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, 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]). 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 selecting digital file formats, refer to the chapter *Records Management in an Electronic Environment* in the Electronic Records Management Guidelines and Appendix A6 of the *Trustworthy Information Systems Handbook*. Also review the requirements of the:

- Public Records Act [PRA] (*Code of Laws of South Carolina, 1976*, Section 30-1-10 through 30-1-140, as amended) available at www.scstatehouse.org/code/t30c001.htm, which supports government accountability by mandating the use of retention schedules to manage records of South Carolina public entities. This law governs the management of all records created by agencies or entities supported in whole or in part by public funds in South Carolina. Section 30-1-70 establishes your responsibility to protect the records you create and to make them available for easy use. The act does not discriminate between media types. Therefore, records created or formatted electronically are covered under the act.

Proprietary, Non-proprietary, Open Standard and Open Source File Formats

- *Proprietary formats*. Proprietary file formats are controlled and supported by just one software developer. Microsoft Word (.DOC) format is an example.
- *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.
- *Open Source formats*. In general, open source refers to any program whose source code is made available for use or modification as users or other developers see fit. Open source software may be developed, modified and distributed by independent software companies for profit. The Linux operating system is an example.
- *Open Standard formats*. Open standard software formats are created using publicly available specifications. Although software source codes remain proprietary, the availability of the standard increases compatibility by allowing other developers to create hardware and software solutions that interact with, or substitute for, other software. The Portable Document Format (.PDF) is based on an open standard.

File Format Types

There are hundreds of file formats used to encode digital information. Below are brief descriptions of the basic files you are likely to encounter. Use the resources in the Annotated List of Resources for more detailed information on specific file formats. Basic file format types include:

MORE →

- ◆ *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 or Rich Text* Format files, are supported by a variety of applications and saved with formatting instructions (such as page layout).
  - — *Portable Document Format (PDF)* files contain an image of the page, including text and graphics. PDF files are widely used for read-only file sharing and printing. Adobe Acrobat is, by far, the most popular PDF file although other types are available. Acrobat reader, available for no charge, is necessary for reading an Adobe 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
    - – Shapefiles (SHP), ESRI GIS applications use vector coordinates to store non-topological geometry and attribute information for features.
  - — *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 uncompressed, 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. TIFF files are either uncompressed or compressed using a lossless algorithm
    - – Graphics Interchange Format (GIF) files, which are widely used for Internet applications. GIF is a lossless compression format but is limited to 256 colors or less.
    - – Joint Photographic Experts Group (JPEG) files, which are used for full-color or gray-scale images. Used primarily for photographs, the standard JPEG format uses a lossy compression algorithm that discards some information to achieve a smaller file size.
    - – Portable Network Graphics (PNG) files. A lossless compression designed to replace GIF files. PNG is completely patent and license free and is of higher quality than GIF.
- ◆ *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.

MORE →

◆ *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, Motion Picture Experts Group (MPEG) formats and Real Video.

◆ *Markup languages*. Markup languages, also called *markup formats*, contain embedded instructions for displaying or understanding the content of the file. They provide the means to transmit and share information over the web. The World Wide Web Consortium (W3C) (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. HTML and XML are derived from SGML.

— Hypertext Markup Language (HTML) is used to display most of the information on the World Wide Web. Because presentation is combined with content through the use of pre-defined tags, HTML is simple to use but limited in scope. Other markup languages such as XHTML and XML offer greater flexibility.

— eXtensible Hypertext Markup Language (XHTML) combines the flexibility found in XML with the ease of use associated with HTML. Strict XHTML rules improve consistency and provide the ability to create your own markup tags. Because they share similar rules, converting XHTML into XML is easier than converting HTML into XML.

— eXtensible Markup Language (XML) is a relatively simple language based on SGML that is gaining popularity for managing and sharing information. XML provides even greater flexibility and control than XHTML while avoiding the complexities associated with SGML.

For additional information on file formats see the *Digital Imaging* guidelines.

MORE →

Table 1 summarizes the common file formats.

Table 1: Common File Formats (contains both proprietary and non-proprietary formats)

File Format Type	Common Formats	Example Applications	Description
Text	PDF, RTF, TXT, DOC, WPD	Letters, reports, memos, e-mail messages saved as text	Created or saved as text (may include graphics)
Vector graphics	DXF, EPS, CGM, SHP	Architectural plans, complex illustrations, GIS	Store the image as geometric shapes in a mathematical formula for undistorted scaling
Raster graphics	TIFF, BMP, GIF, JPEG, PNG	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, Real Networks, WMV, WAV, MP3	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, XHTML, 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 a Windows operating system (e.g., RTF, Word 2000).

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
Saves storage space	**May result in data loss**
More quickly and easily transmittable	**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. Review the decision tree on page 29 in the *Guidelines on Best Practices for Electronic Information* white paper for preliminary planning and use the CLIR workbook in *Risk Management of Digital Information: A File Format Investigation* to assess your unique situation and risk. Thoroughly discuss the "Suggestions for Better File Format Decisions" listed below, to weigh the specific pros and cons of each suggestion for your agency.

Suggestions for Better File Format Decisions

- *Accessibility*. The file format must enable staff members and the public 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. Records can be migrated or converted if you determine a file format is no longer supported. Open source, open standard and non-proprietary formats are preferable to completely proprietary ones.
- *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.

MORE →

Annotated List of Resources

Primary Resources

DLM Forum. *Guidelines on Best Practices for Electronic Information*. Luxembourg: European Communities, 1997.
europa.eu.int/ISPO/dlm/documents/guidelines.html

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.
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.
www-personal.si.umich.edu/~calz/ermlinks/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.

South Carolina Department of Archives and History. *Trustworthy Information Systems Handbook*. Version 1, July 2004.
www.state.sc.us/scdah/erg/tis.htm

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.

State of Australia. "Management of Electronic Records, 4.0 Electronic Records Format." In *Standard for the Management of Electronic Records*. Version 1.0. North Melbourne, Australia: State of Victoria, 2000.
www.prov.vic.gov.au/vers/standards/pros9907/99-7s4.htm
www.prov.vic.gov.au/vers/standards/pros9907/99-7toc.htm

This portion of the Australian standards for electronic records management summarizes the desirable characteristics of a file format and what the file format must be able to support. The second URL provides the table of contents for the entire electronic document that discusses all the standards.

World Wide Web Consortium (W3C)
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.

Cornell University. "Digital Preservation Management: Selecting Short Term Strategies For Long Term Problems"
www.library.cornell.edu/iris/tutorial/dpm/index.html

An online tutorial available from Cornell University Library. The tutorial provides basic information including terms and concepts related to digital preservation.