University of Idaho

Organizing your Research and Data Management

with

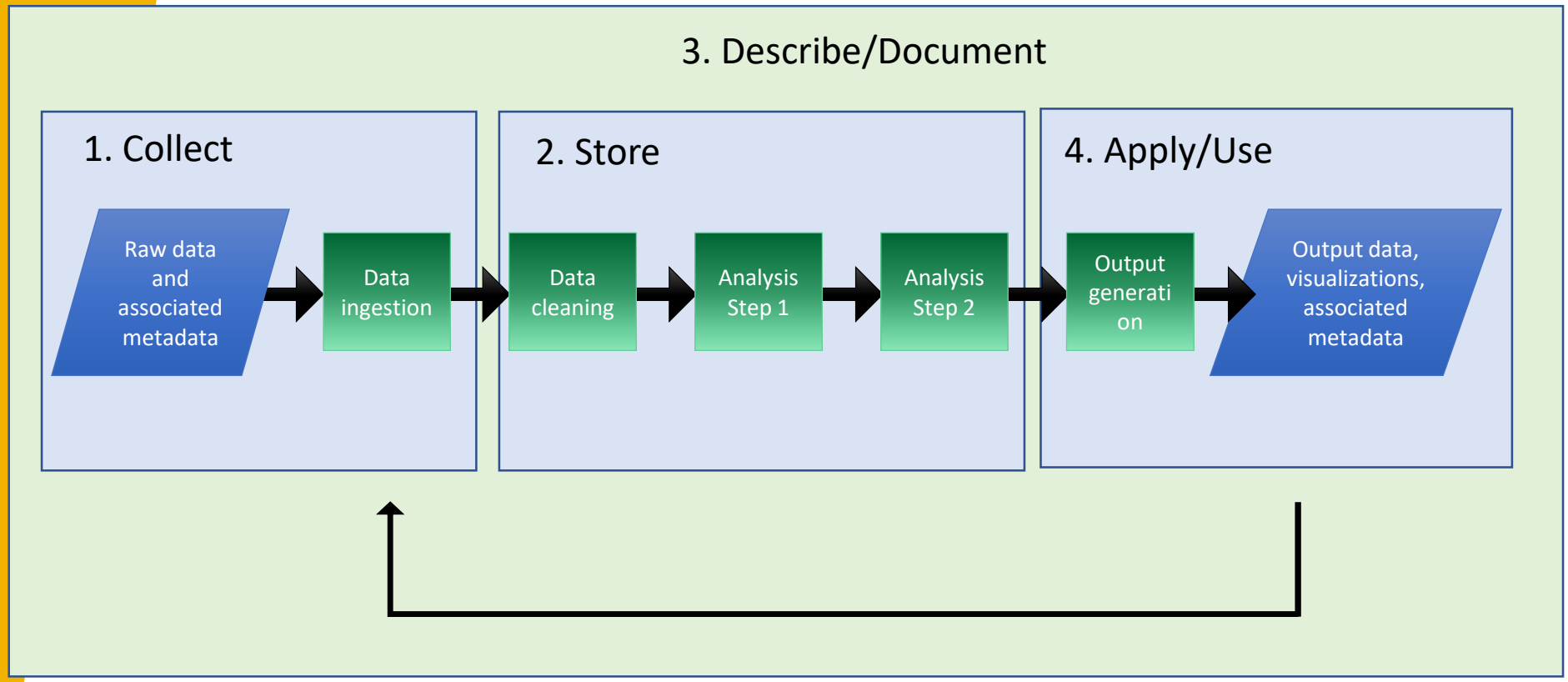
Jeremy Kenyon,
Research Librarian

jkenyon@uidaho.edu

Summary

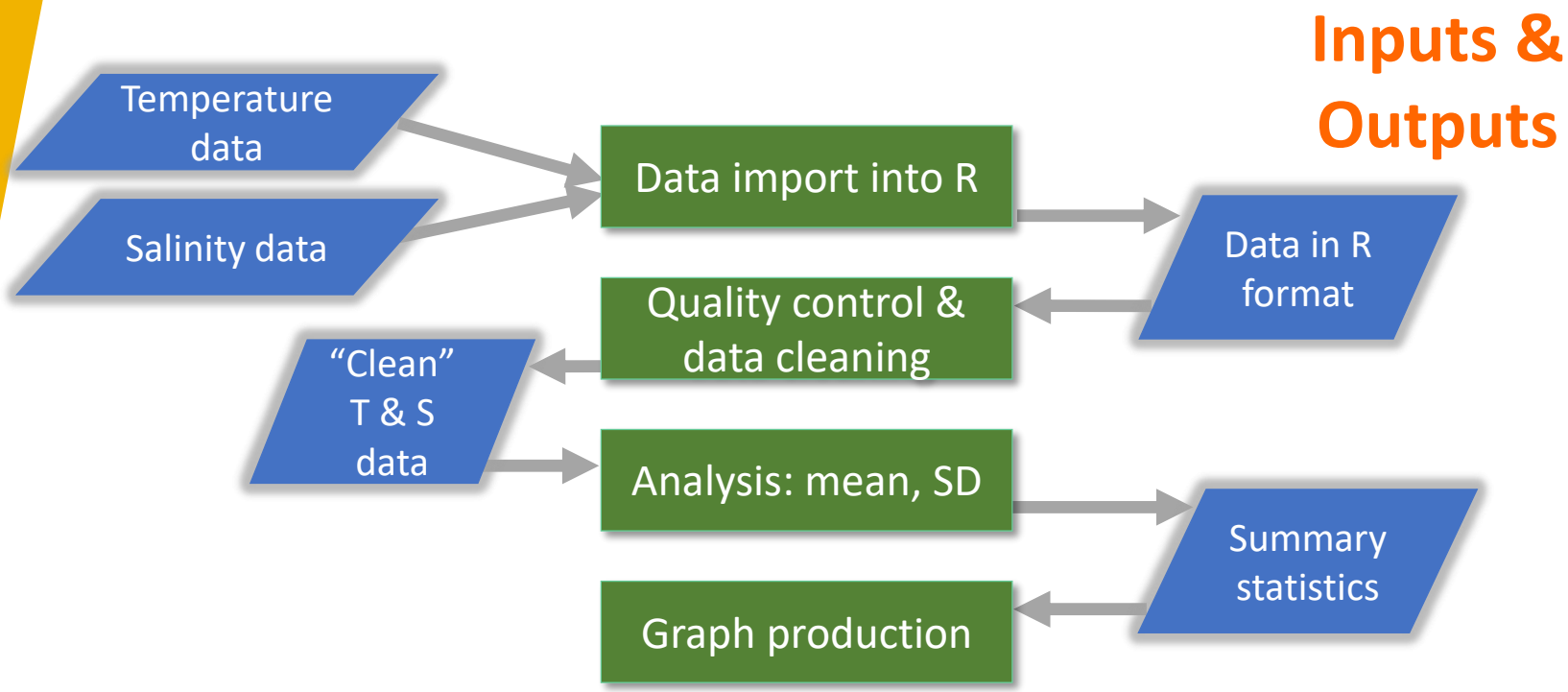
Considering Workflows
Formats/Volume
Data Dimensionality and Format
Storage
Backing Up
File Organization
File Naming

Four Parts of Data Management



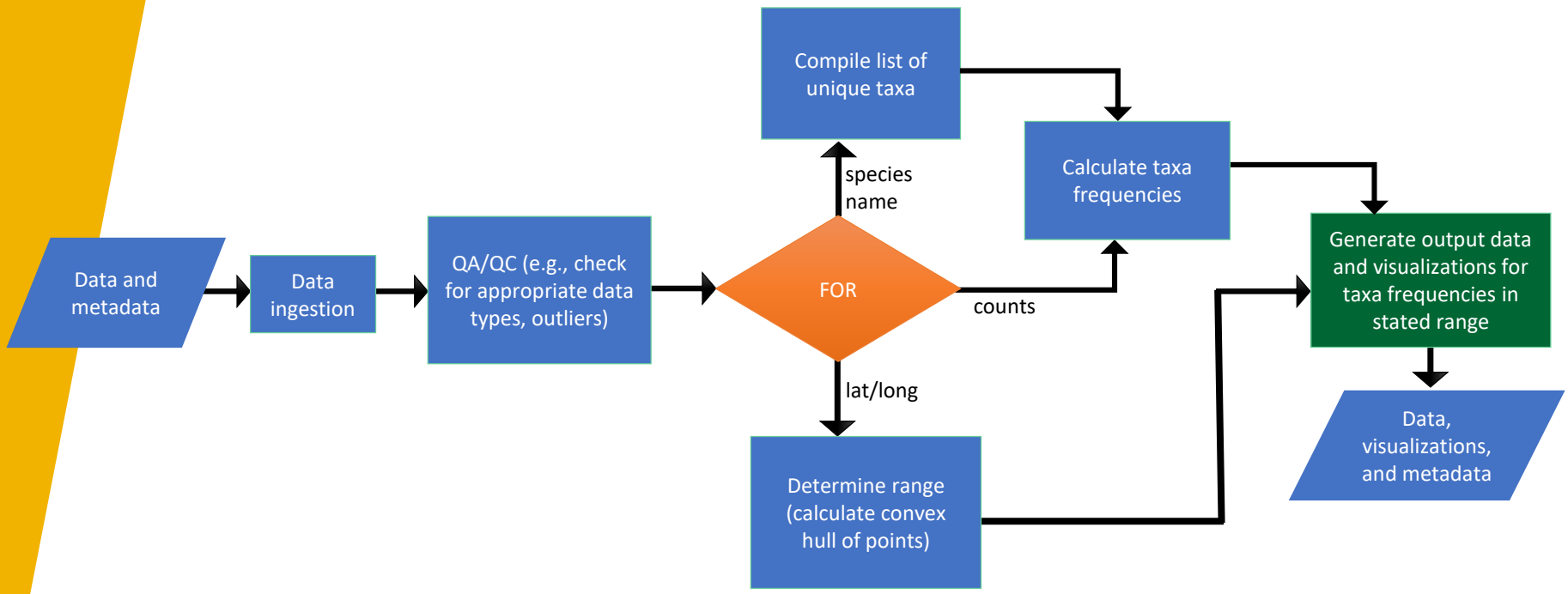
Adapted from: DataONE, 2012

Plan your data strategy



From: DataONE, 2012

Plan your data strategy



From: DataONE, 2012

Data Formats

Two considerations:

- Fancy files (.docx, .dat, .xlsx) – program overhead, valuable/necessary for a specific tool
- Simple files (.txt, .csv) – usually cleaner, fewer encoding issues, more accessible to other tools/systems
  - Some exceptions like .pdf and .shp

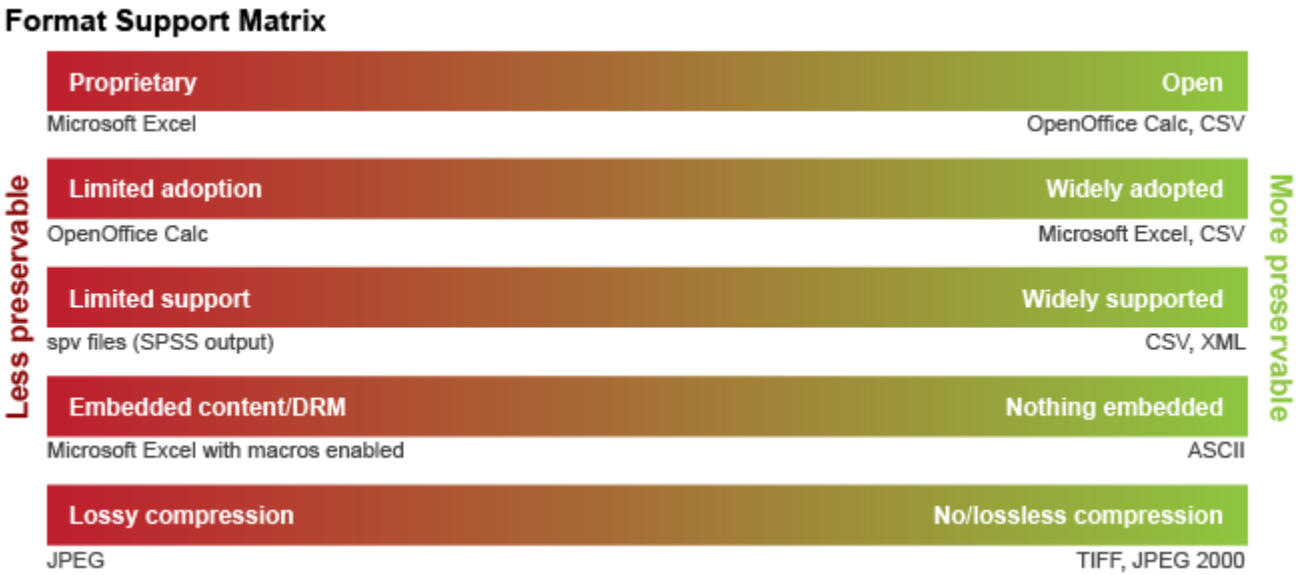


Image from: Univ. Illinois, 2020

Dimensionality in Data Modeling

Start with core “fact(s)” or variables of interest

- e.g. an observation of an animal

You might record a number of facts about the animal

- e.g. weight, length, a behavior

There might be relevant other dimensions

- e.g. time, space, environmental variables, observer/instrument information, proximity to relevant vegetation, etc.

The more of these “dimensions”, the more complex your data may become.

Multi-dimensional Data and Volume

Lower dimensional data can function fine as a spreadsheet

- Easy to add/ingest new data, easy to manage, easy to summarize

Higher dimensional data needs more control and management

- Relational databases, other database systems
- Size, complexity becomes a problem that database software can mitigate

For large 4-dimensional data (x, y, z, t,…*n*)

- Hierarchical Data Format (HDF) and similar (e.g. netCDF)
- Basically, for very large gridded time-series data, it is extremely fast and efficient. Can read/write with it much more quickly than a relational DB, or some other approaches.

Volume:

- Primary issue: large amounts of data will affect your organizational approach, and possibly the structures and storage you use. Figure out the general volume early!

“Not Only SQL” (NoSQL) Databases

If your data is part of the 3 v’s – velocity, volume, variety – a NoSQL database might be the best form for organizing it.

Heterogenous data:

- Images + spatial data + time series + text
- Combining semi-structured and unstructured data are harder in a pure relational environment

Changing, evolving, or “unfinished” data

- If you expect to change the schema frequently, you might need something more flexible than a relational tool

Ultimately, your **data model should be reflective of the problem you are trying to solve.**

- Example: If you are trying to analyze a social network, where the relationships between different people/groups are the key units of analysis, then a graph database might be more suitable than a relational one.

Types of Storage

Local Storage
- Good temporary solution, assuming you have the space

Networked Drive Storage
- Better solution, both for temporary and longer term. Usually due to local support in cases of emergencies or failure.

Cloud Storage
- Convenient storage and sharing platform, but there are issues to consider, including syncing problems (can affect active read/write) and security for sensitive data

Physical Media (Flash Drives, External HDs)
- Meant to move data. Smaller drives are not necessarily reliable for long-term storage.

Storage Options at UI

UI Cloud/Networked Option:

- OneDrive, approx. 5 TB of space
- Both web and “app” (aka mounted drive)

Cloud Options

- Dropbox – 2 GB
- Google Drive – 15 GB
- Other similar providers

Some resources like the Open Science Framework

Occasionally, other local resources through Depts., Labs, and campus units:

- NKN: approx. $325 per terabyte per year

Backing Up Data

3-2-1 Rule
- Have at least 3 copies of your data
- Store them in 2 different media
- Keep 1 copy off-site (geographically differentiated)

Common problems
- Corrupted data, failed hard drive, laptop lost/stolen, mistakes (deletions, user error)

Example plan:
- One copy on local hard drive
- One copy on OneDrive (geographic replication off-site)
- One copy on a physical media device

Image from: Vera, 2020

Adapted from: Univ. of Virginia, 2020

Folder Organization

Hierarchical:

Project Name
- Folder 1
- Folder 2
  - Subfolder 1
    - File 1
  - Subfolder 2
    - File 2
    - File 3

Pros: Clear, Easy to Use, Similar items stored together

Cons:
Items can only go in one place, can become too granular (too deep)
Too many folders can make for a long filepath
(Folder A/SubfolderB/SubfolderC/file1)

Adapted from Malinowski C. 2016

Folder Organization

You ultimately want:

A structure that allows you to quickly and easily find what you're looking for

To include documentation and descriptive information

Organize folders into meaningful categories:

- Primary/secondary/tertiary levels of analysis
- Subject/collection method/time/space/data type
- Code/Reports separate from data itself

File Naming

File naming convention (FNC):

> a framework for naming your files in a way that describes what they contain and how they relate to other files (Brandt, 2017)

Principles:

- Be consistent
- Be descriptive
- Imagine someone else trying to understand your file from the name
- Make it machine readable (computable) and human readable (comprehensible)

Tools:

- Bulk Rename Utility: https://www.bulkrenameutility.co.uk/

File Naming

Time stamps are always useful

YYYY-MM-DD: ISO 8601 date/time format

Use leading zeros, except at the start of the file name

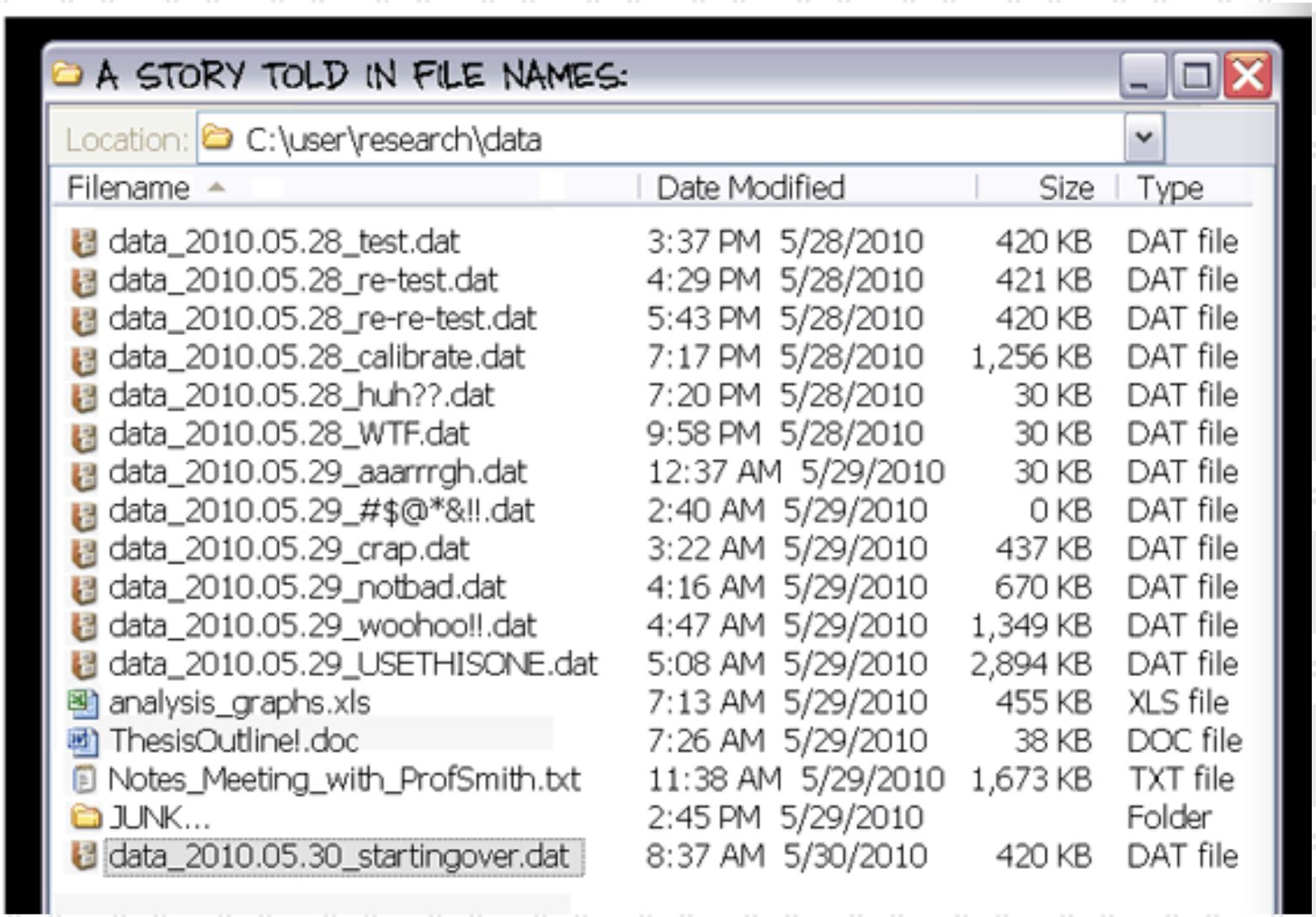
2020-01-10 not *2020-1-10*

Use only one period – it's just clearer to use a period to denote the file extension

Instead of *Smith_metadata.fdgc.xml* use *Smith_metadata_fdgc.xml*

Avoid using generic names like MyData, FirstProject, FinalData

Use filenames that denote significance:
smithLakeObservations or coreSamples_V04



From Strasser, 2011

Questions:

Jeremy Kenyon, Research Librarian

jkenyon@uidaho.edu

208-885-7955

Graduate Student Essentials

September 8: Research Refresher

September 15: Planning and Organizing a Literature Review

September 22: Tools for Building Scholarly Presence

September 29: Citation Management with Zotero

October 6: Tips and Tricks for Microsoft Word, Excel, and OneDrive

October 13: Organizing your Research and Data Management

October 20: Creating a Research Poster

References

DataONE. 2012. “DataONE Education Module: Analysis and Workflows.” Retrieved from: http://www.dataone.org/sites/all/documents/L10_Analysis Workflows.pptx

University of Illinois Library. 2020. “File Format Considerations.” Research Data Service. Retrieved from: https://www.library.illinois.edu/rds/file-formats/

University of Virginia Library Research Data Services + Sciences. 2020. “Data Storage and Backups.” Retrieved from: https://data.library.virginia.edu/data-management/plan/storage/

Malinowski, C. 2015. FileOrg. http://libraries.mit.edu/data-management/files/2014/05/FileOrg_20160121.pdf

Bryan, J. (2015). “naming things.” Presentation at a Reproducible Science Workshop. Retrieved from: http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf

Strasser, C. 2011. “Best Practices for Ecological & Environmental Data Management.” Retrieved from: https://www.dataone.org/sites/all/documents/ESA11_SS3_carly.pdf

Vera. 2020. “Best Practice: 3-2-1 Backup Strategy for Home Users & Businesses [Clone Disk].” Retrieved from: https://www.partitionwizard.com/clone-disk/backup-strategy.html