Acceptable File Formats
Guide

Types and explanation for why they're ranked where they are on the supported listed. (created by ADAH staff member, June 2014)

Text Formats

- **Portal Document Format for Long-Term Preservation (PDF/A)** – PDF/A is a family of ISO standards for constrained forms of Adobe PDF intended to be suitable for long-term preservation of page-oriented documents for which PDF is already being used in practice. The PDF/A standards are developed and maintained by a working group with representatives from government, industry, and academia and active support from Adobe Systems Incorporated. PDF/A attempts to be device independent, self-contained, and self-documented. It does not allow audio/video content, Javascript, and encryption. In addition, all fonts must be embedded, color spaces must be specified in a device-independent manner, and standards-based metadata must be used. Adobe does offer support but it is not guaranteed over time as Adobe could disappear.
  - Illinois State 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. Preferred text preservation format.

- **Plain Text** – from Illinois State Archives 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, and 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. It's an acceptable preservation format for text.

- **Extensible Markup Language (XML)** – is a simple, flexible text format derived from SGML. XML documents fall into two broad categories: data-centric and document-centric. Data-centric documents are those where XML is used as a data transport. Examples include sales orders, patient records, directory entries, and metadata records. One significant use of data-centric XML is for manifests (lists) of digital content; another is for metadata embedded into digital content files. Document-centric documents are those in which XML is used for its SGML-like capabilities, reflecting the structure of particular classes of documents, such as books with chapters, user manuals, newsfeeds and articles incorporating explicit metadata in addition to the text. An XML document's markup structure can be defined by a schema language and validated against a definition in that language.

  - Illinois State Archives -- 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. Other preservation option for text.

- **ODF (OpenDocument Format)** — from Illinois State Archives 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). Preferred text preservation format.

- **Portable Document Format (PDF)** – owned and developed by Adobe. PDF represents formatted, page-oriented documents that can be structured or simple and may contain text, images, graphics and other multimedia content, such as video and audio. It supports annotations, metadata, hypertext links, and books marks. A final state format for delivery to end users, as it freezes the file and is convenient for downloading and printing. Not preferred for images. Full documentation is available at no cost.
  - Illinois State Archives 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.

- **Rich Text (.rtf)** – Bentley Historical Library lists it as a sustainable format, North Carolina Archives has it as accepted but not recommended for long term retention

- **Hypertext Markup Language (HTML)** — from Illinois State Archives 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 is 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.

- **Standard Generalized Markup Language (SGML)** – from KDLA, a common markup language used in government offices worldwide, is an international standard. XML and HTML are types of SGML.

- **Microsoft Word (.doc)** – Bentley Historical Library lists it as an at-risk format that should be converted to a more sustainable format (such as MS World Open XML or PDF)

- **WordPerfect** – mainly only ADAH uses in house. Not recommended file format

Image Formats

- **Tagged Image File Format (TIFF, TIF)** – A tag-based file format for storing and interchanging raster images. TIFF serves as a wrapper for different bitstream encodings for bit-mapped (raster) images. The most recent version is 6.0, published in 1992. TIFF is widely recommended for most digital reproductions. It is a stable, widely supported by applications, and fully documented format. A disadvantage is its file size.
    - **From Illinois State Archives Best choice for preservation of still images** – 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).

- **Portable Network Graphics (PNG)** – Defines both a datastream and an associated file format for a lossless, portable, compressed, raster (bit-mapped) image. Originally designed as an open standard to replace GIF. Documented ISO/IEC 15948:2004. Australia’s national archives and Canada’s national library and archives have adopted it as an acceptable format for images. Good support.

  - Illinois State Archives – 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

- **JPEG** – Widely adopted. TIFF is preferred image preservation format. JPEG is good for access because it’s smaller. Lossy compression can cause problems over time.

- **JPEG2000** – ISO VIEF 15444-1:2004. Compression encoding generally used for full color and grayscale continuous-tone pictorial images. Open standard and good support. Was supposed to be the standard to replace TIFF (???) but hasn’t happened, not widely used.
  - Illinois State Archives – format 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. Not Illinois’ preferred format but will accept it.

- **Graphics Interchange Format (GIF)** – is a bitmapped image format widely used on the Web. Proprietary standard, openly described, widely adopted. Ok for end-user delivery but not for preservation.

- **Microsoft Windows Bitmap Format (BMP)** – is a simple raster graphics image file format designed to store bitmap digital images independently of a display device, originally and primarily on Microsoft Windows and OS/2 operating systems. Used primarily for icons, screen grabs and other purposes within Windows. Not for end-user delivery or preservation. Proprietary, documented as part of programmer support for users of Microsoft.

- **Adobe Digital Native (DNG)** – file format for storing and interchanging camera raw images, usually accompanied by JPEG secondary versions of the image. Many refer to it as an extension of TIFF 6 and its compatible with the TIFF/EP standard. Fully documented.

- **Multi-resolution Seamless Image Database (MrSID)** (.sid) – is a patented, wavelet-based file format designed to enable portability of massive bit-mapped (raster) images. The format employs discrete wavelet transformations (DWT) in a seamless fashion on tile subsets of the image data and stores the wavelet coefficients in a data structure that supports efficient retrieval of the data needed to generate a specified rectangular zone of the image at a chosen spatial resolution. The data structure of a MrSID image is a set of bitplanes designed to support 'transactions' of image data by extracting and delivering exactly and only those bitplanes necessary to construct a view according to desired scene, scale, or image quality, independent of bandwidth constraints. The format was designed to enable instantaneous viewing and manipulation of imagery both locally and over networks without sacrificing quality. New features in Generation 3 of the image format include lossless compression, multiple images in a composite file, and support for selective optimization and decoding by scene or region. Most often a middle-state or final-state format. *Neither preferred nor accepted as a master format by LC.*

- **FlashPix (.fpx)** – Bentley Historical Library lists it as an at-risk format that should be converted to a more sustainable format.

- **Photoshop (.psd)** – Bentley Historical Library lists it as an at-risk format that should be converted to a more sustainable format.

- **Camera Raw (RAW)** – Collective description of a group of proprietary formats employed by digital cameras that use a color filter array or Fovean sensor to capture image data. Raw files contain data captured by the sensors in the array. The data generally receives a very modest amount of processing before being output by the camera. The production of useable images requires additional processing after the data files have been transferred to a computer. The proprietary nature of raw formats means that there is a risk that any given format will not be supported for the long term, especially if the manufacturer goes out of business. LC has some experience but doesn’t prefer it as a file format.

Graphic Formats

- **Scalable Vector Graphics (SVG)** – a language for describing two-dimensional graphics in XML. It allows for three types of graphic objects: vector graphic shapes (paths consisting of straight lines and curve), raster graphics (raster images), and text. Can be used for animations but little evidence of it in practice. Open standard from W3C, documentation is public. Good support.

- **Computer Graphics Metafile (CGM, WebCGM. .cgm)** – developed by the International Standards Organization (ISO) and the American National Standards Institution (ANSI). Open, platform-independent format for the exchange of raster and vector data for technical applications. It is made for use in web browsers. Info from NARA.

- **Encapsulated Postscript (EPS)** – Bentley Historical Library lists it as an at-risk format that should be converted to a more sustainable format.

- **Macromedia Flash (SWF)** (.swf) – Binary file format that delivers vector graphics (especially animations) and other data types, including bitmapped video over the internet to the Flash Player. Developed by Macromedia, Owned by Adobe. Used for final-state, end-user delivery. Fully documented. Not transparent, proprietary binary format. Good support.

Video Formats

- **MPEG-4** – is generally an end-user delivery format, though may also serve as a middle-state format. Open standard, developed through ISO technical program JTC 1/SC 29 for coding of audio, picture, multimedia, and hypermedia information by Working Group 11, aka the Motion Picture Expert 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. From Illinois State Archives.

- **Ogg Theora (.ogg)** – subtype of Ogg File Format. Wrapper format for Vorbis sound data and various other audiovisual bitstreams developed by the Xiph open source project. The Xiph Web site calls the format a container and also and encapsulation format. Fully documented, open source, patent-free. Not the preferred LC format for master copy (LC prefers WAVE_LCPM for master copies).

- **Audio Video Interleaved (AVI)** (.avi) – File format for moving image content that wraps a video bitstream with other data chunks, eg audio. Often a middle-state format, the video source when producing lower-res streaming versions, sometimes a final state format for end user delivery. Fully documented, proprietary format. *Not LC's choice for either access or preservation formats.*

- **QuickTime Movie (.mov)** – File format that wraps video, audio, and other bitstreams. Typically a final state format for enduser delivery, sometimes a middle-state format, eg the source when producing lower-resolution streaming versions. LC's experience is limited to video, American Memory has produced QuickTime files for web service since the mid-1990s. Not LC's preference (prefer MXF_OP1a_JP2_LL for uncompressed or losslessly compressed video). Fully documented. Proprietary format developed by Apple.

- **Motion JPEG2000** – Usually a middle-state (video production) format; after editing, the bitstream is typically compressed in another encoding. The underlying coding algorithms are well documented through the JPEG standards activity. However, additional information about use and wrappers is either informal or exists as proprietary implementations. LC preference and does use the format in its holdings.
    - 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. Illinois State Archives.

- **MPEG-2** – The family of MPEG-2 encodings were initially developed to serve the transmission of compressed television programs via broadcast, cablecast, and satellite, and subsequently adopted for DVD production and for some online delivery systems. Picture, sound, and data elements consist of streams, i.e., the format's sequences of encoded bytes. LC's National Audio Visual Conservation Center has received significant numbers of MPEG-2 files for their holdings. Used as the reformatting format for VHS tapes and other media/formats in the American Folklife Center Veterans History Project. Open standard, developed through ISO technical program for coding of audio, picture, multimedia and hypermedia information by the Motion Pictures Expert Group. Used as a preservation format.

- **MPEG-1** – ISO/IEC 11172. Information technology – coding of moving pictures and associated audio for digital storage media. Generally a final state (end-user delivery) format. Used extensively for access copies of American Memory. Open standard. Developed through ISO technical program.

- **Windows Media Video (.wmv)** – File format based on ASF (Advanced System Format) that wraps a video bitstream. Often final state format for end user delivery; sometimes a middle-state format, such as a high-quality video representation for archiving or as a source when producing lower-resolution streaming versions. Not a LC preference. Microsoft's proprietary ASF wrapper format is fully documented.

- **RealVideo (.rv)** – Bentley Historical Library has it as an at-risk format that needs to be converted to a more sustainable one.

- **Material Exchange Format (MXF)** – object-based file format that wraps video, audio, and other bitstreams, optimized for content interchange or archiving by creators and/or distributors, and intended for implementation in devices ranging from cameras and video recorders to computer systems. The digital equivalent of video tape. Used as archival masters by the Library of Congress National Audio-Visual Conservation Center. Open standard developed by the Society of Motion Picture and Television Engineers, a member of the American National Standards Institute (ANSI).

Audio Formats

- **Audio Interchange File Format AIFF (PCM) (.aif, .aiff)** – File format for sound that wraps various sound bitstreams, ranging from uncompressed waveform to MIDI. Used primarily as an initial or middle state format including use as a master file for audio captured live digitally or reformatted from analog sources. Fully documented, developed by Apple. Not a preference of LC.

- **Waveform Audio File Format (WAVE)** – file format for audio that can incorporate an audio bitstream with other data chunks. Used for content in initial, middle, and final states. Fulle documented. Proprietary format developed by Microsoft and IBM as part of the Resource Interchange File Format (RIFF) for Windows 3.1 with documentation free available.
  ~ 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. From Illinois State Archives
    - **Wave Audio File Format with Linear PCM bitstream (WAVE_LPCM)** – used for initial, middle and final stages. When reformatting analog sound recordings, LC uses the Broadcast WAVE format, wrapping LPCM is used as the archival master format for mono and stereo audio at the Packard Campus for Audio Visual Conservation and by the American Folklife Center.
    - **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. Illinois State Archives' best choice.

- **MP3 (.mp3)** – De facto file format for sound, generally used for final state and end user delivery. Used extensively by LC as non-streaming service format for American Memory. Currently one of the accepted formats for electronic registration of audio by the US Copyright Office. Not documented.

- **Musical Instrument Digital Interface, Standard MIDI (.mid, .midi)** – Bitstreaming encoding format for MIDI messages that in can be thought of as instructions which tell a music synthesizer how to play a piece of music. May be used by composers or arrangers for initial state activities, in middle state exchange of data or archiving, or for final state, end user delivery. Not used or preferred by LC. Fully documented.

- **MPEG4** ISO 14496-1:2003 The second MPTEG-4 file format developed by the Motion Pictures Experts Group. The format's object-based design defines a set of tools that present binary coded representation of individual audiovisual objects, text, graphics, and synthetic objects. Format is intended to serve web and other online applications; mobile devices; and broadcasting and other professional applications. Generally, an end-user delivery, final state format. Open Standard.

- **Advanced Audio Coding (.m4a, .mp4, .aac)** -- Formal name (from MPEG-2 documentation that specifies both the AAC_MP2 bitstream and the ADIF file structure.): ISO/IEC 13818-7:2003. Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC). Common names: AAC and ADIF. Bitstream encoding and file format designed for efficient distribution of sound files over moderate bandwidth connections; may be used at higher data rates for better fidelity. ADIF stands for Audio Data Interchange Format and consists of a brief header that precedes AAC data in a file. Note that the compression approach in AAC_MP2 (used in this format) was subsequently refined as AAC_MP4, which requires a different decoder. Generally used for final-state, end-user delivery. LC no experience with it, prefers Broadcast WAVE format. Open standard.

- **SUN Audio (uncompressed) (.au)** – Bentley Historical Library places it as an at-rosk file format and suggests reformatting to a sustainable one

- **Free Lossless Audio Codec (.flac)** – Open source bitstream encoding format designed for lossless compression of LPCM audio data with many of its default parameters tuned to CD-quality music data. Generally used for final state – end user delivery. LC has no experience and doesn't prefer it.

- **Audio Interchange File Format AIFF** – sometimes referred to as AIFC (.aifc) – format for sound that wraps various sound bitstreams, ranging from uncompressed waveform to MIDI. Used primarily as an initial or middle state format, including use as a master file for audio captured live digitally or reformatted from analog sources. LC no experience and not a preferred format.

- **Windows Media Audio (.wma)** – file format based in Advanced Systems Format (ASF) that wraps an audio bitstream. Usually a final state format for enduser delivery; sometimes a middle-state format, e.g. a high-quality audio representation for archiving or as a source when producing lower-resolution streaming versions. Not LC's preference for recorded sound (prefer WAVE_LCPN.) Within the WMA family, WMA_WMA9_LL (lossless codec) is preferred. Microsoft's proprietary wrapper (ASF) is fully documented.

Spreadsheet Database Formats

- **Comma Separated Values (.csv)** – a simple format for representing a rectangular array (matrix) of numeric and textual values. It is a flat format. It is a delimited data format that has fields/columns separated by the comma character and records/rows/lines separated by characters indicating a line break. May be used at any stage in the lifecycle of a dataset. LC doesn't have experience and doesn't prefer it. It's a de facto format for which no single, official specification exists.
    - Illinois State Archives – 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.

- **OpenDocument Format (.ods)** – 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. (this is all from Illinois State Archives)

- **Delimited Text (.txt)** – recommended by North Carolina Archives

SQL DDL

- **dBase Table File Format (DBF)** – File format used originally by the dBASE database management system to store tables of data and later adopted by similar DBMS packages. The file format is best suited for fixed-field data. LC has no experience as it hasn't collected databases. It's a proprietary format with only the current version documented through copyright deposit.

- **Open Office XML OOXML (ISO/IEC DIS 29500)** -- OPC, Open Packaging Conventions, defines a generic "container" format designed to contain a collection of files (termed "parts" in the OPC specification) that represent a single logical whole. LC has no comment on it. International Open Standard.

- **Microsoft Excel (.xls)** – acceptable for transfer but not recommended for long term preservation by North Carolina Archives

Presentation Formats

- **OOXML (ISO/IEC DIS 29500) (.xlsx)** see above in spreadsheet database formats

- **PowerPoint (.ppt)** – not recommended as a sustainable format by North Carolina Archives.

- **OpenDocument Presentation (.odp)** – recommended sustainable format by North Carolina Archives.

Web Archiving Formats

- **Web Archive File Format (WARC)** – specifies a method for combing multiple digital resources into an aggregate archival file together with related information. Revision of the ARC File Format of the Internet Archive. Open standard, publically documented, developed under the International Internet Preservation Consortium. ISO TC46/SC4/WG12. Supported through Internet Archive's Wayback Machine.

E-Mail

- **Electronic Mail Format (.eml)** – a file extension for an email message saved to a file in the internet message format protocol for electronic mail messages. It is the standards format used by Microsoft Outlook Express as well as some other email programs. The files comply with industry standard RFC 5322 and, therefore, can be used with most email clients, servers, and applications. Partially documented but specific documentation is not readily available. LC appears to not have experience with it.

- **.pst** – an open proprietary data file format used to store local copies of messages, calendar events, and other items within Microsoft software. PST_ANSI earlier version, now it's PST_UNICODE

- **Mailbox File (MBOX)** – preferred sustainable format by Bentley Historical Library