Smithsonian Data Management Best Practices

Storage, Archiving, and Preservation[i] Preparation

STORAGE AND ARCHIVING

Research data and related files require reliable and trustworthy storage at all phases of the research process. Best practices include documenting the information below either in a **Data Management Plan**[ii], or in project protocol documentation.

To provide that trustworthy storage during the **planning and active** phases of a research project be sure you can document:

- Data **ownership** and responsibility.
- Who has **access** to the original or raw data, and how **access is restricted** (password protected, networked server with limited user accounts, etc.) Restricting access to the original data reduces risk of inadvertently (or intentionally!) altering or deleting data.
    - When doing analysis, transformations, or other work, always **use a copy** of the original.
- Estimate of **storage space** needed, including for backup copies, and storage **media**.
    - Preferred storage media include "spinning disc" hard drives, solid state hard drives (SSD), magnetic tape (often found in large data centers), and thumb/jump drives.
    - Optical media, i.e., CDs and DVDs, are not good long-term storage options as they can degrade quickly and fewer machines are able to read them.
- Location and methodology for **backups**, including **schedule** of periodic backups.
    - There should be **two**, preferably three, backup copies.
    - Copies may be physically stored with the researcher, on a networked server, or in the cloud. Each backup copy should be in a **different location**, and/or on different media.
    - Periodically **verify** backup integrity (can you access and read the files).
- Estimate of storage **costs** over the course of the project. Note: some cloud storage systems charge a fee to access and download data.
    - Media represents only a small part of total cost for storage in the long term. Maintaining content for preservation involves human resources as well.

All storage mechanisms are subject to failure and do not last forever. If managing your own data during a long-term project, plan to migrate data from one storage platform to another at least once to ensure that data remains accessible, and to prevent data loss due to media failure. If a third party is storing the data, migration to newer storage should be part of its service.

Before the **conclusion** of a project you should decide and document:

- **Where** the data will be archived, and who will be responsible for data stewardship and curation.
    - The archive or repository must have a **disaster recovery plan**[iii].
    - The organization responsible for long-term archiving must have sufficient resources (staff and funding) to maintain the data for the requisite period of time.
- **Length of time** data must be maintained.
    - Data that cannot easily be reproduced, such as climate sensor data or data from destructive sampling, should probably be retained indefinitely.
    - Data that has legal value should be retained indefinitely or for a length of time stipulated by law or policy.
    - Data that has significant historical or research value should probably be retained for at least 20-50 years.
    - Data that could be re-created with better technologies in the future could be retained only until it can be re-created and replaced.

- **Accessibility** of the data. Is it available immediately or stored in a dark archive offline that might require some retrieval time, or incur costs for retrieval?
- **Completeness** of data. If data is the result of a multi-institution collaboration, with archiving occurring in multiple places, ensure that all data is accounted for. If you only are responsible for part of the overall data, indicate where the remainder is stored.

PRESERVATION

Preservation of data requires **active management** of that data over time to maintain **accessibility**, **authenticity**, and **usability**. Future data managers' ability to maintain your data is influenced by the format of the data, the extent and quality of description (metadata), and how the data was stored and managed during the project.

When thinking about preservation, be aware that *not all data should be retained forever*. Drafts of papers, duplicate copies, superseded versions of datasets, beta versions of software, and other working files are transitory in nature and should probably not be preserved indefinitely. Preservation is a resource-intensive activity that involves much more than just storage and format migration.

To ensure your data is viable in the future:

- **Choose open formats when possible.** Store data in a non-proprietary, commonly used, open format. (see the Best Practice for File Formats)
    - If any of your files are in a proprietary or non-preferred format, deposit a copy saved in an open format with the original file when archiving, e.g., if your metadata file is in MS Word, convert it to PDF/A and archive both the original Word document and the PDF with the data (or just deposit the PDF).
- **Always include sufficient metadata** with your data files (see Best Practices for Describing your Data: Data Dictionaries) including:
    - Any requirements from funders or publishers that mandate sharing the data for a specified period.
    - Documentation of what software (and its version) and hardware were used to create the data, and what quality assurance processes were done on the data. Keep in mind that in the future, data created with special software might require the use of an emulator to view or use it.
- **Follow good storage and data management practices during the project.** Ensure a responsible party is maintaining the data according to best practices, and has created sufficient documentation to provide continuity should they leave the project.

REFERENCES

(accessed 2018-02-28). Smithsonian Institution Archives. Records Management at the Smithsonian Institution Archives - https://siarchives.si.edu/what-we-do/records-management

(accessed 2018-02-28). Digital Curation Centre. Why Preserve Digital Data? - http://www.dcc.ac.uk/digital-curation/why-preserve-digital-data

2017-10. University of Edinburgh. MANTRA – Research Data Management Training https://mantra.edina.ac.uk/

[i] Digital preservation is a process that maintains the integrity and readability of digital data for a specific period of time. This process may include data integrity checks, migration, transformation of the data into new formats, or environment emulation.

[ii] A data management plan is a formal document that outlines what you will do with your data during and after a research project. DMPs help articulate who has responsibility for your data throughout its lifecycle, what can and cannot be done with the data, as well as how to find, re-use, and interpret your data. Many potential data management disasters can be handled easily or avoided entirely by planning ahead.

[iii] A documented process that describes the procedures an organization will follow to recover data and restore IT infrastructure functionality after a disaster.