FILE FORMATS GUIDELINE

Date Effective: December 1, 2007
Approved: November 28, 2007 (Last Revised: February 24, 2010)

Introduction

The purpose of this guideline is to provide guidance and advice to public agencies in the selection and use of file formats that support the interoperability and long-term preservation of government records.

Background

The accuracy and authenticity of a record is dependent on a number of factors, including: (1) the permanence of the record's content and structure; (2) reliable access for as long as the record is required to be retained; and (3) a trustworthy practice or system of recordkeeping.

Electronic records, unlike paper records, are composed largely of two parts: the format, which dictates how the data in the record is stored and displayed; and the actual content and information. Both computer hardware and software are required to process the format and the content and present the record in a human-understandable manner.

Due to technological obsolescence, maintaining accurate and authentic electronic records is a daunting task for most, if not all, public agencies. Therefore it is important to capture electronic records in file formats that better support the preservation of and access to the record's format and content. In addition, capturing records in a recommended format shortly after the time of creation helps ensure that the records are:

1. Complete and without loss of context with other records;
2. Reliable evidence of agency business activities; and
3. Well positioned for future conversions or migrations.

Furthermore, capturing records into appropriate file formats *at* the time of creation: reduces loss of context; produces more reliable records; and is less expensive than retrospective conversion processes. (Agencies using this guideline for imaging should first review the *Imaging Guideline for All Public Agencies*.)

Related State Standards

- *Records Management Best Practice for All Public Agencies* (October 2008)
- *File Formats Best Practice for All Public Agencies* (November 2007)
- *Imaging Guideline for All Public Agencies* (November 2008)
- *Recordkeeping Metadata for All Public Agencies* (October 2008)
- *Information Security Best Practice for All Public Agencies (May 2009)*

Intended Audience

This guideline is intended for any public employee involved in the creation of electronic records which may need to be: (1) shared among other public agencies or the citizens of Vermont or (2) preserved for long term. This guideline is also intended for any public employee responsible for the interoperability or preservation of records.

1 SELECTING A FILE FORMAT

1.1 Business and Recordkeeping Requirements

The selection of a file format should be based on business and recordkeeping requirements, which are documented in strategic plans, project charters, record schedules, and related sources. If business requirements have not been cross-referenced with recordkeeping requirements, this should be done prior to selecting a file format.

1.2 Objective

The file format selection process requires a clear objective. For example, if the objective is simply to increase access and publish State records and information electronically, then it is important to select a file format that is adaptable and designed for increasing access. On the other hand, if the objective is to preserve records electronically (regardless of the records' original format) then it is necessary to choose a file format that also supports long-term access and preservation. If both electronic access and preservation is the objective, it will be necessary to choose one format that is suitable for access and preservation or two formats: one for access and one for preservation.

1.3 Recordkeeping Requirements

It is important to review the records' appraised values and retention requirements prior to converting any files to verify that the appropriate file format is selected.[1] Also, to ensure that records remain legally viable at all times, agencies should fully articulate their plans and document their processes. Audit trails and the systematic application of established policies and procedures will be necessary to demonstrate the accuracy and authenticity of converted records.[2] Procedures should address quality control when preparing records for conversion and verifying and validating the converted files.

1.4 Original Formats

In addition to business and recordkeeping requirements, the selection of a file format is also dependent on the format of the original records. More than 93% of records are created electronically; therefore, agencies should strive to capture their electronic records in a recommended file format rather than printing and scanning records.

[1] If your agency has updated its records program, these values are listed in your agency's record schedule. If your agency has not recently updated its program, these values may be determined by analyzing existing disposition orders. Contact your agency's records officer to obtain appraised values.
[2] For records that contain signatures, agencies should be aware that forensic analysis of signatures is not possible with scanned records that are converted to images.

2 ADDITIONAL CONSIDERATIONS

There are several other issues, such as management needs and associated costs, which need to be considered and addressed beyond selecting an appropriate file format.

2.1 Cost-Benefit Analysis

A cost-benefit analysis should be completed before choosing to implement any conversion plan. In conducting the analysis, tangible savings and benefits and intangible benefits should be equally considered.

2.2 Workflow

In most situations, the records targeted for conversion support current agency functions and activities. Therefore the need to readily and easily convert future records after they are created and received should be introduced into the agency's workflow. Issues to consider include: routing procedures; interfiles and adding new records to existing files; and when records should be captured and converted.

2.2 Computer Hardware and Software

An inventory of current computer hardware and software will offer an overview of what new software or hardware may be necessary on a temporary or long-term basis to convert records or maintain the converted files. When considering alternatives, such as the use of contractors, the inventory will be an invaluable resource.

2.3 Indexing

The ability to retrieve converted records is dependent on how well the records are indexed. It is important to consider indexing by multiple access points (creator name, date of creation, agency name, subjects, record title, etc.) to increase access and retrieval by internal and external users. In addition, agencies should consider embedding metadata in each converted file; records should not be fully dependent on external indexes, such as databases, for contextual information. (See state standard on *Recordkeeping Metadata for All Public Agencies*.)

2.4 Staffing

It is inevitable that workflow patterns or current practices will change, therefore training needs for current staff must be considered. Agencies should also reflect on the need for additional staff or specialized expertise.

3 PREFERRED FILE FORMATS

Below are preferred file formats for capturing, accessing, and preserving records based on original media and appraised value. (Specifications for each format are outlined in *Section 4: Specifications*.) These preferences do not preclude the use of microfilm for preservation proposes. Recommended formats for other records are evolving; contact the Vermont State Archives and Records Administration for assistance.

3.1 Original: Paper Textual Records (Note: Review *Imaging Guidelines*)

Appraised Value	CAPTURE	ACCESS COPY	PRESERVATION
Permanent (Archival)	✓ TIFF; ✓ PDF/A; or ✓ JPEG 2000	✓ JPEG 2000; ✓ PDF; or ✓ PDF/A	✓ JPEG 2000; ✓ TIFF; or ✓ PDF/A
Long-term (Non-Archival)	✓ TIFF; ✓ PDF/A; or ✓ JPEG 2000	✓ PDF; ✓ PDF/A; or ✓ JPEG 2000	✓ TIFF; ✓ PDF/A; or ✓ JPEG 2000
Temporary (Non-Archival)	✓ TIFF; ✓ PDF; or ✓ JPEG 2000	✓ PDF ✓ JPEG 2000	*Not applicable*

3.2 Original: Formatted, Styled Textual Records (such as .doc; .rtf; etc.)

Appraised Value	CAPTURE	ACCESS COPY	PRESERVATION
Permanent (Archival)	✓ PDF ✓ PDF/A	✓ PDF; or ✓ PDF/A	✓ PDF/A
Long-term (Non-Archival)	✓ PDF ✓ PDF/A	✓ PDF; or ✓ PDF/A	✓ PDF/A
Temporary (Non-Archival)	✓ PDF	✓ PDF	*Not applicable*

3.3 Other Records

Section 3.3 and subsequent sections are reserved for additional preferred file formats, such as those for spreadsheets, data, geo-spatial data, video, audio, markup languages, etc.

4 SPECIFICATIONS

4.1 Tagged Image File Format (TIFF)

General Requirements:

- ☐ For all records, regardless of appraised value, image files should conform to Baseline TIFF, as outlined in *TIFF Revision 6.0* (1992).
- ☐ Permanent (Archival) or Long-Term (Non-Archival) records should have one TIFF file for each page of the record, including blank pages for documents are that multi-paged and double-sided.
- ☐ For Permanent (Archival) or Long-Term (Non-Archival) records, compression should be set at "none."
- ☐ Embedded searchable text based on Optical Character Recognition (OCR) for TIFF files of scanned records is acceptable provided that the text is identical in content and appearance to the source document. OCR processes should not alter the original bit-mapped image of the TIFF file.

Specific Requirements:

Original Record	Resolution	Bit Depth
Clearly legible typed or laser printed records on a white background.	✓ 300-600 ppi[3] *600 ppi preferred*	✓ Bitonal (1-bit)
Handwritten records, carbon copies, and other records with poor legibility	✓ 300-400 ppi *400 ppi preferred*	✓ Gray scale (8-bit)
Textual records containing photographs or halftone illustrations.	✓ 300-400 ppi *400 ppi preferred*	✓ Gray scale (8-bit)
Textual paper records containing color that is necessary to retain to understand the context of the record.	✓ 300-400 ppi *400 ppi preferred*	✓ Color (24-bit RGB [Red, Green, Blue])

[3] Pixels per inch

4.2 JPEG 2000 (JP2)[4]

General Requirements:

- ☐ For all records, regardless of appraised value, image files should conform to Baseline JPEG 2000, as outlined in ISO/IEC 15444-1:2000: Part I.
- ☐ Permanent (Archival) or Long-Term (Non-Archival) records should have one JPEG 2000 file for each page of the record, including blank pages for documents are that multi-paged and double-sided.
- ☐ For Permanent (Archival) or Long-Term (Non-Archival) records, compression should be set at "lossless."

Specific Requirements:

Original Record	Resolution	Bit Depth
Clearly legible typed or laser printed records on a white background.	✓ 300-600 ppi[5] *600 ppi preferred*	✓ Bitonal (1-bit)
Handwritten records, carbon copies, and other records with poor legibility	✓ 300-400 ppi *400 ppi preferred*	✓ Gray scale (8-bit) or Color (24-bit RGB [Red, Green, Blue])
Textual records containing photographs or halftone illustrations.	✓ 300-400 ppi *400 ppi preferred*	✓ Gray scale (8-bit) or Color (24-bit RGB [Red, Green, Blue])
Textual paper records containing color that is necessary to retain to understand the context of the record.	✓ 300-400 ppi *400 ppi preferred*	✓ Color (24-bit RGB [Red, Green, Blue])

[4] Despite its name, JPEG2000 is actually not a file format but rather a compression technology.
[5] Pixels per inch

4.3 Portable Document Format (PDF)

General Requirements:

- ☐ All records, regardless of appraised value, may use PDF for access copies where applicable.
- ☐ PDF files should conform to PDF versions 1.0 through the latest version.
- ☐ All fonts within a file should be publically identified as legally embeddable fonts.
- ☐ The "Standard" conversion setting should be applied when converting original files to PDF and the options to "Add links," "Add bookmarks," and "Enable accessibility and reflow with Tagged PDF" should be enabled. The option to "Attach the source file" should be disabled.

Specific Requirements:

There are no specific requirements for this file format at this time.

4.4 Portable Document Format / Archive (PDF/A)

General Requirements:

- ☐ For all Permanent (Archival) and Long-Term (Non-Archival Records), files should conform, at a minimum, to PDF/A-1b Standard as defined in ISO 19005-1:2005 using one of the following standard configuration settings, unmodified: PDF/A-1b:2005 (RGB) or PDF/A-1b:2004 (CMYK). *Files must pass validation.*
- ☐ Permanent (Archival) or Long-Term (Non-Archival) records should have one file per record that captures each page in the record, including those left blank.
- ☐ Permanent (Archival) or Long-Term (Non-Archival) records should be self-contained and external dependencies are not acceptable. The option to "Attach source file" should be disabled.
- ☐ The option to "Enable accessibility and reflow with Tagged PDF" should be enabled.
- ☐ Files should use lossless compression algorithms that are open and not copyrighted. LZW compression should not be used.
- ☐ All fonts should be legally embedded. "Optimize for fast web-viewing" should be disabled.
- ☐ Embedded searchable text based on Optical Character Recognition (OCR) for TIFF files of scanned records is acceptable provided that the text is identical in content and appearance to the source document. OCR processes should not alter the original bit-mapped image of the TIFF file.

Specific Requirements:

Original Format	Tool
Microsoft Word[6]	✓ Adobe PDF Maker Plug-in for Microsoft Word ✓ 2007 Microsoft Office Add-in: Microsoft Save as PDF
Microsoft Office 2007 applications, including Word and Excel	✓ 2007 Microsoft Office Add-in: Microsoft Save as PDF
Other formatted, styled textual records	✓ Acrobat Professional, version 7.0.7 or later
TIFF (unaltered bitmap)	✓ Acrobat Professional, version 7.0.7 or later

[6] Microsoft Word documents do not pass PDF/A validation with Acrobat Professional.

REFERENCES

Adobe Developers Association. (1992). *TIFF: Revision 6.0: Final – June 3, 1992.* Adobe Systems Incorporated: Mountain View, CA. URL: http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf (last accessed: 2010-02-11)

Adobe Systems Incorporated. (2007). *PDF Reference and Related Documentation: April 2007.* Adobe Systems Incorporated: Mountain View, CA. URL: http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference.pdf (last accessed: 2010-02-11)

Buckley, Robert. (2009). JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library, ed. by Simon Tanner. Wellcome Trust Digital Library: London, England. URL: http://library.wellcome.ac.uk/assets/wtx056572.pdf (last accessed: 2010-02-11)

California Digital Library. (2009). CDL Guidelines for Digital Images, Version 2.0. University of California: CA. URL: http://www.cdlib.org/services/dsc/tools/docs/cdl_gdi_v2.pdf. (last accessed: 2010-02-10)

Chou, C. (2006). *Guidelines for Creating Archival Quality PDF Files*. Florida Center for Library Automation: Gainesville, FL. URL: http://www.fcla.edu/digitalArchive/pdfs/PDFGuideline.pdf (last accessed: 2010-02-11)

CENDI Digital Preservation Task Group. (2007). *Formats for Digital Preservation: A Review of Alternatives and Issues*. Federal STI Managers Group: Oak Ridge, TN. URL: http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf (last accessed: 2010-02-11)

International Organization for Standardization. (2003). *ISO/IEC 15444-1:2000: Information Technology – JPEG 2000 Imaging Coding System – Part I: Core Coding System*. International Organization for Standardization: Geneva, Switzerland.

International Organization for Standardization. (2005). *ISO 19005-1:2005: Document Management – Electronic Document File Format for Long-Term Preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*. International Organization for Standardization: Geneva, Switzerland.

JPEG Committee. (2009). JPEG 2000 (webpage). JPEG Committee website: [unknown location]. URL: http://www.jpeg.org. (last accessed: 2010-02-11)

JPEG 2000 Preservation File Format Work Group, Library and Archives Canada. (2008). *JPEG 2000 as a Preservation Format for Digital Raster Images at Library and Archives Canada*. Library and Archives Canada: Ottowa, Canada.

URL: http://www.collectionscanada.gc.ca/digital-initiatives/012018-2100.01-e.html (last accessed: 2010-02-11)

Library of Congress. (2007). *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Library of Congress: Washington, D.C. URL: http://www.digitalpreservation.gov/formats/index.shtml. (last accessed: 2010-02-11)

National Library of Norway. (2007). *Digitization of Books in the National Library*. Oslo: Norway. URL: www.nb.no/content/download/2326/18198/version/1/file/digitizing-books_sep07.pdf (last accessed: 2010-02-10)

National Security Agency. (2005). *Redacting with Confidence: How to Safely Publish Sanitized Reports Converted from Word to PDF: Report #I333-015R-2005*. National Security Agency: Fort Meade, MD. URL: http://www.fas.org/sgp/othergov/dod/nsa-redact.pdf (last accessed: 2010-02-11)

PDF/A Joint Working Group. (2006). *Frequently Asked Questions (FAQs): ISO 19005-1:2005: PDF/A-1, July 10, 2006*. AIIM: Silver Spring, MD. URL: http://www.aiim.org/documents/standards/19005-1_FAQ.pdf (last accessed: 2010-02-11)

VERSION HISTORY

Version 2.0	Version 2.0 of this guideline was approved by iSTART (Information Strategies: Taskforce on Archives, Records & Technology) on February 24, 2010 and is effective upon approval. This is a flexible guideline and can be modified by iSTART as appropriate to meet the needs of public employees and the State of Vermont. Revisions/Additions: **1. Related Standards** a. Related Policies section was changed to "Related State Standards" and section was revised to include references to additional State standards. **2. Preferred File Formats** a. JPEG 2000 was added as a preferred file format for images **3. References** a. References for JPEG 2000 were added and links to existing References were verified. One reference that is no longer available was removed.
Version 1.0	Version 1.0 of this guideline was approved by iSTART (Information Strategies: Taskforce on Archives, Records & Technology) on November 28, 2007 and is effective December 1, 2007. This is a flexible guideline and can be modified by iSTART as appropriate to meet the needs of public employees and the State of Vermont. Revisions/Additions: 1. **Related Policies** (2008-10-17) a. Revised to reflect changes to the *Records Management Best Practice* and *File Formats Best Practice*. 2. **Section 4.3** (2007-12-05): a. "*Files must pass validation"* was added at the end of the first General Requirement. b. *2007 Microsoft Office Add-in: Microsoft Save as PDF* was added as tool for Microsoft. c. Microsoft Office 2007 information as added.