Ceplas FAIRidise

"If analyzing data is half the battle, and getting it is the other half, then managing it once you have it is the third half[1]"

Workshop 2: Data Storage and Backup

The data acts as a backbone for the research. It is important to be aware of the implications of careless data storage, as well as the possibilities to store the data securely with minimal risk of data loss.

Storage

Storage corresponds to the medium (e.g. Hard disk, USB, personal laptop, institute cloud etc) where the data is being kept. It is important to choose a stable medium to store your working data. For example, a USB, an external Hard Disk (HD), can not be classified as a stable storage medium. The same applies to a PC or a Laptop, they cannot be classified as a stable storage medium, as there are risks of hardware failure, software failure, theft, misplacement. These mediums (USB, HD, PC, Laptop) can be used as intermediate or temporary storage whilst one is working on the data but the data should be kept (backed up) in a stable storage. The stable storage mediums can be institutional storage facilities or other cloud solutions that are backed up regularly. These stable mediums could be the optimal storage solution for your working data. For historical data, one could consider using archival services offered by their organization or if unavailable the cloud solutions could also be used.

Before opting for a particular storage medium, it is important to PLAN and answers the questions related to:

- Data size (The amount of storage needed)
- Data sharing and versioning (Who should have access to the data, how often the data will be changed/updated)
- Backup frequency (how to protect against data loss)
- Data security (handling of personal, sensitive data)

Medium	**Recommendations**
PC & Laptop	Store your active data with regular backups and/or version control Store one copy of your data on PC if space permits
Removable storage HD, USB, Tape	One copy of data stored on external drives should also exist elsewhere (e.g. on the cloud, institutional services) Not recommended for Master Copy of your data For sensitive data, encryption and password protection is recommended Regular backups and checkup, maintenance are needed

[1] https://v4.software-carpentry.org/data/index.html

Institutional services	Almost always try to use the services offered by your institute. Keep one copy of your data to the local cloud, with regular synchronization to your working data. (e.g Laptop/PC) Know the security and data protection policies of your data, and that of the institute Know the backup policies
External Cloud Storage	The data which is not susceptible to security or falls under data protection can be stored to external providers OR consider encryption of sensitive data **Store**: Dropbox, gdrive **Collaboration**: gdocs, Openoffice, overleaf

Examples for choosing a data store

Type	Volume	Importance	Change Frequency	Collaboration	Backup Frequency	Track Changes	Data Protection	**Storage Solution**
code	Small	high	high	yes	high	yes	low	Github Gitlab Gdrive Sceibo
manuscript	small	high	high	yes	high	yes	low	Overleaf Gdrive Sciebo
Raw *omics data	intermediate	high	low	no	low	no	low	Institutional Cloud External Storage(HD, Tape)
Analysed omics data	intermediate	high	high	no	high	yes	low	Version controlled institutional services Public cloud
Interviews	intermediate	high	low	no	low	no	high	Local storage solution External Storage (HD, Tape) (encrypted storage

Backup

Backup is creating and storing a copy of your data onto another storage device. Among many reasons, hardware failure is one of the most common reasons for data loss. It is

estimated that 99% of people have encountered hardware failures at least once. Other reasons for data loss are:

- Software malfunction
- Malware/hacking
- Human error
- Theft
- Natural disaster

Hence it is important to maintain a copy of your important data.
Backup should be planned and carried out on a daily/weekly/monthly basis depending on the data (critical, importance, cost, type or sensitivity).
The **3-2-1 backup** principle is usually recommended as an efficient backup strategy. It says that:

1. There should be at least **three** copies of the data
2. At least **two** copies should be on different storage devices
3. At Least **one** of them should be in a remote, different location

Archive

The term *archiving* is used for long term data storage for the data that is not in active use and needs to be stored to historic references and/or to comply with the funding policies (e.g. DFG requires to store data for as long as 10 years beyond the project lifetime). It is important that long term storage of data should be planned and appropriate resources be allocated for the task.
Besides taking care of storage considerations already mentioned above for data store and backup, it is important to store the data in a reusable and sustainable format, alongside proper documentation to enable future reusability and comprehension of the data.
The data archive will grow over time, and it is important to select a trustworthy and authentic provider for this service.

Example File Formats for Data Archiving

File Type	Avoid	Recommendation
any	*Specialised formats, coupled to special softwares*	*Convert to open file formats, if not possible, try to include the software with the files (preferably a working dockerized version)* *Add documentation to the data*
tables	SPSS, PDF, Figure	csv ,tsv, xlsx
text	PDF, Image, doc,	txt, html, rtf, docx
images	Git, jpg	Tiff, jpeg2000, png