Northwestern
INFORMATION TECHNOLOGY

Data Storage 101:

Understanding your data and your options

Tobin Magle

October 8, 2025

https://github.com/nuitrcs/rdm-workshops

Why a workshop about data storage?

Choosing data storage is often an afterthought

- Collecting and analyzing data is interesting.
- Organizing it is not.
- Data is saved without a plan.

This strategy used to work,
BUT...

Why a workshop about data storage?

The storage landscape is changing

- Data is getting bigger and more complex
- Storage is more expensive
- Quotas are lower (Eg: OneDrive)
- Can't keep everything in the same place forever anymore

And so...

We need a plan

Today we’ll cover how to…

- Take stock of your research data
- Note key characteristics that affect your storage decision
- Choose storage that fits your workflow

Taking stock of your data

What is data?

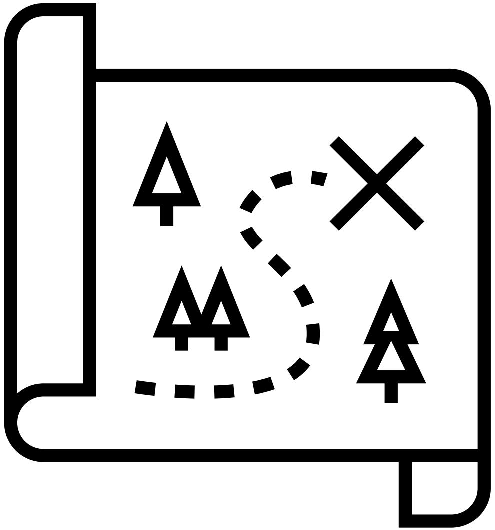
Recorded factual material and evidence collected or generated to validate research findings

- Data collected from scientific instruments
- Survey results
- Measurements collected by hand
- Cleaned and annotated data
- Summaries and visualizations

How do we take stock?

Create a data inventory

- A "map" of what and where data is collected, stored, processed and used
- Includes characteristics like format, sensitivity, who has access, etc.
- Creates a foundation for determining your storage needs



Data Collection

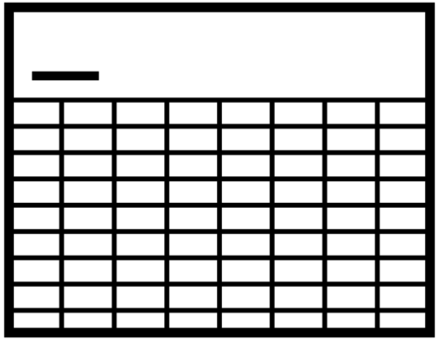
Goal: Keep raw data safe

- Store in an **accessible** location
- Keep **multiple copies** in case one gets corrupted
- **Limit access** to those who need it
- Pro tip: set a copy to **read-only** to prevent accidental alteration or deletion

Types of data produced

Raw

Original, direct from the source



Created by Gene Stroman from Noun Project

Data Processing

Goal: Clean and format data for analysis

- Keep the **raw data** raw: make a copy before making any changes

- What resources do you need to do the processing
  - Software
  - Compute resources

- Consider whether you need to keep a copy of the processed data (is the process automated?

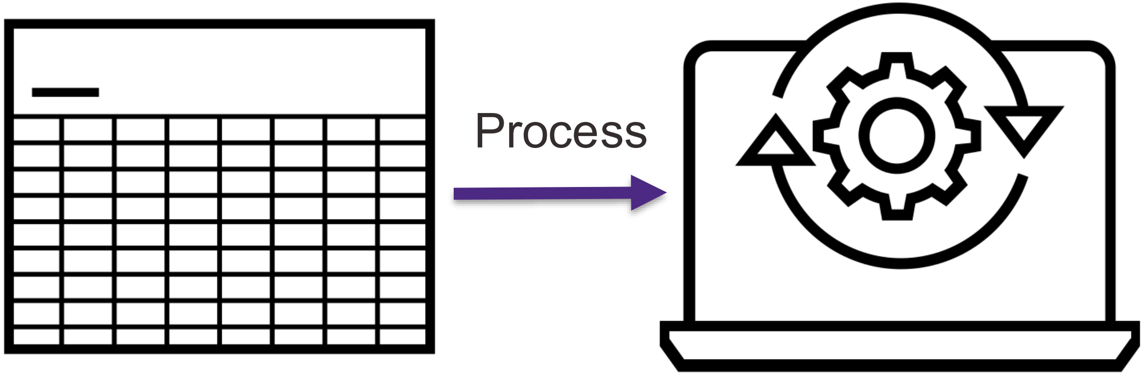
Types of data produced

Raw

Original, direct from the source

Processed

Cleaned and formatted for analysis



Process

Created by Gene Stroman
from Noun Project

Created by Creative Stall
from Noun Project

Data Analysis

Goal: Keep data close to the compute source

- What resources do you need to do the processing
    - Software
    - Compute resources
- How much can be automated?
- Store **results** where it's easy to collaborate on a manuscript

Types of data produced

Raw

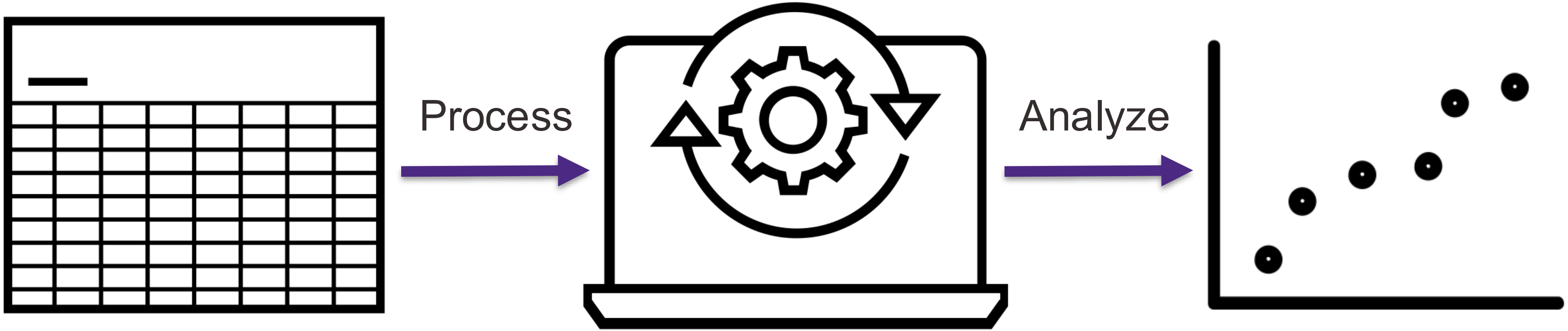
Original, direct from the source

Processed

Cleaned and formatted for analysis

Results

Summaries and visualizations



Process

Analyze

Created by Gene Stroman from Noun Project

Created by Creative Stall from Noun Project

Created by Max Randall from Noun Project

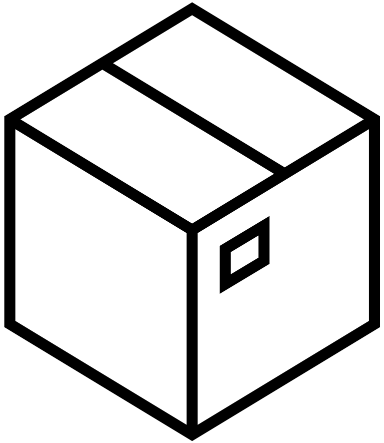
After the Project

Goal: Efficiently preserve your research

- Move your **raw** data to more cost-effective storage

- Think carefully about what **processed data** needs to be kept after results are generated
  - How likely are you to need them again?
  - How hard are they to re-create?
  - Can you automated the process?

- **Results** tend to be small – keep at hand for reference



Taking stock of your data

During your project

- What **raw data** do you produce?
- What **processed datasets** do you produce?
- What **results** do you produce?
- What data classification for each kind?

Homework: Data Inventory

What data do you produce?

- List the data that you (will) produce
- Categorize it as raw, processed, or results

Data	Type
Audio recordings	raw
Transcripts	processed
Anonymized, coded transcripts	processed
Summary tables	results

Key characteristics of your data

Key characteristics

Determine the best place to store your data

- How much data you have
- Who needs access
- What are you doing with the data
- Compliance requirements
- Retention requirements

Answers vary by stage of research project

Size

Where does your data fit?

- Total size of the data (in GB/TB)
- Number of files
- Some storage platforms have size/number limits
- Size affects cost

Access

Who needs access to the data and when?

- People from your research group
- People at Northwestern
- Collaborators external to Northwestern
- The public

Compliance

Know what requirements your data are subject to

- **Northwestern polices**: Use approved storage systems
- **Grant and contract terms**: Controlled Unclassified Information (CUI)
- **Data use agreements (DUAs)**: Specific controls (eg. Encryption) or security standards (NIST 800-171, HIPAA security rule)
- **Federal and state regulations**: HIPAA, BIPA, FERPA

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

Homework: Categorize

For each stage of the research process...

Data	Type	Size	Access	Compliance requirements	Retention period
Audio recordings	raw	GBs	IRB approved	PHI (HIPAA)	7 years
Transcripts	processed	MBs	IRB approved	PHI (HIPAA)	7 years
Anonymized, coded transcripts	processed	MBs	Research team	Northwestern regulations	3 years post project
Summary tables	results	KBs	After: public	pre pub: NU regs After: none	3 years post project

Choosing storage

Types of storage

Storage systems vary by:

- Speed of access
- Access granularity
- Redundancy
- Compliance
- Cost

Northwestern Storage Services

SharePoint

Cloud-based file storage provided by Microsoft

Quest

High performance storage for data processed or analyzed on Quest.

RDSS/FSMResFiles

Mountable storage for research data

Research Data Archival Service
(Coming soon)

Staff mediated archival storage in Amazon S3 Glacier Deep Archive

Storage Access "Tiers"

Warm

- Access instantly
- Not fast enough for all types of analysis
- Redundant: Keep raw data safe

Storage Access "Tiers"

Warm

- Access instantly
- Not fast enough for all types of analysis
- Redundant: Keep raw data safe

Hot

- Fast read-write
- For cleaning and analysis
- Sacrifice redundancy for speed

Storage Access "Tiers"

Warm

- Access instantly
- Not fast enough for all types of analysis
- Redundant: Keep raw data safe

Hot

- Fast read-write
- For cleaning and analysis
- Sacrifice redundancy for speed

Cold

- Less accessible
- Available long-term
- Less costly
- Archive data after a project

Storage by Tier

SharePoint
(Warm)

Access online
Must be synced for analysis

Quest
(HOT)

High speed parallel file system
(GPFS)

RDSS/FSMResFiles

Hot – fast drives for recently accessed files
Warm – slower drives for older files

Research Data Archival Service
(Cold)

Up to 48 hours to access files

Redundancy

Storing multiple copies of files

- Ability to recover from a "disaster" (hardware failure, file corruption, etc)
- Can be built-in to storage systems or DIY
- Store in different physical locations to protect against natural disasters

Storage Redundancy

SharePoint

2 copies in Azure datacenters
Sync to computer

Quest

Projects/Scratch: Single copy
Home: copied to tape

RDSS/FSMResFiles

2 copies: one in Evanston and one in Chicago

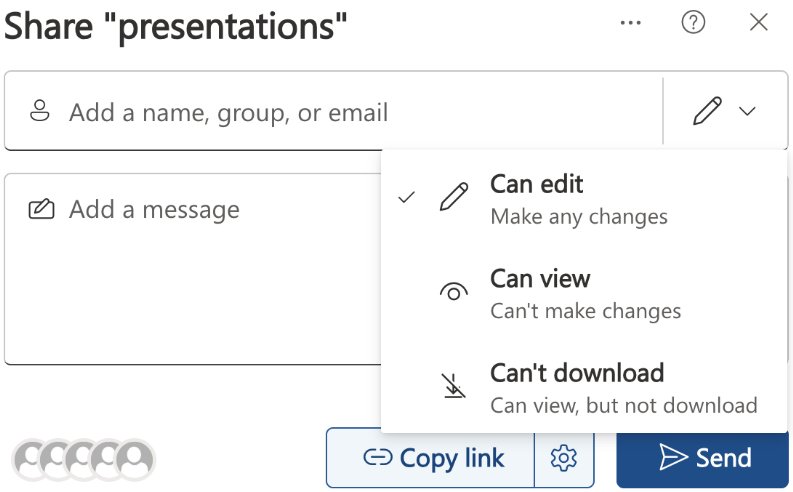
Research Data Archival Service

3+ copies in AWS datacenters in different regions

Access Permissions

Access varies by ...

- What level can you grant access on? (eg: file or folder)
- What type of access can you grant? (read, write, etc)
- What credential do you need to log in? (eg: NetID)
- Can you access data directly or do you need to request access?



Share "presentations"

Add a name, group, or email

Add a message

✓ Can edit
Make any changes

Storage by Access Permissions

SharePoint

By file/folder
Read/write/download
Anyone with a Microsoft account
Self-service

Quest

By file/folder
Read/write/execute
Quest account
Self-service

RDSS/FSMResFiles

All or nothing/By folder
Read or read/write
NetID
Self-service

Research Data Archival Service

Access mediated by RCDS staff

Compliance

Not all storage systems can house all data

- Using University run/approved systems satisfy the requirements for storing research data

- DUAs can require specific features (eg: encryption or audit logging)

- Not all systems can comply with all regulatory frameworks (eg: HIPAA / NIST 800-171)

Storage Compliance

SharePoint

Encrypted, auditing
HIPAA

Quest

Not encrypted, no audit
No PII/PHI

RDSS/FSMResFiles

Encrypted, audit available
HIPAA

Research Data Archival Service

Encrypted
HIPAA

Storage Compliance

SharePoint

Encrypted, auditing

Quest

Not encrypted, no audit

If you need NIST 800-171 compliance,
email researchdata@northwestern.edu

RDSS/FSMResFiles

Encrypted, audit available
HIPAA

Research Data Archival Service

Encrypted
HIPAA

Storage cost

Storage cost is affected by...

- **Redundancy** – how many copies
- **Access speed** – how fast can you access data, read/write
- **University subsidies** – Are you paying the full cost?

Storage by Cost

SharePoint

"No cost" - may change soon

- Warm
- 2 copies
- Fully subsidized

Quest

$195 per TB for five years

- 1 copy
- Hot
- Not subsidized

RDSS/FSMResFiles

$100/TB/year RDSS, no cost FSM

- Warm/Hot
- 2 copies
- ~50%-100% subsidized

Research Data Archival Service

$24/TB/year + retrieval costs

- Cold
- 3 copies
- Not subsidized

Caveats

Everyone's workflow is different

- **Small datasets**: could stay in one place
- **Huge datasets**: Redundancy might be cost prohibitive
- **Highly regulated data**: May only have one option for storage (eg: NIST 800-171)

Exercise: Where to store?

For each step in research process...

Data	Type	Size	Access	Compliance requirements	Retention period	Where to store
Audio recordings	raw	GBs	IRB approved	PHI (HIPAA)	7 years	RDSS (audit)
Transcripts	processed	MBs	IRB approved	PHI (HIPAA)	7 years	RDSS (audit)
Anonymized, coded transcripts	processed	MBs	Research team	Northwestern regulations	3 years post project	SharePoint
Summary tables	results	KBs	After: public	pre pub: NU regs After: none	3 years post project	SharePoint

Take home points

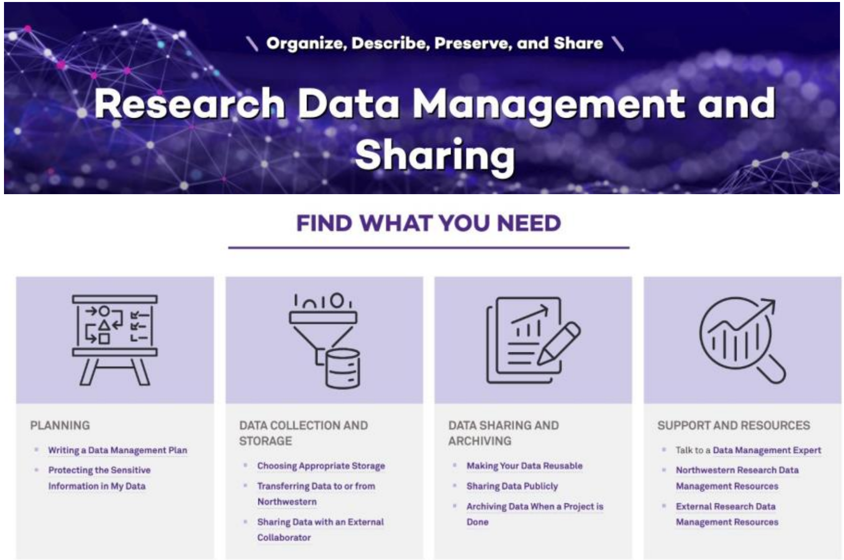
- Everyone's data (and storage needs) are different
- Creating a data inventory can help you identify your needs
- Your needs may vary during different stages of the research process
- Every storage platform has its pros and cons
- Choose options that work with your unique workflow

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:

  Every Monday
  3 p.m. – 4 p.m.
  Mudd Library
  Rooms 2202-2205
  (2nd Floor across from the bridge to Tech).



\ Organize, Describe, Preserve, and Share \

Research Data Management and Sharing

FIND WHAT YOU NEED

PLANNING

- Writing a Data Management Plan
- Protecting the Sensitive Information in My Data

DATA COLLECTION AND STORAGE

- Choosing Appropriate Storage
- Transferring Data to or from Northwestern
- Sharing Data with an External Collaborator

DATA SHARING AND ARCHIVING

- Making Your Data Reusable
- Sharing Data Publicly
- Archiving Data When a Project is Done

SUPPORT AND RESOURCES

- Talk to a Data Management Expert
- Northwestern Research Data Management Resources
- External Research Data Management Resources

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