Make sure your data is Interoperable

File formats

The format in which research data are created usually depends on how researchers choose to collect and analyse data. This is often determined by discipline-specific standards and customs. However, file formats have a lifetime associated with them. Ensuring long-term usability of data requires consideration of the most appropriate file formats.

The formats most likely to be accessible in the future are:

- non-proprietary
- in an open, documented standard
- commonly used by the research community
- in a standard representation e.g. ASCII, Unicode
- unencrypted
- and uncompressed

Popular formats such as those produced by Microsoft Office products (e.g. Word documents or Excel spreadsheets) are very likely to have reasonable longevity, but be aware that they are proprietary (owned by someone) and so will not necessarily exist forever or remain easily readable. We encourage researchers storing important information in open, non-proprietary formats – for example:

- PDF/A rather than Microsoft Word
- CSV rather than Excel
- TIFF rather than Photoshop file
- XML rather than a database.

The UK Data Archive provides recommendations on optimal formats for preservation and there is also a training module file formats on the MANTRA website.

What do I need to consider when creating a file name?

Decide on a file naming convention at the start of your project.

Useful file names are:

- Consistent
- meaningful to you and your colleagues
- allow you to find the file easily.

It is useful if your department/project agrees on the following elements of a file name:

- Vocabulary – choose a standard vocabulary for file names, so that everyone uses a common language
- Punctuation – decide on conventions for if and when to use punctuation symbols, capitals, hyphens and spaces
- Dates – agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD
- Order - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
- Numbers – specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

How should I name my files, so that I know which document is the most recent version?

Very few documents are drafted by one person in one sitting. More often there will be several people involved in the process and it will occur over an extended period of time. Without proper controls this can quickly lead to confusion as to which version is the most recent. Here is a suggestion of one way to avoid this:

- Use a 'revision' numbering system. Any major changes to a file can be indicated by whole numbers, for example, v01 would be the first version, v02 the second version. Minor changes can be indicated by increasing the decimal figure for example, v01_01 indicates a minor change has been made to the first version, and v03_01 a minor change has been made to the third version.
- When draft documents are sent out for amendments, upon return they should carry additional information to identify the individual who has made the amendments. Example: a file with the name datav01_20130816_SJ indicates that a colleague (SJ) has made amendments to the first version on the 16th August 2013. The lead author would then add those amendments to version v01 and rename the file following the revision numbering system.
- Include a 'version control table' for each important document, noting changes and their dates alongside the appropriate version number of the document. If helpful, you can include the file names themselves along with (or instead of) the version number.
- Agree who will finish finals and mark them as 'final.'