Data Management for Graduate Students

André J. Nault, Veterinary Librarian
August 30, 2021
Why plan for data management?

- Most funders require it
- Publishers increasingly require data availability statements
- The U of M makes it your responsibility (2015 Data Policy)
- It’s good for your reputation as a researcher
- Things happen
Hazards of Digital Information

- Deterioration
- Format obsolescence
- Software Version updates (ie. Microsoft)
Data Management occurs throughout a project

Good data management involves practices that:

- ensure integrity of the files/data
- protect the security of data
- enhance efficiency & reliability of the research
- facilitate reproducibility
- plan for long-term access and reuse
Where do you store and backup your data?
## U of M Data Storage Options

<table>
<thead>
<tr>
<th>LOCAL (LAB OR DEPT)</th>
<th>CENTRAL (OIT)</th>
<th>CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work computer</td>
<td>University Server aka. block or active directory</td>
<td>UMN Google Drive UMN Box Secure</td>
</tr>
<tr>
<td>External hard drive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dept/local server</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Security Problems?

<table>
<thead>
<tr>
<th>LOCAL (LAB OR DEPT)</th>
<th>CENTRAL (OIT)</th>
<th>CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal computer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flash drive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CENTRAL (OIT)</th>
<th>CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMN Google Drive UMN Box Secure</td>
<td>Amazon Glacier DropBox</td>
</tr>
</tbody>
</table>
UMN Box Secure Storage

- It is a secure environment delivered by the Center of Excellence for HIPAA Data intended for storing, sharing and accessing sensitive and private-highly restricted files.

- Allows secure sharing of files, and the creation, editing and reviewing of documents with others in real-time.

- Includes encryption, activity logging, Duo Two-Factor Authentication, and access controls such as view-only access. Our implementation of Box Secure Storage is specifically designed to be compliant with HIPAA and security standards for private-highly restricted data.

- Unlimited storage capacity; individual file size limit is set to 15 GB.

https://it.umn.edu/technology/box-secure-storage
3-2-1 Rule for Backup

3) Three copies of your work, of which 1 is a working copy, 2 are backups
2) These are placed on 2 different kinds of storage
1) And at least 1 copy is offsite
Possible Data Backup Workflow

- Work computer
- External hard drive
- Flash drive
- Instrument
- Others?

- University Server (Active directory)
- External hard drive
- Others?

- Google Drive
- Dropbox
- AD backup
Backing up your data

Amazon’s Choice

Highly rated, well-priced products available to ship immediately

Under $50

Seagate Portable 2TB External Hard Drive Portable HDD – USB 3.0 for...

4.5 stars out of 5, 2,311 reviews

Limited time deal

$49.99 $69.99

Between $50 and $100

Toshiba Canvio Advance 4TB Portable External Hard Drive USB...

4.5 stars out of 5, 4,748 reviews

$94.99 $99.99

Above $100

Seagate Expansion Desktop 8TB External Hard Drive HDD – USB 3....

4.5 stars out of 5, 4,236 reviews

$139.99
Do NOT back up a file by copying over another one

PLEASE

DON'T DO IT
Data Corruption

Data corruption or loss can be caused by

- human error
- hardware failure
- malicious intent
- silent data corruption due to software bugs
How do you describe and document the data you collect/manage?
Directory and Folder Structure

A) Organized by File type

Example.A
- Code
  - Step.1
  - Step.2
- Data
  - Processed
  - Raw
- Results
  - Figure.1
  - Figure.2
  - Models
- readme.txt

B) Organized by Analysis

Example.B
- Figure.1
  - Code
  - Data
  - Results
- Figure.2
  - Code
  - Data
  - Results
  - readme.txt

- Use consistent directory names in each person/group’s folder:
  - By FILE type: data, code, text, images, models, etc.
  - By ANALYSIS stage: raw data, processed data, public data
- Keep nested folders to a min (shallow)
- Keep data and documentation files together
### A Story Told in File Names:

**Location:** C:/user/research/data

<table>
<thead>
<tr>
<th>Filename</th>
<th>Date Modified</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>data_2010.05.28_test.dat</td>
<td>3:37 PM 5/28/2010</td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_re-test.dat</td>
<td>4:29 PM 5/28/2010</td>
<td>421 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_re-re-test.dat</td>
<td>5:43 PM 5/28/2010</td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_calibrate.dat</td>
<td>7:17 PM 5/28/2010</td>
<td>1,256 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_huh??.dat</td>
<td>7:20 PM 5/28/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_WTF.dat</td>
<td>9:58 PM 5/28/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_aaarrrgh.dat</td>
<td>12:37 AM 5/29/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_$_%@$&amp;!!.dat</td>
<td>2:40 AM 5/29/2010</td>
<td>0 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_crab.dat</td>
<td>3:22 AM 5/29/2010</td>
<td>437 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_notbad.dat</td>
<td>4:16 AM 5/29/2010</td>
<td>670 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_wooohoo.dat</td>
<td>4:47 AM 5/29/2010</td>
<td>1,349 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_USETHISONE.dat</td>
<td>5:08 AM 5/29/2010</td>
<td>2,894 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>analysis_graphs.xls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ThesisOutline.doc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes_Meeting_with_ProfSmith.txt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JUNK...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data_2010.05.30_startingover.dat</td>
<td></td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
</tbody>
</table>
Be intentional about your file naming

- **Be Descriptive**: interview.txt is not helpful. Instead:
  20150814_interview_site01_respondent04.txt (no more than 255 characters)

- **Don’t embed information in folder structure**:
  2015/august/minneapolis/interviews/reactionmemo.txt

- **Use consistent naming schema**: Create a useful order (for sorting) and decide on shared terminology

- **List versions alphanumerically**, eg. v01, v02, v03 rather than last, final, finalfinal, useTHISone
Data Documentation - Keeping Records

There is more to manage than just the data...

- Traditional lab notebooks
- Electronic lab notebooks (ELNS)
- Documentation Files (readme.txt)
- Tools that facilitate reproducible research
Electronic Lab Notebooks

Evernote

- Full text searchable
- Good for handwritten notes

Open Science Framework: https://osf.io/

- Group collaboration tool
- Overlay to other storage locations
- Wiki tool to capture data narrative

See ELN Comparison Chart
https://libguides.umn.edu/c.php?g=1164012&p=8497442
Readme.txt Files

- General overview of dataset
- Descriptions of data collection methods and conditions
- Characterization of equipment and its response
- Recording any changes made to raw data → analysis, codes, abbreviations, cleaning protocols

Download readme template on library website:
http://lib.umn.edu/datamanagement
http://z.umn.edu/readme
U of M researchers can submit their data for:

- Free
- Open, public access
- Curation services
- Permanent DOIs
- File download analytics
- Long term preservation

Get started at http://z.umn.edu/drum
Preservation File Formats for Long-term Access

<table>
<thead>
<tr>
<th>Type</th>
<th>Format 1</th>
<th>Format 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>MS Word</td>
<td>PDF, TXT, HTML</td>
</tr>
<tr>
<td>Images</td>
<td>Photoshop</td>
<td>TIFF</td>
</tr>
<tr>
<td>Video/Media</td>
<td>Quicktime</td>
<td>MPEG4</td>
</tr>
<tr>
<td>Database</td>
<td>MS Access</td>
<td>DBF</td>
</tr>
<tr>
<td>Tabular Data</td>
<td>MS Excel</td>
<td>CSV</td>
</tr>
<tr>
<td>Presentations</td>
<td>MS Powerpoint</td>
<td>PDF</td>
</tr>
<tr>
<td>Sound/Music</td>
<td>Windows Media</td>
<td>WAV</td>
</tr>
</tbody>
</table>

Note: be conscious about the risks of compressing your files or migrating to a file format that has different affordances than the original