File formats

File formats are the software used to store research data and documentation.

From the perspective of research data management, there are two types of formats: open and closed. Open formats, and free formats, are storage software usable on any operating system. They have a published specification and are maintained by an independent organization or are available without legal restrictions. Closed formats are the opposite; these are commercially owned trade secrets.

The danger with using closed formats in your research is that you may be "locked-in" to a specific manufacturer's product; this makes it difficult to share data either informally or formally, as many closed formats are platform specific. In addition, closed formats by definition do not reveal what information they store in terms of changes made to files, which if you are a researcher dealing with sensitive data can be problematic because you cannot be confident as to what is, or is not, being recorded in the file's metadata. Furthermore, should that manufacturer disappear or decide to withdraw support for a format then your data becomes inaccessible. Closed formats have the potential to become obsolescent and inaccessible within a decade.

Another term you may encounter is proprietary format. Proprietary formats are legally owned and can be either open (published) or closed. However, note that sometimes, the term "proprietary" is used mean closed.

Proprietary or not, consider the long-term availability of, and support for, any hardware and software used to store data and documentation. Attempt to keep copies in open standard formats or at least formats widely used and accepted by the research community. A few common examples of open formats include:

- **Archiving**:

  7z (archiving and compression), MAFF (web page archiving), tar (archiving) ZIP (archiving and compression)

- **Databases:**

  CSV (spreadsheets), NetCDF (scientific data)

- **Multimedia:**

  DjVu (scanned images and documents), JPEG2000 (a standardized image format), PNG (standardized raster image format), SVG (standardized vector image), WebM (video and audio format)

CESSDA AS
CESSDA House

Parkveien 20
5007 Bergen
NORWAY

phone: +47 55 58 21 18
e-mail: cessda@cessda.net
www.cessda.net

- **Text:**

  CSS (websites), HTML (websites), ePUB (open e-book standard), LaTeX (document markup language), Office Open XML (text format) OpenDocument (text format).

While researchers should bear in mind the above criteria, we acknowledge there is a difference between working data and preservation data. We would discourage the use of closed formats for working on your data, but recognize that short-term such formats may have compelling reasons for use. However, we encourage the use of open formats for data collection and preservation copies where possible.

References and further reading