Northwestern
INFORMATION TECHNOLOGY

Spring cleaning

Efficiently manage your research data storage

Presenter: Tobin Magle

Date: May 7, 2025

Slides: Spring-cleaning.pptx

Polls

Where is your data stored?

- RDSS/FSMResfiles
- Quest
- SharePoint
- OneDrive
- Cloud (AWS/Azure/GCP)
- Your computer
- Lab computers
- Other

What OS do you use?

- Windows
- MacOS
- Linux
- Other

Today we'll cover…

- Why spring cleaning?
- Planning your approach
- Migrate your data

Why Spring Cleaning?

Data Storage is getting expensive

- Data is getting bigger and more complex
- Vendors are increasing prices
- Can't keep everything in the same place forever anymore

Spring Cleaning Benefits

- **Conserve storage space** – identify unused data
- **Increase findability** – standardizing structure
- **Promote collaboration** – easier to share with colleagues

Planning your approach

- Create a data inventory
- Label your data as keep, archive, or delete
- Organize your files

Create a data inventory

Create a Data Inventory

Document what you have where

- Where do you store your data?
- What is in each location?
- How much is in each location?

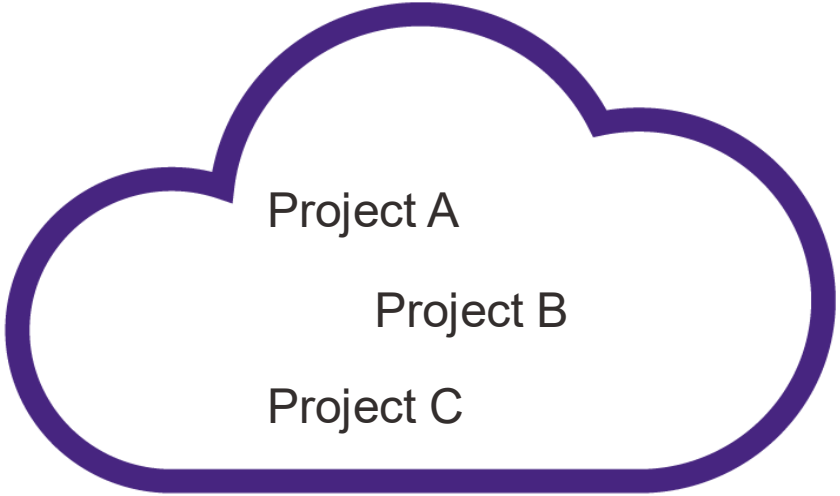
Where does your data live?

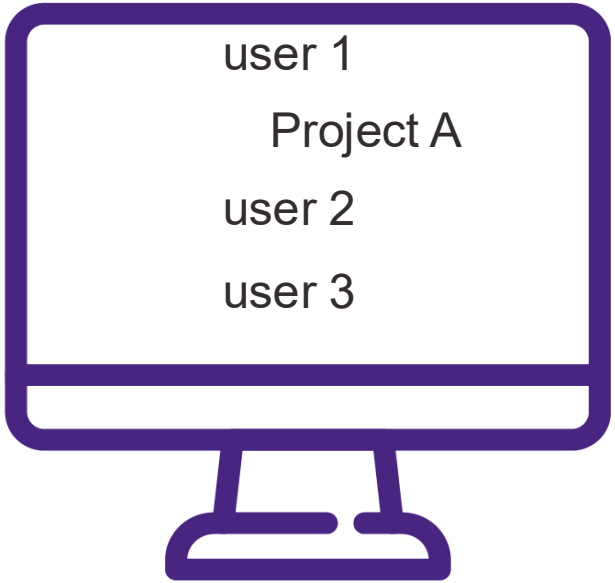
Locations could include...

- Your computer
- RDSS/FSMResFiles
- Quest
- SharePoint/OneDrive
- Lab computers/servers
- Core facility servers
- USB/External Hard Drives
- Cloud

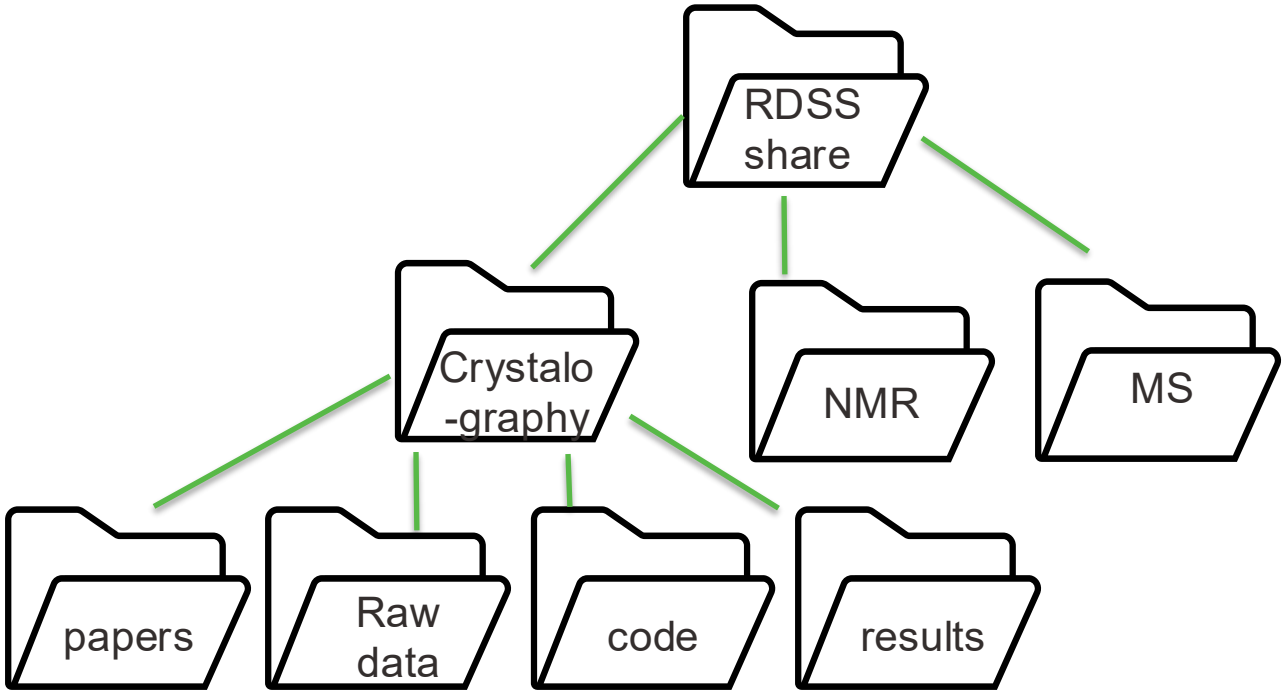
What is in each location?

Different but overlapping datasets

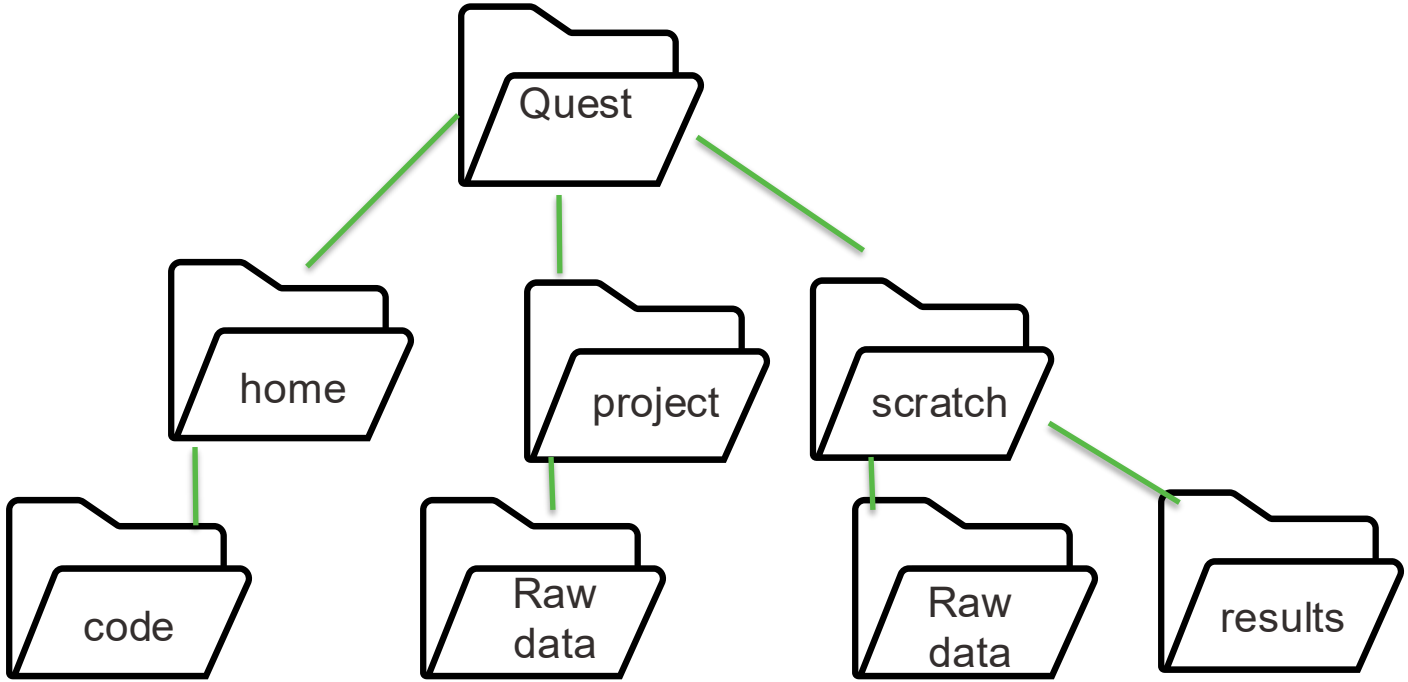




Example location: RDSS



Example location: Quest



How much is in each location?

Method varies by location

SharePoint

In web browser

- **Site owners**: Storage Metrics
- **Users**: Folder Details

Quest

Command line tools

- **Home**: homedu
- **Projects**: <allocation#>
- **Scratch**: dust

checkproject

RDSS/FSMResfiles

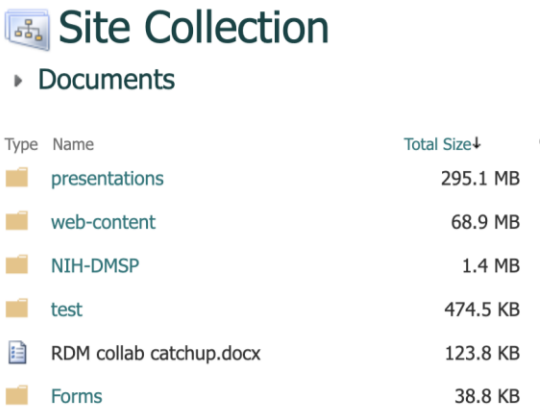
Mount your share

- **Mac**: Finder size field
- **Windows**: Folder Properties
- **Linux**: `du -sh /dir/path/`

How much in SharePoint

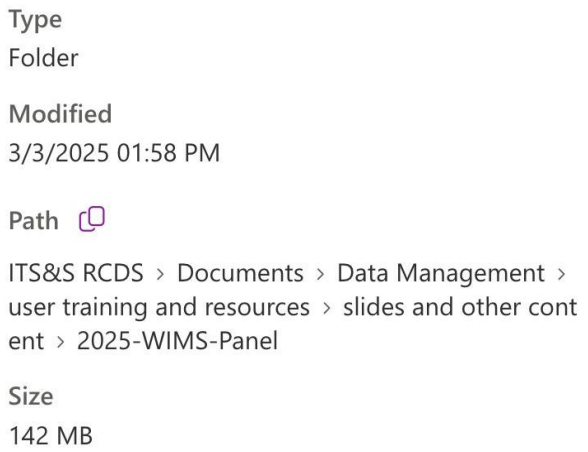
Site Owner

- Click Gear on upper right
- Site information>View all Site Settings
- Site Collection Administration > Storage Metrics



Site User

- Click ••• next to folder
- Click Details
- Scroll to the bottom



How much on Quest

Home

homedu

```
[ctm6768@quser43 ~]$ homedu

Beginning detailed disk usage report for /home/ctm6768.
Please be patient - this can be a time-consuming operation


GPFS quota for /home/ctm6768

42.31 GB used in 851 files (52.89% of 80 GB quota)


43G     /gpfs/home/ctm6768
17K     /gpfs/home/ctm6768/.jupyter
20K     /gpfs/home/ctm6768/R
13K     /gpfs/home/ctm6768/.gsutil
6.5K    /gpfs/home/ctm6768/.ssh
137K    /gpfs/home/ctm6768/.config
5.7K    /gpfs/home/ctm6768/share-quest-hps
501K    /gpfs/home/ctm6768/rserver
1.6M    /gpfs/home/ctm6768/.lmod.d
192K    /gpfs/home/ctm6768/.local
9.9K    /gpfs/home/ctm6768/.dbus
267K    /gpfs/home/ctm6768/ondemand
4.0K    /gpfs/home/ctm6768/space_test
```

Projects

checkproject <allocation>

```
[ctm6768@quser43 ctm6768]$ checkproject a9009
====================================

Reporting for project a9009
------------------------------------
46827 GB in 27374709 files (80.00% of 58360 G
B quota)
Allocation Type: Buy-in Allocation
Expiration Date: Compute and storage resource
s for buy-in allocations expire at different
times.
Please contact quest-help@northwestern.edu fo
r details regarding the expiration of your re
sources.
Status: ACTIVE
Compute and storage allocation - when status
is ACTIVE, this allocation has compute node a
ccess and can submit jobs
------------------------------------


====================================
```

Scratch

module load dust

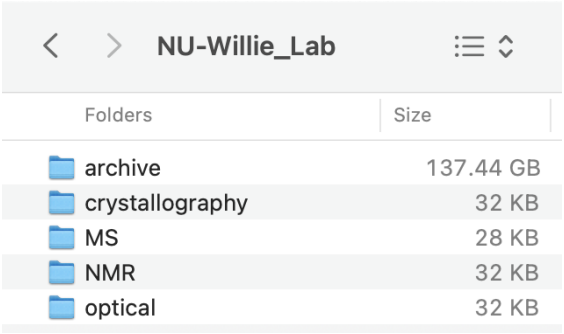
dust /scratch/netid/

```
[ctm6768@quser43 ~]$ cd /scratch/ctm6768
[ctm6768@quser43 ctm6768]$ dust
3.6M     ├── rdss-usage-stats
3.8M     │   ├── spiderT.x86_64_Linux.lua
5.7M     │   ├── lmod
5.7M     ├── .cache
9.7M     ├── Fundamentals of Life Science Tools in Google Cloud.pdf
9.8M     ├── GCP-lifesciences
2.3M     │   ├── places.sqlite-wal
2.7M     │   ├── data.safe.bin
2.7M     │   ├── security_state
5.8M     │   ├── favicons.sqlite
5.0M     │   ├── places.sqlite
7.3M     │   │   ├── 3870112724rsegmnoittet-es.files
8.7M     │   │   ├── 3870112724rsegmnoittet-es.sqlite
16M      │   ├── idb
16M      │   ├── chrome
16M      │   ├── permanent
18M      │   ├── storage
39M      │   ├── nnbrktb5.default-default
39M      ├── firefox
39M      ├── .mozilla
518M     ├── nr.20.tar.gz
1.7G     ├── nr.03.tar.gz
1.8G     ├── nr.05.tar.gz
1.9G     ├── nr.01.tar.gz
1.9G     ├── nr.02.tar.gz
1.9G     ├── nr.04.tar.gz
2.0G     ├── nr.14.tar.gz
2.1G     ├── nr.17.tar.gz
2.1G     ├── nr.08.tar.gz
2.2G     ├── nr.06.tar.gz
2.2G     ├── nr.12.tar.gz
2.3G     ├── nr.07.tar.gz
2.3G     ├── nr.11.tar.gz
2.3G     ├── nr.10.tar.gz
2.4G     ├── nr.09.tar.gz
2.4G     ├── nr.16.tar.gz
2.5G     ├── nr.15.tar.gz
2.5G     ├── nr.13.tar.gz
2.6G     ├── nr.18.tar.gz
2.7G     ├── nr.19.tar.gz
42G      ├── ctm6768
42G      ┌──┴──.
```

How much on RDSS/your computer

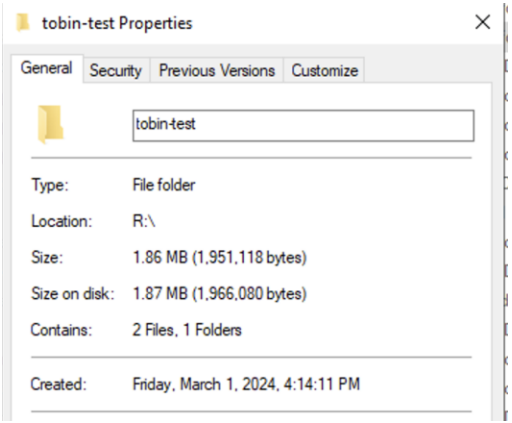
Mac

Open in Finder Window



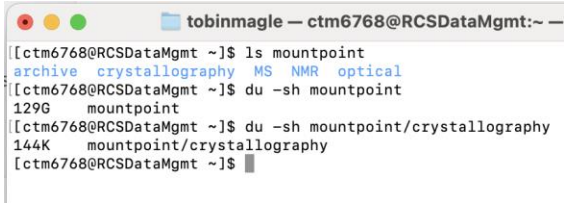
Windows

Right click>Properties



Linux

Command line



Exercise: List files in each location

For each location...

- Choose what level to document: (files, folders, projects, etc.)
- Add a row for each "thing"
- Specify where each "thing" is
- Note how much of it is

file/folder	Location	Amount
crystallography	RDSS	144 k
MS	RDSS	
NMR	RDSS	
a9009	Quest	46 TB
home	Quest	42 GB
scratch	Quest	420 MB

Label your data

Data storage is limited

You can't keep all your data in the same place forever anymore

Options

- **Compression**: zip files to make them smaller
- **Archiving**: move to a more affordable location (often less accessible
- **Deletion**: remove data that isn't useful anymore

shutterstock.com · 86456998

Label your data

Label all your data with one of the following categories

- **To Keep** – Files you need on hand at a moment's notice
- **To Archive** – Files you should keep but don't access often if at all
- **To Delete** – Duplicates, files that can be easily reconstructed or aren't useful anymore

To Keep

Files you want to keep close at hand

Files that...

- you're actively working on
- you might need at a moments notice
- are part of an active collaboration
- are small

To Archive

Files that you need to keep but don't need often (if at all)

Files that are...

- Infrequently accessed (raw data that has been analyzed)
- Too big to store on active storage
- Subject to data retention policies

https://www.it.northwestern.edu/departments/it-services-support/research/data-storage/archiving-data-when-a-project-is-done.html

Data retention policies

Know what policies apply to your data

Data Type	Retention Period
All Northwestern research data	At least three years
Data generated by students	Until the student graduates or leaves Northwestern and all papers are published
Data supporting patent applications	Until the patent process is complete
Data subject to litigation or audit	Until the situation is resolved
Data subject to HIPAA or under a HIPAA waiver	Six years past the end of project completion

https://www.it.northwestern.edu/departments/it-services-support/research/data-storage/archiving-data-when-a-project-is-done.html

To delete

Think critically about what will be useful in the future

- Old drafts of finished documents
- Temporary/Intermediate files
- Results that are easy to reproduce
- Data past retention period
- Data maintained by others (repositories)
- **Duplicates**

Finding duplicates

Gold standard: Checksums

- **Windows**: Powershell

`Find-PSOneDuplicateFile` command from PS One Tools Module

- **Mac**: Terminal

```
find . -type f ! -empty -exec cksum {} + | sort | tee /tmp/f.tmp | cut -d ' ' -f 1,2 | uniq -d | grep -hif - /tmp/f.tmp
```

- **Linux**: Command line:

```
find . ! -empty -type f -exec sha256sum {} + | sort | uniq -w32 -dD
```

Finding duplicates

Gold standard: Checksums

- **Windows**: Powershell

Find-PSOneDuplicateFile command from PS One Tools Module

- **Mac**: Te

This approach can easily take way too long if you have enough data

find . -type f ! -

np

- **Linux**: Command line:

```
find . ! -empty -type f -exec sha256sum {} + | sort | uniq -w32 -dD
```

Finding duplicates

Alternative: look for files with same name/size

Find files with same name and size

- **Windows**: index and sort or Powershell
- **Mac**: Smart Folders - creates virtual folders based on search criteria
- **Linux**: Command line

```
find . -type f -printf "%s %f %p\n" | sort | tee /tmp/files.tmp | cut -d ' ' -f 1,2 | uniq -d | grep -Ff - /tmp/files.tmp
```

Coming soon: Starfish

Data management tool for RDSS/FSMResFiles

- Find candidate duplicates
- Identify un-accessed data that can be archived
- Create usage reports for research groups
- Automate file movement based on tags

Exercise: Label your data

Mark each row as keep, archive or delete

Data set	Location	Amount	Label
crystallography	RDSS	144 k	keep
MS	RDSS		keep
NMR	RDSS		archive
a9009	Quest	46 TB	archive
home	Quest	42 GB	keep
scratch	Quest	420 MB	delete

Organize your files

Organize your data

You can organize your files by...

- Project
- Data Type (.csv, .fasta, .png, etc)
- Type of research activity (survey, assay)
- Subject characteristic (sex, species, etc.)
- Who needs access (internal vs. External)
- Chronologically (Year 1, Year 2)

Good Organization Practices

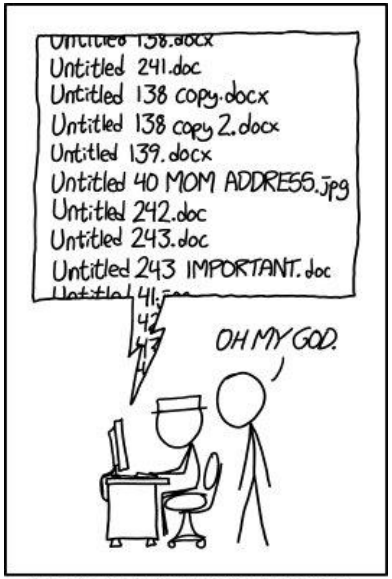
There is no one right answer. Make a plan. Be consistent

- **Portability**: Put all project files in one folder
- **Findability**: Use descriptive file names
- **Navigability**: Flatten your subfolder structure
- **Reproducibility**: Document your approach, be consistent

Descriptive file naming

The file name should tell you what's in the file

- Don’t use default names
- Include info that <u>differentiates similar files</u>
- Be kind to your computer:
  - Don't use spaces (replace with _ or -)
  - Limit special characters
  - **Use sorting to your advantage**

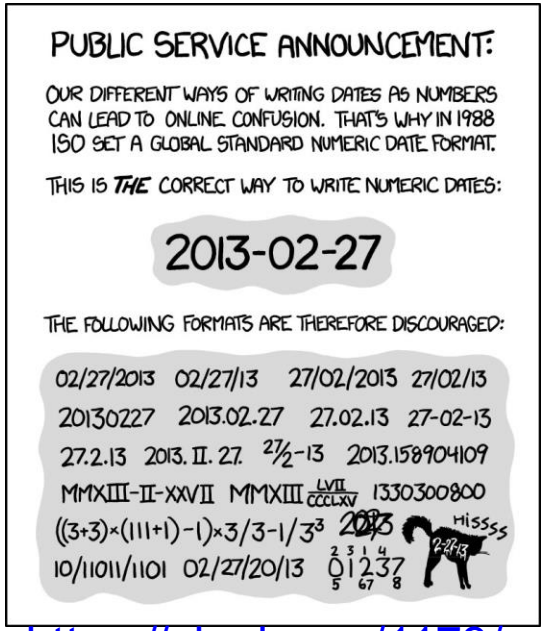


PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

Use sorting to your advantage

Go from general to specific

- Your computer is good at sorting things
- Name your files so that your computer's sorting is useful to you
- Application of the ISO 8601 date standard (YYYY-MM-DD)



https://xkcd.com/1179/

File name examples

So Good

- 2025-05-02-raw_sensor_data.csv
- data_cleaning.py
- 2025-05-02-processed_sensor_data.csv
- Nature_manuscript.docx
- Nature_manuscript_Figure_1.tif

No Good

- Data.csv
- Script.txt
- Final_data.csv
- Paper_V1_ctmedit_FINAL_FINAL.docx
- Jenny_Fig_V3??.tif

Deep Folder structures

No technical limit to how deep folder structure can go, but...

Deep folder structures

- Slows down read/write/list
- make it hard to find things
- Make it tedious to navigate

Best practice: limit to 4 levels

```
Project/
  Data/
      Raw/
          Day1/
          Day2/
      Processed/
          Day1/
          Day2/
  Papers/
      Results/
          Tables/
          Figures/
      Manuscript/
          V1/
          V2/
```

How to make it flatter

Use descriptive file naming

- Good file names can hold all the information you need to make things findable
- PLUS you don't lose the information if the files are reorganized

```
Project/
    Data/
        Raw_data_day1.csv
        Raw_data_day2.csv
        Processed_data_day1.csv
        Processed_data_day2.csv
    Papers/
        Table1.xls
        Figures1.tif
        ManuscriptV1.docx
        ManuscriptV2.docx
```

Document your approach

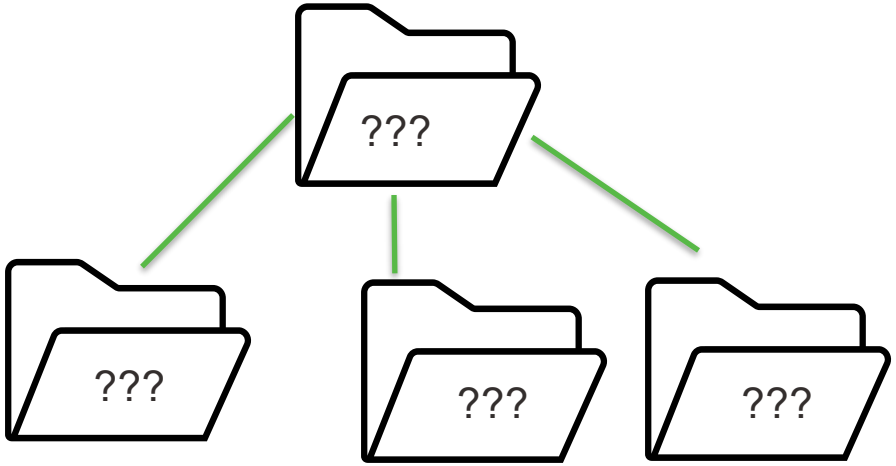
Use a README file to explain...

- How your folders are structured and what they should contain
- Your file naming convention
- Abbreviations or variable names you use
- How to store data for a new experiment (provide a template!)

Exercise: Organize your files

Decide your strategy

- Pick a top-level organizational characteristic
- What subfolders do you need? (minimize your levels)
- Decide on a file naming convention for each type of file

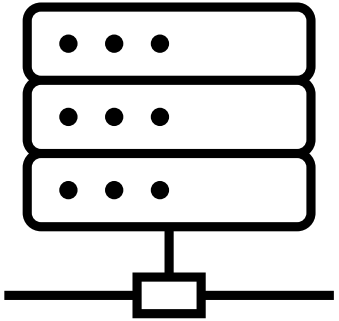


Migrate your data

Where do I put my data

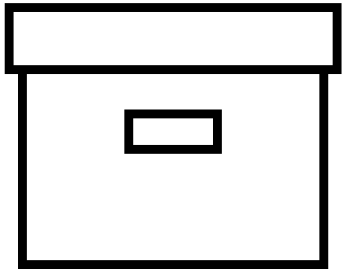
To keep

- Active storage



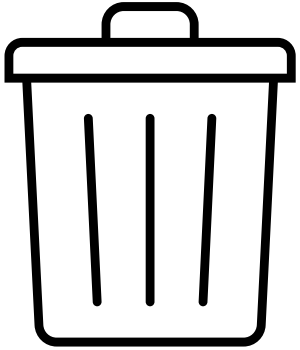
To archive

- Cold Storage



To delete

- Trash Bin



Active Storage Options

SharePoint

- Medium to long-term storage
- Files you want to share outside Northwestern
- Smaller Files (max 250 GB)

RDSS/FSMResFiles

- Medium to long-term data storage (depends on size)
- Files you need at a moment's notice
- Larger files

Quest

- **Home**: store code that you run on Quest
- **Projects**: Short-term storage for data analyzed on quest
- **Scratch**: 30 days max for high I/O work

Archival Storage Options

Small data (< 1TB)

- Keep on active storage

Medium Data

- Consider keeping on active storage
- May need an archive plan as data accumulates
- Consider compression

Big Data (> 10 TB)

- Too expensive to keep in place.
- Consider Cloud Archival storage

Cloud storage

Comes in different tiers

Hot / Cool

- Data you want to analyze in the cloud
- More expensive than Northwestern solutions

Cold

- Data that's too big to store cost effectively on campus, but you might need

Archive

- Big data that you need to keep won't access (<1x/year)

Cloud Storage examples

Cost/accessibility tradeoff

Tier	Access	Cost/TB/ year	Min days	Retrieval cost/TB
Hot	Frequent (multiple times a day), Instant retrieval	~$240	none	
Cool	Infrequently (weekly or monthly), quick retrieval	~$120	~30	
Cold	Rarely (1-2 times a year), slower retrieval	~$48	~90	
Archive	Very rarely (<1x per year), slowest retrieval	~$12	180-365	

Compliance

Make sure storage systems comply with regulations (or vice versa)

- **Northwestern polices**: No unapproved storage systems
- **Storage system policies**: No PII on Quest
- **Data use agreements (DUAs)**: Specific controls (eg encryption)
- **Federal and state regulations**:

Exercise: Where to store?

Mark your spreadsheet

Data set	Current Location	Amount	Pile	New location
crystallography	RDSS	144 k	keep	RDSS
MS	RDSS		keep	RDSS
NMR	RDSS		archive	Cloud
a9009	Quest	46 TB	archive	Cloud
home	Quest	42 GB	keep	Quest
scratch	Quest	420 MB	delete	----

Moving data

Method will depend on:

- How much are you moving? (TB)
- Where are you moving it to?
- How many files are you moving?

Data Preparation

Moving many small files is inefficient

- Bundle files into a file archive (.tar, .zip, etc)
- You can compress files to save space (and $$$) on the destination
- Optimal file size for cloud storage and data transfer is ~1-100 GB

Data Movement Methods

- Drag and drop/ftp/scp work well up to a point
- Larger datasets need more robust data transfer methods
- What happens when your network connection drops?
- How can you identify file corruption?

Globus

Preferred method of data transfer

- Connects to RDSS/FSMResFiles, Quest, SharePoint, Cloud Storage, your computer
- Web or command line interfaces
- Retries if you get disconnected
- Checksum verification

Summary

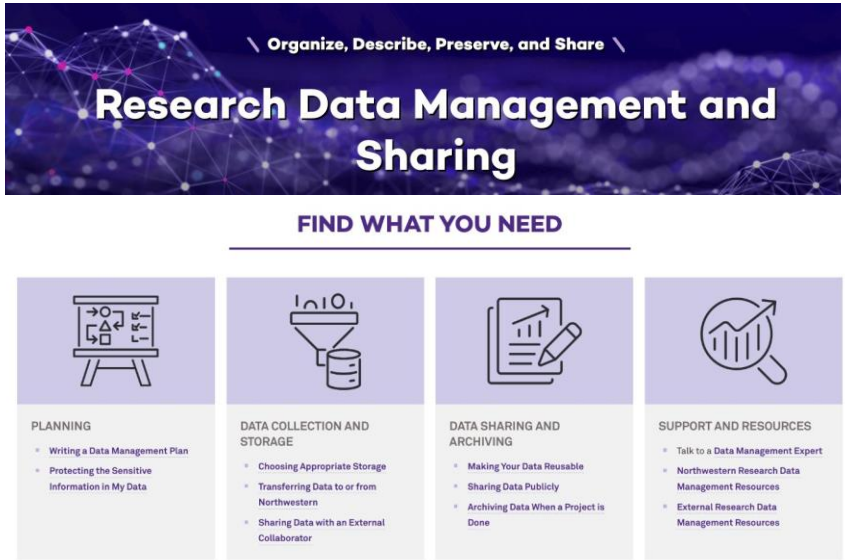
If you make a plan, you're ahead of the game

- Get a handle on what you have
- Decide what you need to keep, archive, or delete
- Get organized
- Implement your plan

RDM Resources

Email researchdata@northwestern.edu for general help

- Northwestern Research Data Management Website
- RCDS RDM Consult form
- RCDS Cloud Consult form
- Galter Data Lab Consult form
- Information Security: Protect your research
- Office hours:



https://www.it.northwestern.edu/departments/it-services-support/research/data-storage/

Questions?