Sustainable File Formats for Electronic Records
A Guide for Government Agencies

Electronic records are produced and kept in a wide variety of file formats, often dictated by the type of software used to create and access a record. Accessibility and user convenience are also common factors that determine the use of one format over another. When dealing with electronic records that have retention requirements past their initial use, however, one must also take into consideration the sustainability of the format used.

Sustainability in this context refers to continued accessibility over time. For example, will a given electronic record be available for users in 10 years? What about 20? Fifty? While no file format can guarantee perpetual accessibility, certain formats have distinct advantages over others in this regard. These formats are often referred to as "sustainable" formats. Sustainable formats often include the below features:

1) Published Documentation and Open Disclosure: Specifications for the format are published and accessible to the public. This means that anyone who wants to create tools to work with the format can do so with no restrictions of copyright. Formats that share these characteristics are commonly called "open-source" or "non-proprietary." Because anyone can create tools to access such formats, they have a low chance of becoming inaccessible in the future, even if the formats themselves become obsolete.

2) Widespread Adoption and Use: The more widely a format is used, the more likely it is to have multiple tools used to access and manipulate it. This reduces the chance of a format becoming inaccessible due to one software publisher going out of business. Widespread adoption also serves as an indicator of general format stability and as a safeguard against loss of accessibility. A wider user base means more stakeholders who have a vested interest in keeping a format going.

3) Self-describing Formats: These formats contain metadata (data about the data) within their structure that interprets the content, context and structure of the file. This means that descriptive information (e.g., the file name, date of creation, identification of data within the file,) can be kept within the file itself, and external documentation is not required. When discussing long-term preservation this is particularly important, since records often become disassociated from their original software environment and accompanying files. The more self-contained a format is, the better the chances of the data contained within being accessible down the road.

4) Unencrypted Files: Electronic records with long-term retention should not be encrypted in any way, as this can severely compromise the future accessibility of those records. Encryption methods change dramatically over time, and the specific software tools needed to access current encrypted records may not exist in the future. A good electronic records management system can handle security, restricting access to records as needed, while leaving the records themselves unchanged.

Below is a list of formats currently recommended for long-term preservation by the Illinois State Archives. The list may be updated or expanded as technology warrants, so be sure to check for newer versions in the future. Questions regarding these or other formats should be directed to Robert Boots, Records Management Section, at 217-782-1082 or rboots@ilsos.net.

Printed by authority of the State of Illinois. September 2015 — 1 — AR D 156

Text

Best Choice:

PDF/A (Portable Document Format / Archives): A variant of PDF that is specifically aimed at long-term preservation, its specifications are published in the standard ISO 19005-1:2005. It sacrifices certain functions, such as the ability to have external hyperlinks or embed audio or video, for the sake of greater reliability. The most notable different between PDF and PDF/A is the latter's ability to embed all necessary fonts within the file itself. This makes the file totally self-extracting, without any need to access external font information to properly present the formatting of the document. PDF/A also embeds descriptive metadata within the file itself, making it self-describing. These two factors make PDF/A the preferred format for long-term preservation of textual electronic records, both born-digital and digitized. Files can be converted to PDF/A by a number of different software tools and plug-ins to existing word-processor software.

Other Options:

PDF (Portable Document Format): A format commonly used to present formatted, page-oriented documents. PDFs can contain text, images, graphics, video and audio, as well as hyperlinks to outside documents. Originally created by Adobe Systems as a propriety format, the source code for PDF and its variants have since been made freely available, making it an open-source format. PDF is widely adopted around the world. Some later versions of PDF can include self-describing metadata. PDFs are acceptable for short to medium-term storage, but are not suitable for long term (20+ years) or permanent preservation. For long-term applications the PDF/A variant is preferred.

XML (Extensible Markup Language): A standard format for structured documents and data on websites, XML is also a preferred format for the preservation of metadata associated with records. XML is maintained and developed by the World Wide Web Consortium (W3C), but is open-source. XML enjoys nearly universal adoption and can be accessed and worked on by scores of freely available software tools. XML is self-describing, but requires association with an appropriate schema (also freely available) in order to properly render all formatting.

HTML (Hypertext Markup Language): A standard format for structured documents and data on websites currently maintained and developed by the World Wide Web Consortium (W3C). HTML is open-source and universally adopted. Unlike XML, HTML does not contain descriptive metadata headings. This limits the machine-readability of HTML, particularly when attempting to perform advanced search functions within files.

Plain Text: The most basic form of text file, Plain Text can be rendered by any software that can read text, across any platform. Plain Text renders only basic characters, spaces and punctuation; however, it does not preserve formatting such as italics or bold letters. It is therefore typically used only for relatively small amounts of information such as software instructions or short notes. Plain Text is open-source and universally adopted. Common file extensions for Plain Text include .txt and .text.

ODF (OpenDocument Format): An XML-based file format used for spreadsheets, charts, presentations and word processing documents. ODF was developed by Sun Microsystems, but is an open format, is freely available to anyone and has been published as an ISO standard

(ISO/IEC 26300:2006). Owing to its relatively recent creation (2005) ODF is not as widely adopted as some other formats, but it is supported by almost all current office suites and word processing programs. File extensions for ODF files vary depending upon the specific type of file, but include .odt (word processing), .ods (spreadsheets) and .odp (presentations).

Still Images

Best Choice:

TIFF (Tagged Image File Format): TIFF was initially created in the 1980s in an effort to standardize file formats created by commercial scanners. The format has gone through a number of revisions since then, becoming an international standard for electronic images. The format is currently owned by Adobe Corporation, but the specifications are open and freely available. Unlike many image file formats, TIFF is uncompressed. This means that the files are larger than a compressed format (such as JPEG) but there is no loss of data. This ensures that the file can be reproduced over time at its full fidelity. TIFF files can contain "tags" that store descriptive metadata about the file. TIFF files may have a file extension of .tif (Windows) or .tiff (Macintosh).

Other Options:

JPEG 2000 (Joint Photographic Experts Group): JPEG-2000 was created by the Joint Photographic Experts Group in 2000 as a next-generation format for electronic images. The format is part of an international standard: ISO/IEC 15444:2004. JPEG-2000 files can be compressed in either lossy or lossless fashion, although only the lossless variety is acceptable for long-term preservation. The format is still relatively new, and thus does not have the same wide-spread use as TIFF. This makes it a slightly riskier choice for preservation, although usage of the format is growing. The lossless compression of JPEG 2000 provides some space savings over TIFF, but it may be better suited as a format for access rather than preservation. The standard file extension for JPEG 2000 is .jp2.

PNG (Portable Network Graphics): A file format initially created with the approval of the World Wide Web Consortium (W3C) as a replacement to GIF (Graphics Interchange Format). PNG is most often used to present images on the web, and can be accessed with a wide variety of web browser and image display software. PNG uses a "lossless" compression algorithm which reduces the size of the file without losing any data. This means that images in PNG format do not suffer from "generation loss," where the quality of an image suffers over time with repeated use. Specifications for PNG are open and freely available, and the format can contain extensive metadata within its structure.

Spreadsheets

Due to the complexity of spreadsheet structure it is challenging to perfectly represent data over time. Different software uses varied means to record formulae and link data, and so advanced functions are not always replicable in more open formats. The below formats represent the best approach for long-term accessibility, but both may be unable to represent certain formatting or functions of spreadsheets originally created in formats such as Microsoft's XLS. Agencies may

want to save copies of spreadsheets with long-term retention in both the native format and in one of the below. This redundant method can preserve the maximum functionality of the spreadsheet while still protecting the core data from format obsolescence.

CSV (Comma Separated Values): A simple format which can be used to represent spreadsheet data. CSV files can be accessed with any spreadsheet software or text editor, but at the cost of potential loss of advanced functionality enjoyed by more proprietary spreadsheet formats. There is therefore a tradeoff with using CSV: universal interoperability is excellent for long-term preservation, but the loss of advanced formulae may compromise the core data of the record. Basic spreadsheets containing tabular data without advanced functions may be better served by CSV than others.

ODF: (See previous entry for general data on ODF) The spreadsheet format of ODF, .ods, is a good choice for preservation of spreadsheets, as it supports more advanced functionality than CSV. However, spreadsheets originally created in other formats such as XLS may suffer some functionality loss upon conversion to ODF due to the non-standardized methods by which different software execute formulae.

Audio

Best Choice:

BWF (Broadcast WAVE Format): A variant of the WAVE format, BWF (sometimes called BWAVE) was developed by the European Broadcasting Union with long-term preservation in mind. BWF takes the existing WAVE file structure and adds additional metadata support. The specifications for BWF are open and freely available, and the format is a *de facto* standard for digital audio for those in the radio, motion picture and television industries. It is also used extensively by audio archives throughout the world. The format is self-describing, as it contains its own structural and descriptive metadata. BWF files are uncompressed, and can be played by any software that is WAVE compatible. In order to display, add or modify metadata in a BWF file, however, one must use software that specifically supports the format. Free software is available that can attach BWF metadata to existing WAVE files. The file extension for BWF is .wav, the same as standard WAVE files.

Other Option:

WAVE (Waveform Audio File Format): WAVE is a format created by Microsoft and IBM in the early 1990s. Though proprietary, the format is fully documented and has been used as the basis for the preservation-oriented variant BWF (see above entry). WAVE files are uncompressed, so they lose no audio data as with some other audio formats. The format also enjoys near-universal adoption, as it is compatible with virtually every audio player available, across computer platforms. Software utilities to convert other formats to WAVE are plentiful and inexpensive (or free). WAVE has limited metadata capabilities, so is a second choice for long-term preservation behind BWF (see above). WAVE can still be an acceptable format for non-permanent audio, provided that appropriate external metadata is associated with the WAVE files.

Video

Whereas best practices typically dictate that only uncompressed formats be used for preservation of electronic content, the area of video preservation becomes more complex. Uncompressed video can take up huge amounts of space in a storage environment, and thus formats utilizing “lossless” or “near-lossless” compression have become more acceptable in some cases. Compression of these types utilizes algorithms to reduce the size of a file without irrevocably losing any data. This can be compared to “lossy” compression, which sacrifices some data to achieve smaller size. Lossy compression is unacceptable for long-term preservation because it permanently alters the structure of digital content and can lead to gradual reduction in quality over time.

MPEG-4 (Motion Picture Experts Group): MPEG-4 is an open-standard format developed by the Motion Picture Experts Group as a format for encoding video content for dissemination on the web. There are two main encoding versions, and numerous subcategories, of the format. Documentation for all varieties of MPEG-4 is extensively published as part of an international standard: ISO/IEC 14496-14:2003. The compression of a given MPEG-4 video file will depend upon the specific software and coding used in its creation and can range from lossy to lossless. For long-term preservation only lossless or near-lossless compression should be used. MPEG-4 supports the embedding of descriptive metadata to help support future access. A number of software tools, both free and paid for, are available to convert existing video files to MPEG-4 format.

Motion JPEG 2000 (Joint Photographic Experts Group): Motion JPEG-2000 is a derivative of JPEG 2000 which codes and displays video. The format is part of an open international standard: ISO/IEC 15444-3:2004. Motion JPEG-2000 files can be compressed in either lossy or lossless fashion, although only the lossless variety is acceptable for long-term preservation. The format is still relatively new, so adoption is not yet as widespread as older video formats. A number of software tools are available that can convert other video formats into Motion JPEG-2000, and it can support a variety of descriptive and structural metadata. File extensions for the format are .mj2 and .mjp2.