How to structure and document research data: Organising a dataset – files and folders

By Leif Longva, UiT The Arctic University of Norway
ORCID: 0000-0001-6638-8317

Date: October 2020

Manuscript:

In this video I will talk about how you should organise your dataset in files and folders. Research data are normally recorded and organised into files of some format. And you may sort the recorded data into a number of files, following some criteria.

These files may then be organised into folders, according to some criterion that unites some files and separates others. The criterion might be date of observation, excursion or experiment number. The folders may further be organised hierarchically.

The criterion by which research data are organised into folders may be apparent from the folder names. That will make it easy for you and for others to get an overview of the content of your dataset, and to browse and investigate the folders and files. More on this later.

Remember the research data management life cycle. Structuring your research data, and deciding how to organise your data into files and folders, is something you need to do as you collect or generate your research data. Or, rather, you should think this through in the planning phase. If you have a clear plan for this before your start collecting and analysing your data, you will avoid the risk of messing things up.

So what about the names of files and folders? Why does it matter what names you choose to use?

Well, it is nice to understand which of your data files holds which data, just by looking at the file names. It may save you and others a lot of time and work if you or others would like to make use of your data at some point in the future.

But more than nice: It is important to keep your data file names tidy. Or else you may even ruin your own data analysis. A consistent and logical file name syntax is the key here.

Before we take a closer look at good routines for structuring research data, I have listed some basic tips on this slide:

First, always keep in mind when structuring your data that anyone, including yourself, should be able to understand your dataset, both now and also many years from now. How to separate your different files through the file name syntax, and how to organise your data files in folders, is something you should think through early in your project – and before you start the data collection process.

Do not postpone the job of giving structure to your dataset until later. And then, later, you will thank yourself.

Another key issue to keep in mind in this process is file description – you might be able to enter some information as an appropriate tag to the file, and you should include descriptions in your Readme file (the Readme file is the topic in a separate video).

Furthermore, you need to consider the file formats you use – avoiding formats that you and others may be unable to open in some years from now.

And how you store your data out in the field or in your lab is also a key issue here. These issues are covered in other videos.

Following good practice in these matters makes it much easier to find and understand your data, not only for you, but also for your peers and colleagues who may want to reuse and build on your data.

So what is recommended when giving names to your files and your folders? Consistent filenames! The file names should explain what basically separates one file from another: Experiment number, or species name, or location name, or whatever will be natural for your dataset.

Use short and descriptive names, without any spaces. Special characters may cause problems. If the time line is important, you may include the date in the names. Use the international dating convention as shown here.

And remember to describe and explain your file naming syntax and dating convention in your Readme file.

Here is a little test for you. Please pause the video and read the question and the four suggested answers. Evaluate each of the answers, and make up your mind whether they are good answers or not.

So the question is: Why is consistent file naming important? Let us now look at each of the suggested answers.

A: ”In order to easily understand what data are found in which data file” – This is a good and correct answer. The file names should be chosen so that they give good information on the content of each file.

B: “In order to easily understand what data the dataset holds” – This is also to some extent a good answer. However, the file names will commonly not hold enough information to fully understand what data the dataset holds.

C: “In order to assess the quality of the data generating method” – No, this will not be possible to assess, based on the file names.

D: “In order to avoid mixing data while analysing the data” – This is a good and correct answer. It is of course very important to avoid mixing data while doing your analysis. The file names should be chosen to help you avoid this.

Here are some examples of good file names: The file naming syntax you choose may be governed by how you want your files to be ordered.

The first example shows files ordered by date.

The second example shows files ordered by the subject of the individual files. It may be the location or some other subject aspect.

The third example shows files ordered by type: All ‘Notes’ ordered together etc.

The last example is forced order with numbers. Note a two-digit system is used. Depending on how large your file collection is, you may need to use a 3-digit system.

All of these examples show you how a little bit of planning can serve well in deciding how to name your files. Note how descriptive the file names are. And they use underscores not spaces.

Using folders will help you structure the content of your dataset. With folders you can organise your data files nicely, and they may be very useful if you have many files. In order to make the structure of your dataset easy to follow, folders should be named consistently and organised logically.

It is good practice to let the main structure in your folders be reflected in your file names. In the example here: data files collected at Jane's Island carry the folder name of Jane's Island, and so forth. This will help you avoid mixing things up while analysing your data. And it will also make it easier for you to organise your data when uploading them to an archive at the final stage of your research project. And you should document this strategy in your ReadMe file.

Finally, be careful to keep track of, and document all changes that create new versions of your files. And this applies to updates to files both before and after the dataset has been archived. When a file is updated, this should be indicated in the file name. But do it in a tidy way and be consistent in how you do it.

So, if you update a file, this should be indicated in the file name, neatly with the version number or the date of the new version.

To include a table of the update history of the file, in the file itself, is a good recommendation.

And don't forget to document all such updates in your Readme file.