Raw data, backup and versioning

What you need to know to preserve your research data

Raw data: some basic assumptions

By raw data we mean the original data that has been collected from a source and not yet processed or analysed. Raw data will provide the foundation for any downstream analyses. In many cases the captured or collected data may be unique and impossible to reproduce, such as time points in weather measurements and interviews. For this reason, they should be safeguarded from any possible loss. Moreover, **raw data** will typically be lossless - i.e. those file formats that are not compressed such as TIFF files for image data as opposed to compressed JPEG file format. Finally, in some cases, raw data may have additional information that may be specific to a brand and/or type of instrument used to capture the data. For example, Leica microscopes use a proprietary data format but is also a container for lossless data - the container contains metadata specific to the Leica microscopes that allows reading, writing and analysis through Leica software.
See also our guide on file formats.

By processed data we mean data that has already undergone some kind of intervention. For instance, the data have been digitised, compressed, translated, transcribed, cleaned, validated, checked and/or anonymised.

By analysed data we mean data already processed, interpreted and analysed. Analysed data can assume several representations (text, tables, graphs, etc.), in order to facilitate a better understanding and communication of the data.

In most cases, one can also consider **raw data** as the official data, that is, the master copy of any given record (see also golden copy). As well as providing the starting point for derivatives generated downstream through analyses, there may be additional branches from which this data is used for other analyses. Therefore, in a typical workflow, **we recommend that you create a copy of the raw data which you use as a *working copy*.** The original data should then be archived in an appropriate manner for long-term preservation. The working copy can then be used for processing and analysing without worrying about overwriting.
For more information on data formats please see this guide.

Data backup

Firstly, one should be aware that backup is not the same as preservation and/or archiving. Once data reaches a final state, preservation **allows easy access** to that data, through, for example, a repository. If there are any data, raw or otherwise, and including those that are final products, that are deemed sufficiently important to be securely stored for a long period of time these should be **archived** - these data can be retrieved but not in the same easily accessible manner offered by preservation.

The preservation of the data may be associated with several reasons: for further analysis or research; for its potential value in terms of re-use; for national/international status and quality; for its originality and/or uniqueness; for its data production costs or innovative nature of the research; its importance for (science's) history; for its relevance of use for non-academic purposes such as cultural heritage or even by funder requirement. By contrast, **backing up data is mainly to prevent data loss** during the active (analysis) stages of the curation lifecycle. Researchers should do this while working with the data, and repositories do it when they preserve data. We recommend that important data are copied at least three times onto at least two storage media and at least one off-site. Moreover, where available, **always use your institution's managed digital services to allow automated backups**. Commercial and non-commercial third-party storage options such as Dropbox are also currently popular, but there is no guarantee that such services will exist in perpetuity while also such options raise questions about ownership.

Here is a checklist to help define a strategy for creating backups:

1. Firstly, find out if your institution has implemented a backup strategy. If so, find out if the backup policy meets your needs. It is also advisable to make direct contact with your IT department if there are any points of contention and also to receive direct advice. Your IT department should be easily contactable and it is in their best interests too to gain knowledge of how their institution's researchers are using their services. If your institution has no appropriate backup strategy, the following steps should be followed.
2. Identify if there are third party tools that can be used to automate backups. Performing automatic backups will provide a better guarantee that backups are made regularly and that they are stored in the right place, reducing the risk of human errors. Microsoft and Apple have software to support automatic backups. There are also cloud storage solutions (CloudStor; Figshare; SpiderOak; StoneFly) that offer backup functionality. However, it is a good idea to create a routine to check that functional backups have actually been created.
3. Identify the type of backup:
   a. Full system and file backup;
   b. Differential backups, where everything that has changed since the last full backup is recorded. If data recovery is required, you will need the last full backup and the last differential backup;
   c. Incremental backups, where only the last changes since the last backup are saved. To restore your data and/or system, you need the last full backup and the entire series of incremental backups.

d. Incremental and differential backups greatly speeds up the process as it only updates the changed files, saving time and disk space when running the full backup every day.

4. Plan how many backups and how often they need to be made. It is recommended that you make three backups to minimise the risk of data loss, even if one of your backups is damaged or even lost. If file sizes are very large or if we are dealing with sensitive data, it may be appropriate to work with fewer copies, but your institution's and/or funder's requirements should always be consulted wherever possible.
5. Define where the backups will be stored. It is recommended that you do this in physically different locations. Backups can be made to network drives, cloud storage, and local or portable devices. These options will depend on the amount of data that needs to be backed up, their frequency, the level of automation and their sensitivity.
6. Forecast the amount of space that will be required for the backup of data and its documentation, and then estimate the required storage capacity on the actual backup media.
7. You will need to determine how long the backups will be kept and how they will be deleted. It is not recommended that you replace one backup with another. However, if you have to backup large amounts of data frequently, it may not be feasible to keep all backups during the entire project. For sensitive data, you will need to ensure that all data has been successfully deleted and cannot be recovered in any way.
8. Verify that personal data will be protected and that backups containing such information are protected from unauthorised access.
9. Make a disaster recovery plan that defines the steps during and after a loss of data to make quick recovery possible. This plan should include contact persons with responsibility to provide support.
10. It is essential that responsibilities are fully identified regarding the performance of both manual backups and verification of automatic backups, as well as the performance of data recovery tests and restoration of any lost data.

Versioning

During the course of analysing data, it is likely that you will create several derivatives of the working copy, sometimes even automatically through scripts. Whether one of these versions is valuable and should be kept for long-term preservation is completely dependent on the data owner. However, it is recommended that versions are frequently monitored to discard those that are not required for verification, reproducibility, or transparency, amongst others.

You might want to keep all versions, but these can be very large files which will take up valuable storage space. Versioning is a key part of any workflow and appropriate measures should be taken to enable this, whether it is simply versioning through slight alterations to file names or using dedicated version control tools. The latter are also commonly used in large projects to allow multiple users to check-out and check-in the same file after making alterations. This also allows provenance, i.e. documenting or inspecting the history of changes.

Just as with backups, the first step is to find out what your organisation provides.

Resources

https://researchdata.nl/en/services/data-management/selecting-research-data/
https://blog.zenodo.org/2017/05/30/doi-versioning-launched/
http://help.zenodo.org/#versioning
https://www.ru.nl/rdm/archiving-data/what-data-should-archived-minimum/
https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/4.-Store/Backup