

File formats

Who should read this?

This is likely to be of interest to researchers, their professional support staff, data centre and repository staff.

What is a file format?

A *file format* is a way of organising meaningful information into a sequence of bits and bytes for storage in a computer system. Most people are familiar with different file formats for documents, images, sound files and perhaps video. The same issues apply to file formats for research data sets.

You can often identify a file's format from the filename extension: .doc for Microsoft Word documents, .html or .htm for web pages, .jpg or .jpeg for JPEG images, and so on.

What's the problem?

There are often many different file formats available for storing the same data. Choosing a suitable file format for data preservation and sharing is more important and more subtle than it might seem. Different formats have advantages or disadvantages, depending on purpose.

Example: Text documents

(Note: The purpose of this example is to illustrate the considerations involved in file format choice, using a type of data that is familiar to as many readers as possible. It is not intended as a tutorial on text file formats, or an endorsement of any of the file formats mentioned.)

Imagine a short text document. There are several options for how to store this document in a computer system. Each has its advantages and disadvantages.

1. Plain text (.txt). In this case, the individual characters in the document (letters, punctuation, newlines etc.) are each encoded into bytes using the ASCII encoding (or another character encoding such as UTF8 or iso8859-1, particularly if the document is not in English), and stored in a simple sequence. This format only stores the text itself, with no information about formatting, fonts, page size, or anything like that. It is portable across all computer systems and can be read and modified by a huge range of software applications. The details of the format are freely available and standardised. If the storage media are damaged, any undamaged sections can be recovered without problems.
2. Microsoft Word document (.doc). In this case the text plus formatting, page size and so on are stored in a complex encoding. The details of this encoding are owned by Microsoft and are not freely available. Only Microsoft Word software can recover and edit the contents with complete fidelity. Word only runs on Windows and Macintosh, not on Linux. Other software applications have been written that can read (and in some cases also edit) the format, but as they don't have access to the full specification, their performance is less than perfect. If the storage media are damaged, it may not be possible to recover anything. While Microsoft has (so far) ensured that new versions of the software can open documents stored in old versions of the format, there is no guarantee that this will always be the case.
3. PDF (Portable Document Format, .pdf). In this case, the text plus formatting, page size and similar information are stored in a moderately complex encoding. While the details of this encoding are freely available, the format is owned by Adobe and can be changed by them at any time, for any reason. The document can be viewed and printed on all major platforms, using free software provided by Adobe (or others). PDF documents cannot be readily edited.
4. HTML (HyperText Markup Language, .html). The text, plus simple formatting, is stored in a simple encoding



that is based on the plain text file format above, with plain text markup interspersed with the text. This format is freely available and controlled by a public-interest standards body. The document can be viewed in any web browser. It can be edited in a text editor by someone who knows HTML, or in any number of “rich text” editors, word processors, HTML editors and so on.

5. A structural markup language like DocBook XML (.xml or .dbk) or the Text Encoding Initiative (TEI). XML is the eXtensible Markup Language, which is a framework for defining markup languages. DocBook is an XML language for describing documents and their logical structure. This information is stored in a simple encoding which can be read and edited on any platform. In this case there is no specification of how the document should appear when viewed on-screen or printed on paper. However the format is well suited to transformation into other formats for viewing (e.g. HTML) or printing (e.g. PDF). Structural markup also assists with indexing for discovery. These are good preservation formats.
6. An image format such as JPEG (.jpg) or TIFF. While the image can be viewed on any modern platform using a wide range of software, editing the content of the document (the sequence of characters, words etc.) will be extremely cumbersome. This format preserves appearance but loses all structure. However this may be an option for documents that only exist on paper, perhaps with a plan to use optical character recognition (OCR) to migrate to a text-based format in the future.

Similar considerations apply to file formats for storing images, sound recordings, video recordings and other types of data such as spreadsheets, lecture slides, numerical data, geographical, spatial and mapping data, databases and so on.

What do you need to consider when choosing a file format?

This list can be used as a checklist for file format choice.

1. Is the format proprietary, or is it controlled by a public standards organisation?
2. Is the format specification publicly available, or does the owner keep it secret?
3. Is format obsolescence a risk?
 - (a) Is it possible that the supporting software could be upgraded and the new version won't be able to open old files?
 - (b) Is it possible that the supporting software could be bought by a competitor and withdrawn?
 - (c) Is it possible that the format could fall into disuse and lack software support?
4. Is the format convenient for extracting the data for further use and indexing for search/discovery, or is it only suitable for viewing the data?
5. Does the format store the data at the required level of fidelity? Will the data be degraded every time it is saved in this format? (In other words, is this a *lossy* format?)
6. Does the format compress the data? (Generally compression makes a file less robust to errors in data transmission or damage or degradation of storage media, but of course it also uses less disk space.)
7. Is the chosen format an accepted standard? (This could be either a formal standard managed by a standards organisation, or a de-facto standard in your field of research.)

Planning implications

File format decisions should ideally be made *before* you start data collection. Migrating data from an unsuitable format to a better one is usually difficult, expensive and may in some cases be impossible, but it may be necessary as part of the ongoing curation of a long-lived data collection.

Further Information

ANDS Guides and other Resources: www.ands.org.au/guides

