Storage & Backup

2021-01-20

Introduction

1. **What is the difference between a backup and an archive?**
    - Original Data
    - Operation Data
    - Backup
    - Archive
2. **Why backup?**
    - hardware failure
    - software or media faults
    - virus infection or malicious hacking
    - power failure
    - human error
    - verification/data requests/FOIA/Lawsuits
    - Funder Mandates

The Back-up Rule: 3-2-1

3-2-1: A secure backup requires a *minimum* of **3 copies** of your data on **2 types of storage media** with **one copy off-site**. Having 1 copy off-site protects your data from local risks like theft, lab fires, flooding, or natural disasters. Using 2 storage media improves the likelihood that at least one version will be readable in the future should one media type become obsolete or degrade unexpectedly. Having 3 copies helps ensure that your data will exist somewhere without being overly redundant. ***Be sure to migrate data from physical storage every 3-5 years***

Storage Media

Successful preservation depends in great part on storage media that are in good physical and operational condition. There are many different kinds of storage media that can be used to save and back-up digital files; each has pros and cons. They include:

- Desktop and Laptop Computers
- Campus servers
- Commercial Cloud Backup (e.g., Dropbox, Google Drive, Amazon, Carbonite)
- External hard drives (range from disks to USB drives)
- Optical Storage (CD, DVD)
- Paper (e.g., printouts of spreadsheets and text files)

All storage media, whether hard drives, discs or data tapes, will wear out over time, rendering your data files inaccessible. To ensure ongoing access to both your active data files and your data archives, it is important to continually monitor the condition of your storage media and track its age. Older storage media and media that show signs of wear should be replaced immediately. **Use the following guidelines to ensure the ongoing integrity and accessibility of your data:**

1. **Test Your Storage Media Regularly:** It is important to routinely perform test retrievals or restorations of data you are storing for extended periods on hard drives, discs or tapes. It is recommended that storage media that is used infrequently be tested at least once a year to ensure the data is accessible.
2. **Beware of Early Hardware Failures:** A certain percentage of storage media will fail early due to manufacturing defects. In particular, hard drives, thumb drives and data tapes that have electronic or moving parts can be susceptible to early failure. When putting a new drive or tape into service, it is advisable to maintain a redundant copy of your data for 30 days until the new device "settles in."
3. **Determine the Life of Your Hard Drives:** When purchasing a new drive unit, note the Mean Time Between Failure (MTBF) of the device, which should be listed on its specifications sheet (device specifications are usually packaged with the unit, or available online). The MTBF is expressed in the number of hours on average that a device can be used before it is expected to fail. Use the MTBF to calculate how long the device can be used before it needs to be replaced, and note that date on your calendar (For example, if the MTBF of a new hard drive is 2,500 hours and you anticipate having the unit powered on for 8 hours a day during the work week, the device should last about 2 years before it needs to be replaced).
4. **Routinely Inspect and Replace Data Discs:** Contemporary CD and DVD discs are generally robust storage media that will fail more often from mishandling and improper storage than from deterioration. However lower quality discs can suffer from delamination (separation of the disc layers) or oxidation. It is advisable to inspect discs every year to detect early signs of wear. Immediately copy the data off of discs that appear to be warping or discolored. Data tapes are susceptible both to physical wear and poor environmental storage conditions. In general, it is advisable to move data stored on discs and tapes to new media every 2-5 years (specific estimates on media longevity are available on the web).
5. **Handle and Store Your Media With Care:** All storage media types are susceptible to damage from dust and dirt exposure, temperature extremes, exposure to intense light, water penetration (more so for tapes and drives than discs), and physical shock. To help prolong its operational life, store your media in a dry environment with a comfortable and stable room temperature. Encapsulate all media in plastic during transportation. Provide cases or plastic sheaths for discs, and avoid handling them excessively.
6. **Multimedia (e.g., photo, video, audio) backup and archiving is a bit more challenging, in part because transcriptions and captioning can be important for interpretation, discovery, and accessibility.**Storage of images solely on local hard drives or servers is not recommended. Unaltered images should be preserved at the highest resolution possible. Store original images in separate locations to limit the chance of overwriting and losing the original image.There are a number of options for metadata for multimedia data, including MPEG standards, PBCore, and the US Library of Congress standards for still images, sound, and moving images.
7. **Finally, consider the short- and long-term security of the originals:** specimens, samples, documents and data sheets, photographs, or other physical items. Options include:
    - Campus Office
    - Lab or Museum/Herbarium
    - Departmental Office
    - Commercial storage facility
    - Home
    - Library

Backup Procedures

Backing up requires discipline and care to ensure nothing is missed and the file back-ups are secure. You can make this easier by doing the following:

1. **Plan ahead, and write down your plan.** A written data back-up and security plan includes such information as:
   - where backups are located
   - who can access backups and how they can be contacted
   - how often data should be backed up
   - what kind of backups are performed
   - who is responsible for performing the backups and their contact information. For large projects or projects with high-volume data streams include a person with secondary responsibility (the back-up back-up) in case the primary person responsible is unavailable
   - what hardware and software are used or recommended for performing backups
   - how and how often to check if backups have been performed successfully
   - the media are used to backup data
   - a list of any data that are *not* archived or backed up
2. **Backup your data at regular frequencies**, with backup strategies (e.g., full, incremental, differential) optimized for the data collection process and data type. For instance, you should (at least):
   - Back up when you complete your data collection activity
   - Back up after you edit / clean data
   - Streaming data (e.g., from data loggers) should be backed up at regularly scheduled points in the collection process
   - High-value data should be backed up at much higher frequencies (daily)
3. **Put procedures in place to make sure you follow your plan.**
   - set up calendar reminders
   - use checklists.
   - Ensure backup copies are identical to the original copy (using checksums, file counts, etc.)
   - Be sure of one of the regular calendar items is to verify successful recovery from a backup copy.
   - **Automate as much as possible.** Automation simplifies frequent backups. but be careful with accidental file loss and that back-up copies are actually independent (e.g. Dropbox and a synced folder on your laptop are *not* independent copies).

File formats for storage

For long term preservation is it necessary to store data in file formats that will be readable in the future. It is also important to provide descriptive information on these data file types and formats. This will facilitate data retrieval and reuse.

1. Document and store data using file-types that are open (ie., non-proprietary), uncompressed, unencrypted (if at all possible), and stable:
   - `ASCII` formatted files will be readable into the future. For tabular data use comma-separated values, `.csv`, for text use `.txt` (alternatively `.rtf`).
   - Images: `TIFF` (uncompressed)
   - Video: MPEG-4 (`.mp4`) or motion JPEG 2000 (`.jp2`)
   - AudioL Free Lossless Audio Codec (`.flac`)
   - Documentation: `.txt` (preferred), PDF/A (`.pdf`)
   - Structured, highly coded data: `.xml`
   - For geospatial (raster) data the following provide a stable format:
     - `GeoTIFF/TIFF`
     - `ASCII Grid`
     - Binary image files
     - `NetCDF`
     - `HDF` or `HDF-EOS`
   - For image (Vector) data use the following file formats (these are mostly proprietary data formats; please be sure to document the Software Package, Version, Vendor, and native platform):
     - ARCVIEW software: store components of an ArcView shape file (`.shp`, `.sbx`, `.sbn`, `.prj`, and `.dbf` files)

  - ENVI – `.evf` (ENVI vector file)
  - ESRI Arc/Info export file (`.e00`)

- ***Note:*** Certain file formats, such as shapefiles, can be made up of as many as 7 individual files. If one of those files is absent from the file assembly the shapefile data utility may be lost. Awareness of adherence to a particular file format standard can also be helpful for determining, for example, if a particular software package can read the data file. Awareness of whether that standard or format is open source or proprietary will also influence how and if the data file can be read.

2. If a particular software package required to read and work with the data file, you need to make sure you have a copy, a backup copy, *and a computer with the proper operating system to use the software.*
3. For audio and image files, back up with the highest-quality file-type possible (i.e., the 'lossless' format rather than the 'lossy' format). For example, `jpeg` is lossy, meaning images saved with this file type will have less detail than images saved with the lossless `TIFF` format if at all possible back-up

Tools

1. Online, generic multimedia repositories and tools include YouTube, Vimeo, Flickr, and Google Photos. These serivces:

- are often low-cost or free
- are open to all
- have functions for provide community commenting and tagging
- some provide support for explicit licenses and re-use
- provide some options for valuable metadata such as geolocation
- may allow for large-scale dissemination
- optimize usability and low barrier for participation However, they are commercial services, and hence have a number of potential drawbacks:
- models for sustainability is profit-based
- may have limits on file size or resolution
- may have unclear access, backup, and reliability policies, be sure to review them carefully Specialized multimedia repositories (e.g. MorphBank, LIFE):
- provide domain-specific metadata fields and controlled vocabularies customized for expert users
- are highly discoverable for those in the same domain
- can provide assistance in curating metadata
- optimize scientific use cases such as vouchering, image analysis
- may provide APIs for sharing or re-use for other projects
- are recognized as high-quality, scientific repositories
- may migrate multimedia to new formats (e.g. analog to digital) But they too have some drawbacks
- rely on research or institutional/federal funding
- may require high-quality multimedia, completeness of metadata, or restrict manipulation
- may not be open to all
- may have restrictions on bandwidth usage

A step back: what needs to be preserved?

To meet multiple goals for preservation, researchers should think broadly about the digital products that their project generates, preserve as many as possible, preserve all that they are required to preserve for as long as required or longer, and plan the appropriate preservation methods for each. Consider how long and how to preserve the following, taking into account **(i)** what would be necessary to reconstruct a complete data set if files downstream were lost **(ii)**, how long each needs to be kept, and **(iii)** the cost of and space required for preserving data for different lengths of time:

1. Raw data (written form, electronic form). *Raw data are almost always worth preserving.*

2. Tables, spreadsheets, or databases of raw observation records and measurements
3. Tables, spreadsheets, or databases of clean observation records and measurements. *If clean data can be easily or automatically re-created from raw data, consider not preserving. If quality control or analysis is time-consuming or expensive, then consider preserving the clean version.*
4. Intermediate products: partly summarized or coded data that are the input to the next step in an analysis
5. Documentation of protocols used to clean, summarize, or code data
6. Software or algorithms developed to prepare data (cleaning scripts) or perform analyses. *Algorithms and software source code cost very little to preserve*
7. Results of an analysis, which can themselves be starting points or ingredients in future analyses, e.g. distribution maps, population trends, mean measurements. *These may be particularly valuable for future discovery and also cost very little to preserve.*
8. Any data sets obtained from others that were used in data processing