How to structure and document research data: Documenting your data – the Readme file

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

Date: October 2020

Manuscript:

So what do we mean when we talk about documenting data? The main goal is to make sure your dataset is understood correctly by anyone who finds it. Therefore you need to include a human readable description that gives the big picture as well as more details on what your dataset is all about.

The metadata you enter is one part of this. But you should also write a Readme file, where you include all information necessary to interpret and understand your dataset correctly.

Remember to start creating the Readme file early, and then enter more information as your project proceeds. Some information may be entered at the planning phase. Some may be entered as you collect or generate the data. And you finalise the Readme file as you make your dataset ready to be archived.

As already mentioned, the purpose of the Readme file is that anyone should be able to interpret and understand your data correctly. So what needs to be included in the Readme file will differ from dataset to dataset, depending on what kind of data they hold. But still there are some generic recommendations on what should be included:

- Make sure people are able to contact you. People interested in your dataset may have questions to ask you. And it may be the beginning of a new fruitful research cooperation for you.
- Make it easy for others to understand what information the dataset holds.
- If your dataset holds lots of files, it will be great help if you explain how the files are structured, and also the logic behind the file names. Make sure the file names are informative, and give information on what data are found in each file.
- It may be obvious from the file names and the folder structure, but otherwise give an overview of what data are found where in the dataset.
- And if your dataset is updated, remember also to update your Readme file

You may also look for examples of Readme files

There are a number of other information elements that should be included in the Readme file. Keep in mind that information that seems obvious to you, may be not be so obvious to others.

And also remember: Knowledge that is crisp and clear in your head today, will fade quite fast. You may want to go back to your own dataset some time into the future. Make sure you don't have to spend lots of time to understand what information your dataset holds.

Here is a little test for you. Please pause the video and read the question. Then evaluate each of the listed elements, and make up your mind whether they belong in the Readme file or not.

So which of these elements belong in the Readme file?

A: “A short description of what the dataset is about” – Yes, this should be included in the Readme-file, to make sure users understand what your dataset is all about.

B: “An overview of the files and the folders – which data are found where” – Yes, this should be included in the Readme file. It will be very helpful in guiding your users.

C: “The author’s assessment of the dataset’s value to others” – No, this is not common to include. And it will not be easy to give an assessment of this.

D: “A description of the data generating method(s) used” – Yes, this should be included in the Readme-file. Information on the method used will be vital in evaluating what information the data holds.

Why do we need the Readme file, in addition to a rich metadata schema? First, you do not have to duplicate the information both in the metadata and in the Readme file. The most important thing is that information in the metadata and in the Readme file together give good documentation and understanding of the dataset.

But keep in mind that the information entered as structured metadata will be much more easily searchable, compared to the information in the Readme file. So for the purpose of making your dataset findable through searching, good metadata quality is the key.

The Readme file is commonly meant to supplement the metadata. You may also view the Readme file as a standalone documentation file, and thus include some duplicated information from the metadata.

And if the dataset is updated, make sure to update the Readme file to reflect this and to document what has been updated.

Archiving your files in preferred formats is important. And this goes of course also for the Readme-file. Do not risk that the Readme file is difficult or impossible to open some years into the future.

You may use PDF, especially if your Readme file includes figures and illustrations. Make sure to use the archival-proof PDF-version, PDF/A.