Research Data Management: Storage Strategies
Hilda Teresa Ayala González

March 26, 2019
Graduate Research and Innovation Center (GRIC)
University of Puerto Rico, Mayagüez Campus

This activity is sponsored by the Transformational Initiative for Graduate Education and Research (TIGER), US Department of Education, Title V, Part B, Promoting Postbaccalaureate Opportunities for Hispanic Americans (PPOHA) Program (#P031M140035).
Goals

• Understand the different storage systems, benefits, and risks

• Discuss strategies to create and maintain readable files thorough a project

• Recognize the importance of redundancy in data management
Pre-test

Link: https://uprm.libsurveys.com/rdmstorage

Scan QR Code
Why would YOU keep data and research documentation?

- So others can use it
- Validate your research findings
- Get a degree
- Further or new research
- Comply with funding agencies
A former graduate student fabricated and falsified the results of his NSF-supported research that appeared in two papers, his dissertation, a draft manuscript, a patent, and proposals submitted to another agency. The data was first called into question shortly before his dissertation defense. Despite not being able to produce the original data, the school allowed him to graduate without resolving the matter. Subsequently, an anonymous allegation to a journal resulted in the retraction of one of the papers. An anonymous allegation to another journal led the co-authors to repeat some of the experiments and publish a correction. The university then initiated an investigation into 28 allegations of fabrication and falsification.

During the investigation, the student was once again unable to provide original data supporting his research, nor was he able to identify the original data from the research materials archived in due course by the university. Ultimately, the investigation committee determined that 26 of the 28 allegations of research misconduct were substantiated. The university found the student committed research misconduct and revoked his Ph.D.

We concurred with the university’s findings and recommended NSF debar the student for 1 year. We further recommended that for 4 years, NSF require certifications and assurances; require submission of a detailed data management plan with annual certifications of adherence for any new awards; and bar the student from participating as an NSF peer reviewer, advisor, or consultant.
Data retention

For how long do you keep your data?

**OMB Circular A-110** states that the retention period is three years from the date the final financial report is submitted.

**Exceptions**

1. If any litigation, claim, or audit is started before the expiration of the 3-year period, the records shall be retained until all litigation, claims or audit findings involving the records have been resolved and final action taken.

2. Records for real property and equipment acquired with Federal funds shall be retained for 3 years after final disposition.

3. When records are transferred to or maintained by the Federal awarding agency, the 3-year retention requirement is not applicable to the recipient.

From "CIRCULAR A-110 REVISED 11/19/93 As Further Amended 9/30/99", 2013, by Office of management and Budget.
Data retention

For how long do you keep your data?

Research misconduct involving NIH funding - **six years after the final resolution date of the case**

Patented inventions *for the life of the patent* in case the patent is challenged or if lawsuits should arise.

Sponsored research, *during the project the researcher must provide access to the data and at the end the institution has the responsibility to provide adequate storage capacity.*

Should a researcher leave the institution, ... *If the researcher takes the original data, a copy must be left at the institution.* In addition, the researcher must agree to retain the original data for the required retention period and to provide access to the original data to the institution as well as other individuals or entities having a legitimate need for access.

From "Retention of data", 2013, by S. Erickson, & K. M.T. Muskavitch
Data at risk: Issues

- Policies: Institutions
- Mandates: Funders, Publishers
- Access control
- Security concerns: protection sensitive data
- Multiple storage
- Multiple copies
- Diversity of formats and media
- Retention and destruction
Data at risk: Issues

- Change in team members
- Theft
- Loss/Erasure
- Natural disaster
- Human accidents
- Hardware failure
- Computer virus
- Corruption of files
- Obsolescence (media, software or format)

(Briney, K., 2015)
Team Digital Preservation to the rescue
Strategies

- Up to date
- Secure

Good hardware + Readable files + Redundant storage

- Not corrupted
- Not obsolete

Multiple copies + Available
Strategies: Hardware

- Floppies: Avoid
- Cassette tape: Avoid
- Old PCs: Avoid

Avoid obsolete media and equipment

For more examples visit MIT’s *The Chamber of Horrors*
Strategies: Hardware

- **CD/DVD**: Use with cautions
- **Cloud Service**: Use with cautions
- **Flash Drives**: Not recommended

- Obsolescence
- Too many
- Third Party
- Terms
- Loss
- Corruption

(Briney, K., 2015)
Strategies: Hardware

- **Personal**
  - Short-term Primary

- **External Drive**
  - Short-term Secondary

- **Shared Drive**
  - Short-term Secondary

- **Thieves**
- **Loss**
- **Degrades**
- **5 years**
- **Maintenance**
- **Personnel**

(Briney, K., 2015)
Strategies: Hardware

• Look for information about
  • Media stability
  • Obsolescence

• Select hardware
  • Modern
  • Technical support

• Renew hardware
  • Revise at least every two years
    • Still works
    • Data inside is also working
  • Migrate content

https://obsoletemedia.org/media-preservation/
Strategies: Hardware

<table>
<thead>
<tr>
<th>Media</th>
<th>Estimated Lifespan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnetic data (tapes)</td>
<td>Up to 10 years</td>
</tr>
<tr>
<td>Nintendo cartridge</td>
<td>10-20 years</td>
</tr>
<tr>
<td>Floppy disk</td>
<td>10-20 years</td>
</tr>
<tr>
<td>CDs and DVDs</td>
<td>5-10 unrecorded, 2-5 recorded</td>
</tr>
<tr>
<td>Blu-Ray</td>
<td>Not certain, probably over 2-5 recorded</td>
</tr>
<tr>
<td>M-Disc</td>
<td>1,000 years (theoretically)</td>
</tr>
<tr>
<td>Hard disk</td>
<td>3-5 years</td>
</tr>
<tr>
<td>Flash storage</td>
<td>5-10 years or more (depends on write cycles)</td>
</tr>
</tbody>
</table>

**Improving life expectancy: Take care of hardware**

- Don’t expose to heat
- Careful with physical force
- DO: Keep in cool and free of dust spaces. Carry only when necessary.

Adapted from "Data storage lifespans: How long will media really last?" by Casey Morgan.
Strategies: Hardware

Third parties: Cloud storage

• **It means**
  • Someone else can access your data
  • Bankrupt, disappear, sold

• **Cost varies**
  • Storage
  • Data access/transfer

• **Read their terms and services**
  • Data encryption (transfer or storage)
  • Servers location
  • Data ownership
  • Data deletion

• **Don’t have it ONLY up there!**

---

**MySpace admits losing 12 years’ worth of music uploads**

By Zoe Kleinman
Technology reporter, BBC News

16 March 2019

Katie Nash started out by uploading her music on MySpace

MySpace, one of the first online social networks, has apologised after a server migration caused a huge loss of data.

From BBC, 2019
Strategies: Hardware

Third parties: Cloud storage

- National Digital Stewardship Alliance Levels of Preservation

- AVPreserve Cloud Storage Vendor Profiles
  - Amazon Glacier, Amazon, Chronopolis, Dternity (formerly Permivault), DuraCloud, Oracle Archive, Oracle Standard, Preservica, Rackspace Cloud Files

From "Cloud Storage for Preservation NDSA Levels of Digital Preservation Vendor Comparison Grid", 2014, by AVP
Strategies: Hardware security

• Control access to data
  • Access privileges
    • Folders/Files
  • Limit access to raw data folder
  • Don’t provide access to all folders in shared drive, only essential
  • Keep your stuff secure & safe
    • Use strong passwords
    • Use different passwords
Strategies: Hardware security

• Control access to data
  • Don’t leave computer unattended
  • Don’t share external drives
  • Avoid leaving you account open
  • Encrypt important files
    • Remember the password
    • Unencrypt at end
Strategies: Hardware selection

- Estimate of data to be generated
- Cost of storage system
- Availability of storage system

(Adapted from "File size: BITS, BYTES, KB, MB, GB, TB" by Techpikk.)

(Briney, K., 2015)
Strategies: Readable files

A file format is a standard way that information is encoded for storage in a computer file.

• **It tells the computer how** to display, print, and process, and save the information.
• **It is dictated by the application program** which created the file, and the operating system under which it was created and stored.
• Some file formats are designed **for very particular types** of data, others can act as a **container for different types**.
• **A particular file format is often indicated by a file name extension** containing three or four letters that identify the format
Strategies: Readable files

When you open an electronic file you expect:

• An **application program** will render its content
• Content is what is **suppose to be**
• All the content is there
  • Text
  • Images
  • Video
  • Audio
Strategies: Readable files issues

What happens when you get these instead?
Strategies: Readable files issues

There is not an application program to open it.

Encoding is not recognizable by an application program.

What happens when you get these instead?

Image got corrupted because it was not saved correctly.
Strategies: Readable files issues

- Numerous files through the life of a project
  - Also in different storage media
- Multiple file formats
  - Proprietary/open
- Software dependable
  - Constant new versions/updates
  - Obsolescence
  - Loss of content
  - Change in features or style

We can’t always keep track...
Strategies: Readable files issues

When data is at risk?

• Sitting on a computer
  • Can be erased
  • Attacked by a malware or virus

• Transfer
  • Can be hacked or corrupted during the communication from one media to another or between clients

• Format change (.doc to .pdf)
  • Can lose content, style, embedded fonts or images, etc.
Strategies: Readable files issues

...many things can go wrong

- Failure of the file on storage device due to
  - Closing the file incorrectly
  - Not enough time to save / download
  - Obsolescence

- Storage media may have degraded
  - Making some parts not readable

- Some of the bits in a bitstream might become lost or damaged
  - Bit corruption
Strategies: Readable files

- **Identify**
  - Type
  - Longevity

- **Control**
  - Select
  - Migrate

- **Confirm**
  - Content
  - Integrity
Strategies: Identify

• Identify file formats you produce or depend on
  • Help you gain an understanding of the application program or software need it to render it
    • Do you keep the software too?
    • Is there another format that can be use?

• Read about the life expectancy of the format
  • Supported
  • Versions
  • Outdated
  • Superseded

You can identify the format
- Looking at the extension file
- Application program to open
Strategies: Identify

Identify formats

- Automated batch identification of file formats
- Life-expectancy/support

**DROID / PRONOM**
- Open source
- Software
- Runs on Java or Command line

**NOTE:** Info about formats needs to be looked manually
Strategies: Control

• Select the file formats
  • Most supported
  • Creator
  • Various applications or software
  • Up to date
  • Use in field

• Migrate content
  • Keep a version of the original
  • Revise content for accuracy

NOTE: Migration might change content in the file or you can lose essential information
Strategies: Control

When choosing a new file format:

• Open
• Well Documented
• Non-proprietary
• Non-compression
• Metadata friendly
• Platform-independence
• Widely accepted by research community
Strategies: Confirm

- Corroborate CONSTANTLY that the content
  - Is there
  - Is available
  - Can be open

- Ensure the integrity of the file is **not compromised**
  - Is accurate
  - Nothing has change
  - The names are the same
  - Files are not missing

**Fixity check** a method for ensuring the integrity of a file and verifying it has not been altered or corrupted.

It is most often accomplished by computing checksums such as MD5, SHA1 or SHA256 for a file and comparing them to a stored value.

A **checksum** is a unique numerical signature derived from a file.

Digital Preservation Coalition, 2015
Strategies: Confirm

**When to use checksums**

- Corroborate that objects in storage have not change through time
- Confirm that a file that was transferred did not change

### Most commons checksums

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Level of effort</th>
<th>Level of detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD5</td>
<td>Cryptographic hash function</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>SHA1</td>
<td>Cryptographic hash function</td>
<td>Moderate</td>
<td>Very high</td>
</tr>
<tr>
<td>SHA256</td>
<td>More secure cryptographic hash function</td>
<td>High</td>
<td>Very high</td>
</tr>
</tbody>
</table>

Checksums are compared from "Checking Your Digital Content: How, What and When to Check Fixity?" by NDSA, 2014

---

**How does it work**

1. You create a file
2. Generates checksum
3. Checksum is a fingerprint
4. Generates checksum
5. Checksums are compared
Strategies: Confirm

Identify changes
- Integrity verification
- File monitoring

AVP Fixity
- Open source
- Application
- MAC / Windows

NOTE: Still requires to manually check the errors
Strategies

Definition of redundant

1. **a**: exceeding what is necessary or normal: **SUPERFLUOUS**
   
2. **b**: characterized by or containing an **excess**
   
3. **specifically**: using more words than necessary

4. **c**: characterized by similarity or repetition
   
5. **//**: a group of particularly **redundant** brick buildings

6. **d**: **chiefly British**: no longer needed for a job and hence laid off

2. **PROFUSE, LAVISH**

3. : serving as a duplicate for preventing failure of an entire system (such as a spacecraft) upon failure of a single **component**
Strategies

*Redundancy = Lots of copies keep stuff safe*

- Avoid loss
- Restore access to corrupted/destroyed data and documentation

**Backup strategy**

3 Copies 2 Mediums 1 Offsite

(Briney, K., 2015)
Strategies: Backup

Select best backup method

• Manual or automatic
  • Full
  • Incremental
  • Differential

• Cloud provider
  • Backup, NOT syncing
  • Easy to access?

• Frequency = daily, weekly, monthly
  • Easy to replace?

Storage ≠ Backup

Make sure it works

• Test the backups!
• Keep at least two copies
• Generate a checksum
• Ensure they are available

(Briney, K., 2015)
Strategies: Long-term

Did your project was funded by external funds?

*Do you want to publish your findings or create a patent?*

Want to keep your data after the project completion?

*Would you like others to use your data?*

Consider a Digital Repository and join us for the next workshop!
Exercise

Get content for exercise

- Read “RDM_IntruccionesFixity”
- Download the RDM_FileNaming & ADDFiles folder for exercise

• Practice
  - Run a report
  - Make some changes to folder
  - Run report and compare

• Erase project at the end of the exercise

15 minutes
Questions and comments
Post-test

Link: https://uprm.libsurveys.com/RDMStoragePost

Scan QR Code
Evaluation: http://uprm.libsurveys.com/tiger
Password: 2018

Title - RDM: Storage Strategies

Hilda Teresa Ayala González
March 26, 2019
References


Find Us

http://libguides.uprm.edu/gric

gric@uprm.edu

@GRICUPRm

@GricUprm

GRIC UPRM