

File Formats

Summary

Rapid changes in technology mean that file formats can become obsolete quickly and cause problems for your records management strategy. A long-term view and careful planning can overcome this risk and ensure that you can meet your legal and operational requirements.

Legally, your records must be trustworthy, complete, accessible, legally admissible in court, and durable for as long as your approved records retention schedules require. For example, you can convert a record to another, more durable format (e.g., from a nearly obsolete software program to a text file). That copy, as long as it is created in a trustworthy manner, is legally acceptable.

The software in which a file is created usually has a default format, often indicated by a file name suffix (e.g., *.PDF for portable document format). Most software allows authors to select from a variety of formats when they save a file (e.g., document [DOC], Rich Text Format [RTF], text [TXT] in Microsoft Word). Some software, such as Adobe Acrobat, is designed to convert files from one format to another.

Legal Framework

For more information on the legal framework you must consider when developing a file format policy, refer to the *Introduction* and Appendix D of the *Trustworthy Information Systems Handbook*. Also review the requirements of:

- Official Records Act [Minnesota Statutes, Chapter 15.17] (available at: <<http://www.revisor.leg.state.mn.us/stats/15/17.html>>), which mandates that government agencies must keep records to maintain their accountability and specifies that the medium must enable the records to be permanent. It further stipulates that you can copy a record and that the copy, if trustworthy, is legally admissible in court.
- Records Management Act [Minnesota Statutes, Chapter 138.17] (available at: <<http://www.revisor.leg.state.mn.us/stats/138/17.html>>), which establishes the Records Disposition Panel to oversee the orderly disposition of records using approved records retention schedules.
- Minnesota Government Data Practices Act (MGDPA) [Minnesota Statutes, Chapter 13] (available at: <<http://www.revisor.leg.state.mn.us/stats/13/>>), which mandates that government records should be accessible to the public unless categorized as not-public by the state legislature.

- Uniform Electronic Transactions Act (UETA) [Minnesota Statutes, Chapter 325L] (available at: <<http://www.revisor.leg.state.mn.us/stats/325L>>) and Electronic Signatures in Global and National Commerce (E-Sign), a federal law (available at: <<http://thomas.loc.gov/cgi-bin/query/z?c106:S.761:>>). Both UETA and E-Sign address the issues of the legal admissibility of electronic records created in a trustworthy manner and the application of the paper-oriented legal system to electronic records.

Key Concepts

As you consider the file format options available to you, you will need to be familiar with the following concepts:

- Proprietary and non-proprietary file formats
- File format types
- Preservation: conversion and migration
- Compression
- Importance of planning
- File format decisions and electronic records management goals

Proprietary and Non-proprietary File Formats

A file format is usually described as either proprietary or non-proprietary:

- *Proprietary formats.* Proprietary file formats are controlled and supported by just one software developer.
- *Non-proprietary formats.* These formats are supported by more than one developer and can be accessed with different software systems. For example, eXtensible Markup Language (XML) is becoming an increasingly popular non-proprietary format.

File Format Types

Below are brief descriptions of the basic files you are likely to encounter. You can use the resources in the Annotated List of Resources for more detailed information on specific file formats. Basic file format types include:

- *Text files.* Text files are most often created in word processing software programs. Common file formats for text files include:
 - Proprietary formats, such as Microsoft Word files and WordPerfect files, which carry the extension of the software in which they were created.

- RTF files, which are supported by a variety of applications and saved with formatting instructions (such as page layout).
- Portable Document Format (PDF) files, which contain an image of the page, including text and graphics. PDF files are widely used for read-only file sharing. However, only Adobe Acrobat can make a PDF file, and Acrobat is necessary for reading a PDF file.
- *Graphics files.* Graphics files store an image (e.g., photograph, drawing) and are divided into two basic types:
 - Vector-based files that store the image as geometric shapes stored as mathematical formulas, which allow the image to be scaled without distortion. Common types of vector-based file formats include:
 - Drawing Interchange Format (DXF) files, which are widely used in computer-aided design software programs, such as those used by engineers and architects
 - Encapsulated PostScript (EPS) files, which are widely used in desktop publishing software programs
 - Computer Graphics Metafile (CGM) files, which are widely used in many image-oriented software programs (e.g., Photoshop) and offer a high degree of durability
 - Raster-based files that store the image as a collection of pixels. Raster graphics are also referred to as bitmapped images. Raster graphics cannot be scaled without distortion. Common types of raster-based file formats include:
 - Bitmap (BMP) files, which are relatively low-quality files used most often in word processing applications
 - Tagged Image File Format (TIFF) files, which are widely usable in many different software programs
 - Graphics Interchange Format (GIF) files, which are widely used for Internet applications
 - Joint Photographic Experts Group (JPEG) files, which are used for full-color or gray-scale images
- *Data files.* Data files are created in database software programs. Data files are divided into fields and tables that contain discrete elements of information. The software builds the relationships between these discrete elements. For example, a customer service database may contain customer name, address, and billing history fields. These fields may be organized into separate tables (e.g., one table for all customer name fields). You may convert data files to a text format, but you will lose the relationships among the fields and tables. For example, if you convert the information in the customer database to text, you may end up with ten

pages of names, ten pages of addresses, and a thousand pages of billing information, with no indication of which information is related.

- *Spreadsheet files.* Spreadsheet files store the value of the numbers in their cells, as well as the relationships of those numbers. For example, one cell may contain the formula that sums two other cells. Like data files, spreadsheet files are most often in the proprietary format of the software program in which they were created. Some software programs can import and export data from other sources, including software programs designed for such data sharing (e.g., Data Interchange Format [DIF]). Spreadsheet files can be exported as text files, but the value and relationship of the numbers are lost.
- *Video and audio files.* These files contain moving images (e.g., digitized video, animation) and sound data. They are most often created and viewed in proprietary software programs and stored in proprietary formats. Common files formats in use include QuickTime and Motion Picture Experts Group (MPEG) formats.
- *Markup languages.* Markup languages, also called *markup formats*, contain embedded instructions for displaying or understanding the content of the file. The World Wide Web Consortium (W3C) (<http://www.w3c.org>) supports these standards. Common markup language file formats include the following:
 - Standard Generalized Markup Language (SGML), a common markup language used in government offices worldwide, is an international standard.
 - Hypertext Markup Language (HTML) is used to display most of the information on the World Wide Web.
 - Extensible Markup Language (XML) is a relatively simple language based on SGML that is gaining popularity for managing and sharing information.

Table 1 summarizes the common file formats.

Table 1: Common File Formats

File Format Type	Common Formats	Sample Files	Description
Text	PDF, RTF, TXT, proprietary formats based on software (e.g., Microsoft Word)	Letters, reports, memos, e-mail messages saved as text	Created or saved as text (may include graphics)
Vector graphics	DXF, EPS, CGM	Architectural plans, complex illustrations	Store the image as geometric shapes in a mathematical formula for undistorted scaling
Raster graphics	TIFF, BMP, GIF, JPEG	Web page graphics, simple illustrations, photographs	Store the image as a collection of pixels which cannot be scaled without distortion
Data file	Proprietary to software program	Human resources files, mailing lists	Created in database software programs
Spreadsheet file	Proprietary to software program, DIF	Financial analyses, statistical calculations	Store numerical values and calculations
Video and audio files	QuickTime, MPEG	Short video to be shown on a web site, recorded interview to be shared on CD-ROM	Contain moving images and sound
Markup languages	SGML, HTML, XML	Text and graphics to be displayed on a web site	Contain embedded instructions for displaying and understanding the content of a file or multiple files

Preservation: Conversion and Migration

Your most basic decision about file formats will be whether you want to convert and/or migrate your file formats. If you convert your records, you will change their formats, perhaps to a software-independent format. If you migrate your records, you will move them to another platform or storage medium, without changing the file format. However, you may need to convert records in order to migrate them to ensure that they remain accessible. For example, if you migrate records from a Macintosh operating system to a Microsoft Windows operating system, you need to convert the records to a file format that is accessible in the new one (e.g., RTF, Word 2000). For more information on conversion and migration, refer to the *Electronic Records Management Strategy* and *Long-Term Preservation* guidelines.

You will face three basic types of loss determining your course of action:

- *Data*. If you lose data, you lose, to a varying degree, the content of the record. Bear in mind that, legally, your records must be complete and trustworthy.
- *Appearance*. You also risk loss of the structure of the record. For example, if you convert all word processing documents to RTF, you may lose some of the page layout. You must determine if this loss affects the completeness of the record. If the structure is essential to understanding the record, this loss may be unacceptable.
- *Relationships*. Another risk is the loss of the relationships of the data in the file (e.g., spreadsheet cell formulas, database file fields). Again, this loss may affect the legal requirement for complete records.

Keep in mind that a copy of a record is legally admissible only if it is created in a trustworthy manner and is accurate, complete, and durable.

Compression

As part of your strategy, you may choose to compress your files. The pros and cons are summarized in Table 2 below.

Table 2: Pros and Cons of File Compression

Pros	Cons
<ul style="list-style-type: none">• Saves storage space• More quickly and easily transmittable	<ul style="list-style-type: none">• May result in data loss• Introduces an additional layer of software dependency (the compression software)

The greatest challenge in compressing files is that you may lose data. Compression options vary in their degree of data loss. Some are intentionally “lossy,” such as the JPEG format, which

relies on the human eye to fill in the missing detail. Others are designed to be “lossless.” You may choose to compress some files and not others.

Importance of Planning

The challenges of preservation can be overcome with good planning. Use the resources in the Annotated List of Resources, and thoroughly discuss the issues raised in the Key Issues to Consider section, to weigh the specific pros and cons of each option for your agency. Review the decision tree in the *Guidelines on Best Practices for Electronic Information* white paper for preliminary planning and use the workbook in *Risk Management of Digital Information: A File Format Investigation* to assess your unique situation and risk.

File Format Decisions and Electronic Records Management Goals

The goals of electronic records management that may be affected by file format decisions include:

- *Accessibility.* The file format must enable staff members and the public (as appropriate under the MGDPA) to find and view the record. In other words, you cannot convert the record to a format that is highly compressed and easy to store, but inaccessible.
- *Longevity.* Developers should support the file format long-term. If the file format will not be supported long-term, you risk having records that are not durable, because the software to read or modify the file may not be available.
- *Accuracy.* If you convert your records, the file format you convert to should result in records that have an acceptable level of data, appearance, and relationship loss.
- *Completeness.* If you convert your records, the file format you convert to should meet your operational and legal objectives for acceptable degree of data, appearance, and relationship loss.
- *Flexibility.* The file format needs to meet your objectives for sharing and using records. For example, you may need to frequently share copies of the records with another agency, use the records in your daily work, or convert and/or migrate the records later. If the file format can only be read by specialized hardware and/or software, your ability to share, use, and manipulate the records is limited.

Key Issues to Consider

Now that you are familiar with some of the basic concepts of file naming, you can use the questions below to discuss how those concepts relate to your agency. Pay special attention to the questions posed by the legal framework, including the need for public accessibility as appropriate, completeness, trustworthiness, durability, and legal admissibility. Consider the degree of acceptable data, appearance, and relationship loss. Take a long-term approach so that your file formats will meet your operational and legal requirements now and in the future.

Discussion Questions

- What are our goals for electronic records management?
- How is our agency affected by the legal requirements?
- What current file formats do we use? Will the developer support these formats long-term?
- Are we planning on converting and/or migrating our records?
- What levels of data, appearance, and relationship loss are acceptable?
- What resources do we have for processing and maintaining records?
- How will our decisions affect other groups that may need current and future access to our records (e.g., other government agencies, the public)?

Annotated List of Resources

Primary Resources

Clausen, Lars R. *Handling File Formats*. Denmark: The State and University Library, The Royal Library, May 2004.

<<http://www.netarchive.dk/publikationer/FileFormats-2004.pdf>>

This report is a publication of the Netarchive.dk project, which seeks strategies for archiving the Danish part of the World Wide Web. The report offers a succinct and intelligent analysis of the issues surrounding file format preservation, including the categorization of formats, aspects of preservation quality, assessment criteria for future usability, and preservation strategies.

DLM Forum. *Guidelines on Best Practices for Using Electronic Information*. Luxembourg: European Communities, 1997.

<<http://dlmforum.typepad.com/>>

<<http://dlmforum.typepad.com/gdlines.pdf>>

This white paper was published by the DLM Forum, an organization of records management experts from the Member States of the European Union and the European Commission. The paper provides a basic overview of the file formats in use worldwide. Topics include the information life cycle; the design, creation, and maintenance of electronic records; short-term and long-term access; and accessing and sharing information.

Lawrence, G.W., W.R. Kehoe, O.Y. Rieger, et al. *Risk Management of Digital Information: A File Format Investigation*. Washington, D.C.: Council on Library and Information Resources, 2000.

<<http://www.clir.org/pubs/abstract/pub93abst.html>>

This publication provides an overview of file format issues related to records management strategies. The publication also provides a comprehensive workbook for users to help them develop a records management strategy.

Additional Resources

Electronic Recordkeeping Resources.

<<http://www.kshs.org/government/records/electronic/ermlinks.htm>>

This web site provides a comprehensive list of links to other Internet resources related to electronic records management. The site is managed by Cal Lee, who originally constructed it while employed at the Kansas State Historical Society. Topics include security, preservation, access, and technology infrastructure.

Minnesota Historical Society, State Archives Department. *Trustworthy Information Systems*

Handbook. Version 4, July 2002.

<<http://www.mnhs.org/preserve/records/tis/tis.html>>

This handbook provides an overview for all stakeholders involved in government electronic records management. Topics center around ensuring accountability to elected officials and citizens by developing systems that create reliable and authentic information and records. The handbook outlines the characteristics that define trustworthy information, offers a methodology for ensuring trustworthiness, and provides a series of worksheets and tools for evaluating and refining system design and documentation.

PRONOM: The File Format Registry.

<<http://www.nationalarchives.gov.uk/pronom/>>

PRONOM is maintained by the Digital Preservation Department of the UK National Archives. Visitors to the site can search within five areas (File Format, Product, Vendor, Support Period, and Release Date), each of which offer more options. Choosing "File Format," for instance, allows visitors to search just by extension to get a straightforward list of associated software or by compatible products, which returns a list of products, versions, release dates, vendors, read/write capabilities, and invariance. Links on vendor and product lead to a wealth of additional detail. Reports can be easily printed or exported into XML or CSV (Comma Separated Value file) for further use.

Wotsit's Format: The Programmer's Resource.

<<http://www.wotsit.org>>

This online catalog of file formats is broken down into categories such as "Graphics Files," "Text Files/Documents," and "Spreadsheet/Database." Visitors can browse each section or can use the provided search engine to zero in on their mark. Each format carries a one-line description and a link to further information either online or in a download file.

World Wide Web Consortium (W3C)

<<http://www.w3.org>>

W3C is a consortium of organizations around the world that develops and promotes common web protocols. The site contains news, specifications, guidelines, software, and tools for web development on a wide variety of topics, including markup languages and transfer protocols.