How to structure and document research data: Preparing the data for archiving – file formats

Transcript of video

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

Date: October 2020

Manuscript:

We all know and have experienced that technology is developing and changing. Software and corresponding files formats may become more advanced, giving more options to the user on what to include in the file, and how. But the problem with this is that the content in a data file which is a few years old may not be correctly presented when opened. Or the file may not be possible to open at all.

To make sure your dataset is accessible also many years from now, it is important to choose file formats where this is guaranteed not to happen. This is the topic of this video.

Making sure your dataset is accessible in the future, has to do with the file formats you use when archiving your data. But you may also need to consider how you want to or need to handle your data in the collecting or generating phase. As well as how you may analyse your data. Perhaps you need to use designated software for this? So maybe you need to convert the files with your analysed data into formats suited for archiving.

These are all issues you should think through in the planning phase of your project. How can you best accommodate all your needs in the various phases of your project? And then you need to make sure that the formats you use do not prevent conversion to a preferred format for archiving.

Commonly, preferred formats are characterised by being:

- Non-proprietary. That is, there are no commercial interests controlling the format and the design and development of the format's encoding scheme.
- The format specification should be open and well documented, following international standards.
- The encoding of the characters should also follow well documented standards. This is to avoid characters becoming unreadable.
- And file formats should be uncompressed, since something may be lost while compressing the size.

There is a nice list of preferred file formats maintained by DANS (Data Archiving and Network Services) in the Netherlands, which will help you in determining whether a file format does indeed comply with the requirements of being preferred.

Often, you may find it convenient to use a non-preferred format for your work files. For tabular data, Microsoft Excel is much used. This is perfectly all right. But make sure you convert such files into preferred formats before archiving.

Make a note of these basic tips I show here.

While preparing your data for archiving, you also need to consider _what_ to archive. During your research project you may have created lots of temporary files, or you may have done some failed experiments in the lab. It is not necessary to archive all such data. So you may need to do a careful selection of what to archive.

But be careful not to exclude null-observations or negative data. Those are observations too.

And data that are of little value to you in your project, may be of significant value to others. Therefore it is good practice to archive all raw data, if practically possible.

But of course, be careful not to archive openly personal or sensitive data. Everything that makes it possible to identify persons must be removed from your data.

Remember what we said about file formats for archiving: They should be preferred formats for long time preservation. However, non-preferred (original) formats may be archived _in addition to_ the preferred format. This may be convenient for those who find your dataset while this original file format is still a much used and convenient format.