SIA

Guidance Update

November 2004

ELECTRONIC RECORDS

Recommendations for Preservation Formats

Highlights

Highlights of SIA_EREC_04_03, an update on electronic record preservation format guidance.

At a Glance

The SIA selection of preservation formats for electronic records follows current best practices in the area of electronic records management balanced with the electronic records and IT environments at the Smithsonian Institution.

This guideline update comprises the full recommendation with notation of the correspondence of the recommended formats to the SI ***Technical Reference Model***, Version 2.0.

Records (non-digital and digital) are commonly transferred to an archive once the records have become inactive. The formats of the records transferred are therefore likely to be out of alignment with the TRM because the TRM continues to be an active record. Identification of preferred preservation formats looks to bridge this disparity while applying the best practices of digital preservation.

This document describes the practices within SIA and serves as an educational tool and guidance to other SI units.

ELECTRONIC RECORDS

Recommendation for Preservation Formats

Overview

This policy document is intended for SI staff responsible for the organization and management of electronic records. It describes Smithsonian Institution Archives (SIA) guidelines regarding file formats used for the long term preservation of electronic records. All electronic records transferred to SI Archives *requiring permanent retention* will be handled according to the information contained in this document and related procedural documents. It is recommended, but not required, that electronic records which will be retained for a designated period ("temporary" records) also conform to these guidelines.

This document addresses file format concerns only. For guidance on the full set of practices necessary to ensure reliable, authentic electronic records are created and maintained by the SI units prior to transfer to SI Archives, please contact the IT Archivist (ferranter@si.edu or 202-357-1421 x45) at SI Archives.

General questions about record maintenance from SI units requiring assistance should be directed to contact OSIAREF@si.edu or call (202) 357-1420. An SIA archivist can assist staff to determine whether records should be permanently maintained in the archives, temporarily stored in the records center, or discarded on-site.

Contents

Abbreviations

SI	Smithsonian Institution
SIA	Smithsonian Institution Archives
TRM	Smithsonian Institution ***Technical Reference Model***

Introduction

SI staff and volunteers use a wide variety of equipment and software in the course of creating electronic records. Digital preservation "best practices" recommend specific file formats for long term archival use. This policy document outlines the formats preferred by SI Archives for long-term record preservation. The electronic record formats most commonly used at SI are described below along with the corresponding SIA primary (preferred) preservation format. Secondary preservation formats will be indicated where available and appropriate. Secondary formats are to be used only when the primary preservation format cannot be accomplished.

These formats have been chosen because of their documented acceptance by the information technology, archival, and digital record professional communities. Factors leading to this acceptance include format longevity, format maturity, level of use in relevant professional communities, incorporated primary and related information standards, and long term accessibility of any required viewing software.

Note: *Both primary and secondary preservation formats have been aligned with the SI Technical Reference Model Version 3in the great majority of the cases. It is important to keep in mind that SIA is obliged to preserve files which may not comply with the current version of the TRM, files which may have been created prior to the existence of the TRM. SIA will follow digital preservation best practices, seeking to align preservation formats with the TRM whenever possible.*

SI Archives will request SI units publish or finalize their electronic records in either a primary preservation format (preferred) or a secondary format prior to transfer to SI Archives. Ideally, this occurs when the record is originally created. Still, this does not always occur. In this case, SI Archives Electronic Records staff will attempt to preserve the records received by conducting the format conversion prescribed in this document on a copy of the received records. Successful transformations will be verified and subsequently archived in place of the original record.

Preservation Formats for Electronic Records

In accordance with best practices, SI Archives prefers to preserve transferred electronic records in the formats described in the table below. This table lists the original/creating application by its native format(s), the corresponding primary preservation format (preferred) and the secondary format. The secondary format will be used only when a record of sufficient quality cannot be created in the primary preservation format. A glossary of acronyms used is appended to the end of this document.

Where contributing SI units have prepared their electronic records in a preservation format, SIA will preserve those records in the received format unless extenuating circumstances apply. Records originally transferred to SIA in the secondary preservation format may be later converted to the appropriate primary preservation format by SIA staff as part of the long term maintenance of the record.

This document does <u>not</u> address related considerations and procedures required in the conversion from original formats to preservation formats. It is essential that individuals responsible for these activities refer to forthcoming format-specific Electronic Record publications as appropriate.

Original/ Creating Application	**Primary Preservation Format**	**Secondary Preservation Format**	**TRM References**
(various) .txt	ASCII (keep original extension)		
Corel WordPerfect .wpd, .wp*x*, .doc, .rtf	PDF 4.0 or higher	RTF (text)	Pg 4-202 (PDF) Pg 4-93 (RTF)
Microsoft Word .doc, .dot, .rtf	PDF 4.0 or higher	RTF (text)	Pg 4-202 (PDF) Pg 4-93 (RTF)
Lotus 1-2-3	*Same as original*	CSV	No reference
Microsoft Excel *.xls	PDF	CSV	Pg 4-202 (PDF)

Original/ Creating Application	Primary Preservation Format	Secondary Preservation Format	TRM References
Microsoft Powerpoint *.ppt	PDF		Pg 4-202 (PDF)
Graphics Applications			
(various graphics applications) *.bmp, *.tif, *.jpg, *.gif	TIFF (24-bit RGB) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)	JPEG (uncompressed) (for JPG and GIF formats only.)	Pg 4-98 (TIFF) Pg 4-101 (JPEG)
Adobe Photoshop, Illustrator .psd, .eps	TIFF (24-bit RGB) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)		Pg 4-98 (TIFF)
Macromedia Fireworks/ Freehand .png	TIFF (24-bit RGB) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)	TIFF (8-bit, grayscale or b/w for black and white images only) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)	Pg 4-98 (TIFF)

Original/ Creating Application	Primary Preservation Format	Secondary Preservation Format	TRM References
Microsoft Image Composer .mic	TIFF (24-bit RGB) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)	TIFF (8-bit, grayscale or b/w for black and white images only) (minimum 3000 pixels along long dimension or 400 dpi, whichever is greater)	Pg 4-98 (TIFF)
Database Management Systems (DBMS)			
Dbase III+ .dbf, .dbt	*Same as original*	CSV	CSV not listed Dbase not listed
Lotus Notes (database)	*Same as original*	CSV	CSV not listed Lotus Notes not listed
Microsoft Access (database) *.mdb	*Same as original*	CSV	Pg 4-42, 4-74, 10-7
Microsoft SQL Server *.bak	*Same as original, full backup made with MS SQL*	CSV	Pg 10-10 (MS SQL Server)
Sybase SQL Server, Adaptive Enterprise Server, SQL Anywhere	*Same as original, full backup made with Sybase dbms. Also requires full backup of Sybase master database.*	CSV	Pg 4-74 (Sybase)

Original/ Creating Application	Primary Preservation Format	Secondary Preservation Format	TRM References
WebPage Developers and HTML Editors			
Adobe PageMill, Microsoft Frontpage, Macromedia Dreamweaver (other various) .htm, .html, .shtml	XHTML 1.0	PDF (Isolated, single webpage only)	Pg 4-86 (XHTML) Pg 4-89
Active Server Page(various) .asp	XHTML 1.0	ASP (depending on parameter complexity)	Pg 4-86 (XHTML) Pg 4-25 (ASP)
Macromedia Cold Fusion .cfm	XHTML 1.0		Pg 4-86 (XHTML) Pg 4-33 (CFM)
Macromedia Flash/Shockwave .swf	SWF (No change)	*Analyze carefully for embedded URLs or other file dependencies.*	Pg 4-27, 4-30, 4-105, 10-15, 10-23
Real Media .rm, .ram	*Same as original*		Pg 4-27, 10-6
Architectural Design/Engineering Graphics			
AutoCAD *.cad	PDF (v6.0 with layer retention)		Pg 4-202 (PDF) Pg 10-15 (CAD)
Microsoft Visio *.vsd	PDF (v6.0 with layer retention)		Pg 4-202 (PDF) Pg 10-16 (Visio)

Glossary

Acronym	Description
8.3	The MS-DOS file naming convention of eight characters followed by a period (.) and three final characters. The three final characters are popularly used as acronyms for the file format of the electronic document. For example, “demo.ppt” is a Microsoft Powerpoint document. PPT would be the “.3” expression, or the acronym for a Powerpoint file.
ASCII	A text file where each character or space is represented by one byte encoded according to the ASCII (American Standard Code for Information Interchange) code.
ASP	Active Server Page. This web page format uses scripting, normally VBScript or JavaScript code in combination with HTML to dynamically generate a complete HTML page for display on the requesting web browser. The complete HTML is not generated until that page is requested by a web browser, allowing webmasters to deliver customized content without dramatically increasing the web content they must manage.
CFM	Cold Fusion template/page. Cold Fusion is a Macromedia web development application used to create dynamic web pages in a fashion similar to the ASP format developed by Microsoft.
CSV	Comma Separated Values. Another name for comma-delimited text format and usually the 8.3 extension value used for this type of format.
DPI	Dots per inch. A means of expressing the amount of information recorded in a digital image correlating to the resolution of the image.
JPEG	JPEG is a lossy compression technique for color images developed by the Joint Photographic Experts Group. File sizes can be reduced with a loss in detail.
JPG	Alternate representation of JPEG.
PDF	Portable Document Format developed by Adobe Systems.
RGB	Red, Green, Blue components of a color TIFF image.

RTF	Rich Text Format. A format standard which embeds basic formatting instructions in an essentially ASCII document. Margins, font style, indentation and other formatting instructions are supported.
SWF	Shockwave file format used by Macromedia's Flash player application. An increasingly popular plug-in, or supplemental application, used with web-browsers. Such a file is commonly referred as a Flash component.
TIFF	Tagged Image File Format. Very popular format for storing bit-mapped images. Supports black-and-white, grayscale, and color images. Orders the bytes of the image file in either Intel (PC) order or Macintosh order. SIA TIFF records use Intel byte order.
URL	Universal Resource Locator. The "address' of an Internet-accessible document. Most frequently begins with 'http://…' but also includes 'ftp://…' and 'telnet://…' Unfortunately, the longevity of a given URL has been documented to be only six months, on the average, in today's Internet culture.
XHTML	Extensible Hypertext Markup Language. This information standard essentially expresses HTML code in an XML syntax. XHTML 1.0 has been recognized by the Internet-related vendors as the successor to HTML 4.0 and is the equivalent of the most recently adopted HTML 4.1 protocol.
XML	Extensible Markup Language. A flexible text format derived from the Standard Generalized Markup Language (SGML) to meet the challenges of large-scale electronic publishing. Its uses have since spread into database environments as well.