SERI
State Electronic
Records Initiative
COUNCIL OF STATE ARCHIVISTS

CoSA
Council of State Archivists

CoSA's File Format Guidelines

These guidelines are intended to offer insight into file format decision-making processes for state, territory, and the District of Columbia archives. Archives are encouraged to use this document to inform their own requirements documents for file format preferences and policies.

What Is a File Format?

A file format specifies how the computer uses bits to encode information for storage in a computer file. File formats can be designed for specific kinds of data or designed to generally house a wide variety of data types. File formats often have specifications describing the methods of encoding. Some file formats have proprietary specifications and some have open source specifications. There are widely circulated, freely published specifications and there are formats for which the specifications are never published (and some which never have specifications written at all). A file format is typically identified by a file extension that the computer uses to quickly identify the contents of the file and what software to use in opening the file. Examples of common file extensions include: .doc, .jpg, .gif, .pdf, .html, and many others.

How Should I Evaluate File Formats for Preservation?

Factors that influence a specific file format's suitability for long-term preservation may include:

Complexity—The more easily a file format can be understood and managed, the less risk it poses to long-term accessibility.

Adoption—If a file has been widely adopted by users throughout the world, there will be more tools and resources to ensure that file format's longevity. If the file format is not used worldwide, but has wide usage in the archival community, the community will more likely have determined ways to continue to support it.

Transparency—File formats that allow the file to be identified and have its contents checked support the work necessary to preserving electronic records throughout their lifecycle.

Documentation—There is a spectrum of documentation for file formats: fully documented, partially documented, little-to-no documentation. The more quality documentation about a file format, the more information practitioners have regarding how to manage the files.

Interoperability—File formats that function with a variety of services, hardware, and software are easier to maintain over time.

Openness—If there are few restrictions on interpretation of the file format it will be easier to preserve that file format. Digital rights protections, licensing, patents, and intellectual property issues complicate preservation. If the specifications to the file format are in the public domain it increases the ability for archivists to manage the format throughout the preservation lifecycle.

Why do File Formats Matter for Preservation?

File formats can impact how successfully content can be archived, preserved, and used in the future. Some file formats are well suited to preservation because they are more likely to remain readable over time, remain compatible through technological change, and best maintain the qualities of the content. Depending on the priorities of the records (is look-and-feel most important? Or functional use of the data contained in the record?) some file formats may be more suitable than others.

What Are Common File Format Choices for Preservation?

This table is an aggregation of common file format choices for preservation based on content type (i.e. text, image, audio, GIS, etc.). Additional information can be found on file format choices in materials listed in the Additional Resources section of this document. Archives can use this table and the evaluation factors listed above to develop their own lists of formats they prefer to support for long-term preservation.

Content Type	Preferred formats	Acceptable formats
Text and word processing documents	PDF/A PDF	RTF TXT
Spreadsheets and structured data	PDF/A PDF CSV	Tab delimited TXT XML with schema
Digital images	Uncompressed TIFF	JPG PNG JPG2 DNG
Audio files	WAV	AIFF FLAC
Video	MOV AVI Motion JPEG 2000	MPEG-4
Websites and social media	WARC	ARC Content management system file
Email	XML email preservation format	PST
Database Management Systems (DBMS)	XML with schema CSV	Keep original file format
Geospatial	SHP DBF GeoTIFF NetCDF	Keep original file format
CAD	PDF/A PDF/E PDF with original file	Keep original file format

Find additional resources about file formats and electronic records at the Council of State Archivists (CoSA) website, www.statearchivists.org.

Additional Resources:

For more information about file formats and their sustainability, visit:

National Archives and Records Administration Tables of File Formats

Library of Congress Sustainability of Digital Formats

Federal Agencies Digital Guidelines Initiative

Smithsonian

Digital Preservation Coalition