# 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]"*

# Session 1: Directory Structure and File Naming

Researchers work with data continuously growing in size and heterogeneity. Researchers can benefit from good data management practices to help them navigate through their own data in a time efficient and secure manner. While most of the discussion around FAIR principals target the research output more than the data of an individual, in initial workshops, we focus more on personal, or individual data management practices and key points.

## Organization (Directory Structure)

Well-organised folder structures with descriptive names make it easier to find and keep track of all the files. You can design a directory structure that suits your needs and covers your or your group's working style. Think in terms of a hierarchy, or a taxonomy of terms/projects, that is being handled in your research. Decide when to create a new folder (e.g. administration, literature review, data, meetings) and when to subdivide into further folders (project->data). An interesting directory structure could be that of ISA[2], organizing files in a hierarchy of Investigation->Study->Assay. Investigation is the highest level with the information about the project, partners, tasks etc., the subdirectory study can keep record of several studies handled in this project and assay directories within each study can keep record of the data, and their analysis - e.g. we can think of another level (analysis) to keep track of analysis and computations applied on the raw data at the assay level.
However, instead of organizing by projects, you could also organize your data by the data types or methods. Individual preferences can be taken into account e.g. who created the data, when it was generated, or the experimental conditions. However, the key objective of the organization should be that the schema remains logical and easily understandable when you (or someone else) decides to reuse this data after a substantial time has passed.
It is also important to consider different or all aspects of the project and develop a directory scheme that includes important descriptors e.g. "project/study/assay/run/sampletype/date/datafile". It is strongly advised to document the naming conventions and place a readme file in the root directory for future reference, this assists in sharing the folder hierarchy with contributors effortlessly.

## File Naming (checklist)

The file name is the main identifier for the research data. A proper and descriptive naming convention can help in quickly finding the data and help in easily understanding the

[1] https://v4.software-carpentry.org/data/index.html
[2] https://isa-specs.readthedocs.io/en/latest/isamodel.html

context(content or purpose) of the particular file. There are two major aspects of file naming; The first aspect, handles readability and understandability of the file names,.e. MyThesis.pdf Vs 123.pdf. It is important to assign file names that are self-explanatory. The file names should be descriptive, and meaningful. However, it should be noted that in many cases, it is not advisable to rename raw files generated directly from equipment. In this case one can consider organizing the raw data inside a folder whose name describes the data. The second aspect of file naming is the characters used for naming, e.g. it is not advisable to use special characters in filenames. The names should, ideally, not exceed 25 characters. They should not differentiate between lower and smaller cases. The ISO 8601 format (YYYYMMDD) is recommended for dates. For example, the file name could contain date, description, project, location, work, sample-name, analysist-type, or version number of a file. One can consider drafting a context dependent naming convention for their files. It is also good to think about sorting e.g. sort by name, with all the files starting with name “Project” may not be very useful.

The developed naming convention must be documented for future reuse and comprehension (more on this later). A good file naming, not only helps in findability, but also helps in providing an auditing trail for your data development. Proper naming can help prevent confusion when working in a collaborative environment and helps to ensure that the files are not accidentally deleted or overwritten.
A succinct rationale for file naming is provided by the university of Edinburgh “Naming records consistently, logically and in a predictable way will distinguish similar records from one another at a glance, and by doing so will facilitate the storage and retrieval of records, which will enable users to browse file names more effectively and efficiently.”[3]

[3] https://www.ed.ac.uk/records-management/guidance/records/practical-guidance/naming-conventions