FLORIDA VIRTUAL CAMPUS

PDF/A Validation and Conversion in Florida Digital Archive
September 19, 2013

Background

Portable Document Format (PDF) is one of the most popular document formats in existence today and is widely used in academic, government, corporations and various other institutions as a final document format. In an effort to ensure the long-term preservation for those PDF documents, International Standard Association (ISO) commissioned a working group to put together PDF for Archiving (PDF/A) standards. The first standard, PDF/A-1 with conformance level 1a and 1b, was published in December 2005 [1] with several additional standards, PDF/A-2 and PDF/A-3 with conformance level 2a, 2b, 3a and 3b, published a few years later. [2] [3]

Since 2005, the Florida Digital Archive (FDA) has archived roughly 255,000 PDFs, the versions of which range from 1.1 to 1.7. Though the FDA has been encouraging its affiliates to submit PDF documents conforming to the PDF/A standard,[4] [9] the majority of the PDFs in the FDA's archive are not validated as PDF/A-compliant for two main reasons: 1) many of the PDFs archived in the FDA were produced prior to the publishing of PDF/A standard, and it takes many years for PDF/A validation and conversion tools to be mature enough for an institution-wide adoption; and, 2) the DAITSS (Dark Archive In The Sunshine State) software's use of the built-in PDF/A-1 validation provided in JHOVE [5] does not parse the contents on streams, and cannot determine PDF/A conformance to the degree required by ISO 190005-1. (JHOVE, free open-source software produced by Harvard Digital Library, was adopted by the FDA as a format validation tool for various formats such as TIFF, JPEG, PDF, etc.). In an effort to provide better PDF/A validation, and also to convert existing PDFs into PDF/A format, FDA evaluated several software packages in the summer of 2012. [6] [7] Results of that evaluation showed that pdfaPilot is the best PDF/A validator and converter on the market today that also satisfies the FDA's requirements for a PDF/A validator and converter.

New Changes Now In Effect

Effective August 1, 2013, the Florida Digital Archive (a service managed by the Florida Virtual Campus) has purchased and incorporated pdfaPilot into its preservation software to deliver an ISO-compliant PDF/A validation and PDF to PDF/A conversion solution for all PDFs. This means that all incoming PDFs submitted to the FDA will be treated as follows:

1. If a PDF is identified as a PDF/A document with conformance level: 1a, 1b, 2a, 2b, 3a and 3b, the PDF document will be validated according to its declared PDF/A conformance level. Any non-conformance reported from pdfaPilot will be reported as a warning in its ingest report. For example:

    - **Anomaly**: pdfaPilot:PDF/A entry missing
    - **Anomaly**: pdfaPilot:Syntax problem: Indirect object "obj" keyword not followed by an EOL marker
    - **Anomaly**: pdfaPilot:XMP property not predefined and no extension schema present

2. If a PDF is not identified as a PDF/A document, the FDA will convert the PDF into a PDF/A-1b document by applying fixes to the PDF, such as embedding un-embedded fonts, converting device-dependent color spaces to device- independent ones, etc., and save the PDF/A-1b document as a normalized version. (Please note that the original PDF will remain unchanged and is always kept in the archive.)

3. Some PDFs may not be able to be converted to PDF/A document. Among the reasons that a PDF cannot be converted may include, but are not limited to:

    a. The PDF is encrypted. This includes any password-protection.

    b. The PDF contains un-embedded fonts. However, PDFs containing un-embedded popular fonts (e.g., Microsoft TrueType fonts) are most likely able to be converted to PDF/A since the FDA has installed those fonts on its production server for font embedding purpose during PDF to PDF/A conversion.

    c. The PDF contains embedded files. Embedded files cannot be automatically converted into PDF/A-1b since PDF/A-1b prohibits embedded files. However, those PDFs may be able to be converted to PDF/A-2b which allows embedded files, with the manual configuration by the FDA.

    d. The PDF contains properties that exceed the implementation limit allowed by Adobe. [8]

4. Errors reported during PDF to PDF/A conversion are recorded in the FDA preservation database and will be monitored by FDA staff for future improvements. The errors are also reported in FDA ingest, disseminations, or refresh reports. As an example, here is a snippet of an ingest report showing the conversion errors:

    daitss-test://EUO0W09XF_Z6F1KZ/file/2-norm-0

    - describe
    - normalize <error>Embed missing fonts:Tahoma-Bold</error><error>Convert to PDF/A-1b</error><error>Remove additional encoding entries in cmap of symbolic TrueType fonts</error>

    In addition, any PDF file that fails to convert to PDF/A format will result in a zero-length normalized file being listed in the “Archival Attributes” section of the FDA reports. For example:

    Archival Attributes

Id	Name	Size	Origin	Message Digests	Events	Broken Links	Warnings
daitss-test://EUO0W09XF_Z6F1KZ/file/0	sip-files/ateam-bad-pdf.xml	1575	DEPOSITOR	2	3	0	0
daitss-test://EUO0W09XF_Z6F1KZ/file/1	sip-files/ateam.tiff	921972	DEPOSITOR	2	2	0	0
daitss-test://EUO0W09XF_Z6F1KZ/file/2	sip-files/00001.pdf	3649429	DEPOSITOR	2	2	0	0
daitss-test://EUO0W09XF_Z6F1KZ/file/2-norm-0	aip-files/2-norm-0	0	ARCHIVE	2	2	0	0

5. When packages containing PDFs are disseminated, they will be reprocessed by DAITSS and PDF/A files will be created whenever possible. The normalized PDF/A files will be contained in the “aip-files” directory of the Dissemination Information Package and will contain the “.pdf” file extension.

The FDA is currently considering the possibility of mass refreshing existing PDFs in its archive to: 1) re-validate those PDFs with pdfaPilot for better validation results; and, 2) convert PDFs into PDF/A format and save the resulting PDF/A in its archive.

For more information

If you have any questions about the new PDF/A validation and conversion, please contact the Florida Digital Archive at FDA-TECH-L@lists.ufl.edu.

References

[1] International Organization for Standardization, “ISO 19005-1:2005 Document Management – Electronic Document File Format for Long-Term Preservation – Part 1: Use of PDF 1.4 (PDF/A-1).” December 1 2005.

[2] International Organization for Standardization, “ISO 19005-2:2011 Document Management – Electronic Document File Format for Long-Term Preservation – Part 2: Use of ISO 32000-2 (PDF/A-2).” July 31 2011.

[3] International Organization for Standardization, “ISO 19005-3:2012 Document Management – Electronic Document File Format for Long-Term Preservation – Part 3: Use of ISO 32000-3 (PDF/A-3).” Oct 15 2012.

[4] Chou, Carol, “Guidelines for Creating Archival Quality PDF Files.” Florida Virtual Campus, Version 1.1, June 2006, Florida Digital Archive, http://fclaweb.fcla.edu/uploads/Lydia%20Motyka/FDA_documentation/PDFGuideline.pdf

[5] Harvard University, “JHOVE – JSTOR/Harvard Object Validation Environment – PDF Module,” http://jhove.sourceforge.net/pdf-hul.html

[6] Jamin Koo & Carol C.H. Chou, “PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow,“ Oct 2012, International Conference on Preservation (iPRES), http://fclaweb.fcla.edu/uploads/iPRES_PAPER86_Abstract.docx

[7] Jamin Koo & Carol C.H. Chou, “PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow,” Journal of New Review of Information Networking, May 16 2013, 18:1, 1-15, http://www.tandfonline.com/doi/abs/10.1080/13614576.2013.771989#.UhtoJWR4bZg

[8] Adobe System Incorporated, PDF Reference third edition, “Adobe Portable Document Format,” Page 706, http://partners.adobe.com/public/developer/en/pdf/PDFReference.pdf

[9] Chou, Carol C.H., Florida Virtual Campus, Recommended Data Formats for the Preservation Purpose in the Florida Digital Archive, http://fclaweb.fcla.edu/uploads/recFormats.pdf