Storage

Preservation and Curation of ETD Research Data and Complex Digital Objects

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE

Photo by Annie Spratt on Unsplash

Storage

As you work toward your degree, you store your files on your laptop. You appreciate its portability as you move around from your apartment to various sites on campus, and you back it up occasionally to an external drive. This arrangement works fine until, one day, your laptop is stolen. You go to your external drive and find that your last backup is from four months—and two dissertation chapters—earlier. ■

Rationale and Motivations – Why

Where and how you choose to store your research materials and writings will determine how long they survive. The risks to your digital files are many and varied, from file corruption to viruses, environmental disasters to theft, and from storage device malfunctions to accidental or malicious deletions.

By mindfully planning, choosing, and documenting where you store your digital materials, you can make sure that you and others will have access to them in the future.

Photo by Max Lakutin on Unsplash

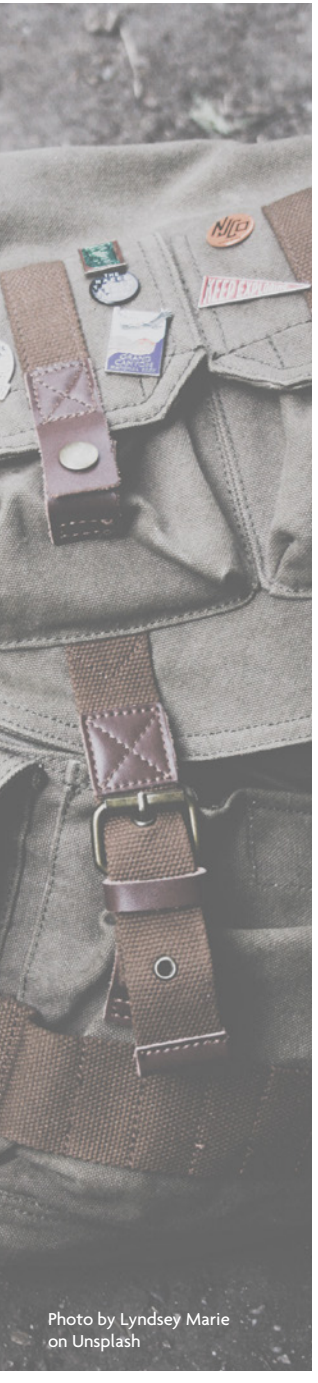

Photo by Lyndsey Marie on Unsplash

The Basics – How to Do It

Taking simple steps now to create a systematic storage process, whereby you make multiple copies and store them in geographically dispersed locations, will secure your content against a wide range of mishaps and disaster scenarios.

Below, we consider some common storage environments you may currently use or be familiar with. We then describe one easy scenario for creating and maintaining copies of your work in multiple storage locations over time.

Storage Options

Your laptop or desktop: This may be your primary work location, where you save documents and files as you work on them, either on a computer's hard drive or internal flash storage (either on a spinning disk or a solid state drive, SSD).

External hard drives (spinning or SSD): You may use an external drive that permits creating backup files of the documents you need to store. Ideally, this is a dedicated drive that is not used for any other purpose and that stays in one location.

Flash drives and other solid state media: You may use a thumb drive, flash drive, or memory card to store a variety of your research outputs, including multimedia files and datasets. These are inexpensive, easily obtainable, and portable.

The cloud: You may store your research files and outputs in "the cloud." This means that they are stored on a server that is managed by a third party that you can access via an Internet connection. "Cloud" environments include a range of options, including a solution that your own academic institution may offer or a service offered by a for-profit company (e.g., Google Drive, Amazon S3, Flickr, YouTube, Dropbox, Apple iCloud).

Backup Methods and Techniques

Creating a "backup" or additional copy of your content as soon as you begin creating it should be a regular part of your research process. Establishing a formalized routine through which you create new backups on a regular schedule (e.g., daily or weekly) ensures you capture regular snapshots of your work as you progress.

Backup practices might include:

- Copying your files to a departmental or university-based storage network.
- Copying your files to an external hard drive or solid state media, like a thumb drive.
- Copying your files to a third-party storage solution, like Google, iCloud or others.

This "backup" process does not need to be arduous or complicated; simply connecting to your backup site regularly and copying your current content over into a dated folder will suffice. Some external drives and cloud services will guide you through this process via software that automatically dates each snapshot it takes of your content. If you are not using one of these software packages (e.g., "Time Capsule" for Mac or PC), you will need to establish a separate, dated folder for each backup you make. For example, if you back your files up weekly, you would create folders corresponding to each week:

- 20160411
- 20160418
- 20160425
- ...

As you create new backups, make sure you do not delete all of your old backups. At times, you may need to return to an older backup in order to fully recover your content, especially in cases where you have accidentally deleted content or overwritten a file without realizing it, or in cases where a virus infects your files. A good rule of thumb is to maintain at least bi-monthly backups of your files; your safest bet is to maintain all of the backups you make over time.

Security Practices

Another key component of storage is the ongoing protection of your data and its integrity. This involves limiting access to your data, not only via establishing and maintaining passwords and protections for your own devices, but also for specific documents, files or folders. It also includes keeping up-to-date antivirus software protection on your computer. In the case of sensitive research content, you might also use

encryption or watermarking as a means of ensuring no one can tamper with your original work.

In order to secure your content, we recommended the following pathway:

- Maintain at least one local (i.e., non-cloud-based) copy of your content.
- Maintain at least three separate complete copies of your research content.
- Maintain at least one copy in a different geographic location.
- Maintain a history of changes in at least one location (e.g., using a "Time Capsule" software package to automatically back up your content without deleting older copies).
- Document your basic, regular practices (e.g., produce a text file that specifies how, when, and where you store and back up your materials).

Long-Term Preservation

Digital preservation is a term that represents a more complex set of activities that goes well beyond storage. Digital preservation is defined as the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (Digital Preservation Coalition). An operative term in this definition is "managed activities." Individuals seeking to preserve digital materials must understand that preservation requires planning, care, and coordination over time.

Picture, for example, your perfectly stored/preserved Microsoft Word file created using Microsoft Word for Mac, version 15.26. In twenty years, will you be able to read it, even if the bits and bytes are perfectly stored and intact? Will you know what they are and how to render them into current operating systems and software environments?

Digital preservation is an ongoing process, and one that changes as our technical environments change. Elements of digital preservation are worth considering as part of your ongoing storage routine; others may best be accomplished by the repository into which

you ultimately submit your research content (e.g., your university library's ETD repository or a data repository such as ICPSR). Some possible elements you might incorporate into your routine include the following:

- Produce and maintain a spreadsheet-based inventory of all of your content, documenting file names, file sizes, file locations, and file types.
- Systematize your folder- and file-naming conventions, ensuring that your file names provide human-identifiable information, such as your editing date and a title code (e.g., if you are working on a photograph on March 4, 2017, and the photograph is of the Boston Globe newsroom, you might name the file 20170304_BostonGlobeNewsroom.jpeg).
- Make sure your file names are followed by the correct file extension (e.g., .txt, .pdf, .tiff, .jpeg, .xls, etc.).
- Avoid using special characters in all file and folder names as these may keep your file or folder from opening in some contexts (e.g., \?:*?‹›{}[]&$,;.!).
- Document the formats you are managing and the potential sustainability issues associated with these formats (e.g., Flash recently was the main supported format for video; today, it is nearly obsolete).
- Save a copy of your research files in non-proprietary formats so that you do not have to have a software license to render and use the files. For example, if you are using SPSS (an expensive software program that you may not have access to later in your career), save a copy of your dataset in an open format, such as tab-delimited or comma-separated values (e.g., .csv).
- Consider creating and regularly checking "checksums" or digital signatures for your most important research files. Checksums can be generated by several open source tools and utilities and they can be stored in your inventory. Fixity is a current tool that is both simple to use and freely available (see "Resources" below).

Photo by Annie Spratt on Unsplash

Tools – What to Use

Some hardware and software options for the creation and storage of backup copies include:

- Hardware: external hard drives, flash drives and other solid state media (thumb drive or memory card); a rewritable disc; or an online (server-based) storage environment (university or cloud-based, including Google Drive, Amazon Cloud, Flickr, YouTube, Dropbox, iCloud, Rackspace, CrashPlan, BackBlaze, etc.).
- Software: There are a number of well-regarded software solutions that can help you to manage the backup process, including Time Capsule, Genie Timeline Home, StorageCraft ShadowProtect, Acronis True Image, iDrive, and SOS Online Backup.

Some ways to produce and manage checksums include:

- Hashdeep is a lightweight open source command line application that provides technicians with features and commands for creating and comparing checksums for digital objects at both file and batch levels. It includes a reporting function that explains the reason for a comparison test's failure.
- Fixity is a lightweight software program to automate checksum monitoring. It allows users to select a regular interval at which the tool will generate checksums and compare values for a set of files, and then have a report sent to them upon completion.

Advice for creating and updating an inventory include:

- The Library of Congress has a basic example of a personal digital archiving inventory. This sample inventory gives a clear example of how to structure a basic inventory that identifies file types, locations, and types of storage media related to a collection or set of collections.

Resources to help you identify obsolete or near-obsolete formats include:

- The PRONOM registry provides an operational public file format registry that can help you to understand what formats are endangered and what formats are well supported. For more on PRONOM, see http://www.nationalarchives.gov.uk/PRONOM/Default.aspx and https://en.wikipedia.org/wiki/PRONOM.

Migration tools that can help you migrate near-obsolete formats include:

- Adobe Systems' "Image Processor" requires a paid license, but it is widely used for bulk image format migrations.
- Most software programs give you multiple options to "save as," and some allow bulk format migrations as well. Use these internal tools while you have access to the software.

Photo by Frank McKenna on Unsplash

Photo by Lize Luero on Unsplash

Resources

All links provided were last checked 10/3/2017, and the content we reference here was saved on that date in the Internet Archive's Wayback Machine. For links that no longer work, please visit the Wayback Machine and enter the url to surface the resource.

- For basic advice on backing up your content, please see Jesus Vigo, "World Backup Day: Best practices to back up your data," Tech Republic, last updated March 31, 2015.
- To understand more about the use of cloud environments for backups, please see Charles Beagrie Ltd., "How Cloud Storage can address the need of public archives in the UK," The National Archives, last updated April 2014.
- For general information on how to archive and back up your content, see the growing set of resources associated with the Personal Digital Archiving movement.

Activities

1. Take one project you are working on now, and develop a spreadsheet-based inventory for the associated files indicating file names, sizes, types, and storage locations. (Use http://blogs.loc.gov/thesignal/files/2016/05/pda_inventory.pdf as a guide.)

2. Establish a regular routine for backing up your content in at least one additional location. Make sure the routine includes a regular schedule, a way of storing content organized by the date of a backup, and a way to maintain multiple backups simultaneously.

Photo by Gregoire Jeanneau on Unsplash


INSTITUTE of
Museum and Library
SERVICES


EDUCOPIA
INSTITUTE