Navigating the PDF/A Standard

A Case Study of Theses in Oxford's Institutional Repository

Anna Oates[1], J. Stephen Downie[1], Edith Halvarsson[2], and Michael Popham[2]

[1] University of Illinois at Urbana-Champaign, Urbana, IL 61821, USA
[2] University of Oxford, Oxford OX1 2JD, UK
annaio2@illinois.edu

Abstract. The PDF/A (Portable Document Format–Archival) was established by the International Organization of Standardization as the ISO 19005 standard for long-term preservation of electronic documents. In a case study of the Oxford institutional repository theses collection, PDF/A was evaluated as a possible format for standardizing theses disseminated online. While the ISO requirements of a well-formed PDF/A promises sustainability and easy recovery of content, the case study uncovered that the standard restricts some document features from being incorporated into a well-formed PDF/A. Non-conformances to the standard are found across electronic theses and dissertations, from non-Latin glyphs used in scientific and language papers to embedded content, such as images. A further complication for achieving ISO 19005 compliance is that, despite non-conformance to the ISO standard, validation tools do not always catch non-conformance errors in documents which claim to conform to PDF/A. While PDF/A is a logical solution for long-term digital preservation, the stringent standard prevents some content which is frequently used in academic research from conforming to the ISO 19005 standard.

Keywords: ISO 19005, Institutional Repositories, Digital Preservation.

1 Introduction

Institutional repositories' priorities are largely twofold: 1) to provide access, and 2) to ensure preservation. While many facets of digital preservation have been standardized, those standards change with the digital environment, even for seemingly uncomplicated digital objects: text documents. The primary research question of this case study is: is PDF/A an adequate file format for documents deposited in an institutional repository? To implement PDF/A into the case study workflow, this secondary question was asked: what software is best suited for creation and conformance of PDF/A (ISO 19005)?

1.1 PDF/A

The ISO 19005 standard was developed to ensure long-term preservation of electronic documents. However, the standard has not achieved its goal to serve as a suitable format

for long-term preservation for all document types. Rather, it should be used as a container format [1] for digitized materials, and, as PDF was originally purposed, a format for dissemination and exchange. PDF/A differs from standard PDF by restricting features that have a high risk for digital preservation.

Versions. The three versions of the standard—ISO 19005-1:2005, ISO 19005-2:2011, and ISO 19005-3:2012—are distinguished not by their suitability for use or for preservation but by their possibility for and appropriateness in different use cases.

Conformances. In addition to the versions of PDF/A are three levels of conformance to the standard:

1. Level A (Accessible) provides the highest level of conformance with the ISO standard. Due to the stringent requirements, conformance with Level A is often met only when created from born-digital documents. Implemented in ISO 19005-1:2005, ISO 19005-2:2011, and ISO 19005-3:2012.
2. Level B (Basic) provides the lowest level of conformance with the ISO standard, only placing requirements on the visual appearance of a document. Level B conformance is most suitable for digitized documents. Implemented in ISO 19005-1:2005, ISO 19005-2:2011, and ISO 19005-3:2012.
3. Level U (Unicode) is similar to Level B but increases accessibility by requiring Unicode mapping of fonts. As with Level A, Level U should be used for born-digital documents. Implemented in ISO 19005-2:2011 and ISO 19005-3:2012.

2 Case Study: Oxford University Research Archive

This research was conducted on a set of 56 born-digital and digitized theses composed of 107 unique files that are held in the Oxford University Research Archive (ORA). A Research Archive Assistant familiar with the ORA collection scope selected theses according to their file content complexity, taking into consideration the presence of images, vector graphics, Optical Character Recognition text, mathematic formulas, and non-Latin scripts.

2.1 Methodology

The dataset was migrated from the deposited file type (i.e., PDF; or source file, preferred for deposit and preservation: .doc, .docx, .rtf, .odt, or .tex) to PDF/A, totaling 678 derivative files. Findings from this research are representative of migrated files rather than files created with the intent to be saved as PDF/A.

Seven software were used for creation and conformance: Adobe Acrobat DC 2015, callas pdfaPilot Desktop v. 7, Intarsys PDF/A Live! v. 6.2, LibreOffice v. 5, pdfforge PDFCreator v. 2.5.1, PDF Studio v. 12, and PDFTron PDF/A Manager CMD v. 1.x. After creation or conformance, the PDF/A files were validated using veraPDF v. 0.8, a validation software for PDF/A created under the EU PERFORMA project [2].

3 Analysis

Out of 678 total documents, creation and conformance software had an average success rate of 70.2% (see Fig. 1).

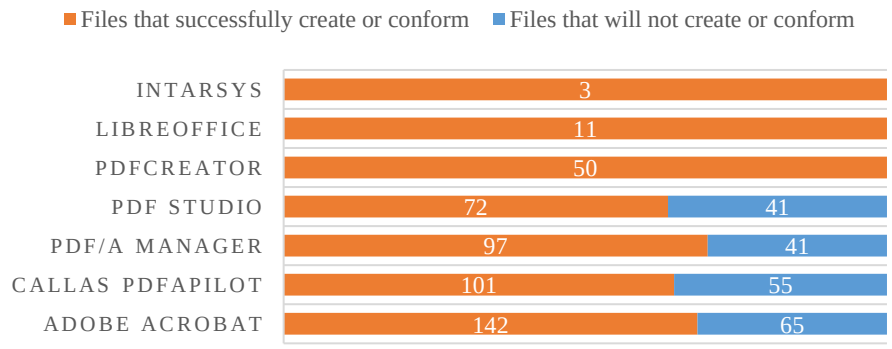

Fig. 1. Comparison of creation and conformance software success.

Software that claimed ISO 19005 conformance of file output consistently produced files that failed validation with veraPDF (see Fig. 2).

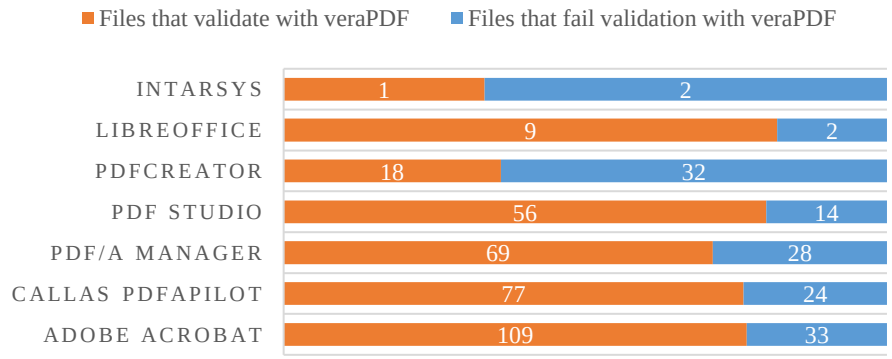

Fig. 2. Comparison of validation success or failure with veraPDF, selected from files that were successfully created or conformed.

For the purposes of this research, we considered some areas of non-conformance, including vector graphics and non-Latin fonts, as significant properties, and thus vital for preservation. The creation and conformance software often failed to produce files that maintained these significant properties, creating files that did not visually or semantically represent content present in the deposited file.

Adobe and PDF/A Manager failed to conform files to a version and conformance level of PDF/A if the underlying document features were non-conforming (e.g., non-complying image decode features). callas was unique in that it attempted to change the

semantic and structural features of documents to achieve ISO compliance. callas changed the image decode filter if it was non-conforming. For example, a PDF source file contained images with the JPXDecode filter, i.e. JPEG2000 compression, which callas changed to DCTDecode to achieve ISO 19005-1:2005 compliance when conforming the source file to PDF/A-1b. callas also changed non-Unicode embedded fonts and captured the visual appearance of the original font as a pixelated raster image:

$$\hat{O}_{\mathrm{rot\ frame}} = \mathrm{e}^{+\mathrm{ip}\omega_0 t}\hat{O},$$

Fig. 3. Image of text after conformance to PDF/A-2a using callas.

4 Discussion

This case study found that, if achieved, ISO 19005 conformance is not the *best* preservation format but rather one of a number of viable formats for preservation. Rather than migrating files to a distorted but ISO-conforming PDF/A, we recommend that institutions perform a risk assessment of their collections to identify document features that are harmful for digital preservation.

While the ISO 19005 standard offers several outlets for document creation, not every document will conform to the best-fit version and conformance level of the standard. For example, born-digital documents should be created in accordance with Level A conformance; digitized documents should be created in accordance with Level B conformance. Institutions must both understand the requirements of each variance of ISO 19005 and the components of the source file. Furthermore, policies should be in place to require that theses be deposited as source files to avoid retroactive recreation to achieve conformance as PDF/A.

5 Future Work

To provide useful recommendations for working with electronic theses and dissertations, there must be further investigation of significant properties and risk assessments of theses. This research project has been extended to continue the exploration of the PDF/A (ISO 19005) implications for long-term preservation.

References.

1. Han, Y.: Beyond TIFF and JPEG2000: PDF/A as an OAIS submission information package container. Library Hi Tech, 33(3), 409-423 (2015).
2. Wilson, C., McGuinness R., and Jung J.: veraPDF: building an open source, industry supported PDF/A validator for cultural heritage institutions. Digital Library Perspectives 33(2), 156-165 (2017).