Data Management 101

Love Data Week

February 14, 2018

Danielle Kane

Data Management & Curation Librarian

What is data?

Data takes many forms including:

- Observational
- Experimental
- Simulation
- Derived or Compiled
- Reference

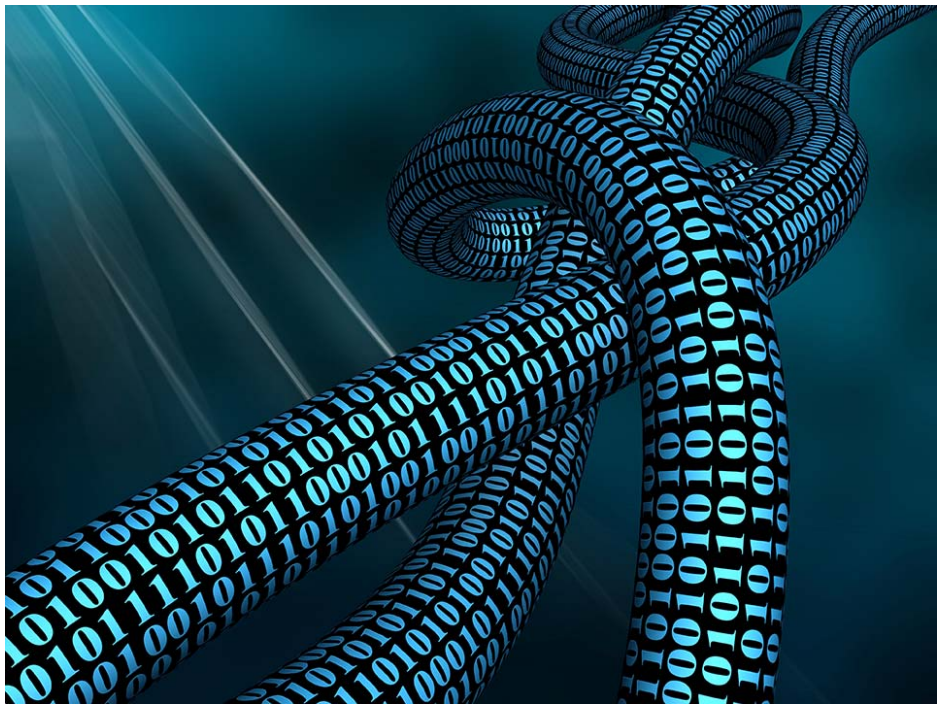


Image: http://community.aras.com/en/promise-of-the-digital-thread/

A data management horror story by Karen Hanson, Alisa Surkis, and Karen Yacobucci. This is what shouldn't happen when a researcher makes a data sharing request! Topics include storage, documentation, and file formats.

What would you do if:

Your hard drive crashes?

Your computer gets stolen?

Your backup fails?

You need to reuse your old data?

Your building burns down?

Your collaborator suddenly quits?

You are accused of fraud?

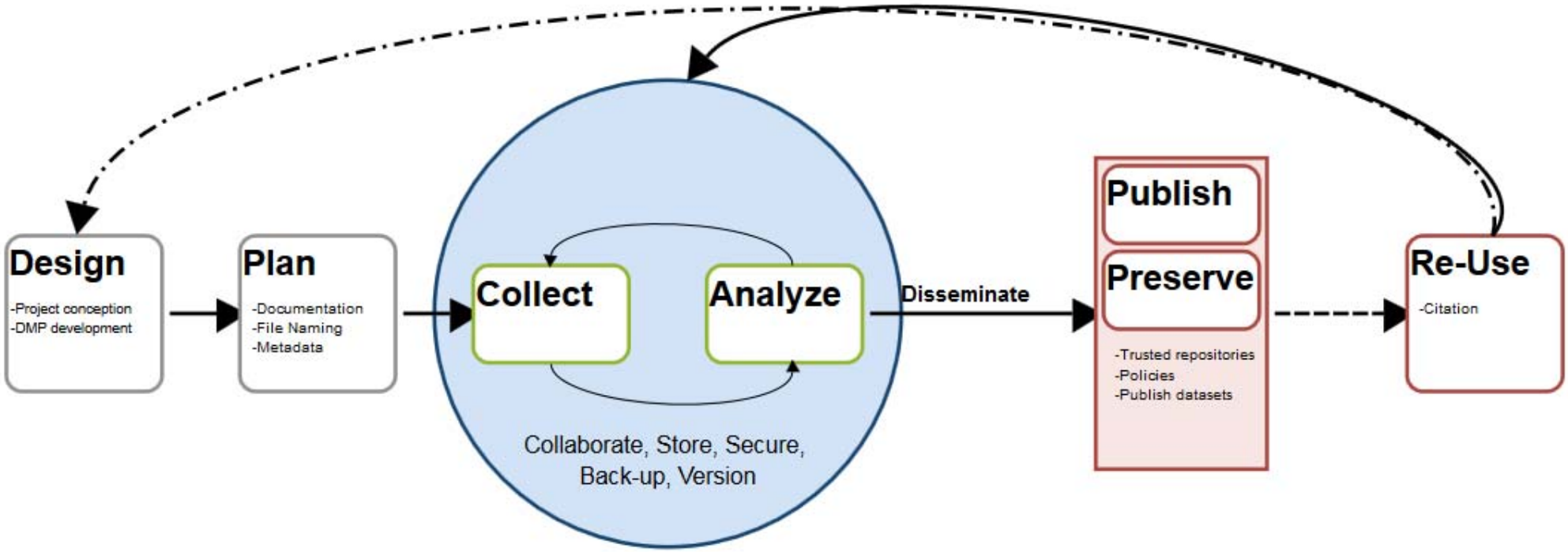


Image from: https://www.library.cmu.edu/datapub/dms/data/101

Why Data Management?

- Find data more easily
- It is easier to analyze organized and documented data
- Get credit for your data
- Avoid accusations of fraud and misconduct
- Don’t loose data!

File Organization & Naming

Why should I organize my files and use file naming conventions?

- It's easier to find files
- Avoid duplicate files
- It's easier to wrap up projects when you know which files belong to it!

File Organization

Any system is better than none!

- Small projects
  - One project, one folder
- Larger projects
  - Separate folders for project stages or for data
  - Separate folder for different types of data
  - Date-based folders

File Naming Conventions

- Name your files consistently
- Keep them short – less than 25 characters
- Use underscores instead of spaces
- Avoid specialized characters
- Use the file dating convention: YYYY-MM-DD

File Versioning: why?

- What if you only had one copy and you made a mistake you can't undo?

- How much work would it take if you had data stored in multiple locations and you needed to find the most recent and complete file?

File Versioning: what to do

- For analyzed data – use version numbers
- Save files often to a new version
- Label versions
- For code, consider GIT or SVN

Exercise: File Naming Conventions

Develop a file naming convention for your most common data type.

Document, document, document!

- No point in having a system if you don't document
  - README.txt
    - Use .txt over .doc because it's more durable
  - Use the front cover of a research notebook
  - Put a printout by the computer
  - Etc.

What should you document?

- In a project-wide README.txt
  - Basic project information
    - Title
    - Contributors
    - Grant info
    - Etc.
  - Contact info for at least one person
  - All locations where data lives, including backups
  - Useful information about the files and how they're organized

Documentation

- Consider the difference between
  - Someone inside your lab
  - Someone outside your lab but in your field
  - Someone outside your field
- Two parts: methods and metadata

Methods

- How the data were gathered
- How the data should be interpreted
- What you did
  - Limitations on what you did
- …build trust in your data

Metadata

- What you're looking at
- Who made it and when
- How it go there
- What it means and
- What you can do with it
- …before you even look at the file

Methods

- Examples of methods to document
  - Code
  - Survey
  - Codebook
  - Data dictionary
  - Anything that lets someone reproduce your results

Metadata

- Informal and formal descriptions of data
- Informal standard can fit your unique research
- Benefits of a formal standard
  - Completeness and aids in sharing
  - Often required for deposit into a repository

Metadata Standard: Dublin Core

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type

Metadata Example

- Contributor
  - Jane Collaborator
- Creator
  - Kristin Briney
- Date
  - 2013 Apr 15
- Description
  - A microscopy image of cancerous breast tissues under 20x zoom. This image is my control, so it has only the standard staining described on 2013Feb2 in my notebook
- Format
  - JPEG
- Identifier
  - IMG00057.jpg
- Relation
  - Same sample as images IMG00056.jpg and IMG00055.jpg
- Subject
  - Breast cancer
- Title
  - Cancerous breast tissue control

Metadata

- Decide on a metadata standard before you collect the data!
  - Easier to record metadata when collecting data than to convert later
- Keep metadata CONSISTENT whenever you can

Metadata

- Tons of formal standards available across many, many disciplines
- Consult
  - Disciplinary repository
  - Your peers
  - Subject librarian
  - Data services librarian

Exercise: Documentation

For your most common data type, make a list of the most important information to record for each dataset

Storage

- Library Motto: Lots of Copies Keeps Stuff Sage!
- Rule of 3: 2 onsite, 1 offsite
- Storage run by experts is more reliable than storage you run yourself
  - It costs more, but that's for a reason

Storage Options

- Computer
- USB/flash drive
- CDs/DVDs
- External hard drive
- Shared drives/servers
- Tape backup
- Cloud storage

Your computer

- You're using it, but should you keep data on it?
- Don't be disorganized
- Don't keep sensitive data here
- **Verdict:** by itself it is not enough

USB/Flash Drive

- Pros:
  - Small, convenient package
  - Big enough for a wide variety of datasets
- Cons:
  - Will you remember to back your data up onto it?
  - Easy to lose and to perpetuate out-of-date copies
- **Verdict:** good for data transport, but not for backup

CD-ROMS/DVD Roms

Pros

- More reliable
- Portable

Cons

- Will you remember to back your data up onto it?
- Hassle to deal with and slow to write to
- Difficult to keep track of old copies

Verdict: Not good for quick backup, and just okay for periodic offsite backup

External Hard Drive

Pros

- Relatively cheap
- Large storage capacity

Cons

- You have to set up, maintain, and audit it yourself
- Some brands are less reliable
- Disorganization a problem

Verdict: Coupled with automatic-backup software, an okay choice for onsite backup

Shared Drives/Servers

Pros

- Keeps data off your easily-stolen laptop
- Not your problem to manage
- Shared costs typically mean lower costs

Cons

- Who's managing the thing? Are they competent?
- Can have storage quotas
- Can be hard to get to outside the lab or the office

Verdict: if well-managed, a good choice for regular use, onsite, or offsite backup

Tape Backup

Pros

- Can happen near-invisibly
- Highly reliable
- Tolerably secure (not always on network)

Cons

- Can be hard or slow to get data back
- Not always audited as often as it should be

Verdict: good for onsite or offsite backup

Cloud Storage

Pros

- Convenient syncing
- Cheap
- If client-side encryption is involved, decently secure

Cons

- Required network connection
- Ongoing (and out of your control) costs
- Your backup is hostage to their business risks
- Reliability, security, privacy not guaranteed

Verdict: for savvy shoppers, fine for offsite backup. A little risky for your only backup.

Exercise: storage

- Conduct a quick inventory of your data
- Inventory where your files are currently stored, including backups.

Backups

- Any backup is better than none
- Automatic backup is better than manual
- Your research is only as safe as your backup plan
  - Insert horror stories here

Ideal Backup Characteristics

- Low effort
- High reliability
- As secure as necessary
  - Tradeoffs between security and convenience
- As open as possible to collaborators
- Well organized

Check Your Backups

- Backups only as good as ability to recover data
- Test your backups periodically
  - Preferably a fixed schedule
  - 1 or 2 times a year may be enough
  - Bigger/more complex data should be checked more often
- Test your backup whenever you change things

A final note

- Must retain data for at least 3 years
  - Better to retain for 5 or 6 years
- Consider letting someone else worry about it
  - A disciplinary repository
  - DASH

Exercise: Backups

Sketch out your ideal backup system, and identify the first step in getting to there from your current system.

What to avoid

- Data hoards
- Data scattered over several machines
- Storage doesn't mean "ownership"
  - If it's communal, it belongs in a communal place
  - If data collection happens on an individual's machine, that doesn't mean the data should stay there!

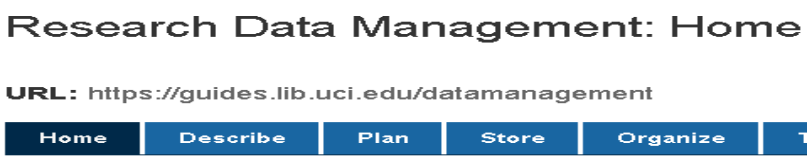
Security

- Does your data fall under the following?
  - HIPPA - Health information
  - FERPA - Student information
  - FISMA - Government subcontractor
  - Human subject research, etc

Where to go from here

- Talk to your coworkers
  - But be aware you might not be able to change things
  - Discuss:
    - Common schemes for metadata and file naming
    - Centralized documentation
    - Robust backup
- Use good practices and be a model for others

Digital Scholarship Services (DSS) Resources



https://guides.lib.uci.edu/datamanagement

dash | UCI

dash.lib.uci.edu

UNIVERSITY OF CALIFORNIA EZID *Identifiers made easy*

https://ezid.cdlib.org/

DMPTool

https://dmptool.org/

How DSS can help!

Provide assistance with:

- Writing grant winning Data Management Plans
- Depositing data into repositories for access and preservation
- Capturing metadata to allow re-use
- Creating permanently resolvable hyperlinks
- Connecting your data with your publications

Need Help?

https://www.lib.uci.edu/dss

libdss@uci.edu

Follow us on Twitter #UCILIBDSS

Schedule DSS Consultation

Security Resources

- Research Computing Support from OIT
- UCI Information Security and Privacy
  - https://security.uci.edu/