PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow

Jamin Koo
University of Michigan
4322 North Quad
105 S State St
Ann Arbor, MI 48109
jaminkoo@umich.edu

Carol C.H. Chou
Florida Digital Archive,
Florida Virtual Campus - Gainesville
5830 NW 39th Ave.
Gainesville, FL 32606
002-1-352-392-9020
cchou@ufl.edu

ABSTRACT

PDF/A is a version of Portable Document Format backed by ISO standard that is designed for archiving and preservation of electronic documents. Many electronic documents exist in PDF format. Due to its popularity, the ability to convert an existing PDF into a conforming PDF/A file is as important, if not more, as being able to produce documents in PDF/A format in digital preservation. In recognition of this fact and encouraged by growing interest from its affiliates, the Florida Digital Archive (FDA) conducted an evaluation of several of the PDF to PDF/A converter applications, the result of which is reported in this paper. There is room for interpretation in the ISO standards concerning PDF/A, which can be manifest in the development of software. In selecting a PDF to PDF/A converter product, reliability of the outcome in terms of PDF/A compliance must be established along with functionality. The goal of this paper is not to rank or promote the software evaluated, but rather to document the FDA's evaluation process and present the results in such a way that they provide insight into challenges and potential drawbacks during similar evaluation or implementation.

1. INTRODUCTION

The FDA has been in production since 2005. As of 2012, the FDA has over a hundred thousand PDFs in its archive with the presence of all PDF versions from 1.1 to 1.7 where 90 percent of them are version 1.4. Though FDA has encouraged its affiliates to submit PDF/A, less than 1 percent of its PDF archive is identified to be PDF/A-1b using JHOVE's PDF/A-1 validation[1].

To ensure the long-term preservation of its PDFs in the archive, FDA conducted a study to select a PDF to PDF/A conversion application as part of its PDF format normalization strategy in the summer of 2012. The ultimate goals will be 1) to provide better PDF/A validation than the existing one provided by JHOVE; and 2) to normalize all non-PDF/A PDFs in the archive into at least PDF/A-1b.

Eight products currently available in the market were identified from the PDF/A Competence Center on the PDF Association website, of which three were selected for in-depth evaluation after a thorough review of product specifications. Most selection criteria have general applicability, such as the ability to fix un-embedded fonts and device-dependent color spaces; however, some requirements, such as Linux support and command line operation, were FDA specific. This paper evaluates PDF/A validation and conversion features of the three products selected, which are pdfaPilot CLI v3.1.159, 3-Heights PDF to PDF/A Converter v4.0.9.0 and PDF/A manager v5.80. The desktop version of pdfaPilot was also used but for troubleshooting purposes only.

2. VALIDATION

The Bavaria Report [1] is a thorough analysis of PDF/A validation products published in 2009, which included two of the three products assessed in this study. Given the age of the report, the FDA decided to do a preliminary validation testing on the most recent version of all three products using the same test files on which the Bavaria Report was based. The Isartor testsuite[2] was excluded as the two products already showed 100% compliance in the Bavaria Report on Isartor testsuite.

Table 1: Validation Testing

	Total	False alarm	Miss	Accuracy
pdfaPilot	80	0	8	90%
3-Heights	80	17	4	74% or 95%
PDF/A Manager	80	0	7	91.3%

Note that 3-Heights flagged 17 conforming PDFs as invalid due to embedded fonts declared in the form fields when no form field was visible in the document. PDF Tools, the maker of 3-Heights, confirmed this as a bug that would be addressed in future releases. With this corrections, the accuracy of 3-Heights goes from 74% to 95%.

The differences in accuracy were not enough to indicate superior performance by any of the products on PDF/A validation. However, pdfaPilot produced notably better and more detailed error reporting out of the three.

3. CONVERSION, CROSS-VALIDATION

The conversion testing for each product was based on 203 PDFs chronologically sampled from the FDA archive, which all three products identified as not PDF/A compliant during initial validation. The conversion testing includes pre-conversion validation, conversion, self-revalidation on output files, and cross-revalidation by the other two products. All conversion operations were performed per the PDF/A-1b compliance level.

The Initial Conversion Success Rate and Actual Conversion Success Rate in Table 2 represent the percentage of successful conversions based on post-conversion self-validation and the success rate after an in-depth review of conversion logs and error

[1] JHOVE does not parse the contents on streams, so it cannot determine PDF/A conformance to the degree required by ISO 19005-1.

[2] Isartor testsuite is a set of files by PDF/A competence center to check software conformance on PDF/A-1 standard.

reports, respectively. False positives (non-compliant output files that passed self-validation) were identified through verification of errors and, in some cases, visual inspection of the files.

Table 2 Conversion Success Rate by Product

	Initial Conversion Success Rate	Actual Conversion Success Rate
pdfaPilot	79.7%	79.7% (--)
3-Heights	89.6%	84.2% (🡻)
PDF/A Manager	92.1%	83.7% (🡻)

The slightly higher conversion success rates shown by 3-Heights and PDF/A Manager can be attributed to the way these products handle encryption and embedded files. While pdfaPilot required the input files be free of these inhibitors, 3-Heights and PDF/A Manager "fixed" the problem by simply removing such items. However, in the case of non-working bookmarks, 3-Heights and PDF/A Manager flagged them with invalid destination errors, whereas pdfaPilot ignored them and completed the conversion without fixing the bookmarks.

Table 3: Conversion Failures by Product

	pdfaPilot	3-Heights	PDF/A Mgr
Environment Issues	14 (33%)	12(38%)	0
Embedded files	6(17%)	0	0
Encrypted	4(10%)	0	0
Problem PDF	17(40%)	9(28%)	16(38%)
False Positive	0	11(34%)	17(52%)

The conversion errors were grouped into four categories: 1) environment issues, such as fonts and color profiles availability; 2) embedded files in input PDF files; 3) encryption; and 4) other problems in input PDF files including but not limited to syntax and metadata issues. The false positive results from 3-Heights and PDF/A Manager were due to the products failing to detect mostly font-related (environment) and syntax/metadata (other) issues. Both products converted a few files with mis-rendered characters due to a Symbol-Italic font that was un-embedded and unavailable in the system for a fix, resulting in visual differences between the original and the output files (e.g. "beta" italic character appearing as a rectangle). Many of the false positives by PDF/A Manager resulted from the product failing to detect and/or fix XMP issues (e.g. missing XMP packet headers) per XMP Specification 2004 [4] referenced by ISO 19005-1 [2].

4. CHALLENGES

The environment issues are directly tied to the rendering and usability of the files. Even a single missing or mis-rendered glyph, as seen in some false positive files by 3-Heights and PDF/A Manager, can be difficult to detect without proper flags and warnings and have a devastating impact especially in PDFs with scientific data. One of the biggest potential roadblocks in dealing with fonts and color profiles is the rights issues. There are ways to circumvent possible copyrights infringement through font substitution but some specialized fonts may prove to be difficult not only to procure but also to use in PDF/A conversion, as their makers can prohibit embedding of fonts.

Handling of inhibitors like embedded files and encryption also needs to be considered in PDF to PDF/A conversion. While embedded files can become non-issue per later PDF/A standards, encryption of any type can hinder long-term preservation efforts including the conversion to PDF/A. Indiscriminate removal of encryptions or embedded files should be employed with caution because of potential adverse effects that may not be immediately evident, although the ability to remove non-critical encryptions may indeed prove useful to some institutions.

As thorough as the standards and documentations for both the PDF and PDF/A formats are, there is room for interpretation in determining the PDF/A compliance, between different documentations in particular. A pertinent example concerns the opposite positions that PDF Tools (maker of 3-Heights) and callas software (maker of pdfaPilot) take regarding non-working bookmarks. While the invalid destination error is a legitimate error per PDF 1.4 reference [3], there is no specific provision about bookmarks and destinations in ISO 19005-1 [2], which is why callas software does not consider the invalid destination error severe enough to stop or fail conversion for even when pdfaPilot cannot fix or restore the bookmark functionality.

5. CONCLUSION

Establishing reliability and accuracy of PDF/A converter software is not as clear-cut as one might wish, due to the variables involved and challenges demonstrated above. Purely quantitative assessment of the product performance has proven difficult even with adjusted statistics based on extensive analysis of errors. Given the complexity of PDF/A compliance requirements and the automatic fixes applied by the products during the conversion process, which will only grow more sophisticated as technology advances, the two most apparent differentiators are 1) the level of documentation and reporting capabilities of the product; and 2) the access to knowledgeable support staff. For these reasons, this study found pdfaPilot more reliable than the other two products.

6. FUTURE WORK

PDF/A-2 accommodates more features such as embedded files, JPEG 2000, transparency, etc. In addition, to yield higher successful conversion, pdfaPilot also provides a "force-conversion" feature that can convert problem pages into images with invisible text, still allowing marking, searching and copying. The FDA hope to find some resources in the future to continue the PDF to PDF/A conversion testing with a focus on PDF/A-2 and the pdfaPilot's force-conversion feature.

7. ACKNOWLEDGEMENTS

The authors wish to thank Priscilla Caplan and Lydia Motyka for providing insightful feedback and support on this project.

8. REFERENCES

[1] PDFlib. May 4, 2009. Bavaria Report on PDF/A Validation Accuracy, http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf

[2] International Standard Organization, Dec 1, 2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)

[3] Adobe Systems Incorporated, Dec 2001, PDF Reference version 1.4

[4] Adobe Systems Incorporated, Jan 2004, XMP Specification, https://www.aiim.org/documents/standards/xmpspecification.pdf