ANDS Guide

AUSTRALIAN NATIONAL DATA SERVICE

File Formats

Level: Awareness

Last updated: 10 January 2017

Web link: **www.ands.org.au/guides/file-formats**

In simple terms, a file format describes the way information is organised in a computer file. File formats apply to documents, images, audio files, video files and research data sets for example .doc or .pdf. It is important that organisations implement data management policies that conform to standards that manages risk of file format obsolescence or degradation of information storage. A comprehensive list of differing file formats can be found by a web search using the key term 'file formats list'.

A key related concept is Preservation of research data: ANDS Introduction to Preservation

Choosing a suitable file format for data preservation and sharing is vital for the sustainability of future access and reuse of that data. This may require careful analysis of the advantages of proprietary or open standards software to ensure that access, reuse and future storage of the data meets future reuse of that data stored.

Institutional planning implications

File format types should ideally be considered and decided upon *before* the commencement of data collection. e.g. Information lost by storing data using a lossy image, sound or video format cannot be recovered. Migrating data from an unsuitable format to a more sustainable option is always difficult and expensive, and may in some cases be impossible. Uncompressed non-lossy file formats take up a lot more storage space that needs to be taken into account when budgeting for storage.

- University of Western Australia: Research Data Preservation Formats
- University of Sydney: File Formats
- Monash University: Durable Formats

Tools to manage file formats

- FIDO (Format Identification for Digital Objects): command-line tool to identify the file formats of digital objects, and is designed for simple integration into automated workflows
- BitCurator Access: open-source software that supports the provision of access to disk images Webinar on using BitCurator
- Apache Tika: toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF)
- BWFMetaEdit: free, open source tool that supports embedding, validating, and exporting of metadata in Broadcast WAVE Format (BWF) files

File format obsolescence

File formats can become obsolete for various reasons:

- Software / file formats are upgraded and the new version no longer works with the old version
- Software that supports the format is bought out by a competitor and withdrawn
- Format falls into disuse or no-one writes software to support/implement it
- Format is no longer compatible with current software or is not backwards compatible with older software

The result of this obsolescence means that it may no longer be possible to access the file, read the file or reuse the data, either entirely or partially. Risks also emerge for users if the software required resolving the format is restricted or the developer changes licensing or costed use of that software.

The ICPSR Digital Preservation Management Tutorial provides a useful overview of obsolescence for file formats and software.

Migration

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 preserve the entire environment needed to access and/or use the data. This approach either involves:

1. maintaining old computer hardware, together with the operating system and all the required software, or
2. writing/obtaining special emulation software that recreates the software-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 common examples of proprietary formats are: AutoCADs DWG drawing format, the MP3 MPEG Audio Layer 3 format and Adobe Photoshop's PSG native image format. Most proprietary formats are closed, meaning that the neither definition nor development of the format is available to the public.

This means that data stored in the format can only be accessed using the format owner's software. Some formats are both open and proprietary e.g. Adobe PDF Microsoft OOXMLAn *open format* is one where the description of the format is available to the public.

Open formats typically are developed and maintained by communities of interest. Examples include:

1. Standard image formats: JPEG 2000, PNG and SVG
2. For text: ASCII, PDF, Open Document Format and Office Open XML format (the native format for recent versions of Microsoft Word)
3. For the web: HTML, XHTML, RSS and CSS
4. NetCDF for some scientific data

Lossy and Lossless formats

A lossless format retains the original detail of the data file e.g. TIFF for images.

A lossy format discards information permanently in order to reduce the scale and size of file in effect lowering the quality of that data e.g. JPEG for images

"Lossy" compression is a data encoding method that compresses data by discarding (losing) some of it. The procedure aims to minimize the amount of data that needs to be held, handled, and/or transmitted by a computer. Images become progressively coarser as the data that made up the original one is discarded (lost). Typically, a substantial amount of data can be discarded before the result is sufficiently degraded to be noticed by the user.

For audio file formats, the ubiquitous MP3 format is lossy, while WAV format is lossless. The implications of re-saving or converting data from one format to another becomes apparent when the quality of that data is compromised in quality due to this removal of information.

Note - metadata such as file title, description, date etc. is not removed during this process.

Compression

Compression refers to ways of making data take up less storage space without losing any of the content. For long-term preservation, uncompressed formats are less risk prone.

A lossless file that has been compressed can be restored to its original state, completely unchanged. In the case of lossy formats the reduction in size is achieved effectively by throwing select data away (losing).

The compression process makes data more susceptible to "bit-rot". Bit rot is the small electric charge of a bit as memory disperses, possibly altering program code. The risk is that a change of one bit in a compressed text file may cause major changes across the entire document, rendering it useless. The advantage of retaining a low-resolution lossy format file set is quick navigation or ease of transmission.

The importance of standards

Standard file formats are essential for effective data sharing. In many cases a research discipline will have a mandatory or preferred standard for saving and storing research data e.g. SPSS data files for social science data sets and FITS ('Flexible Image Transport System') which is a standard data format used in astronomy.

Retaining multiple formats

Retaining multiple formats and instances of data may add to the scale of data being stored or sync difficulties however doing so can reduce the risk of loss of the original high-resolution file.

An alternative to keeping multiple formats is to use content management system software eg. Alfresco 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.

Future file formats

Development of file formats in the near future will likely incorporate information that pertains to geospatial platforms and environments.

Mobile technologies and the onset of virtual reality based data creation mean there are a number of consortia such as the Open Standards for Real-Time 3D Communication that conform with the International Organisation for Standardization .

The onset of geospatial data also presents new challenges as raster, vector and grid formats develop alongside other formats such as Worldfile used for geo-referencing a raster image such as a JPEG or BMP file. The importance of organisations to remain vigilant of, and responsive to, future proofing format sustainability is a critical consideration for the access, re-use and compatibility of data.

An introduction to geospatial resources and formats.

Preservation formats and display formats

High-resolution data e.g. a lossless uncompressed bitmap may require conversion to a .jpeg format for ease of visualisation online or transmission via email messaging. Another example is a standard XML format, which is best rendered to HTML or PDF for ease of viewing or printing purposes

Consideration must therefore be made for the long-term preservation of data taking into account the storage, display, visualisation, conversion or re-use of data.

The US Library of Congress Sustainability of Digital Formats website provides a dynamic overview of preservation and sustainability of digital formats.

Further information

ANDS Guides and other Resources

ANDS file wrangling further reading

Feedback?

We welcome your feedback on this guide. Please email contact@ands.org.au with any comments or questions.

About ANDS

The Australian National Data Service (ANDS) makes Australia's research data assets more valuable for researchers, research institutions and the nation.
ANDS is a partnership led by Monash University in collaboration with the Australian National University (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). It is funded by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative