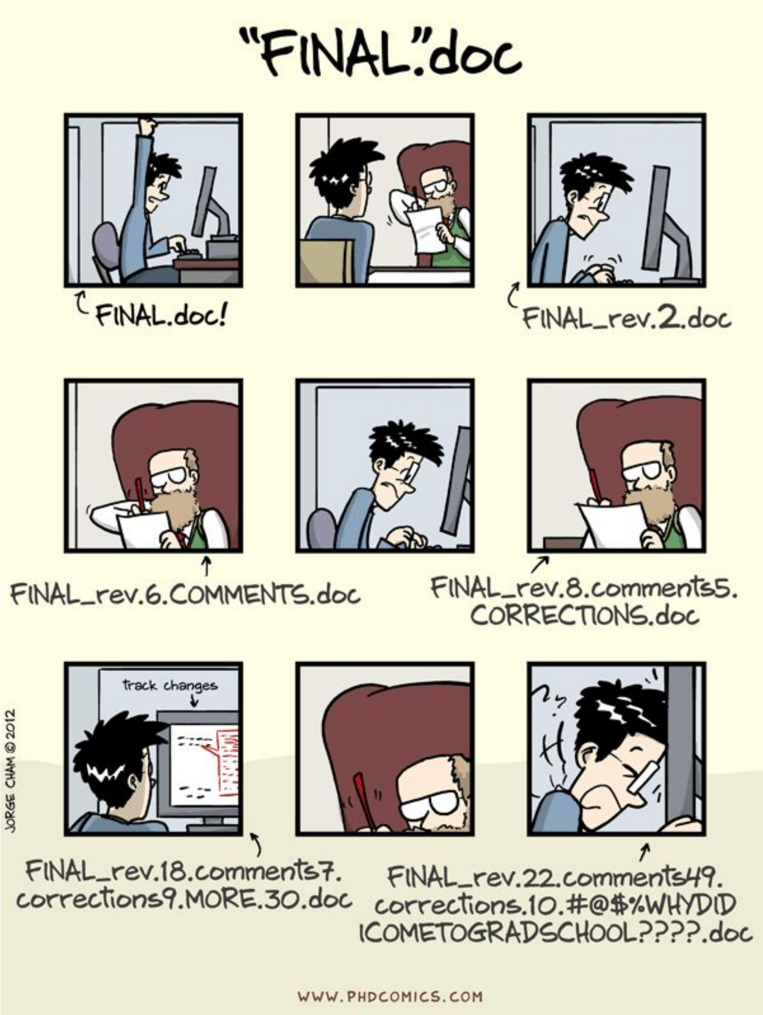
Language

English [eng]

License

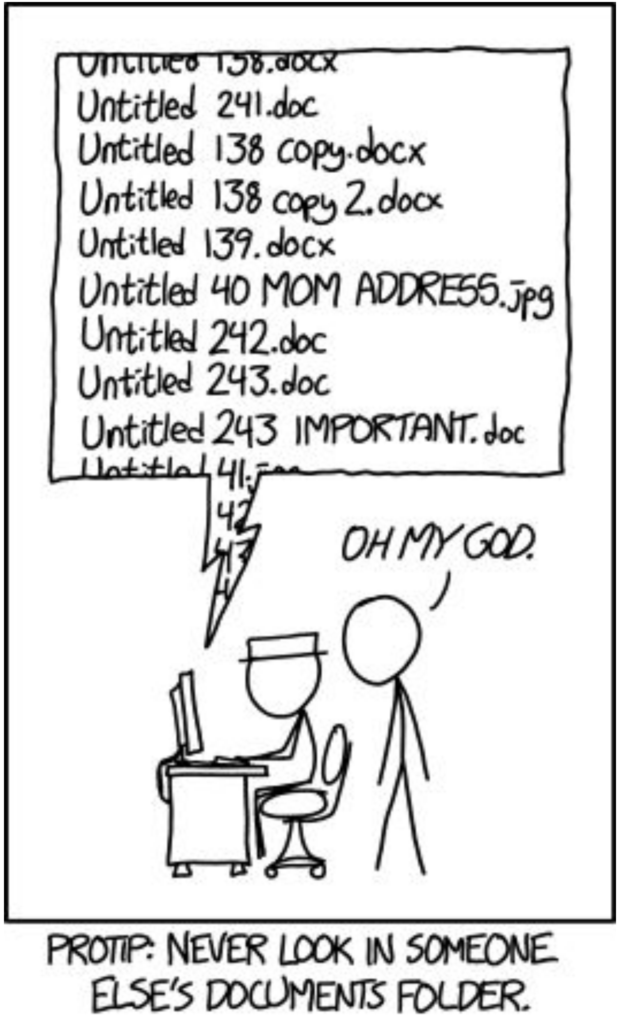
Creative Commons BY Attribution 4.0 International

File Names & Structures


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDID ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

PHD Comics: NotFinal.Doc. Jorge Cham, 2012. https://phdcomics.com/comics.php?f=1531 .

xkcd: Documents. Randall Munroe. https://xkcd.com/1459/



File Naming Best Practices

- Be consistent
  - Do: Use the same format for all files
  - Don’t: Keep changing file names
- Be descriptive
  - Do: Avoid generic terms
  - Don’t: Use “Final” in your file name
- Limit file name length
  - Do: Use (some) abbreviations
  - Don’t: write-out-every-word-in-your-data-file.xlsx

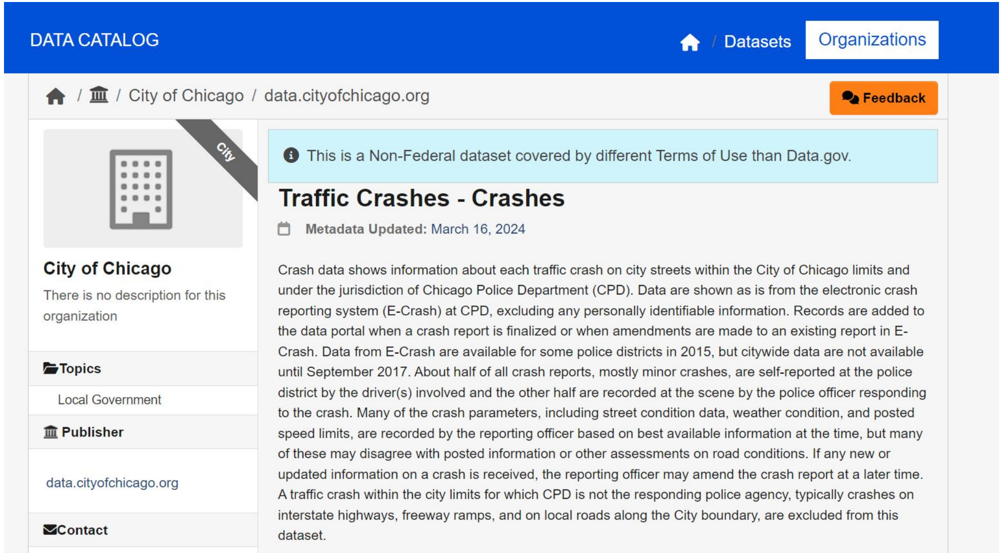
Abbreviations

- Use meaningful abbreviations
- Document your decisions
  - Don’t assume you’ll remember what an abbreviation means
- Have group & individual identifiers
- Use version numbers

File Naming Best Practices

- Use CamelCase (not all systems preserve case)
  - Do: FileName-2023.pdf
  - Don’t: use spaces in your file names.doc
- Use standardized numbers and dates
  - Do: Use leading zeros (001.png)
  - Don’t: Have files named 9-12-11.csv
- Use the Latin alphabet
  - Don’t: Use punctuation or special characters
  - - and _ are okay!

Don’t use spaces in your file names



This is a Non-Federal dataset covered by different Terms of Use than Data.gov.

Traffic Crashes - Crashes

Metadata Updated: March 16, 2024

Crash data shows information about each traffic crash on city streets within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD). Data are shown as is from the electronic crash reporting system (E-Crash) at CPD, excluding any personally identifiable information. Records are added to the data portal when a crash report is finalized or when amendments are made to an existing report in E-Crash. Data from E-Crash are available for some police districts in 2015, but citywide data are not available until September 2017. About half of all crash reports, mostly minor crashes, are self-reported at the police district by the driver(s) involved and the other half are recorded at the scene by the police officer responding to the crash. Many of the crash parameters, including street condition data, weather condition, and posted speed limits, are recorded by the reporting officer based on best available information at the time, but many of these may disagree with posted information or other assessments on road conditions. If any new or updated information on a crash is received, the reporting officer may amend the crash report at a later time. A traffic crash within the city limits for which CPD is not the responding police agency, typically crashes on interstate highways, freeway ramps, and on local roads along the City boundary, are excluded from this dataset.

City of Chicago

There is no description for this organization

Topics

Local Government

Publisher

data.cityofchicago.org

Contact

Traffic_Crashes_-_Crashes.csv

- This also breaks certain software (or makes it harder to use)

https://catalog.data.gov/dataset/traffic-crashes-crashes

Dates

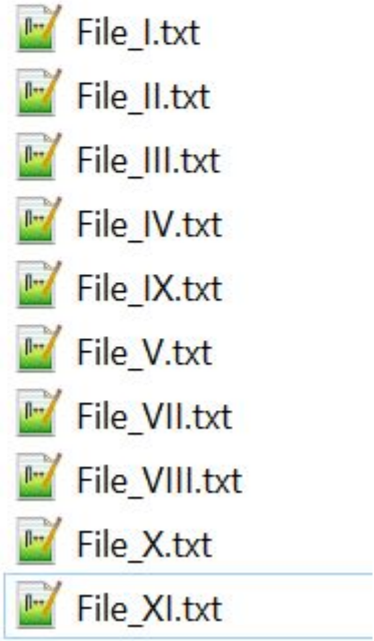
- Don’t use terms like “Quarter 1” or “spring”
- ISO 8601
- 2023-10-31 or 20231031 (YYYY-MM-DD)
- Remember that historically dates were not consistent
    - 1712-02-30 (February 30th, 1712) was a real date in Sweden

Use the Latin alphabet

- This sucks
- I’m sorry
- If your data features non-latin characters reach out to us and we can provide advice
- When you use them, make sure they’re encoded properly

Don’t use Latin numerals

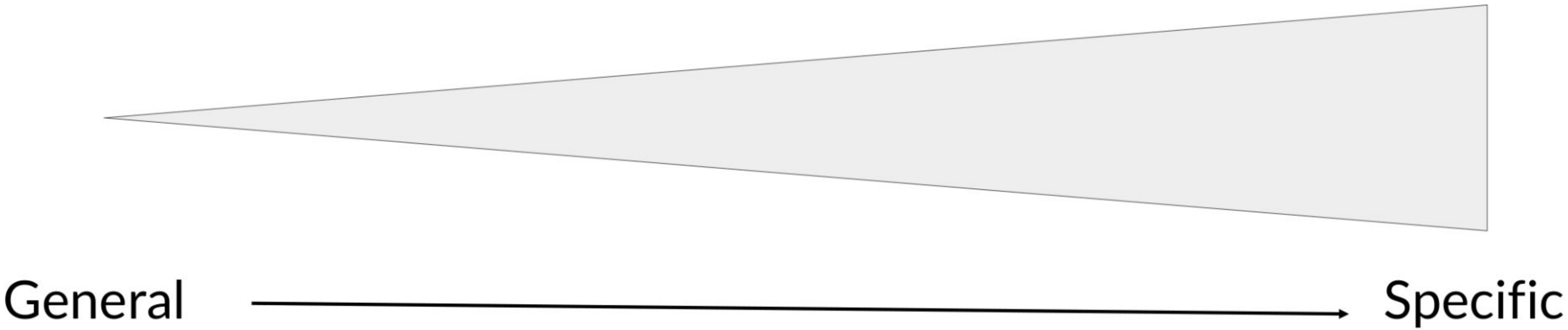
- Don’t use Latin numerals to name your files (they won’t sort properly)
- You can use them in documentation (sometimes)



File_I.txt
File_II.txt
File_III.txt
File_IV.txt
File_IX.txt
File_V.txt
File_VII.txt
File_VIII.txt
File_X.txt
File_XI.txt

20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx

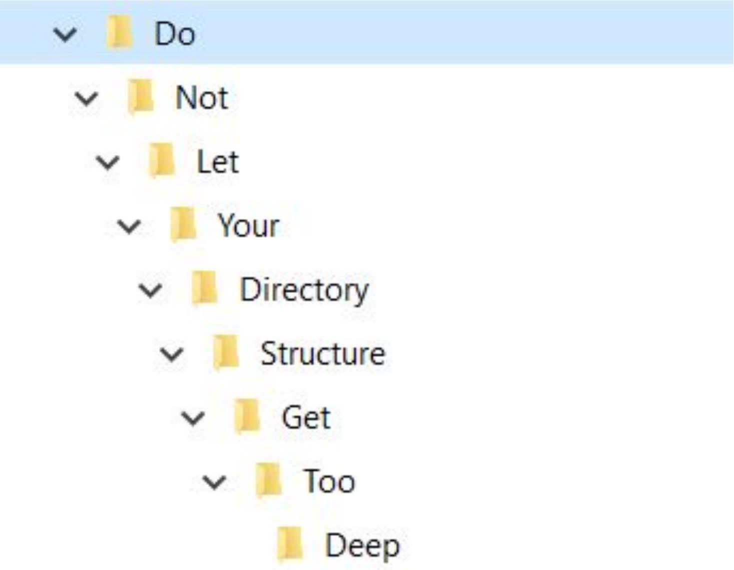
Date **Project** **Experiment** **Type** **ID** **Version**

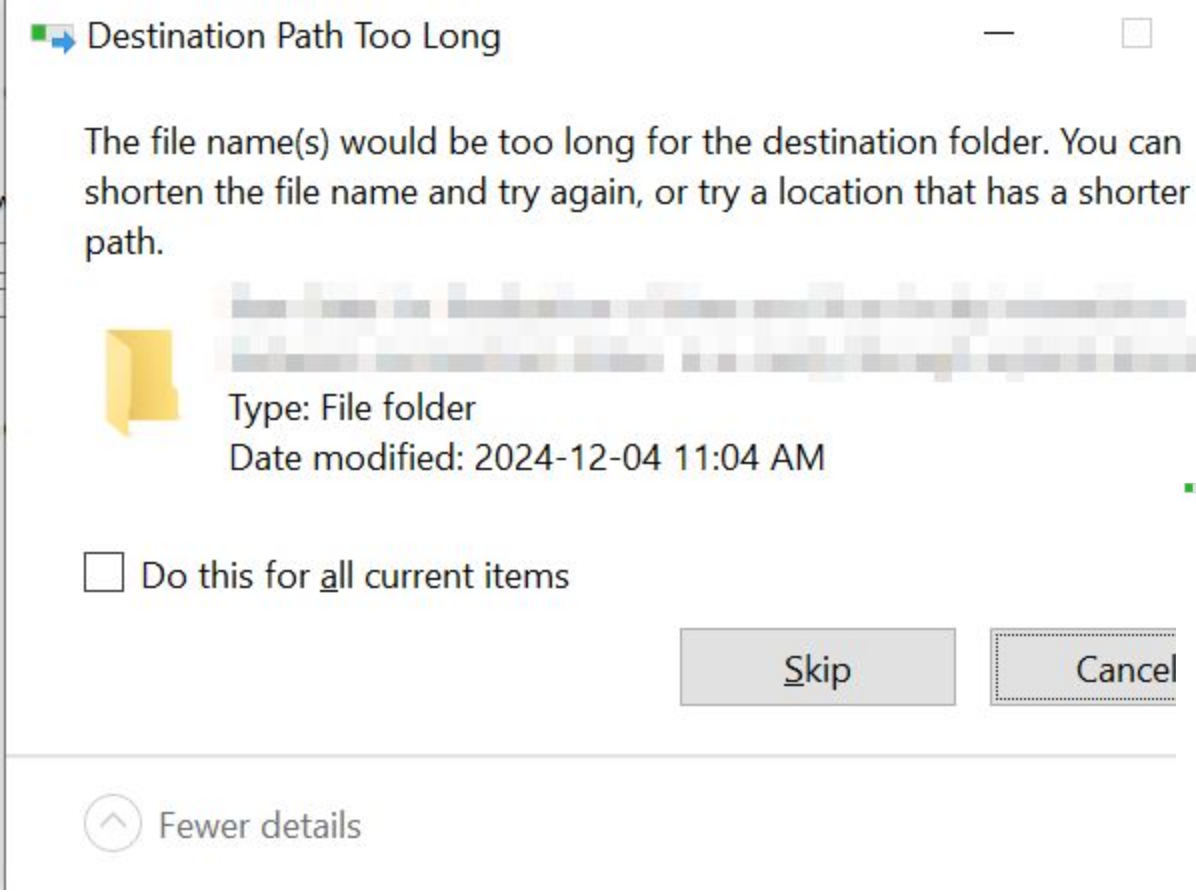


General → Specific

File Structuring

- Don’t save everything to the desktop
- Don’t let your structure get too deep





Destination Path Too Long

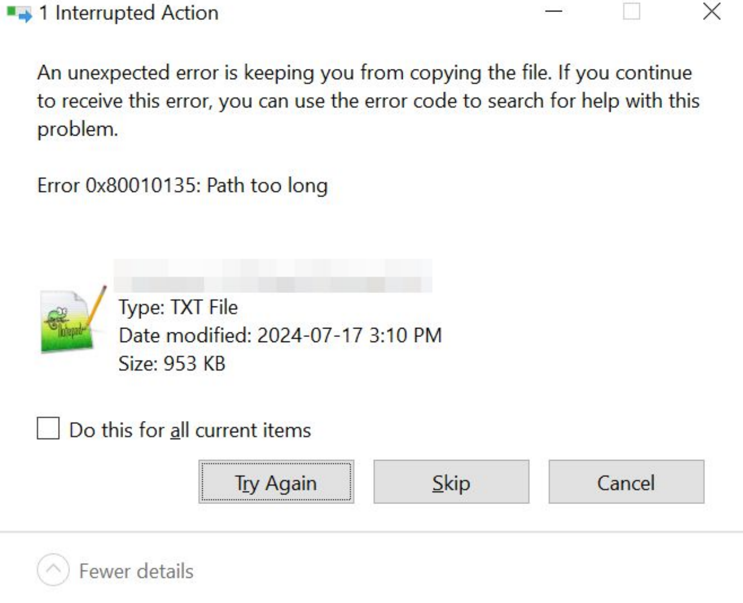
The file name(s) would be too long for the destination folder. You can shorten the file name and try again, or try a location that has a shorter path.

[redacted]

Type: File folder

Date modified: 2024-12-04 11:04 AM

□ Do this for all current items



1 Interrupted Action

An unexpected error is keeping you from copying the file. If you continue to receive this error, you can use the error code to search for help with this problem.

Error 0x80010135: Path too long

[redacted]

Type: TXT File

Date modified: 2024-07-17 3:10 PM

Size: 953 KB

□ Do this for all current items

Project directory structure

Project_1

- methods
- raw_data
  - readme
- analysis
  - analysis_method_1
    - 2017
    - 2018
  - analysis_method_2
- scripts
- manuscript
  - text
    - version_1
- readme and/or ELN link

Always keep your raw data!

(Raw data should be “read-only” if possible.)

File Archiving and Compression

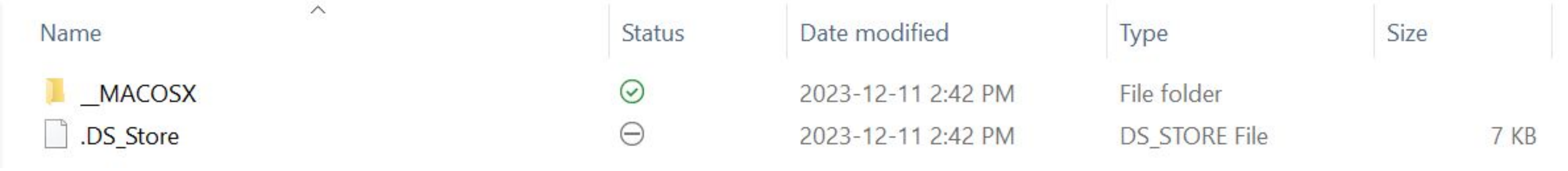
- Usually called “zipping”
- Means collecting different files and directories into one file
- Many programs/methods of doing this
  - .zip, .rar., .7z, .tar, .gz, etc.
- Done to minimize size of files (compression)
- Done to collect many files and directories into one file (archiving)
- Makes it easier to share many files

File Archiving and Compression

- Consider how the files will look once they’re extracted (unzipped)
- Create a directory structure that will make it easy for people to understand and use your files
- Ensure all file paths will allow associated code to work

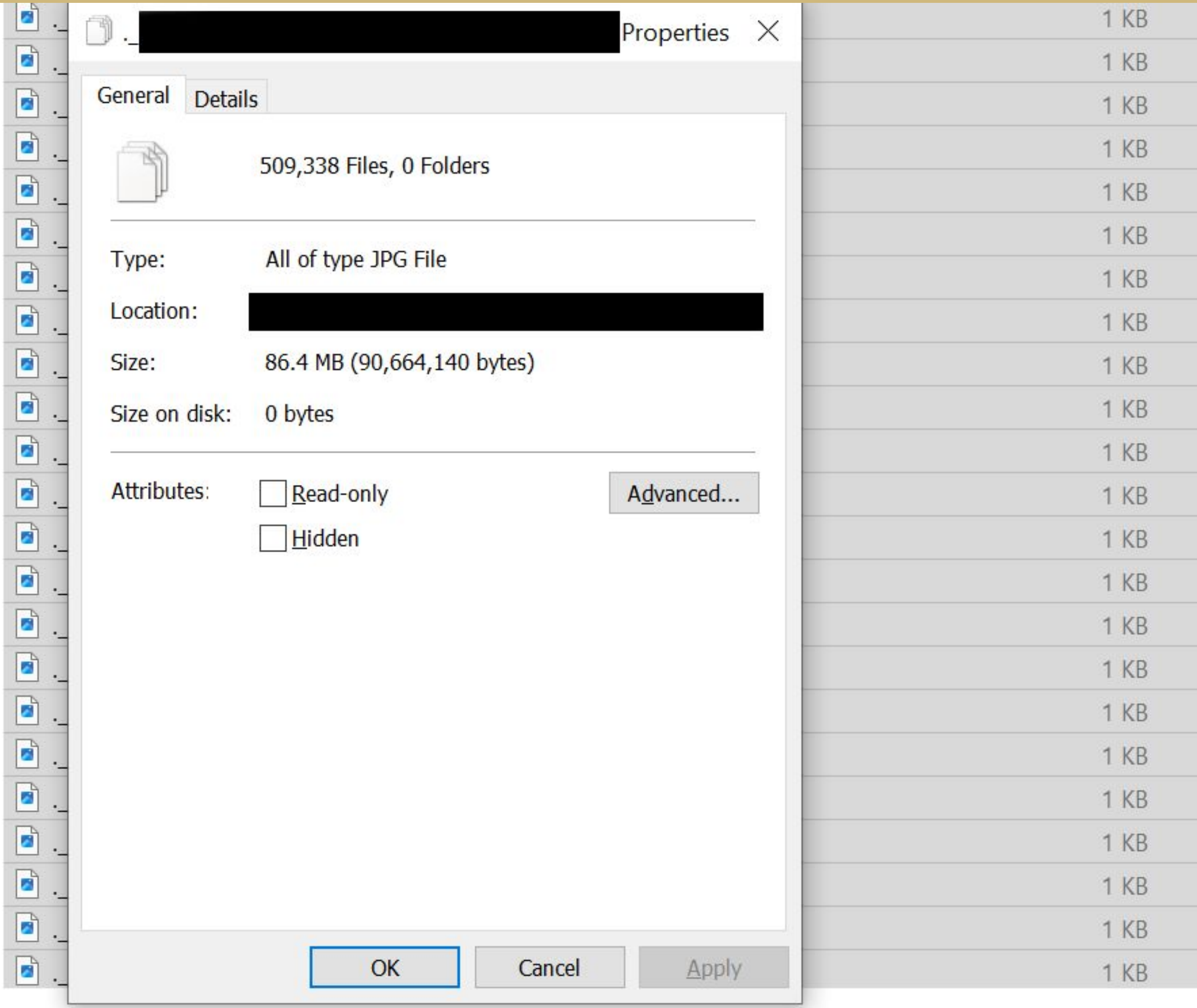
MacOS

- __MACOSX, .DS_Store, & ._ files are created automatically in MacOS
- They’re invisible on Mac, but visible on Windows and Linux systems
- They will be included in .zip files



Name	Status	Date modified	Type	Size
_MACOSX	✓	2023-12-11 2:42 PM	File folder	
.DS_Store	⊖	2023-12-11 2:42 PM	DS_STORE File	7 KB

This .zip file contained over 500,000 ._ files



._[redacted] Properties

509,338 Files, 0 Folders

Type: All of type JPG File

Location: [redacted]

Size: 86.4 MB (90,664,140 bytes)

Size on disk: 0 bytes

Attributes: □ Read-only □ Hidden

1,018,676 items | 509,338 items selected

File Formats & Units of Measurement

File Formats

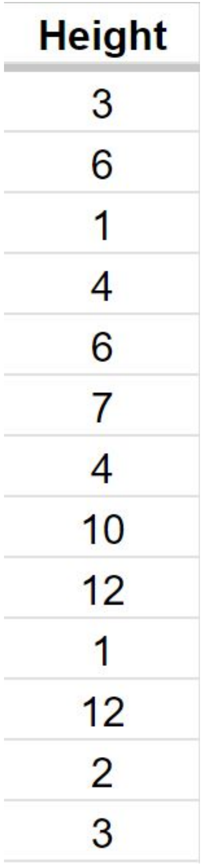
- Non-proprietary/Open
    - Can still be used even if original software is inaccessible
    - e.g. Use CSV files instead of .vc
- Unencrypted
- Uncompressed/lossless

File Formats

- Use the default file extension
    - Do: Save Excel files as .xlsx
    - Don’t: Save Excel files as .spreadsheet
- Yes, I have seen people use non-standard file extensions

Units of Measurement

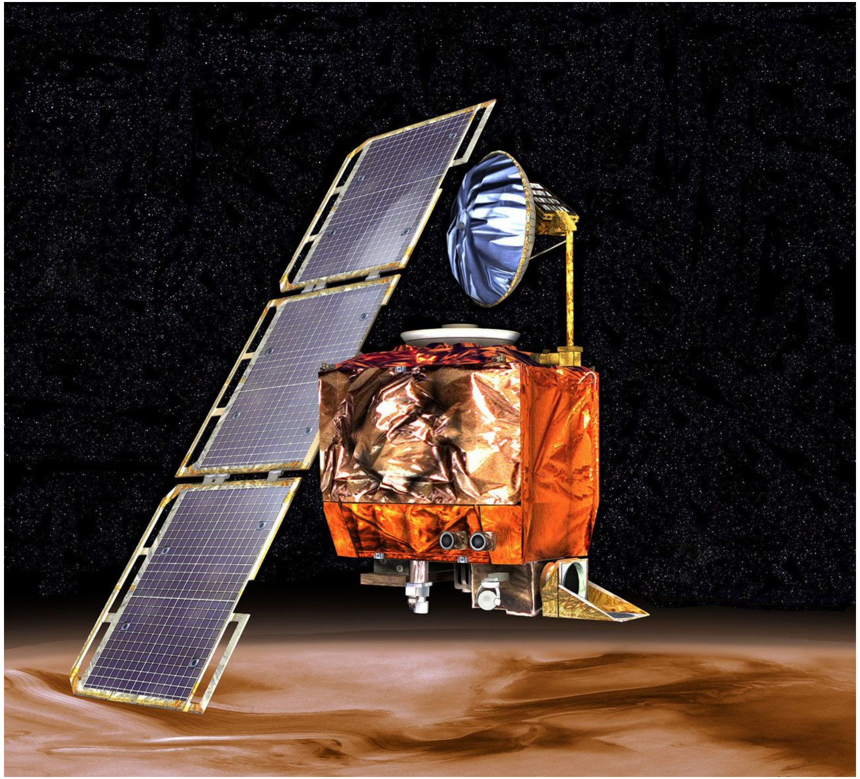
- Be consistent
- Use standardized measurements
- Don’t keep changing between them
- Ensure everyone is using the same system
- Define them in your documentation



Height
3
6
1
4
6
7
4
10
12
1
12
2
3

Learn from the Mars Climate Orbiter

- Spacecraft that was probably destroyed in the Martian atmosphere
- The failure was due to a measurement mismatch between SI units (metric) and US customary units causing numbers to be incorrect by a factor of 4.45



Data Management Plans
&
Data Management and Sharing Plans

What is a DMP/DMSP?

- A framework for how you’ll manage your data
- Describe your plans for collecting, organizing, storing, and sharing your data
- About two pages long

When should you make a DMP?

- When you have to for grant requirements
  - Different grants have different requirements
- When you don’t have to, but want to, keep track of your data

Two types of DMPs

- The “final” version you’ll submit with a grant proposal
- The “living” version that you’ll continue to update as your research project progresses

What goes in a DMP?

- Types of data generated in this project
- Estimated size of data
- Software and file formats that will be used
- Where data will be stored, who can access it, and any security considerations
- Any privacy, legal, or ethical constraints
- Metadata standards
- How the data will be preserved/shared
- How the data can be reused
- A description of roles and responsibilities

DMPTool

DMPTool

- Has pre-formatted DMP templates from various funding agencies
- Walks you through the process of completing the DMP

DMPTool Activity

- Go to https://dmptool.org/
- When signing in, indicate that you're from CU Boulder
- Select a DMP Template relevant for your field (if you're having trouble choosing, pick the generic NSF template)
- Look through the prompts
- Attempt to answer the prompts in the context of your proposed project

DMPTool Activity

- Were any of the prompts challenging or confusing?
- Is there anything you would like clarification on?

- We’re happy to read drafts and provide feedback on your draft DMPs, so please send them our way!

Other Useful Tools

Data Management Tools

- Open Science Framework
  - One-stop shop for project management
- Open Refine
  - Clean data
- File Renamers (various)
  - Rename files so they’re consistent
  - You can also batch rename files in the command line
- Zotero
  - Reference management software

Consultations

How CRDDS can assist with Data Management

- One-on-one or small group consults
- Review draft DMPs and README files
- Help navigate data policies
- Find data repositories
- Advice on file formats, etc.
- Email us: crdds@colorado.edu

Acknowledgements

- Slides adapted from previous presentations by Matthew Murray (CU Boulder)