

File formats

Who should read this?

This is likely to be of interest to researchers, their support staff, data centre and repository staff, and in general to anyone who has to make decisions about file formats for data storage, transmission and sharing.

Preliminary reading

ANDS Guide: [File formats \(awareness\)](http://ands.org.au/guides/file-formats-awareness.html) (<http://ands.org.au/guides/file-formats-awareness.html>)

What is a file format?

In simple terms, a file format describes the way information is organised in a computer file. More formally, the ICPSR *Digital Preservation Management Tutorial* defines a file format as follows:

“Without a format specification, a file is just a meaningless string of ones and zeros. The specification indicates the proper subdivision, encoding, sequence, arrangement, size, and internal relationships that uniquely identify the particular format and allow it to be properly interpreted and rendered.”

Most people are familiar with different file formats for documents, images, sound files and perhaps video. There are also many different file formats for research data sets, and the same issues apply.

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 than it might seem. Different formats have advantages or disadvantages, depending on purpose. File formats can become obsolete, they can be tied to particular software, and they can restrict how people can access the content, or even degrade the information being stored. They can also be owned by corporations whose primary interest is profit, not the preservation and sharing of research data.

File format obsolescence

File formats can become obsolete for various reasons:

- the software and the format get upgraded and the new version of the software no longer works with the old version of the format;
- the software that supports the format gets bought by a competitor and withdrawn;
- the format falls into disuse or no-one writes software to support/implement it; or
- the format is no longer compatible with modern environments.

If a file format becomes obsolete, it may no longer be possible to read the file or use the data. In some cases, this data loss may only be partial: for example it may still be possible to recover the text from an obsolete word processing format, but the formatting may be lost. In other cases data loss may be complete.

If data is stored using a format that is, or is about to become, obsolete, then it may be necessary to *migrate* to a more suitable format. The alternative is to somehow preserve the entire environment needed to access and/or use the data. This approach either involves maintaining old computer hardware, together with the operating system and all the required software, or writing special *emulation* software that recreates the software’s operating environment within more recent systems.



Open and proprietary formats

A *proprietary* format is one that is owned by an individual or a corporation. Some examples of proprietary formats in common use are AutoCAD's .dwg drawing format, the MP3 MPEG Audio Layer 3 format and Adobe Photoshop's .psd native image format.

An *open* format is one where the description of the format is available to the public. Examples of open formats include the JPEG 2000, PNG and SVG standard image formats; ASCII, PDF, Open Document Format and Office Open XML format (the native format for recent versions of Microsoft Word) for text; HTML, XHTML, RSS and CSS for the web and NetCDF for some scientific data.

Most proprietary formats are *closed*, meaning that the definition of the format is not available to the public. This means that data stored in the format can only be accessed using the format owner's software. This presents a number of risks:

- The company might go out of business, leaving the format unsupported.
- The company might change the format without warning, creating a conflict between data stored by different versions of the software. This forces users to purchase an upgrade to the software in order to access data. This sort of format obsolescence, leading to forced upgrades, is part of the business model for some software companies.
- The company may attempt to block all attempts to access data stored in their format using any other software, in order to "protect their investment". Essentially this means that your data is being held hostage, and you have to pay a ransom in order to access it.

Some formats are both open and proprietary. Examples are Adobe PDF format and the Microsoft OOXML format. While this is less of a risk than closed, proprietary formats, there is still a risk associated with the format being owned and controlled by a for-profit corporation.

Another risk associated with proprietary formats is that the owner decides to restrict or charge for access to data stored using that format.

Open formats owned and controlled by public standards bodies avoid these risks. As the format definition is freely available, anyone can in principle write software to access data stored using that format.

On the other hand, not many individuals or not-for-profit institutions have the capability or the resources to write such software for complex file formats. Open-source formats also risk falling into disuse unless they have a sufficiently large and active body of users.

In general, all else being equal, open formats are preferable to closed, proprietary formats. In some situations however, proprietary formats are unavoidable. One such situation is where a scientific instrument produces data that is already in a proprietary format, perhaps owned by the instrument manufacturer.

In other cases a proprietary format is the de-facto standard in a particular discipline. In this case you will need to balance the disadvantages of having your data in a proprietary format against the advantages of easy data sharing with colleagues who use the same format.

Preservation formats and display formats

Sometimes the format that is best for long-term preservation is not convenient for displaying, visualising or re-using data. For example, a high-resolution satellite image might be best stored as a lossless, uncompressed bitmap, but for viewing through a web application, parts of the image might be converted to JPEG on the fly in order to save on transmission costs. Similarly a document might be stored in a standard XML format, but for viewing or printing would be rendered to HTML or PDF as these formats are easier for humans to read. Storing a document in PDF locks it to a particular page size and makes reading on-screen difficult. Similarly some data formats restrict how the data can be accessed.

Lossy formats

Some standard file formats save space by throwing away detailed information that is assumed to be unimportant. For example, JPEG images smear out fine detail in photographs. Once lost, this information can never be recovered. A lossless format like TIFF keeps all the detail.



At least at the higher quality levels, this loss of information is very difficult for the human eye to detect, which is why such methods are used in most small digital cameras for example. There are similar issues with sound and video file formats. For sound files, the ubiquitous MP3 format is lossy, while WAV format is lossless.

Every time you edit an image that is stored in a lossy format, and then save the result, you will usually lose more information. So it is particularly important to avoid lossy file formats when data is being changed repeatedly.

For scientific data, it is important not to lose information, so you should generally prefer a lossless format that preserves all the data, rather than a lossy format. Of course this may not always be practical. For sound recordings and video you may have to make a decision about what level of quality is acceptable. This depends on the purpose: for music recordings and perhaps for linguistics research, excellent sound quality is essential; for interviews with politicians, it is the words that matter, and a lower sound quality may be acceptable.

Compression

Compression refers to ways of making data take up less disk space without losing any of the content. A file that has been compressed can be restored to its original state, completely unchanged. This is not the case for lossy formats, where the reduction in size is achieved effectively by throwing data away.

Although compressing data doesn't destroy it, the compression process makes data more susceptible to "bit-rot". For example, if you randomly change one bit in a plain text file, you alter at most one character of the text, and such an error can usually be corrected by a human reader. On the other hand, if you change one bit in a compressed text file, you may cause major changes across the entire document, rendering it effectively useless.

The reason that a single-bit error in a text file can usually be corrected by a human reader, is that written text contains a great deal of redundancy. This is not usually the case for numerical research data, so it is even more important to ensure correct transmission.

This means that for long-term preservation, uncompressed formats are usually preferred.

Many file formats include data compression (either lossy or lossless) as part of the format itself. This includes almost all image, sound and video file formats. Alternatively any file in any format can be compressed (losslessly) by standard methods like Zip, Gzip or Bzip. Some file formats combine both approaches. For example, a file in Open Document Format is a Zip compressed folder of (mainly) XML files.

The importance of standards

Using standard file formats is essential for effective data sharing. Always prefer a file format that is defined by a standard. As people say, "The good thing about standards is that there are so many to choose from", so it is usually possible to find a suitable standard format for your data.

Some disciplines have standards for research data. (For example, the use of SPSS data files for social science data sets.) If all your colleagues are using a particular format, then it makes sense for you to use it too.

Keeping multiple formats

This can be a good idea. In particular, you should always keep the original data file, exactly as it was first acquired, even if it is in a less-than-ideal format.

Of course, if you are keeping multiple copies of the same data in different formats, you will need processes to ensure that they stay in synch: if you modify one version, you must be sure that you make the same modification to all the versions.

An alternative to keeping multiple formats is to keep your data in a good preservation format, and equip your repository with front-end software that can convert it to multiple alternative formats on the fly. For example, a repository might store a text document in a gold-standard preservation format like DocBook XML, but provide a service that can also disseminate the document as HTML, PDF or Word, depending on the preference of the reader. Similarly for image data, as in the astronomy example above, you might store your image as lossless, uncompressed TIFF, but have software that can serve it up to others in a variety of image formats at different resolutions, depending on their needs. For tabular numerical data, you might store it as comma-separated values in a plain text file, but be able to serve it up to colleagues as an Excel spreadsheet or an SPSS data file, depending on their preference.



Storage implications

Uncompressed non-lossy file formats take up a lot more space than the alternatives. You will need to take this into account when budgeting for storage.

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 always difficult and expensive, and may in some cases be impossible. For example, you can never recover information lost by storing data using a lossy image, sound or video format.

Further reading

'File Formats for Long-Term Access' from the MIT Libraries' Data Management and Publishing guide: <http://libraries.mit.edu/guides/subjects/data-management/formats.html>

The US Library of Congress 'Sustainability of Digital Formats' site: <http://www.digitalpreservation.gov/formats/>

Digital Curation Centre's *Digital Curation Manual*, Instalment on File Formats: <http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats/>

'Obsolescence: File Formats and Software', from the ICPSR Digital Preservation Management Tutorial: <http://www.icpsr.umich.edu/dpm/dpm-eng/oldmedia/obsolescence1.html>

The UK Data Archive's 'Data Formats Table' page of optimal data formats that are used for long-term preservation of data: <http://www.data-archive.ac.uk/create-manage/format/formats-table>

Format registries

The UK National Archives' PRONOM format database: <http://www.nationalarchives.gov.uk/PRONOM/>

The Harvard-based Global Digital Format Registry: <http://www.gdfr.info/>

The proposed Unified Digital Format Registry, which will bring the two together: <http://www.udfr.org/>

Further Information

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

