Draft
Not for citation or quotation without written permission of the Kentucky Department for Libraries and Archives.

Recommended Data Formats for Preservation Purposes in the KDLA Digital Archive

The following table represents the digital formats that KDLA has recognized and is encouraging agencies to use when transferring records to the archives. These formats may also be used by agencies when maintaining records with long-term retention (retention period of more than 10 years) in-house. The formats, and corresponding confidence levels, represent KDLA's preferences for long-term preservation. Agencies are free to use other formats (including those not listed) for active business use as long as they meet with state approved standards and architecture. However, systems employed by agencies should support these formats or be able to export records to these formats.

The confidence levels identified in the table below are ranked from High (most conducive for long-term preservation) to Low (least conducive for long-term preservation.) The confidence levels are determined by a combination of sustainability factors including:

1. **Disclosure.** Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. Non-proprietary, open standards are usually more fully documented and more likely to be supported by tools for validation than proprietary formats. However, what is most significant for this sustainability factor is not approval by a recognized standards body, but the existence of complete documentation, preferably subject to external expert evaluation.
2. **Adoption.** Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.
3. **Transparency.** Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.
4. **Self-documentation**. Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.
5. **External Dependencies.** Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.
6. **Impact of Patents.** Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.
7. **Technical Protection Mechanisms**. Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository.

The High and Medium confidence levels represent the formats that KDLA feels are the most sustainable over time. Agencies should avoid using formats listed in the Low confidence field, or make sure that the records in question can be converted to the formats in the Medium and/or High fields.

Media	High Confidence Level	Medium Confidence Level	Low Confidence Level	Notes/Comments
Text	- Plain text (encoding: US ASCII, UTF-8, UTF-16 with BOM) - PDF/A-1 (*.pdf) - XML (XSD/XSL/XHTML, etc.; with included or accessible schema and character encoding explicitly specified)	- Plain text (ISO8859-1 encoding) - PDF (*.pdf) (embedded fonts) - Rich Text Format (*.rtf) version 1.x - OpenOffice (*.sxw) - **Microsoft Word (*.doc)*** - WordPerfect (*.wpd)# - HTML 4.x (include a DOCTYPE declaration) - SGML * MS Office is the state approved standard and supported by the state. # WordPerfect is the federal court standard.	- PDF (external font) - All other text formats not listed here - DjVu (alternative format to PDF. Uses a different compression to make a smaller file. Published standard. Created by AT&T; owned by producers of MrSID GIS format. – Used by USGS and other GIS and Washington State Digital Archives	
E-mail	- Plain Text - Outlook Message format (*.msg) Any of the High Confidence level text formats listed above.	- Any of the Medium Confidence text formats listed above - Outlook Archive (*.pst) For general correspondence maintained in the agency with proper backup and security	-	A sub-type of text file formats.

		controls.		
Raster Image	- TIFF (uncompressed) - PNG (*.png) - JPEG (raw)?	- BMP (*.bmp) - **JPEG/JFIF (*.jpg)** - JPEG2000 (prefer uncompressed) (*.jp2, *.jpx) - **TIFF (CCITT Group 3/4,** JPEG, PackBits compression)	- MrSID (*.sid) - TIFF (with LZW compression or in Planar format) - GIF (*.gif) - FlashPix - PhotoShop (*,psd) - All other raster image formats not listed here	- “Raw” JPEG are those images that have not been resized. - Depends on compression format. - Uncompressed is obviously better than some – lossless better than lossy.
Vector Graphics	- SVG 1.1 (*.svg)	- CGM - WebCGM - DWF * * AutoCAD is the state approved product.	- - Encapsulated PostScript (EPS) - - Macromedia Flash (*.swf) - - All other vector image formats not listed here	
Audio	- AIFF (uncompressed) (*.aif, *.aiff) - WAVE (LPCM only) (*wav)	- Standard MIDI (*.mid, *.midi) - Windows Media Audio (*.wma) * - MP3 (MPEG 1/2, Layer 3) (8.mp3) - SUN Audio (uncompressed) (*.au) - Ogg Vorbis (*.ogg)	- AIFC (*.aifc) - NeXT SND (*.snd) - RealNetworks ‘Real Audio’ (8.ra, *.rm, *ram) - WAVE (compressed) (*.wav) - All other audio formats not listed here	- MP3 is a non-documented compressed version of MPEG – the bare MPEG is open (v. 1 & 2 are ISO standards)

		*Same as Word files, Windows is the supported state standard.		
Video	- **MPEG-1, MPEG-2 (*.mpg, *.mpeg)** - Motion JPEG2000 (*.mj2) - AVI (*.avi) (uncompressed) - Motion JPEG (*.avi, *.mov)	- Ogg Theora (*.ogg)	- AVI (compressed) (*.avi) - QuickTIme Movie (*.mov) - MPEG 4 (*.mp4) - RM (RealNetworks; ‘Real Video’) (*.rv) - Windows Media Video (*.wmv) - All other video formats not listed here	-MPEG v. 1 & 2 are open ISO standards but the compression types vary.
Spreadsheet Database	- Delimited Text (*.txt, *.csv) - SQL DDL	- DBF (*.dbf) - OpenOffice (*.sxc) - **Excel (*.xls)*** * Excel part of the MSOffice group and supported by state architecture standards.	- All other spreadsheet/database formats not listed here	
Presentation	-	- OpenOffice (*.sxi) - PowerPoint (*.ppt)	- All other presentation formats not listed here	

Notes:
1. File formats listed under Low Confidence Level will be converted to a High or Medium Confidence format or preserved at the bit level only.
2. Fully or partially encrypted files must be unencrypted prior to transfer to KDLA.
3. Password protected files must be opened with protections removed prior to transfer to KDLA.
4. Any files produced with Digital Right Management controls must have all controls removed prior to transfer.
5. As a general rule, use platform independent, vendor independent, nonproprietary, stable, open and well supported formats.