NAVIGATING THE PDF/A STANDARD: A CASE STUDY OF THESES IN THE UNIVERSITY OF OXFORD'S INSTITUTIONAL REPOSITORY

BY

ANNA OATES

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Advisers:

Professor Jerome McDonough
Dean Allen Renear

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. While the ISO requirements of a well-formed PDF/A ensure sustainability and easy recovery of content, 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 preservation of electronic documents, the stringent standard prevents some content which is frequently used in academic research from conforming to the ISO 19005 standard. This thesis evaluates the PDF/A and its potential use as a preservation file format for electronic theses and dissertations.

ACKNOWLEDGEMENTS

I would like to thank Professor J. Stephen Downie, Ryan Dubnicek, and the HathiTrust Research Center for providing the opportunity to begin this research project as part of the Oxford-Illinois Digital Libraries Placement Programme. I extend a special thanks to Edith Halvarsson for her constant support during the placement and for encouraging me to extend this research into a thesis. I would also like to thank Michael Popham, Sarah Mason, and all the folks at the Bodleian Digital Library Systems and Services, and the Oxford e-Research Centre.

I would also like to thank my advisors, Professor Jerome McDonough and Professor Allen Renear, who enabled me to extend the summer research project into a Master's thesis. Both Dr. McDonough and Dean Renear patiently worked with me as I fixated until the last minute of nearly every deadline on aspects of the thesis, that as a continuous piece of research would never be complete. To Jerome, I would like to extend special thanks for encouraging me to consider every aspect of importance and for identifying complex files for my dataset.

I would like to thank the Illinois School of Information Sciences at the University of Illinois at Urbana-Champaign for developing a program that encourages its students to engage in research.

Finally, I would like to thank my family, friends, and colleagues for their support and flexibility when it seemed that I would never have enough time or energy. Wayne Ryan, thank you for always encouraging me to strive for things that, to me, seem impossible to achieve.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

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 (International Organization for Standardization [ISO], 2005). In 2002, information professionals in libraries, archives, government, industry, and federal agencies formed a working group to establish a "purpose-built file format for standardised archiving" (PDF Association, n.d.a, para. 5). While the ISO requirements of a well-formed PDF/A ensure sustainability and easy recovery of content, the standard restricts some document features from being incorporated into a well-formed PDF/A. Non-conformances to the ISO 19005 standard are found across electronic theses and dissertations, from non-Latin glyphs used in scientific and language papers, to other embedded content, such as images. A further complication to 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 preservation of electronic documents, the stringent standard prevents some content which is frequently used in academic research from conforming to the ISO 19005 standard. To better understand the requirements of ISO 19005 conformance and to provide broad recommendations for institutional file format policies, this thesis seeks to answer the following questions:

- [*primary RQ*] Is PDF/A an adequate file format for creation or conformance of electronic theses and dissertations?
- [*secondary RQ*] What areas of non-conformance to the ISO 19005 standard are impractical to avoid for theses and dissertations deposited in a non-PDF/A file format?

- [*secondary RQ*] Do those areas of non-conformance precipitate considerable preservation risks?

From June 26, 2017 to August 4, 2017 the author of this document undertook a studentship with the Bodleian Libraries as part of the Oxford-Illinois Digital Libraries Placement Program (OIDLPP). During the placement, the author investigated ISO 19005 conformance of a set of born-digital and digitized theses in the Oxford University Research Archive (ORA). This thesis extends the placement research, investigating potential preservation risks present in electronic theses and dissertations. This thesis further contextualizes and evaluates the use of PDF/A in electronic theses and dissertations repositories.

According to a 2007-2008 survey conducted by MetaArchive Cooperative and Networked Digital Library of Theses and Dissertations (NDLTD) on 96 institutions, "72% of responding institutions reported that they had no preservation plan for the ETDs [Electronic Theses and Dissertations] they were collecting" (Halbert, Skinner, & Schultz, 2012, p. 263). Conducted over ten years ago, these survey results are no longer indicative of institutional practice. However, a decade later, the storage and maintenance of ETDs continues to be a topic of importance.

Chapter 2: Literature Review begins with an overview of the PDF and PDF/A file formats. Review of the file formats lends to understanding the methodology later discussed in Chapter 3. The literature review also introduces the concept of relationships enabled and imposed by technical systems. This discussion is used to contextualize PDF/A as a digital object, in addition to its position as a standardized specification.

The discussion in Chapter 4 explores two features. First, it evaluates the effectiveness of PDF/A using the results of an experiment that tested the success of migration from an original

source file to PDF/A or from PDF to PDF/A, respectively. This experiment provides concrete and computational results of the practicality of including PDF/A as a recommended or even required file format in regulatory policies dictating which formats should be included in an institutional repository. The second half of the discussion evaluates the implicit recommendations of the ISO standard through a sociotechnical lens. This portion of the discussion contends that the ISO 19005 standard simply cannot be considered a valid format for long-term preservation of electronic documents. The distinction between and success of both PDF and PDF/A has proven to be a continuous struggle throughout this research. If PDF is branded as the "archival" format, then a PDF/A file should be representative of the most pristine capture of the digital record. This thesis suggests, however, that it is not.

CHAPTER 2. LITERATURE REVIEW

Integral to the discussion of ISO 19005 and its resulting PDF/A file format is an understanding of document features allowed in PDF/A. Furthermore, to comprehensively understand PDF/A, there is requisite knowledge of PDF. Several individuals have contributed to the history of PDF, often specific to their domains of practice (e.g., Prepress, graphic design). In this thesis, the in-depth analysis of PDF and PDF/A assist in the greater discussion of information systems and sociotechnical implications of those systems, which have also been contextualized in this chapter.

2.1. History of PDF

In 1991, Adobe co-founder, John Warnock introduced the idea of the Interchange PostScript (IPS) file format, which would later become the PDF (Portable Document Format) file format. Originally purposed as an internal project, he hoped that the IPS format would solve interoperability issues imposed when digitally disseminating the Adobe logo. In his statement on the "Camelot" project, Warnock wrote:

> Imagine being able to send full text and graphics documents (newspapers, magazine articles, technical manuals etc.) over electronic mail distribution networks. These documents could be viewed on any machine and any selected document could be printed locally. This capability would truly change the way information is managed. (Warnock, n.d., p. 5)

At the Seybold Conference hosted in San Jose, CA in 1991, Warnock publically introduced his project for the IPS (Jaeggi, 2016). By 1992, PDF Version 1.0 and its accompanying Adobe Acrobat component were officially announced at the COMDEX Fall conference, with the software winning a "Best of COMDEX" award (Adobe Systems Incorporated, 1992, p. 29). On

June 15, 1993, Adobe released PDF Version 1.0. The file format was foundational in steps toward the exchange of information that could be viewed across operating systems, but it did have limitations. Among those, PDF Version 1.0 only supported RGB color space, and it did not support video and audio data embedding or graphics, as Warnock had originally anticipated (Adobe Systems Incorporated, 1993). Furthermore, embedding of symbolic fonts, such as Cree, was complicated by requiring additional information about the character shape and "a compressed version of the Type 1 font program for the font [to be] included in the PDF file" (Adobe Systems Incorporated, 1993, p. 10). This limited the dissemination of information that was not represented with StandardEncoding fonts (i.e., textual representation was limited to standard Latin glyphs without embedding fonts from an additional font package). Adobe Acrobat 1.0, internally referred to as "Carousel," was the first Adobe software to support PDF viewing and editing.

In November 1994, Adobe released PDF Version 1.1, which introduced features that improved document elements. Among these new features were password protection, device-independent color space, and smaller file sizes (Adobe Systems Incorporated, 1996a). With this release, Adobe also introduced a revised PDF software: Adobe Acrobat 2.0.

November 1996 made way for another slew of improvements to the format and software. Adobe released Adobe Acrobat 3.0, "Amber," and PDF Version 1.2. PDF 1.2 supported audio and video embedding, Han characters, and enhanced features for prepress (Adobe Systems Incorporated, 1996b). Despite Adobe's attempt to increase viability of the format for the prepress industry, it lacked uptake among prepress professionals because "there were simply too many ways in which a perfectly valid but unusable PDF-file could be created" (Leurs, 2017, p. 2).

Thus, the prepress industry pushed for a variant of PDF that better suited prepress documents, which later manifested as PDF/X.

In April 1999, Adobe released Adobe Acrobat 4.0, internally referred to as "Stout," and an accompanying release of PDF Version 1.3. PDF 1.3 initiated support for embedding any file type within a PDF file and improved CIDFont, digital signatures, and JavaScript (Adobe Systems Incorporated, 2000).

In May of 2001, Adobe released Adobe Acrobat 5.0, "Brazil" and PDF Version 1.4. With PDF Version 1.4 came support for transparency, JBIG-2 compression decoding, embedded metadata, and PDF tagging, among several other features (Adobe Systems Incorporated, 2001).

In 2003, Adobe released two versions of Acrobat 6, "Newport," Adobe Acrobat 6.0 and Adobe Reader 6.0. Adobe Acrobat 6.0 supported enhanced features of the software for professional uses, such as integration of a Preflight function, PDF/X support, and transparency flattening (Adobe Systems Incorporated, 2003b). PDF Version 1.5, released the same year, saw increased support for compression algorithms with inclusion of JPEG2000 compression (ISO/ICE 15444), in addition to many other enhanced features, including encryption, digital signatures, and Tagged PDF (Adobe Systems Incorporated, 2003).

In January 2005, Adobe released Adobe 7, "Vegas," as Adobe Acrobat 7.0 and Adobe Reader 7.0, in addition to PDF Version 1.6. Both the software and file format included support for embedding of OpenType fonts, which was previously limited to embedding of TrueType or PostScript Type 1 fonts. The software and file format also supported embedding of 3D data (e.g., CAD files), making the PDF format viable for graphic designers and architects (Adobe Systems Incorporated, 2004). In addition to the new features were improvements for encryption, annotations, and Tagged PDF.

By 2006, Adobe released PDF Version 1.7 and versions of the Adobe software that better supported user-needs: Adobe Acrobat 8.0 and Adobe Reader 8.0. Adobe Acrobat 8.0, "Atlas," defaulted PDF creation to PDF Version 1.6 and improved the ability to save to other versions of PDF, and improved usability of the software. Adobe Acrobat 8 included support for PDF/A, as well as an improved Preflight that introduced the ability to apply "fix-ups." With PDF Version 1.7 came better support for 3D embedding, including the ability to comment on 3D objects, as well as increased control of 3D animations. The format also supported embedding of printer settings to define aspects such as scaling and paper selection. (Adobe Systems Incorporated, 2006)

After the release of PDF Version 1.7, "Adobe announced its intent to release the full…specification to AIIM [Association for Information and Image Management]…for the purpose of publication by the International Organization for Standardization" (Adobe Systems Incorporated, 2017b). In January 2008, PDF Version 1.7 was instituted as the ISO 32000-1:2008 standard for "Document management – Portable document format – Part 1: PDF 1.7." As an ISO standard, PDF no longer sat under the aegis of Adobe, and as such, prevented Adobe from releasing new versions of the format from which solely they would profit. To circumvent this and continue releasing variants of PDF, Adobe introduced PDF extensions. To date, there have been two extensions: BaseVersion 1.7 ExtensionLevel 3, released in 2008 and supported by Adobe Acrobat 9.0 and Adobe Reader 9.0; and BaseVersion 1.7 ExtensionLevel 5, released in 2009 and supported by Adobe Acrobat 9.1 and Adobe Reader 9.1. Adobe's release of Acrobat 9, "Nova," supported the Adobe ExtensionLevel 3 for increased file embedding and enhanced data extensions for embedding geospatial data.

Without releasing new versions of the file format, Adobe continued to update its software with enhanced features. In 2010, Adobe released Acrobat X (10.0) and its three instantiations: Acrobat X Standard, Acrobat X Pro, and Adobe Reader X. In addition to the desktop versions for Windows and Mac operating systems, Adobe introduced smartphone compatibility for Adobe Reader X on Android devices (Jain, 2010). That year, Adobe also published PDF/VT as an ISO standard (ISO 16612-2:2010). PDF/VT, or PDF–Variable Transactional, is an extension of PDF/X and is purposed specifically for the exchange of variable data and transactional printing. The following version of Acrobat, Acrobat XI (11.0), was released in October 2012 to enhance PDF editing and support as a cloud service, in addition to improve interoperability with tablet devices and Windows 7 and Windows 8 operating systems. The current version of Acrobat was released in April 2015 as Acrobat DC and Acrobat Reader DC. Among the new features in Acrobat DC were increased mobile compatibility, editing tools that enabled features such as spellcheck, and increased accessibility with VoiceOver support and high contrast text (Adobe Systems Incorporated, 2017b).

Most recently, in August 2017, International Organization for Standardization published guidelines for PDF 2.0 as ISO 32000-2. ISO 19005-4, announced for release in 2018, will continue the implementation of PDF/A based upon the standardized PDF, following the most recent ISO 32000-2.

See table 1 for a comprehensive illustration of the differences between each version of PDF, which has been mapped by Betsy A. Fanning.

Table 1. Features introduced in PDF, found in "Preservation with PDF/A (2nd Edition)" from *DPC Technology Watch Report* (Fanning, 2017, p. 5).

Table 1 – Features introduced in PDF

	PDF 1.1	PDF 1.2	PDF 1.3	PDF 1.4	PDF 1.5	PDF 1.6	PDF 1.7
External links	✓	✓	✓	✓	✓	✓	✓
Article threads	✓	✓	✓	✓	✓	✓	✓
Security features	✓	✓	✓	✓	✓	✓	✓
Device-independent colour	✓	✓	✓	✓	✓	✓	✓
Notes	✓	✓	✓	✓	✓	✓	✓
Support for OPI (Open Process Interface) 1.3		✓	✓	✓	✓	✓	✓
Support for CMYK (colour model for cyan, magenta, yellow, and key black)		✓	✓	✓	✓	✓	✓
Maintenance of spot colours in PDF		✓	✓	✓	✓	✓	✓
Halftone functions could be included as well as overprint instructions		✓	✓	✓	✓	✓	✓
2-byte CID fonts			✓	✓	✓	✓	✓
OPI 2.0 specifications			✓	✓	✓	✓	✓
DeviceN, a new colour space to improve support for spot colours			✓	✓	✓	✓	✓
Smooth shading, a technology that allows for efficient and very smooth blends (transitions from one colour or tint to another)			✓	✓	✓	✓	✓
Annotations			✓	✓	✓	✓	✓
Transparency support that allows text or images to be seen through				✓	✓	✓	✓
Improved security				✓	✓	✓	✓
Improved support for JavaScript				✓	✓	✓	✓
Improved compression techniques including object streams and JPEG2000 compression					✓	✓	✓
Support for layers					✓	✓	✓
Improved support for tagged PDF					✓	✓	✓
Improved encryption algorithms						✓	✓
OpenType fonts embedded						✓	✓
Ability to embed files to be a container file format						✓	✓
Ability to embed 3D data						✓	✓
Improved support for commenting and security							✓
3D support improvements							✓

Beyond standard PDF are 5 subset standards of the format: PDF/A, PDF/E, PDF/UA, PDF/VT, and PDF/X. Each of the subsets supports document features necessary for a specific discipline. PDF/X was developed primarily to support exchange of documents; PDF/E was developed for engineering documents; PDF/VT, similar to PDF/X, was developed to support graphic technology; PDF/UA was developed to improve accessibility; and PDF/A was developed to function as an archival form of PDF.

2.2. History of PDF/A

In 2005, the International Organization for Standardization released the ISO 19005-1:2005 standard, which "specifies how to use the…PDF 1.4 for long-term preservation of electronic documents" (ISO, 2005). This specification resulted in the Portable Document Format–Archival (PDF/A) file format. In 2011, ISO released a second part to the standard—ISO 19005-2:2011. The following year, ISO 19005-3:2012 was established to support embedding of any file type.

In addition to the three versions of PDF/A are three levels of conformance to the standard (see Table 2 for versions and conformance levels with their respective naming as PDF/A):

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.

Table 2. ISO standards, levels of conformance, and their respective PDF/A flavors.

	ISO 19005-1:2005	**ISO 19005-2:2011**	**ISO 19005-3:2012**
Level A	PDF/A-1a	PDF/A-2a	PDF/A-3a
Level B	PDF/A-1b	PDF/A-2b	PDF/A-3b
Level U	N/A	PDF/A-2u	PDF/A-3u

PDF/A differs from standard PDF by limiting features that should be included in a well-formed PDF/A. Features restricted from PDF/A are those that have been cause for concern for long-term preservation, and thus PDF/A is intended to function as a more stable and sustainable file than standard PDF. Below are key stipulations for all ISO 19005 conforming files:

- Fonts and images must be embedded;
- Device-independent color space specified;
- Standards-based metadata stored in XMP;
- No file encryption;
- No external content references; except for annotations, such as hyperlinked text;
- No embedded audio or video;
- No JavaScript; and
- No JPXDecode[1] or LZW compression (i.e., JPEG2000, and GIF and some TIFF images, respectively). (ISO, 2005)

These specifications distinguish PDF/A as suitable for long-term preservation. Each validating PDF/A file avoids external linkages so that the document can be rendered without relying upon external, often OS-dependent information. For example, fonts and color space must be defined and embedded within the file and cannot be referenced to an external entity. Specific requirements of an application capable of rendering a PDF/A file without inflicting damage upon the file are detailed in section 2.2.2. Survey of Lifecycle Software. Furthermore, the standard does not allow features that are considered unstable and thus unsuitable for long-term preservation.

[1] Revised in ISO 19005-2:2011 and ISO 19005-3:2012.

With the release of ISO 19005-2:2011, permitted document features were revised to include transparency and layers—accommodating PDF export tools supported by OpenOffice and Microsoft Office 2007, —JPXDecode for JPEG2000 compression, OpenType fonts to enhance support of symbolic fonts, digital signatures, and embedding of other PDF/A-1 and PDF/A-2 files (ISO, 2011).

PDF/A-3 attracted extensive criticism across archival and preservation communities for the standard's liberal approach to file embedding. In addition to the revisions introduced with the ISO 19005-2:2011 specifications, ISO 19005-3:2012 permits embedding of any file type (i.e., ISO 19005-3:2012 does not restrict embedding of files to PDF/A, as specified in the ISO 19005-2:2011 standard). Despite the criticism and widespread rejection of the format, the National Digital Stewardship Alliance (NDSA) recognizes that PDF/A-3 is appropriate in situations that require manipulation to files. A response to PDF/A-3 is detailed in their report on "The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions: An NDSA Report" (Caroline Arms et al., 2014, February). However, the working group concluded that, for long-term preservation, the use of either PDF/A-1 or PDF/A-2 is recommended over the use of PDF/A-3.

2.2.1. Survey of Memory Institution Use of PDF/A

The greatest capacity in which PDF/A is a recommended format for memory institutions has been for Electronic Theses and Dissertations (ETD) repositories. Several ETD repositories recommend or require students to deposit their thesis or dissertation as a PDF/A file (see appendix 2 for a partial list of these entities). These institutions often provide creation guidelines, instructing students how to create their source files (i.e., .docx, .doc, .rtf, .otd, .tex) as PDF/A. These directions may include instructions for conversion on Windows and Mac operating

systems. However, the creation directions do not discuss the requirements for creating a valid PDF/A.

2.2.2. Survey of PDF/A Lifecycle Software

In this section, the term "lifecycle" is used to reference two moments throughout the life of a PDF/A file: 1) the creation of a PDF/A file and 2) PDF/A file validation to check ISO 19005 conformance. The first action will occur only once, while the second action should be an ongoing preservation action to ensure that a given file conforms to the standard. The tools mentioned below contribute to either or, sometimes, both of these moments in the PDF/A lifecycle.

Since the establishment of PDF/A as an ISO standard, Adobe has ensured that their Acrobat products support the conformance and validation of PDF/A. In addition to Adobe products, there are several other proprietary and GNU GPL (General Public License) tools that create or conform text files to PDF/A, as seen in appendix 3. Because ISO 19005 requires that all information necessary to render the document contents be embedded, PDF/A files will inherently be larger than PDF files that render the same content. PDF/A documents are required to follow the specifications detailed in the ISO standard, as such, validation of the file structure is required for a PDF/A file to be considered PDF/A. Thus, PDF/A files should be validated to ensure PDF/A compliance. In 2007 and 2008, the PDF/A Competence Center developed the Isartor Test Suite, which served as a validation system for conformance of PDF/A-1b files (PDF Association, n.d.b). Since then, Isartor Test Suite has guided the development of other PDF/A validation systems, including the PDF/A validation tool used for this research, veraPDF. veraPDF was a project that began as the Open Preservation Foundation's (OPF) response to the EU

Commission's PREFORMA project[2] (Wilson, McGuinness, & Jung, 2017). veraPDF validates PDF/A files in accordance with the version and conformance specifications. The PDF Association (2013) recommends that files be validated after creation, on receipt, prior to transmission or distribution, before archiving, and at the end of certain processes, such as after adding additional files to a conforming PDF.

For the PDFlib (2009) "Bavaria report on PDF/A Validation Accuracy," a test was created to determine the most successful PDF/A validation tool. The test was constructed using the original 204 documents implemented by the Isartor Test Suite, in addition to 85 documents selected specifically for the Bavaria Test Suite. They validated these 289 documents with the following software: Adobe Acrobat 9.0, Adobe Acrobat 9.1, Adobe LiveCycle PDF Generator, Apago PDF Appraiser, callas pdfaPilot, Intarsys PDF/A Live, rPDF Tools: 3Heights PDF Validator Shell, Seal Systems: PDF Longlife Suite/PDF Checker, and Solid Documents: Solid Framework. Of the software used for validation, PDFlib found that callas pdfaPilot had the highest success rate against their validation criteria with 91% validation accuracy.

Over the course of several years, the Florida Virtual Campus evaluated the PDF/A format, as well as tools for PDF/A creation, conformance, and validation. In her "Guidelines for Creating Archival Quality PDF Files," Carol Chou (2006) mentions tools for patron conversion support, in addition to features of the ISO 19005-1:2005 standard that should be considered when creating a PDF/A-1a or PDF/A-1b file. In 2012, the Florida Virtual Campus began testing tools; the results of which were published in an article titled, "PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow" (Koo & Chou, 2013).

[2] http://preforma-project.eu/

This article provides an overview of the Florida Virtual Campus software test, in which the following validation software were tested: callas pdfaPilot, 3-Heights, and PDF/A Manager (the software version was not indicated in the dissemination of this research). Of the three software selected for testing, they found that callas pdfaPilot displayed the fewest errors post-conversion to PDF/A (Koo & Chou, 2013, p. 9). The results of this test informed the Florida Virtual Campus's decision to purchase callas pdfaPilot for validation of and conversion to PDF/A (Florida Virtual Campus, 2013, p. 1).

In addition to the creation of and validation of PDF/A, the ISO standard has set forth requirements for PDF/A viewers. Perhaps the most important to this research is that ISO 19005 specifies that PDF/A-compatible viewers must render documents using only the embedded fonts and specified color space profile (ISO, 2005).

2.3. Implications of Long-term Sustainability

At the 2007 annual American Libraries Association (ALA) conference, a working group for the Association for Library Collections & Technical Services (ALCTS) Preservation and Reformatting Section (PARS) created a document to define digital preservation, in which they offer an extended definition:

> Digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change. Digital preservation applies to both born digital and reformatted content.
>
> Digital preservation policies document an organization's commitment to preserve digital content for future use; specify file formats to be preserved and the level of

preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information.

Digital preservation strategies and actions address content creation, integrity and maintenance. (American Library Association [ALA], 2008)

This definition has been chosen as representative for describing the objectives of digital preservation, due to its embodiment within cultural heritage institutions, specifically libraries in the United States. Other definitions of digital preservation similarly consider the three primary attributes defined by the PARS working group: policy, access, and maintenance. Similar to ALA's definition, the Library of Congress (LC) simply defines digital preservation as "the active management of digital content over time to ensure ongoing access" (Library of Congress, n.d.a). While the ALA and LC set foundational definitions for practitioners in the United States, the two definitions largely overlook one of the more difficult conversations of digital preservation: the "Authenticity" of digital objects (Consultative Committee for Space Data Systems [CCSDS], 2012).

In Paul Conway's (2000) "Overview: Rationale for Digitization and Preservation," he discusses the importance of representation for digitally-reformatted materials (e.g., ¾" U-matic tape digitized to .avi), pressing the requisite to Protect, Represent, and Transcend source materials. In this article, Conway touches upon the importance of capturing "significant properties," a term that has been under considerable debate in more recent years. There are multifaceted definitions and conceptions of "significant properties," "significant characteristics," "essence," "essential properties," or "authenticity." In 2009, Angela Dappert and Adam Farquhar presented "Significance Is in the Eye of the Stakeholder," in which they explored and defined the differences between "significant properties" and "significant characteristics." The latter they

define as encompassing a "property / value pair," in which the property is an abstraction (e.g., file size) and the value is that property in relation to an object (e.g., property = file size; value = 121342 bytes) (Dappert & Farquhar, 2009b, p. 299). They state that this pairing can be preserved and that a significant property, then, cannot logically be preserved because it is an abstraction. This thesis does consider Dappert and Farquhar's (2009b) difference between significant properties and significant characteristics but refers to the entire entity as significant properties when discussing the authenticity of digital objects (CCSDS, 2012). The concept of significant properties as representative of "authenticity" of born-digital and digitized theses and dissertations is discussed in Chapter 3.

2.4. Sociotechnical Implications

Digital preservation is accomplished through a system, and that system must be sustainable over time. As the definition established by the ALA working group states, "Digital preservation strategies and actions address content creation, integrity and maintenance" (ALA, 2008). Here, the ALA working group suggests that preservation is not something that can be achieved but rather something that requires stewardship and long-term maintenance. A reference model of a sustainable digital preservation system has been described in CCSDS 650.0-M-2 and established as ISO 14721:2012.

2.4.1. Risk Assessment of File Formats

The *Reference Model for an Open Archival Information System (OAIS)* (2012) discusses "significant properties" in the context of a term coined as "Transformational Information Property," where "'significant property'…is sometimes used in a way that is consistent with its being a Transformational Information Property" (CCSDS, p. 1-16). Where "significant property" is often conflated to include characteristic types and property values (Dappert & Farquhar,

2009b), a "Transformational Information Property" specifically defines the system and "actors" (Akrich, 1992). The documentation considers "Transformational Information Properties" as requisite for transformations that do not exhibit one-to-one mapping,[3] noting that "[Transformational Information Properties] could be important as contributing to evidence about Authenticity" (CCSDS, 2012, p. 5-7). One of the objectives of OAIS is to, "Make the preserved information available…and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity" (CCSDS, 2012, p. 3-1). The supporting evidence of authenticity then, as inferred from CCSDS (2012), must be retained as significant properties.

Despite the relationship of significant properties to digital objects and even specific file formats, the CCSDS (2011; 2012; 2014) has not created documentation to provide recommendations for performing risk assessment of file formats. In their *Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories* (TDR) (2014), it is noted that "inadvertent loss of data or personnel are beyond the scope of this document" (CCSDS, p. B-2). The Rog and van Wijk (2002) article on "Evaluating File Formats for Long-term Preservation" establishes a metric for file format assessment to include evaluation of the following criteria: "openness," "adoption," "complexity," "technical protection mechanism (DRM)," "self-documentation," "robustness," and "dependencies" (p. 3-4). Similar to the Rog and van Wijk (2002) method, Brown (2003) specifics evaluation of "open standards," "ubiquity," "stability," "metadata support," "feature set," "interoperability," and "viability" (p. 5). Brown's (2003) document goes into further depth, establishing a metric for evaluation of

[3] For example, migrating from .docx to PDF/A will result in the loss of the original (.docx) bit stream as a consequence of the "Non-Reversible Transformation" (CCSDS, 2012, p. 1-13).

migration file formats, which assesses factors of "authenticity," "processability," and "presentation" (p. 5). In this metric, authenticity is achieved when "the [file] format…[preserves] the content (data and structure) of the record, and any inherent contextual, provenance, referencing and fixity information" (Brown, 2003, p. 7). With authenticity embedded within the conversation of file formats, it is vital to consider the risk of implementing a particular file format for migration, whether the objective of file migration is for preservation of a digital object or for providing access to a digital object.

2.4.2. Using Technology

In addition to the maintenance systems to support long-term preservation efforts, the usage of preservation "tools" is impacted by sociotechnical forces. For the purposes of this section, PDF/A is referred to as a "tool" that functions as a single mechanism for assisting, or attempting to assist, in the ongoing preservation of digital objects (Noonan, McCrory, & Black, 2010). Considering the factors of digital preservation in conjunction with the continuously debated term of significant properties, long-term sustainability becomes impossible to achieve in the entity of a single file format.

2.4.3. Differing inscriptions for PDF and PDF/A

PDF was developed as a format for "exchange"—or to use a term more commonly referenced in memory institutions, "dissemination" of information. Of the many subset standards for institutional-specific or discipline-specific usage, PDF/A was not developed for the purpose of dissemination, but for preservation. This presents an inherent point of confusion for users or "actors" (Akrich, 1992).

2.5. Summary

As indicated in this brief history of the evolution of PDF, the format has conformed to multifaceted user-groups' needs through the introduction of standard subsets. From PDF/X for facilitating graphic exchange to PDF/A for sustainability of electronic documents, PDF and its many versions and flavors support myriad use cases. The ISO 19005 standard's explicit recommendation for the "Long-term preservation of electronic documents" implies that PDF/A was developed for the purpose of preservation of electronic documents. Preservation is a multidimensional, ongoing process that relies upon a comprehensive system that considers the authenticity of digital objects. Thus, PDF/A, a singular tool, cannot function as the quintessence of preservation.

CHAPTER 3. METHODS

3.1. Overview of Datasets

This research is composed of three unique datasets: 1) original dataset, 2) extended dataset, and 3) secondary dataset. The original dataset was created as part of the OIDLPP placement. The extended dataset was created solely to test use-case scenarios that were not present in the original dataset. Finally, the secondary dataset was created toward the end of the OIDLPP placement to inform recommendations for the use of PDF/A. Overviews of specific content present in each dataset are provided in sections 3.1.1. and 3.1.2..

3.1.1. Primary Dataset: Creation and Conformance to PDF/A—Case Study of Theses in ORA

During the OIDLPP placement, the dataset used in this thesis was collected. The methodology for collecting this data is described below. (A complete flowchart of the research can be found in appendix 4.)

About the original dataset.

The dataset consisted of 56 theses, totaling to 104 unique files. Theses were selected by a Bodleian Digital Library Systems and Services (BDLSS) Research Archive Assistant, who was familiar with the ORA collection scope. Selection criteria included file content complexity, such as digitized theses with embedded images and Optical Character Recognition (OCR) generated text, mathematical formulas, embedded graphs and tables, and non-Latin script.

About the extended dataset.

The second dataset consisted of 19 theses and other textual documents containing unique features. These theses and textually complex documents were identified by the author of this thesis and the thesis research faculty advisor. As with the original dataset, documents were selected for their complexity, focusing specifically on the following features: embedded LZW

encoded TIFF images, embedded images with hyperspectral data, Han unification fonts, and Native American language fonts.

3.1.2. Secondary Dataset: Usage of PDF and PDF/A in Institutional Repositories—Interviews with Individuals working with Institutional Repositories

To better understand uptake and usage of the PDF and PDF/A file formats, it was recommended that representatives of institutions using the file formats be interviewed. This research was reviewed by the University of Illinois at Urbana-Champaign Institutional Review Board (IRB) and was approved as exempt test protocol #18056 (see appendices 5-8). A call for participation in the study was sent to several listservs serving individuals working in digital preservation and digital repositories across the globe. Individuals or groups of individuals responded to the call for participation on behalf of their institution, stating their interest in participating in the survey and their preferred mode of participation: (1) written questionnaire, (2) phone interview, or (3) video conference.

3.2. Methodology

Findings of the PDFlib Bavaria report (2009) and the research completed for Florida Virtual Campus (2013) discussed in Chapter 2 guided the selection of migration and validation tools used for this research. In addition to considering the software success of previous research, the cost of software factored into tools selected for testing. Only open source software or software that offered free trials were used for testing of the original dataset and the extended dataset.

3.2.1. Original dataset workflow

The 56 theses were divided into 5 testing batches by the date they were received from the ORA selection team member. The 104 unique theses files consisted of Microsoft Word for

Windows (.docx), Microsoft Word Document (.doc), LaTeX (.tex), and Portable Document Format (.pdf) files. Source files were conformed to or created as PDF/A,[4] resulting in 636 total derivative PDF/A files. Each batch underwent the same workflow and testing criteria as follows:

1. Collect source file metadata
    a. Input batch source files in DROID
    b. Run file identification
    c. Save report as .csv
    d. Record DROID output of file size (bytes), file format, format version
2. Create as or Conform to PDF/A
    a. Creation tools (source file→PDF/A): 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]
    b. Conformance tools (PDF→PDF/A): Adobe Acrobat DC [2015], callas pdfaPilot Desktop [v. 7], LibreOffice [v. 5], PDFTron PDF/A Manager CMD [v. 1.x]
3. Check for PDF/A conformance
    a. If file did not successfully conform/create, check non-conformance with conformance/creation tool Preflight (only available with Adobe Acrobat DC [2015], callas pdfaPilot Desktop [v. 7], PDF Studio [v. 12] (produced incomplete Preflight), and PDFTron PDF/A Manager CMD [v. 1.x])
    b. If file successfully created/conformed, validate with veraPDF [v. 0.8]

[4] Throughout this thesis, the terms "create" and "conform" are referenced. "Creating" a file as PDF/A requires that the process begin with a source file (i.e., .doc, .docx, .odt, .rtf, .tex), which is then processed to conform to the ISO 19005 standard. "Conforming" a file to PDF/A begins with a PDF file, which is then processed to conform to the ISO 19005 standard.

4. Record migrated file metadata
    a. Input batch of migrated files in DROID
    b. Run file identification
    c. Save report as .csv
    d. Record DROID output of file size (bytes)

3.2.2. Extended dataset workflow

Following a similar approach to that described for the original dataset workflow, the 19 theses and other complex documents were conformed to PDF/A. Original source files were not included in the extended dataset, and thus, PDF creation tools were not necessary for testing. The key difference in the secondary dataset workflow was aimed at creating more PDF/As of various flavors. As such, all born-digital documents have been conformed to PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3a, PDF/A-3b, and PDF/A-3u. All digitized documents have been conformed to PDF/A-1b, PDF/A-2b, and PDF/A-3b. Another revision to the workflow has been to use a later release of veraPDF [v. 1.6.3]. A total of 403 derivative PDF/A files were created.

3.2.3. Secondary dataset

Participants in the survey were asked a series of questions, which can be found in appendix 8. Phone interviews and video conferences were encouraged to invoke more in-depth discussion than written responses would allow.

3.3. Criteria for Failure or Success

This research considers two factors for determining the conformance of a PDF/A file: validation of ISO 19005 conformance and evaluation of significant properties.

3.3.1. ISO 19005 Validation

For the purposes of this research, veraPDF v. 1.6.3 was used to validate PDF/A files. veraPDF returns an output stating that a file either *Passed* or *Failed* validation against a PDF/A specific flavor validation profile (PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3a, PDF/A-3b, PDF/A-3u). If the file fails, veraPDF returns a list of *Validation information* that details which rules have been violated, the number of violation occurrences, and the document location of the violation, as seen in figure 1.

Statistics

Version:	1.6.3
Build Date:	2017-06-12T10:44:00+01:00
Processing time:	00:00:47.520
Total rules in Profile:	103
Passed Checks:	555432
Failed Checks:	1721

Validation information

Rule	Status
Specification: ISO 19005-1:2005, Clause: 6.3.5, Test number: 3	
For all CIDFont subsets referenced within a conforming file, the font descriptor dictionary shall include a CIDSet stream identifying which CIDs are present in the embedded CIDFont file, as described in PDF Reference Table 5.20	Failed
4 occurrences	Hide
PDCIDFont	
fontName.search(/[A-Z]{6}\+/) != 0 || CIDSet_size == 1	
root/document[0]/pages[283](1701 0 obj PDPage)/contentStream[0](2117 0 obj PDContentStream)/operators[746]/font[0](QEFSYS+TimesNewRomanPS-ItalicMT)/DescendantFonts[0](QEFSYS+TimesNewRomanPS-ItalicMT)	
root/document[0]/pages[245](1467 0 obj PDPage)/contentStream[0](2152 0 obj PDContentStream)/operators[233]/font[0](SDSDKP+TimesNewRomanPS-BoldItalicMT)/DescendantFonts[0](SDSDKP+TimesNewRomanPS-BoldItalicMT)	
root/document[0]/pages[246](1479 0 obj PDPage)/contentStream[0](2154 0 obj PDContentStream)/operators[206]/font[0](LQBQSC+ArialMT)/DescendantFonts[0](LQBQSC+ArialMT)	
root/document[0]/pages[283](1701 0 obj PDPage)/contentStream[0](2117 0 obj PDContentStream)/operators[787]/font[0](WNWVPR+TimesNewRomanPSMT)/DescendantFonts[0](WNWVPR+TimesNewRomanPSMT)	

Figure 1. veraPDF failed validation output for PDF/A-1b validation profile.

While veraPDF validation provides a baseline for identifying non-conforming features in documents which claim to be PDF/A according to the migration software, some failed output features are simply impossible to overcome unless a document was created with the intent to

conform to PDF/A. In the instance of deposits to an institutional repository, the depositor's document creation process may not be conducive to creating a PDF/A file. Thus, veraPDF establishes an impractical standard for PDF/A conformance because not all source files or PDF files will create as or conform to PDF/A. The discussion found in Chapter 5 explores the challenges of creation and conformance tools. The veraPDF Consortium was cognizant of institutions' limitations and, thus, developed the ability to implement a *Policy Profile* in the software so that veraPDF will validate only against clauses required by an institution for a "valid" PDF/A (Wilson, McGuinness, & Jung, 2017). The *Policy Profile* was not implemented in this research.

3.3.2. Embedded Features Validation

Despite veraPDF's thoroughness and flexibility, it does present limitations, of which the veraPDF Consortium was aware when developing the software. ISO 19005 extends across hundreds of pages of specifications and the standard only continues to grow. There are assumed specifications not detailed in the ISO 19005 standards that overextend the objectives of veraPDF; thus, "the consortium decided to restrict the scope of the proposed development to the clauses contained in the PDF/A standards themselves" (Wilson, McGuinness, & Jung, 2017, p. 161). In their overview of the software, Carl Wilson, Rebecca McGuinness, and Joachim Jung (2017) identify four cases for which veraPDF does not validate: JPEG2000, Fonts, ICC profiles, and Tagged PDF. For example, as noted in Chapter 2, PDF/A-1 does not consider the presence of JPEG2000 embedded images because the standard was released while JPEG2000, standardized as ISO 15444, was still fraught with criticism as an unsustainable format (Adams, 2013). Therefore, creation and conformance software cannot output a conforming PDF/A-1 if JPEG2000 images are present. PDF/A-2 and PDF/A-3, however, provide clauses of support for

JPEG2000, referencing the ISO 32000 specification for PDF 1.7. ISO 32000, and describe how a JPEG2000 should be embedded in PDF 1.7 files. Still, the ISO 32000, ISO 19005-2:2011, and ISO 19005-3:2012 standards do not identify the requirements of a valid JPEG2000.

3.3.3. Significant Properties

In addition to consideration of the formal requirements of ISO 19005 that are validated with veraPDF, documents were visual inspected for their significant properties.

From 1998-2002, JISC funded the CURL (Consortium of University Research Libraries)[5] Exemplars in Digital Archives (Cedars) Project, which was led by the University of Cambridge, the University of Leeds, and the University of Oxford. In addition to the general conversation of digital preservation, results of the Cedars project have become foundational in the ongoing discussion of significant properties. Among many digital preservation projects since Cedars, are four that have contributed considerably to the understanding of what is referred to throughout the literature and this paper as "significant properties."

- CAMiLEON (Creative Archiving at Michigan and Leeds: Emulating the Old on the New), University of Michigan and University of Leeds, funded by NSF and JISC, -2003[6]
- InSPECT (Investigating Significant Properties of Electronic Content Over Time), Arts and Humanities Data Service, Centre for e-Research, and The National Archives [UK], funded by JISC, 2007-2009[7]

[5] http://web.archive.org/web/20041011141405/http://www.curl.ac.uk/projects/cedars.html http://www.ukoln.ac.uk/metadata/cedars/

[6] CAMiLEON "[Investigated] the feasibility of using emulation as a digital preservation strategy" (Hedstrom & Lee, 2002, p. 222).

[7] https://web.archive.org/web/20160303182529/http://www.significantproperties.org.uk

- Paradigm (**P**ersonal **Ar**chives **A**ccessible in **Dig**ital **M**edia), University of Oxford and University of Manchester[8]
- PLANETS (Preservation and Long-Term Access via Networked Services), EU project, ended in 2010[9]

Of these, InSPECT has been perhaps the most influential in defining significant properties. The InSPECT Framework Report notes five factors for considering "significance":

1. "Significance is relativistic, rather than being universal and unchanging;
2. Interpretations of significance will differ dependent upon the intended purpose and the criteria that is applied
3. Meaning may be intrinsic in the construction of an item.
4. Meaning is conveyed through a process of communication from a source
5. Meaning may be interpreted differently by stakeholders, dependent upon their knowledge base, environment in which they operate and other factors." (Knight, 2013, para. 2)

Knight's first and second points are of particular note because they suggest that there can be no permanent definition for significant properties of a digital object. Depending upon the context both of use and of use at a particular time, the significant properties that an institution considers requisite will vary. The metric for defining significant properties for this research, thus, is not extensible for all institutions. That said, while some institutional repositories may not bear concern about specific clauses in the ISO 19005 specification, other elements that have been overlooked in the specification (Wilson, McGuinness, & Jung, 2017) may be crucial to retaining the authenticity of a digital object as created by the author.

[8] http://www.paradigm.ac.uk
[9] http://www.planets-project.eu

Park, Zou, and McKnight (2007) define a thesis as "a set of digital objects which contain a variety of different objects including images, scientific formula, bibliographic records and other semantic information" (p. 89). Their requirements for archiving are that, "it must be flexible, and can be used to preserve original information as much as possible" (Park, Zou, & McKnight, 2007, p. 89). Thus, For the purpose of this research in the context of electronic theses and dissertations, the following features have been considered factors of authenticity and thus vital significant properties:

- Embedded fonts
- Embedded vector graphics
- Embedded raster images
- Embedded non-image files, non-PDF/A[10]

Migration from a source file to PDF/A materializes the OAIS principle of "Repackaging: A Digital Migration where there is some change in the bits of the Packaging Information" (CCSDS, 2012, 5-5). By migrating a file from one format to another, the digital object bit stream is altered to the extent that the most effective relationship to the original object is recorded in the significant properties.

For this research, significant properties were tested with binary responses ("yes, the migrated file differs from the original file"; "no, the migrated file is the same as the original file"). The testing was performed first by comparing the visual appearance of the source file with the migrated file. If there was a visual change in appearance, the embedded metadata of the

[10] The dataset did not consist of non-image embedded files. Thus, this significant property was not tested.

original and migrated files' features (e.g., fonts and images) was compared using JHOVE (JSTOR/Harvard Object Validation Environment).[11]

[11] http://jhove.openpreservation.org/

CHAPTER 4. RESULTS

4.1. PDF/A Lifecycle Software

The analysis of the primary dataset consists of quantitative measures as well as qualitative measures. Qualitative measures include the author's examination of specific anomalies of non-conformance or software failure. The criteria for investigation of significant properties have been detailed in Chapter 3.

For the purposes of the analysis of migration[12] to PDF/A, the original dataset has been combined the with the extended dataset. Differences in the workflow that may have impacted the outcome of success or failure in migration are noted. The combined migration success rate of Adobe Acrobat DC,[13] callas pdfaPilot, Intarsys PDF/A Live!, LibreOffice, PDF Studio, PDFForge PDFCreator, and PDFTron PDF/A Manager CMD migration software was 65%—of the 1073 completed attempts to migrate to PDF/A 698 were successful.[14] Figure 2 separates the total migration success rate by software to illustrate a comparison of the success of software used for testing. Migration success is determined by the software, which creates a notification of success or failure of migration to PDF/A.

[12] Throughout the results, creation of PDF/A or conformance to PDF/A is reduced to the term "migrated." These terms are combined under the aegis of "migration" unless a differentiation between creation and conformance is requisite to the analysis.

[13] As noted in the methods, two versions of Adobe Acrobat Pro were used for the testing. However, because the version release is the same (i.e., Adobe Acrobat Pro DC), the analysis does not indicate a difference in the versions.

[14] This value is a combination of both the original dataset and the extended dataset.

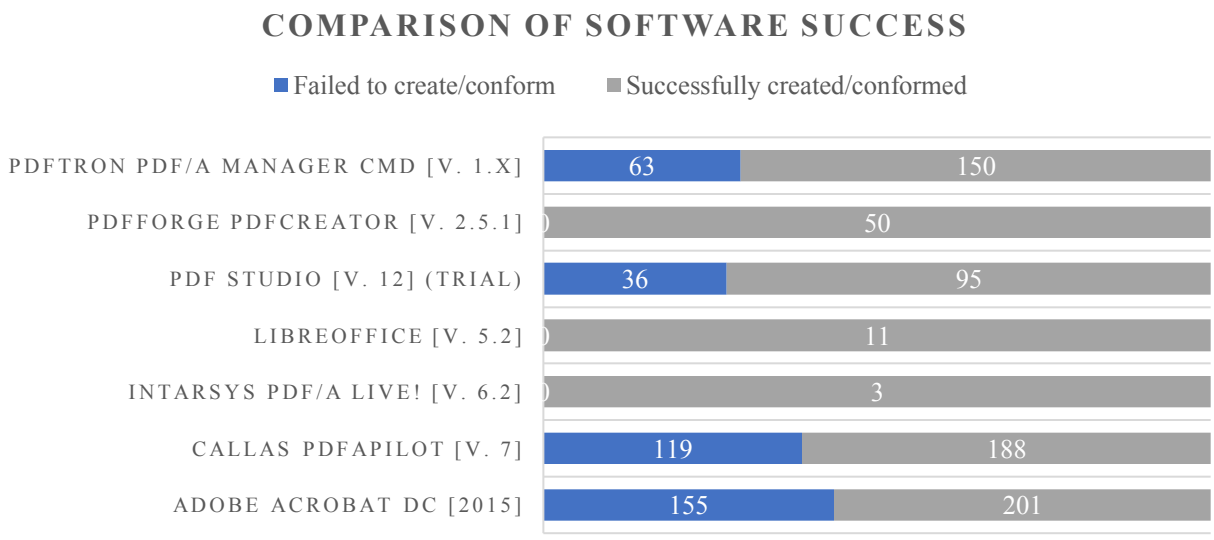


Figure 2. Comparison of software success.

As indicated in Chapter 3, files that successfully migrated were validated with veraPDF. Of the 698 files that migrated successfully, 483 files passed validated with veraPDF, producing a 69% validation success rate. Figure 3 provides a comparison of the validation success by migration software.

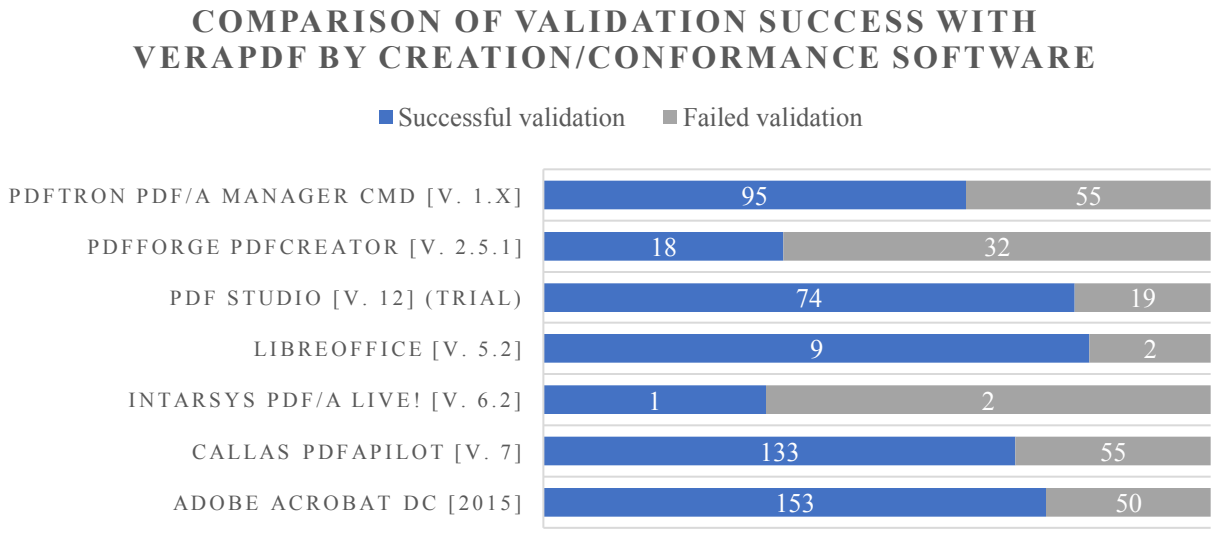


Figure 3. Comparison of validation success by software.

Figure 4 provides a comparison of validation success by PDF/A migration format.

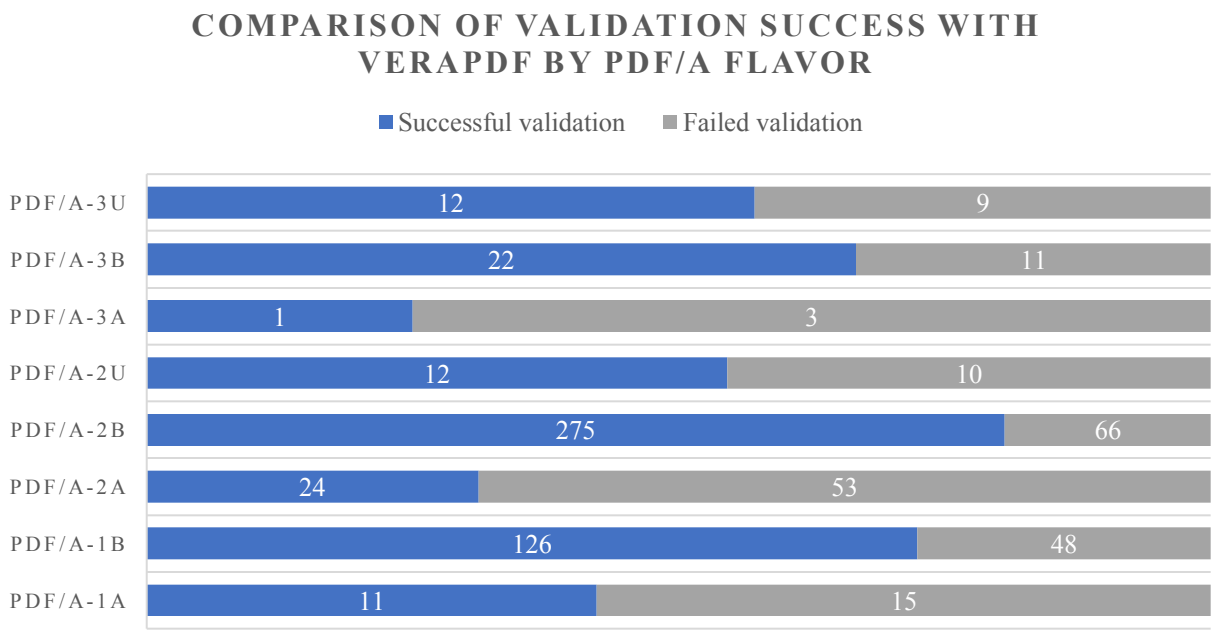


Figure 4. Comparison of validation success by PDF/A flavor.

As mentioned previously, when migrating a file to PDF/A, the software will either complete with success—i.e., creating what the software claims to be a valid PDF—or failure—i.e., software claims inability to create a valid PDF. Of the software used for testing, PDF Studio 12 would not output a file if it failed to create what it claimed a valid PDF. Thus, these results for failed PDF Studio 12 migration do not provide any file-specific information, such as migrated file size. Figures 5 and 6 provide a comparison of migrated file sizes by migration software and PDF/A migration flavor, respectively. These data have been separated by PDF/A creation and conformance software used for testing to assess whether a particular software is more likely to cause an increase or decrease in file size.

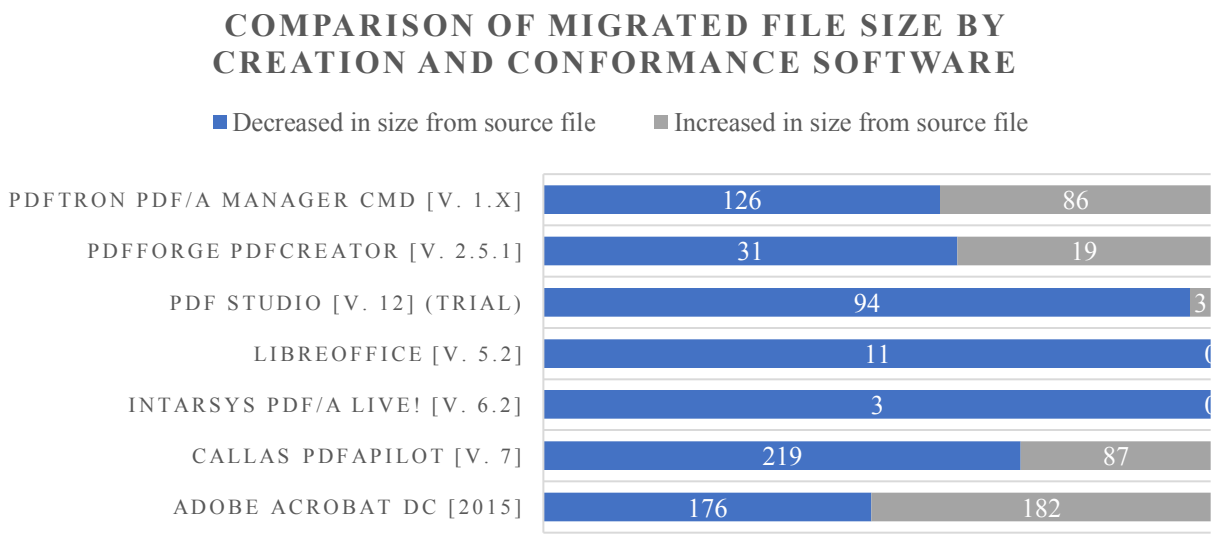


Figure 5. Comparison of migration success rate by software.

As with the figure 5, figure 6 illustrates whether a migrated file increased or decreased in size when migrated to PDF/A. The results have been separated by PDF/A flavor to assess whether a particular flavor is more likely to cause an increase or decrease in file size.

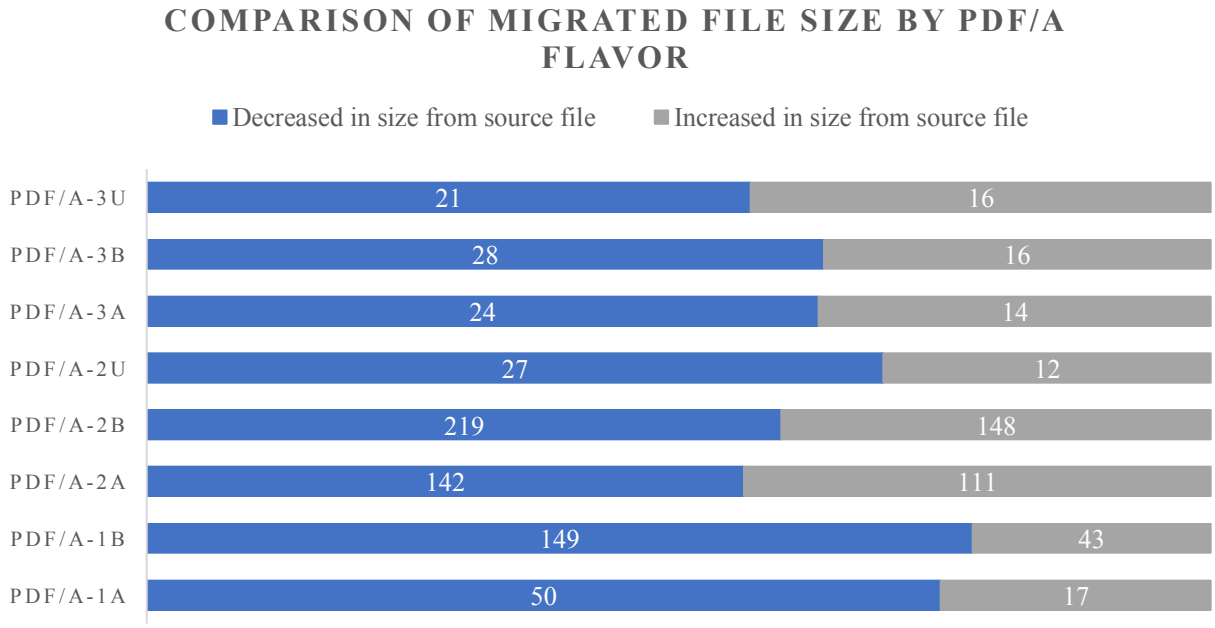


Figure 6. Comparison of migrated file size by PDF/A flavor.

4.2. Significant Properties

Of the 1073 migration attempts, the font-related significant properties of 86 files altered, whether the migration was successful or unsuccessful. This resulted in a 92% authenticity success rate (i.e., the significant properties were not impacted by migration). Changes to embedded fonts identified included migration to a different font; and embedding of "CALS_InvisibleTTFont," while capturing the original font as an embedded image. Not all embedded images were evaluated, so a comprehensive authenticity rate of embedded images cannot be provided. That said, changes to embedded images identified included migration to a different image decode filter and removal of interpolation. Specific examples of these changes are presented in sections 4.2.1. Fonts and 4.2.2. Images.

It should be noted that the failure of significant properties testing and the failure of either migration or validation shows some correlation. It is assumed that, if a software attempts to change an embedded feature to conform to ISO 19005, the migration attempt may encounter failures not caught by the software. The correlation is not found between the testing methods but rather by the metrics involved in the testing. To achieve authenticity, the original embedded features should be present in the migrated file. This accounts for both conforming and non-conforming features. The subset of non-conforming features encompasses those that cause either migration or validation to fail.

4.2.1. Fonts

When attempting migration to Level A conformance, files containing fonts and glyphs that do not map to Unicode either failed migration or were altered by the software.

Figures 7 and 8 show the transition of appearance displayed in files with non-conforming fonts when migrated with callas pdfaPilot. The software replaced the visual appearance of the

original file with a raster image (TIFF) capture, and replaced the non-Unicode embedded fonts with “CALS_InvisibleTTFont.”

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

Figure 7. Image of text from Microsoft Word source file.

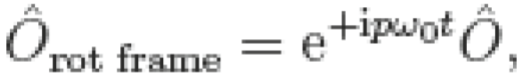


Figure 8. Image of text from PDF/A-2a migrated with callas pdfaPilot.

Adobe Acrobat DC implemented a similar change to the embedded text. Adobe embedded the original fonts—“CMMI,” “CMR,” “CMSY,” “TimesNewRomanPSMT,” “Times-Roman,” “Times-Italic,” and “Times-Bold”—as “CALS_InvisibleTTFont.” Unlike callas pdfaPilot, Adobe did not consistently capture the original appearance (figure 9), resulting in an incomprehensible visual rendering (figure 10).

The series of discrete saccades that are required to see a “whole scene,” are time-consuming.

While the eye is moving, image blur makes the visual system effectively blind. Saccadic suppression imposes an additional sensitivity loss for a period that outlasts the saccade by 50 ms or more [Tri02].

Figure 9. Image of text from PDF source file.

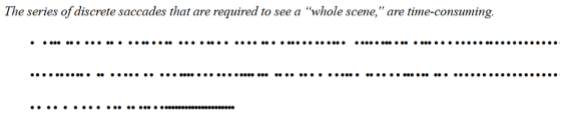


Figure 10. Image of text from PDF/A-2u migrated with Adobe Acrobat DC.

Because Adobe Acrobat DC cannot create a source file directly as PDF/A, it will first create a document as PDF and then conform the PDF file to PDF/A. Upon opening the source file in Adobe Acrobat DC, the source file font “GR Oxford” (figure 11) was embedded as “Times New Roman” and “TimesNewRomanPSMT2” (figure 12).

τὸς ἀγαθὸς ἔστερξεν Ἄρης, ἐφίλησε δ’ ἔπαινος, |
καὶ γήραι νεότης οὐ παρέδωχ’ ὑβρίσαι· |
ὦγ και Γ[λ]αυκιάδης δήιος ἀπὸ πατρίδος ἔργων |
ἦλθ’ ἐπ[ὶ] πάνδεκτον Φερσεφόνης θάλ<α>μον (= *IG* II² 10998, *GVI* 1637).

Figure 11. Image of text from Microsoft Word source file.

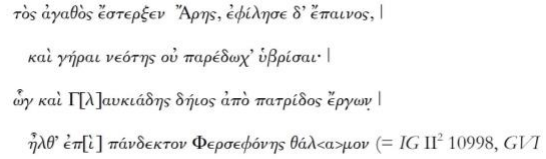


Figure 12. Image of text from PDF/A-2a migrated with Adobe Acrobat DC.

4.2.2. Images

Adobe Acrobat DC, callas pdfaPilot, and PDFTron PDF/A Manager removed interpolation from embedded images, conforming to the ISO 19005 specification which states, "If an Image dictionary contains the **Interpolate** key, its value shall be false" (ISO, 2005, p. 8). All other embedded image features, including width (553 px), height (621 px), colorspace (DeviceRGB), and compression (DCTDecode [JPEG compression]) remained the same after migration. While the embedded features remained the same, the pixel array was not consistent across migrations. Figures 14 (callas pdfaPilot) and 16 (PDFTron PDF/A Manager) exhibit the same pixel array, while the pixel distribution in figure 14 (Adobe Acrobat Pro) is different.

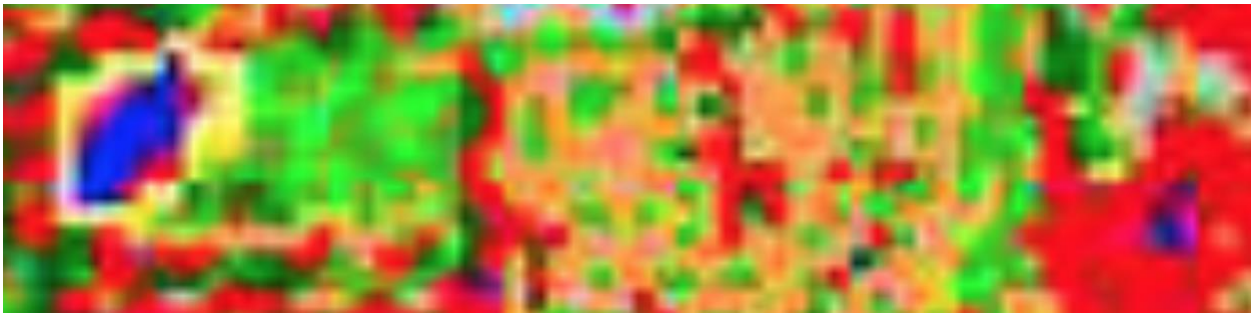


Figure 13. Image of embedded hyperspectral image in source PDF (version 1.5) file, where Interpolate key = "true."

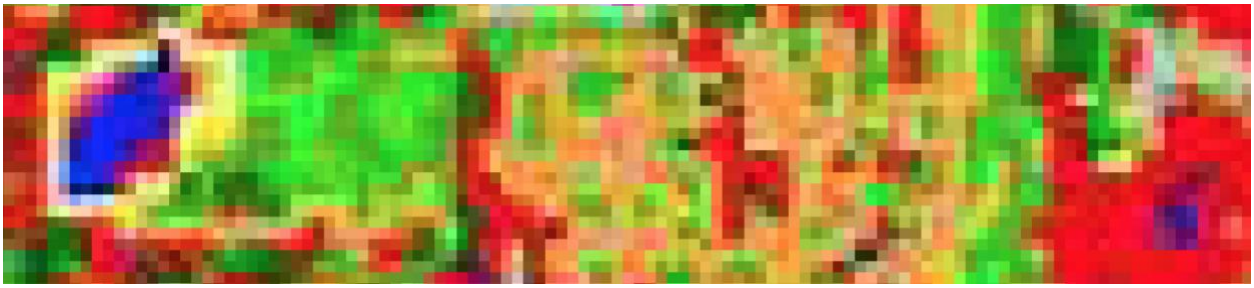


Figure 14. Image of embedded hyperspectral image in PDF/A-2a migrated with Adobe Acrobat DC.

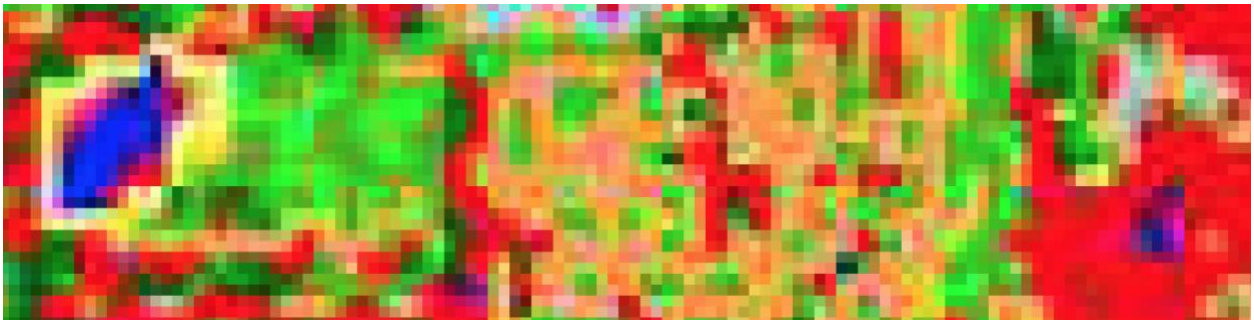


Figure 15. Image of embedded hyperspectral image in PDF/A-2a with callas pdfaPilot.

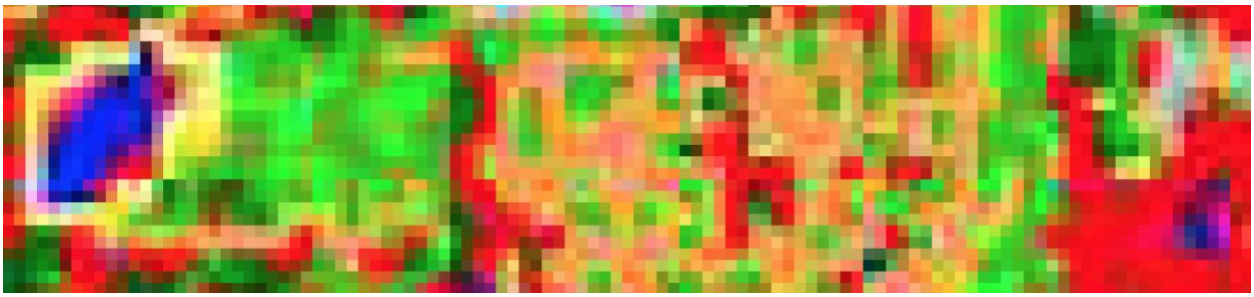


Figure 16. Image of embedded hyperspectral image in PDF/A-2a with PDFTron PDF/A Manager.

4.3. Secondary Dataset: PDF/A Survey

Three respondents formally participated in the survey—two by written questionnaire and one by phone interview.[15] All respondents were representatives of cultural heritage institutions, with a range of representation in size of institution and institutional missions. All institutions reported that they use PDF/A (appendix 8, Question A) and all institutions recommended the use of PDF/A (appendix 8, Question J). However, when asked which flavor of PDF/A their institution would recommend, two responded with PDF/A-1b and the third responded with PDF/A-2b. The complete results of the responses are presented in Chapter 5: Discussion.

[15] There was also informal participation in the survey by four participants representing their institutions. Those responses have not been included in the survey results but fortuitously have affected the discussion of this thesis. These informal participants contributed in an informal capacity—two by written response, one by phone interview, and one by video conference. Due to the variability of the responses and the nature of the institutional structure, informal responses have not been included in the Secondary Dataset: PDF/A Survey results.

CHAPTER 5. DISCUSSION

5.1. Analysis of Results

As demonstrated in Chapter 4, the migration software imposed myriad changes to the embedded and visual appearance of data. These changes are discussed in sections 5.1.1., 5.1.2., and 5.1.3.. Additionally, the complete results of the survey are provided and discussed in section 5.1.4..

5.1.1. Software and Source Format

As found in the Florida Virtual Campus (2013) study on PDF/A and implementation tools, identifying a robust and reliable migration software is requisite to successfully recommend PDF/A as a preferred file format. The results of this thesis suggest that software present two primary challenges for institutions: 1) software are not always successful in achieving ISO 19005 conformance, and even when they are successful, 2) software may generate a loss of significant properties in the migrated file.

After examining success and failure rates of migration and validation, it was discovered that there may be a correlation of success between the source file and the migration flavor. When conforming a PDF file from one version to a flavor of PDF/A, the PDF file must be altered 1) to meet the criteria of the ISO standard and thus, 2) to meet the criteria of the underlying PDF (i.e., version 1.4 or version 1.7). Given the popularity of PDF as a dissemination format, students will deposit their thesis or dissertation as a PDF without also providing the source file. Such practice limits institutions from creating optimal files. The research completed for this thesis indicates that files directly created as PDF/A have a slightly higher success rate than files that are conformed from PDF to PDF/A (see appendices 9 and 10): 43% success rate for .pdf, 67% success rate for .doc, and 53% success rate for .docx.

Unlike the figures in appendices 9 and 10, where versions of both PDF Microsoft Word processing files are separated, here, PDF versions have been condensed into one group but Microsoft Word processing files are distinguished and analyzed separately as .doc and .docx. Versions of files are given unique PUIDs (PRONOM Unique Identifier). While Microsoft Word Document 97 - 2003 [PUID fmt/40] (i.e., .doc) and Microsoft Word for Windows 2007 onwards [PUID fmt/412] (i.e., .docx) are commonly referred to as different "versions" across file format policies, .doc and .docx are distinct files in their core.[16] "Versions" of Microsoft Word documents differ by their underlying structure, where a .doc is a single binary file and a .docx is a zipped folder consisting several elements stored as separate .xml files that contain information about font encoding. "Versions" of PDF differ by admitting or improving file functions, as seen in Chapter 1: table 1, but the underlying file structure remains consistent.

If textual representation is retained in the image layer of a PDF file but the embedded text layer is stripped of the text written in document creation and replaced with a ISO 19005 conforming text that fails to capture the glyphs, the content is lost (Klindt, 2017). This affects the accessibility of content. Thus, Tagged PDF, a requirement for ISO 19005 Conformance Level A, becomes another factor for consideration (ISO, 2005). In addition to "declaring and describing the logical structural aspects of document content" (ISO, 2005, p. 19), Tagged PDF ensures accessibility for assistive reading technology. In not requiring Conformance Level A for Tagged PDF, the accessibility of content cannot be guaranteed for assistive reading technology (Adobe Systems Incorporated, 2008).

[16] Inasmuch, referring to them as different versions of Microsoft Word processing files misrepresents them simply as different versions rather than as distinct file structures.

As discussed in Chapter 2, PDF/A does not allow document features that are considered harmful for long-term preservation, such as audio and video, JavaScript, and 3D vector graphics. Such objects are considered harmful for long-term preservation due to their unsustainable and volatile nature. Thus, they are prohibited from inclusion in a conforming PDF/A. This implies two standards for document creation: 1) files which will be created as PDF/A should not contain non-conforming objects, or 2) non-conforming objects embedded in files which will be created as PDF/A should be changed to conforming objects. When considering the first option, non-conforming features that are requisite for document creation should be saved as a separate file. The separate, non-conforming file, then, may be referenced in the conforming PDF/A to retain all content. The second option would allow the embedding of non-conforming objects as long as those objects are migrated to a conforming object. For example, a 3D vector graphic may be migrated to a raster image with a conforming compression scheme.

5.1.2. Fonts

When source files failed either to migrate to PDF/A or failed ISO 19005 validation, the underlying error was due to the presence of non-conforming fonts or glyphs. As shown in figures 7-12 (Chapter 4: Discussion, Fonts), there were several variations of migration failure identified. In the first instance (figures 7-8), the source file contained both conforming and non-conforming fonts. However, one of the software used for migration, callas pdfaPilot, generalized the non-conformance over the entire document, embedding all fonts as "CALS_InvisibleTTFont." This error speaks to a failure of the software, which attempts to overcompensate, no matter the required sources, in order to achieve conformance. Figures 9-10 illustrate a similar problem when migrated with Adobe Acrobat DC. Unlike callas pdfaPilot, Adobe did not capture the visual representation of the text as an image. When embedding "CALS_InvisibleTTFont," the

visual representation was then lost. Finally, it was found that authenticity was most extensively damaged when migrating from a source word-processed file to PDF and then to PDF/A. Figures 11-12 illustrate the change imposed by Microsoft Office Word 2007 PDF export, where the embedded font was changed. Both the source Microsoft Office for Word file and the exported PDF contained Unicode-mapping fonts. Thus, it is unknown what caused this change to occur when the author exported as PDF.

Another instance of damage caused by PDF export was found in a non-thesis containing Klingon glyphs. However, the font set was not embedded when the source author exported the document as PDF using LibreOffice 5.0. In the PDF files used for this research, the font existed in the PDF files only as a flat vector image with an incomprehensible embedded Latin character set. Thus, the content was lost prior to migration from PDF to PDF/A. This particular instance speaks to the failure of technology systems, in which the user is unaware of the implications of migrating a word-processed document (e.g., .docx) to a page description format (e.g., PDF). Without continuous interaction with a document creator throughout the lifecycle of a digital object, content may unintentionally be lost.

5.1.3. Images

The loss of interpolation negatively affects hyperspectral image data. As shown in figures 13-16 (Chapter 4: Results, Images), content was misrepresented in each migration to PDF/A. In addition to loss of interpolation, the distribution of pixels was different throughout PDF/A flavors,[17] indicating that the software made different changes in migration depending upon the migration flavor. In addition to inhibiting interpolation, migration to ISO 19005-1:2005

[17] Only images PDF/A-2a are included in the illustration of interpolation redaction in figures 13-16.

conformance (PDF/A-1a and PDF/A-1b) does not admit embedding of JPEG2000 images (JPXDecode). Upon identifying JPXDecode in migration to PDF/A-1, callas pdfaPilot automatically changed the decode filter, which resulted in a different distribution of pixels.

5.1.4. Survey

It is vital to understand that the survey was conducted in support of the OIDLPP project and to increase the author's understanding of PDF/A usage across memory institutions. The survey results are not intended to indicate a comprehensive representation of PDF/A uptake and PDF/A policies. The resulting data presented below informed the extended dataset and are useful to the discussion.

Question A: Does your institutional repository use PDF/A?

All respondents' institutions accept PDF/A. One respondent remarked that, "We use PDF/A for most of the textual archival materials we digitize from paper, but we don't require PDF/A for born-digital collections" (Respondent 2).

Question B: If your institution has not integrated PDF/A as an institutional repository standard file format, do you have other technical requirements for your PDF files (e.g., structural composition, visual rendering, semantic properties, embedded fonts, colour space, linearization, etc.)?

This question was not applicable to any of the respondents; however, one commented that, "We treat our born-digital materials as a balance between value and preservability…we deem [born-digital materials] valuable enough to accession, we commit to a minimum of bit-level preservation and aim for long-term usability of the digital content through migration/normalization if needed" (Respondent 2).

Question C: If your institution uses PDF/A, did your institution research the format before implementing its use into the repository?

All respondents commented on researching the format, but they did not describe the level of their research.

Question D: Does your institutional repository have policies regarding the use of PDF/A for long-term preservation of electronic documents?

While none of the respondents were able to share a file format policy for their institutional repository, one institution noted that they believed PDF/A may still be absent from in-depth discussion of policy recommendations because it is still considered a new format (Respondent 3). This assertion is evident in the recommendation of PDF/A-1b usage (see Question E). Another respondent reported its inclusion in their digitization policies (Respondent 2). The final respondent provided an in-depth overview of their approach to PDF/A:

> PDF/A files are created from PDFs collected from the 'wild' as a last-resort data normalization effort, not a preservation effort, nor do we serve the PDF/A to the public. The use case being that should the original PDF fail to function we have some comfort that the PDF/A would retain the original's intellectual content but not all of its functional content or look-and-feel…. conversion to any flavor of PDF/A entails some risk of failing to capture all of the original's significant properties. If required, we could remediate failures at the file level, but currently we do not have the staffing to perform such in-depth work. (Respondent 1)

Question E: If your institution uses PDF/A, what** Versions **(i.e., ISO 19005-1:2005 PDF/A-1, ISO 19005-2:2011 PDF/A-2, ISO 19005-3:2012 PDF/A-3) and** Conformance Levels **(i.e., Level A (Accessible), Level B (Basic) Level U (Unicode)) does your institution recommend?

Respondents 1 and 2 recommended PDF/A-1b. Respondent 3 recommended PDF/A-2b. Respondent 3 also noted that Conformance Level A would be ideal for born-digital materials so that Tagged PDF data would be available for future usage. The interviewee noted, however, a lack in robust software available for achieving Conformance Level A. This response was not surprising. In their file format policies, NARA (Todd, 2009) and the LC (Library of Congress, n.d.d) recommend the use of PDF/A-1b. Respondent 3 reported having done extensive research

on the versions and conformance levels, and chose PDF/A-2b as the preferred version due to its inclusion of JPEG2000 embedding (JPXDecode filter).

Question F: Has your institution integrated veraPDF into the workflow for PDF/A validation?

All respondents were aware of the veraPDF software. Respondent 2 noted having considered veraPDF as a risk assessment tool for born-digital collections. Respondent 3 implemented veraPDF for validation upon finding that veraPDF catches more non-conformances than the Adobe Acrobat Preflight.

Question G: If not, are there specific reasons for not using veraPDF as the primary mode of validation of PDF/A files?

Respondent 2 noted that they "use [Adobe] Acrobat Pro [Preflight]…without a validation step that uses additional software."

Question H: If not, what is your institution's PDF/A validation process?

Respondent 1 wrote that, "We do not have any specific validation criteria at the moment. Our only criteria is being able to successfully write a PDF/A-1b using Ghostscript. We would investigate if that process failed."

Question I: If a PDF/A does not validate, what non-conformances does your institution allow? (Please provide a list of non-conformance exceptions; e.g., glyph-related, image related.)

Respondent 1, which "[uses] Ghostscript via a scripted batch process to generate PDF/A files as part of a repository pre-ingest workflow," identified two errors that occasionally returned when using their Ghostscript module:

1. When the color space parameter was set to sProcessColorModel=DeviceCMYK, Ghostscript returned: *"GPL Ghostscript 9.20: PDFA doesn't allow images with Interpolate true."*
2. "*GPL Ghostscript 9.20: Annotation set to non-printing, not permitted in PDF/A, reverting to normal PDF output*"

They found that, "Changing the parameter to '-sProcessColorModel=DeviceRGB' resolves the [first] warning, but we have not explored the consequences to the resulting PDF/A-1b"

(Respondent 1). Respondent 2 remarked upon "striking a balance between value and the resources required for continued preservation and access" of materials.

Question J: Speaking on behalf of your institution, would you recommend the use of PDF/A?

All respondents recommend the use of PDF/A. In their response to this question, Respondent 2 reverted to their response to Question I, recognizing that institutions will always be bound by resource constraints. However, they recommended, when possible, the use of PDF/A.

Question K: What are some barriers that institutions might encounter when implementing PDF/A into their workflows?

Respondents made note of software constraints for batch processing (Respondent 3) and institutional constraints, which require "workflows that are both realistic and flexible" (Respondent 2). One respondent's answer resonated clearly with the primary dilemma explored in this thesis: "Identifying the **risk management consequences** of choosing a specific implementation, e.g. lowest-common-denominator approach or individual file-level inspection and remediation" (Respondent 1).

5.2. Evaluation of Sociotechnical Context

PDF/A "Forbids dynamic content" (PDF Association, n.d.c) and prohibits editing, making the document static—unable and unwilling to change—dead. In the instance of PDF, the technology was created to circumvent limitations of operating systems in order to disseminate information. By creating a file format that accurately represented information as the creators—or "developers" (Akrich, 1992)—intended, Adobe enabled dissemination of information. The various standard subsets of PDF, including PDF/E, PDF/UA, PDF/VT, and PDF/X follow the original intent of PDF—to exchange information in accordance with the original representation.

PDF/A is the exception to this technological purpose. As a format developed to serve the purpose of long-term preservation, dissemination is not an apparent priority.

In this discussion of exchange, the word "dissemination" is used rather than "access" for their minute but fundamental differences. According to the Oxford English Dictionary (OED) Online (2017), dissemination is "The action of scattering or spreading abroad seed, or anything likened to it; the fact or condition of being thus diffused; dispersion, diffusion, promulgation." In this definition, dissemination may be understood as a platform that enables the use of an object or the policies that allow an object to be or not to be used. The OED Online's (2017) sixth entry for "access" defines access as "The fact or possibility of being approached or reached….Frequently with reference to the ease or difficulty of this." This word, then, refers to the innate ability to use or not to use an object. In the context of technological objects, "Access" is used to reference the technological limitations of a digital object for providing access to a group of defined users.

The truism, "preservation is access [dissemination]" suggests that PDF/A might fulfill the requisite as a dissemination format; however, the ISO 19005 standard implicitly suggests that accurate representation of and dissemination of information is not a priority. For example, if a document contains an embedded 3D vector graphic drawing, if migrated to PDF/A, the content will no longer be completely represented. Rather, it will be a still and rasterized capture of a previously transmutable data structure, and information will be lost as the 3D vector graphic is forced into a flat, raster image. While the visual representation of content is largely preserved and disseminated, the original access to the content is lost.

5.3. Evaluation of Significance in Theses and Dissertations

Even after creating informed file format policies for digital preservation, institutions are still limited to fulfilling the recommendations of those policies by their institutional infrastructure. As found from the survey (Question I, Respondent 2), the institution was limited by the donations of the content holders and resources constraints. Archival repositories, for example, may exclude specific file formats in their acquisition policies, but doing so limits the information retained for historical research. In the cases of ONB (Austrian National Library) and BDLSS, which accept materials produced under the institution (e.g., ETDs), the institutions "[are] not able to legally enforce these guidelines [for creating preservable PDFs]" (Strodl et al., 2007, p. 243). With these constraints, it is requisite for institutions to consider what constitutes authenticity.

For several decades, born-digital documents have continued to introduce challenges for digital preservation, where "electronic surrogates created by digital preservation projects are now effectively equal to the original items"[18] (Teper & Kraemer, 2002, p. 65). While Electronic Theses and Dissertations (ETD) are primarily text documents, generally created with a word-processing software, they are complex text documents, often containing digital objects in addition to the text, such as images (raster images and vector graphics) and mathematical formulas. These features introduce additional complexities for digital preservation that may not be present in other textual documents only containing structured text. As "unique university works" (McMillan, 2004, p. 161), the significant properties of ETDs cannot solely be identified

[18] Here, Teper and Kraemer (2002) use the term "digital preservation projects" in the context of digitization as a means of preservation (p. 65).

for their word-processed properties.[19] In addition to the somewhat agreed upon understanding of those "significant characteristics," there must also be considerations of a word-processed document that has been purposed as a thesis or dissertation. These are unique pieces of work created under the aegis of a university or college as a fundamental element proving the success of a students' academic and research endeavors (Teper & Kraemer, 2002). The purpose of preserving ETDs, then, falls within a greater ecosystem of the university,[20] which considers ETDs as definitive of student scholarship, as consummate of the institutional expectations of excellence, and as representative of scholarly standard for students.

From Dappert and Farquhar's (2009b) research and the extended conversation, it is evident that significant properties are not the only consideration for digital preservation. In their article, "Modeling Organizational Preservation Goals," Dappert and Farquhar (2009a) include significant properties in their model of preservation guiding factors. However, as seen in the model, these are factors that inform requirements. They are not explicit requirements. In the context of PDF/A, this means that the file format (PDF/A) is not the only action of preservation. Furthermore, the file format is only a guideline that will likely not be the only factor considered for preservation.

Cited in McMillan's (2004) article is Teper & Kraemer's (2002) "Long-term Retention of Electronic Theses and Dissertations," which references Jeff Rothenberg to define ETDs by the following four principals:

> First, preserving the item required its ability to be copied perfectly. Second, preservation required that individuals had the ability to access the information without geographic

---

[19] Archivematica provides a list of suggested significant characteristics of word processing digital objects: https://wiki.archivematica.org/Significant_characteristics_of_word_processing_files

[20] The term "university" is used to describe entities whose mission is to educate and increase scholarship.

> restraint. Third, the preservation of digital information required that the item be machine-readable. Finally, the preservation of born-digital information required that an institution preserve the unique functionality of the original item. (Rothenberg, 1999, p. 65)

While much of this definition has become obsolete for requirements of digital accessibility, the final requirement to "preserve the unique functionality of the original item" (Rothenberg, 1999, p. 65) speaks to the subliminal question underlying this thesis: what constitutes a significant property? And thus, what content must be retained for successful long-term preservation?

Cited across the literature, including Dappert and Farquhar (2009a; 2009b) and NARA (2009), is Knight's (2009) definition of significant properties:[21] "The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record" (Knight, 2009, p. 4). While defining significant properties assists in understanding the concept of discussion, Wilson (2007) recognizes a deficit of discrete definitions for significant properties and suggests that they cannot be defined universally. Significant properties, then, must be evaluated in context. Institutions should identify possible significant properties of a type of digital object and predict use-case scenarios of those types of objects. Then, those institutions will be able to define the significant properties across their collections.

The significant properties selected for this research included properties that can be tested (e.g., file size) and properties that may require evaluation that informs the testing (e.g., difference in appearance of image; retrieving embedded properties of the image with JHOVE). While both instances of significant properties are retrievable from files, determining the weight of their

[21] Significant properties are contextualized and defined for this research in Chapters 2 and 3.

significance has proved challenging throughout the research for this thesis. As experienced in Hedstrom, Lee, Olson, and Lampe's (2006) research on user's reactions to migration and emulation of digital objects, (videotape and text document), PDF/A does not have the capacity to retain all "look and feel." For the purpose of this research, significant properties refer to the "look and feel" (Hedstrom et al., 2006) of digital objects, in addition to more generalized features,[22] such as word count or color space, which may not always impact the "look and feel" of the document. Hedstrom, et al. (2006) recommended future research on "aspects of 'look and feel' that warrant preservation" (p. 187). However, as discussed throughout this thesis, creating a list of significant properties that purposes as the holy grail for all digital objects is inadequate. The range and function of digital objects is so extensive that it requires specific assessment of each object and those objects' institutional use cases in order to make effective decisions regarding which significant properties—"look and feel"—are requisite for preservation of authenticity.

Emulation and migration inherently change the representation of a file and the user's experiences with that file. As Guttenbrunner and Rauber (2012) astutely note,

> the program 'rendering' the [digital] object is neither necessarily the program originally used to render it nor the one that will be used to render it and thus the results [of migration] are not necessarily authentic to the original rendering once the object is rendered in a different environment. (p. 159)[23]

---

[22] See Archivematica's list of significant properties of word processing documents (footnote 19).

[23] The author has not cited this definition as it was originally printed in Wilson's report on significant properties because the document located did not include the final phrase cited in Knight. Wilson defines significant properties as "The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects" (Wilson, 2007, p. 8). Note that sources, including Dappert and Farquhar (2009) have incorrectly accredited this definition to Wilson (2007), who does not include the final phrase in the InSPECT Significant Properties Report.

Here, Guttenbrunner and Rauber (2012) suggest that the interaction with a file will not be the same if carried out in a different environment than the original environment used for rendering, making the rendering environment a significant property. Born digital theses and dissertations are generally first created in a word-processing software and then migrated to a version or subset standard of PDF. Thus, Guttenbrunner and Rauber's (2012) recommendation cannot be achieved because the original rendering software environment cannot effectively be recreated after the file has been migrated.

While the visual data may be retained with conformance of ISO 19005, there still exist negative impacts on the long-term use of embedded textual data and on accessibility. Migration software will force ETDs which contain non-Unicode glyphs or non-mapping fonts to conform to the standard, thus mapping non-conforming fonts and glyphs to those which will conform. This damage is twofold: 1) the glyph may be mapped to a string of Latin glyphs that do not represent the visual representation and meaning of the text in the source file, or 2) the font may be mapped to a conforming font which does not capture the historical, semantic, and visual evolution of the written language.

Hence, an exploration of authenticity has been so integral to this research and discussion. A common digital preservation action is to check the fixity of a digital object. If a checksum changes during the lifecycle of a digital object, it can be assumed that the object has been corrupted to some extent. Whether the corruption was imposed by unintentional human intervention or due to a non-human moderated change (e.g., bit rot), the document will be assumed as changed—no longer the original. Other common digital preservation actions include migration of file formats. When migrating from one format, the fixity will inevitably change. While the metric of fixity has been stripped from consideration of authenticity, ultimately, the

file must be considered the same as the previous, un-migrated version. It is still no longer the original—only a representation of the original, and as such, the authenticity must be documented in the significant properties.

Ultimately, this question must be asked: Is the preserved document a dead digital object? When migrated to PDF/A, a document should no longer admit editing. If the document changes, it no longer conforms to the ISO requirements of a conforming PDF/A. Thus, the removal of editable objects does not affect the purpose of usage as prescribed in ISO 19005. Rothenberg's (1999) definition of ETDs suggests that ETDs are not a dead digital object, but rather objects that possesses functionality. Then, by taking any measures to achieve compliance, the institutional objective, as suggested by Rothenberg, has failed. The digital object is dead.

5.4. Future Research

As found from the responses to Survey: Question I, an in-depth analysis of non-conformances may be useful to bring awareness of the meaning of non-conformance and whether non-conformance is actually harmful for digital preservation. Among the data collected for this research were the non-conformances for each migrated document as identified by either the migration software built-in Prelight or veraPDF. Due to the large capacity of data output, this data was considered out of scope and thus has not been evaluated in the results and discussion. While this data was not fully analyzed, it was used to identify whether loss of authenticity was due to an attempt to conform to the criteria as required by the standard. It would be useful to repurpose this data to develop a risk assessment metric for measuring validity (ISO 19005 conformance) and authenticity (significant properties) of file migration to PDF/A.

CHAPTER 6. CONCLUSION

Throughout this thesis, the author has attempted to remain objective in the discussion of PDF/A. While it is recommended that institutions retain the original submission format and migrate to the most similar format (word-processed format to word-processed format), it is understood that not all institutions will have the infrastructure or resources to achieve these preservation actions. If an institution chooses a singular file format for storage of all ETDs, it is requisite to make informed decisions before establishing file format policies for digital preservation. When considering PDF/A, whether an institution chooses ISO 19005-1:2005, ISO 19005-2:2011, 19005-3:2012, or 19005-4:[2018] as their preferred version of PDF/A compliance, they must consider whether each criteria of the standard restrict or are damaging to document features that persist throughout their collections. Thus, a consideration of the collection scope and risk factors to the collection should be considered when choosing a preferred version of ISO 19005 conformance.

The data collection for this project began before the publication of ISO 32000-2, which introduces additional complexities for PDF that are expected to be expanded for PDF/A in ISO 19005-4. Even if ISO 19005-4 mitigates the failure of validation and loss of authenticity discovered in this research, the primary point of discussion of this thesis persists. As PDF becomes embedded across memory institutions as a formal standard, PDF/A's existence as a *de jure* standard for long-term preservation of electronic documents becomes ever more present. PDF is a standard for an electronic document; PDF/A is a standard for a sustainable electronic document. Neither, however, can perform as sufficient solutions for exchange of all electronic documents.

REFERENCES

"access, n." (2017, June). *OED Online*, Oxford University Press. Retrieved from www.oed.com/view/Entry/1028

Adams, C. (2013). Is JPEG-2000 a preservation risk? [Blog post]. Retrieved from http://blogs.loc.gov/thesignal/2013/01/is-jpeg- 2000-a-preservation-risk/

Adobe Systems Incorporated. (1992). *1992 annual report.* Retrieved from http://corphist.computerhistory.org/corphist/documents/doc-447f58fec06f8.pdf

Adobe Systems Incorporated. (1993). *Portable document format reference manual*. Addison-Wesley Publishing Company.

Adobe Systems Incorporated. (1995). *Adobe Acrobat version 2.1.* Retrieved from http://hepunx.rl.ac.uk/~adye/acrobat-facts.pdf

Adobe Systems Incorporated. (1996a, March). *Portable document format reference manual version 1.1*. Retrieved from https://www.scribd.com/document/98936276/Portable-Document-Format-Reference-Manual-Version-1-1

Adobe Systems Incorporated. (1996b, November). *Portable document format reference manual version 1.2.*

Adobe Systems Incorporated. (2000). *PDF reference: Adobe portable document format version 1.3.* (2nd ed.). Addison-Wesley.

Adobe Systems Incorporated. (2001). *PDF reference: Adobe portable document format version 1.4*. (3rd ed.). Addison-Wesley.

Adobe Systems Incorporated. (2003a). *PDF reference: Adobe portable document format version 1.5*. (4th ed.). Addison-Wesley.

Adobe Systems Incorporated. (2003b). Adobe Acrobat 6.0 professional: For prepress and print professionals. Retrieved from http://www.planetpdf.com/planetpdf/pdfs/acro_aag_ue.pdf

Adobe Systems Incorporated. (2004). *Adobe PDF reference guide: Portable document format version 1.6.* (5th ed.). Adobe Press. Retrieved from http://proquest.safaribooksonline.com.proxy2.library.illinois.edu/0321304748

Adobe Systems Incorporated. (2008). *Accessing PDF documents with assistive technology: A screen reader user's guide.* Retrieved from https://www.adobe.com/content/dam/acom/en/accessibility/pdfs/accessing-pdf-sr.pdf

Adobe Systems Incorporated. (2017a). PDF reference and Adobe extensions to the PDF specification. Retrieved from http://www.adobe.com/devnet/pdf/pdf_reference.html

Adobe Systems Incorporated. (2017b, June 5). New features summary: 2015 releases of Adobe Acrobat DC. Retrieved from https://helpx.adobe.com/acrobat/using/whats-new-dc-2015.html?t3

Akrich, M. The de-scription of technical objects. (1992). In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205-224). Cambridge, MA: MIT Press.

American Library Association. (2008, February 21). "Definitions of digital preservation." Retrieved from http://www.ala.org/alcts/resources/preserv/defdigpres0408

Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleishhauer, C., Lazorchak, B., Morrissey, C., & Murray, K. (2014, February). The benefits and risks of the PDF/A-3 file format for archival institutions: An NDSA report. Retrieved from http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014.pdf

British Library. (2015.) PDF format preservation assessment. Retrieved from http://wiki.dpconline.org/images/e/e8/PDF_Assessment_v1.3.pdf

Brown, A. (2003, June 19). Selecting file formats for long-term preservation. *The National Archives.* Retrieved from http://webarchive.nationalarchives.gov.uk/20060820092744/http:/www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_file_formats.pdf

Caplan, P. (2008). Support for digital formats. *Library Technology Reports, 44*(2), 19–21.

CENDI Digital Preservation Task Group. (2007). Formats for digital preservation: A review of alternatives and issues. Retrieved from https://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf

Chou, C. (2006, June). Guidelines for creating archival quality PDF files. *Florida Center for Library Automation.* Retrieved from

https://web.archive.org/web/20100527230704/http://fclaweb.fcla.edu/uploads/Lydia%20Motyka/FDA_documentation/PDFGuideline.pdf.

Cochran, E. (2012). Rendering matters: Report on the results of research into digital object rendering. *Archives New Zealand.* Retrieved from http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering

Consultative Committee for Space Data Systems. (2011, September). *Audit and certification of trustworthy digital repositories. CCSDS 652.0-M-1, Magenta Book.*

Consultative Committee for Space Data Systems. (2012, June). *Reference model for an archival information system (OAIS). CCSDS 650.0.M-2, Magenta Book.* Issue 2.

Consultative Committee for Space Data Systems. (2014, March). *Requirements for bodies providing audit and certification of candidate trustworthy digital repositories. CCSDS 652.1-M-2, Magenta Book.*

Conway, P. (2000). "Overview: Rationale for digitization and preservation." *Handbook for Digital Projects: A Management Tool for Preservation and Access.* Andover, Massachusetts: Northeast Document Conservation Center. Retrieved from https://www.nedcc.org/assets/media/documents/dman.pdf

Dappert, A., & Farquhar, A. (2009a). Modelling organizational preservation goals to guide digital preservation. *International Journal of Digital Curation 4*(2), 119-134. 10.2218/ijdc.v4i2.102

Dappert, A., & Farquhar, A. (2009b). Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries* (pp. 297–308). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04346-8_29

"dissemination, n." (2017, June). *OED Online*, Oxford University Press. Retrieved from www.oed.com/view/Entry/55401.

Dreger, M. (2006). PDF/A. *DttP: A Quarterly Journal of Government Information Practice & Perspective*, *34*(2), 8–9.

Dryden, J. (2008). PDF/A-1: A Ray of Light in the Digital Dark Age? *Journal of Archival Organization*, *6*(1/2), 121–124.

Fanning, B. (2008). Getting to Know PDF/Archive (PDF/A). AIIM E-DOC, 22(3), 58–59.

Fanning, B. (2017, July). Preservation with PDF/A (2nd Edition). *DPC Technology Watch Report 17-01.* http://dx.doi.org/10.7207/twr17-01

Florida Virtual Campus. (2013, September 19). PDF/A validation and conversion in Florida digital archive. Retrieved from https://libraries.flvc.org/documents/181844/502298/PDFA+Validation+and+Convers,ion+in+FDA/4bbd8264-19e9-44ba-8a19-81e45cc9527e

Grace, S., Knight, G., Montague, L. (2009). InSPECT final report. *InSPECT.* Retrieved from https://web.archive.org/web/20151024064901/http://www.significantproperties.org.uk/inspect-finalreport.pdf

Guttenbrunner, M., & Rauber, A. (2012). Evaluating emulation and migration: Birds of a feather? In *The Outreach of Digital Libraries: A Globalized Resource Network* (pp. 158–167). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34752-8_22

Halbert, M., Skinner, K., & Schultz, M. (2012). Preserving electronic theses and dissertations: Findings of the lifecycle management for ETDs project, presented at iPres, Toronto, Canada, October 5, 2012. Retrieved from https://educopia.org/presentations/preserving-electronic-theses-and-dissertations-findings-lifecycle-management-etds

Han, Y. (2015). Beyond TIFF and JPEG2000: PDF/A as an OAIS submission information package container. *Library Hi Tech, 33*(3), 409–423.

Harvey, R. (2012). *Preserving Digital Materials.* Berlin: De Gruyter.

Hedstrom, M., & Lee, C. (2002). Significant properties of digital objects: Definitions, applications, implications. In *Proceedings of the DLM-Forum 2002, Parallel session 3*.

Hedstrom, M., Lee, C., Olson, J., & Lampe, C. (2006). "The old version flickers more": Digital preservation from the user's perspective. *The American Archivist*, *69*(1), 159–187. https://doi.org/10.17723/aarc.69.1.1765364485n41800

Hodge, G., & Anderson, N. (2007). Formats for digital preservation: A review of alternatives and issues. Information Services & Use, 27(1/2ter), 45–63.

Hovde, S., & Harling-Henry, C. (2012). University of Maryland early dissertations for doctor of medicine (1813-1889): Challenges and Rewards of a Digitization Project. *Journal of Electronic Resources in Medical Libraries, 9*(4), 261–271.

International Organization for Standardization. (2005). Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1) (ISO 19005-1:2005).

International Organization for Standardization. (2008). Document management—Portable document format—Part 1: PDF 1.7 (ISO 32000-1:2008).

International Organization for Standardization. (2011). Document management—Electronic document file format for long-term preservation—Part 2: Use of ISO 32000-1 (PDF/A-2) (ISO 19005-2:2011).

International Organization for Standardization. (2012a). Document management—Electronic document file format for long-term preservation—Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3) (ISO 19005-3:2012).

International Organization for Standardization. (2012b). Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model (ISO 14721:2012 (CCSDS 650.0-P-1.1)

International Organization for Standardization. (n.d.). Document management—Electronic document file format for long-term preservation—Part 4: Use of ISO 32000-2 (PDF/A-NEXT) (ISO/CD 19005-4).

Jaeggi, S. (2016, October 2). First presentation of the PDF concept 25 years ago. Retrieved from https://pdf-aktuell.ch/pa/language/en/first-presentation-of-the-pdf-concept-25-years-ago/

Jain, G. (2010, November 18). Now available: Adobe Acrobat Reader X for Android. *Adobe*. Retrieved from https://web.archive.org/web/20101206083149/http://blogs.adobe.com/readermobile/2010/11/18/introducing-adobe-reader-x-for-android/

Klindt, M. (2017). PDF/A considered harmful for digital preservation, presented at iPres, Kyoto, Japan, September 25-29, 2017. Retrieved from https://ipres2017.jp/wp-content/uploads/15.pdf

Knight, G. (2009, October 13). InSPECT framework report. *InSPECT.* Retrieved from https://web.archive.org/web/20150113002105/http://www.significantproperties.org.uk:80/inspect-framework.html

Knight, G., and M. Pennock. (2009). Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation 4*(1):159-174. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/110/87

Koo, J. & Chou, C.C.H. (2013) PDF to PDF/A: Evaluation of converter software for implementation in digital repository workflow. *New Review of Information Networking 18*(1), 1-15. doi:10.1080/13614576.2013.771989.

Leurs, L. (2017, March 5). The history of PDF [Blog post]. Retrieved from https://www.prepressure.com/pdf/basics/history

Library of Congress. (n.d.a). Digital preservation: About. Retrieved from http://www.digitalpreservation.gov/about/.

Library of Congress. (n.d.b). "DOCX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5." *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Retrieved from https://www.loc.gov/preservation/digital/formats/fdd/fdd000397.shtml

Library of Congress. (n.d.c). "PDF (Portable Document Format) Family." *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Retrieved from https://www.loc.gov/preservation/digital/formats/fdd/fdd000030.shtml

Library of Congress. (n.d.d)."PDF/A, PDF for Long-term Preservation." *Sustainability of Digital Formats: Planning for Library of Congress Collections.* Retrieved from https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml

Lynch, C.A. (2003, February). Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL Bimonthly Repor*t *226*, 1-7.

McMillan, G. (2004). Digital Preservation of Theses and Dissertations Through Collaboration. *Resource Sharing & Information Networks*, *17*(1/2), 159–174.

Moore, R., & Evans, T. (2013). Preserving the grey literature explosion: PDF/A and the digital archive. *Information Standards Quarterly, 25*(3), 20–27.

Noonan, D.W., McCrory, A., & Black, E. L. (2010). PDF/A: A viable addition to the preservation toolkit. *D-Lib Magazine*, *16*(11/12). doi:10.1045/november2010-noonan

Perrin, J.M., Winkler, H.M., & Yang, L. (2015). Digital Preservation Challenges with an ETD Collection — A Case Study at Texas Tech University. *Journal of Academic Librarianship*, *41*(1), 98–104.

PDF Association. (2013). "Validation: Is it really PDF/A?" *PDF/A in a Nutshell: 2.0.* Retrieved from https://www.pdfa.org/validation-is-it-really-pdfa/.

PDF Association. (n.d.a). A short history of PDF/A. Retrieved from https://www.pdfa.org/a-short-history-of-pdfa/.

PDF Association. (n.d.b). Isartor test suite. Retrieved from https://www.pdfa.org/isartor-test-suite/.

PDF Association. (n.d.c). PDF/A FAQ. Retrieved from https://www.pdfa.org/pdfa-faq/.

PDFlib. (2009, May 4). Bavaria report on PDF/A validation accuracy. *PDFlin GmbH*. Retrieved from http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf.

Rimkus, K., Padilla, T., Popp, T., & Martin, G. (2014). Digital preservation file format policies of ARL member libraries: An analysis. *D-Lib Magazine 20*(3/4). Retrieved from http://www.dlib.org/dlib/march14/rimkus/03rimkus.html.

Rog., J. & van Wijk, C. (2002). "Evaluating File Formats for Long-term Preservation." National Library of the Netherlands. Retrieved from https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf

Rothenberg, J. (1999). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council of Library and Information Resources*. Washington, D.C.: Council on Library and Information Resources. Retrieved from https://www.clir.org/pubs/reports/rothenberg/introduction/

Rothenberg, J. (2000). Preserving authentic digital information. In *Authenticity in a Digital Environment* (pp. 51-68). Washington, DC: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/abstract/pub92abst.html

Seadle, M. (2009). PDF in 2109. *Library Hi Tech, 27*(4), 639–644.

Strodl, S., Becker, C., Neumayer, R., Rauber, A., Bettelli, E. N., Kaiser, M., Hofman, H., Neuroth, H., Strathmann, S., Debole, F., & Amato, G. (2007). Evaluating Preservation Strategies for Electronic Theses and Dissertations. In *Digital Libraries: Research and Development* (pp. 238–247). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77088-6_23

Sullivan, S.J. (2006). An archival/records management perspective on PDF/A. *Records Management Journal, 16*(1), 51–56. https://doi.org/10.1108/09565690610654783

Swartz, N. (2004). Coming in 2005: International PDF-archive standard. *Information Management Journal, 38*(4), 8–8.

Teper, T. H., & Kraemer, B. (2002). Long-term retention of electronic theses and dissertations. *College & Research Libraries*, *63*(1), 61–72.

Termens, M., Ribera, M., & Locher, A. (2015). An analysis of file format control in institutional repositories. *Library Hi Tech, 33*(2), 162–174.

Todd, M. (2009). "File formats for preservation." *The National Archives.* Retrieved from http://www.dpconline.org/docman/technology-watch-reports/375-file-formats-for-preservation/file

Turró, M.R. (2008). Are PDF documents accessible? *Information Technology & Libraries, 27*(3), 25–43.

UNESCO. (2003). Guidelines for the preservation of digital heritage, prepared by the National Library of Australia. Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

The U.S. National Archives and Records Administration. (2009). *Significant properties.* Washington, DC: The U.S. National Archives and Records Administration. http://www.archives.gov/era/acera/pdf/significant-properties.pdf

Warnock, J. (n.d.). The Camelot project. Retrieved from https://blogs.adobe.com/acrobat/files/2013/09/Camelot.pdf

Wilson, A. (2007, April 10). InSPECT significant properties report. *Arts and Humanities Data Service.* Retrieved from

https://web.archive.org/web/20141019070530/http://www.significantproperties.org.uk/wp22_significant_properties.pdf

Wilson, C., McGuinness R., and Jung J. (2017). veraPDF: Building an open source, industry supported PDF/A validator for cultural heritage institutions. *Digital Library Perspectives 33*(2), 156-165.

APPENDICES

Appendix 1: Glossary of Terms

ALA — American Library Association
ALCTS — Association for Library Collections & Technical Services
BDLSS — Bodleian Digital Library Systems and Services
CAMiLEON — Creative Archiving at Michigan and Leeds: Emulating the Old on the New
CCSDS — Consultative Committee for Space Data Systems
CURL — Consortium of University Research Libraries
DPC — Digital Preservation Coalition
GNU GPL — GNU's Not Unix General Public License
ETD — Electronic Theses and Dissertations
InSPECT — Investigating Significant Properties of Electronic Content Over Time
IPS — Interchange PostScript
IRB — Institutional Review Board
ISO — International Organization for Standardization
JHOVE — JSTOR/Harvard Object Validation Environment
LC — Library of Congress
NARA — The U.S. National Archives and Records Administration
NDLTD — Networked Digital Library of Theses and Dissertations
NDSA — National Digital Stewardship Alliance
OAIS — Open Archival Information System
OCR — Optical Character Recognition
OED — Oxford English Dictionary
OIDLPP — Oxford–Illinois Digital Libraries Placement Programme
ONB — Austrian National Library
OPF — Open Preservation Foundation
ORA — Oxford University Research Archive
Oxford — University of Oxford
Paradigm — **P**ersonal **Ar**chives **A**ccessible in **Dig**ital **M**edia
PARS — Preservation and Reformatting Section
PDF — Portable Document Format
PDF/A — Portable Document Format–Archival
PDF/E — Portable Document Format–
PDF/UA — Portable Document Format–
PDF/VT — Portable Document Format–
PDF/X — Portable Document Format–
PLANETS — Preservation and Long-term Access through Networked Services
PUID — PRONOM Unique Identifier
U of I — University of Illinois at Urbana-Champaign

Appendix 2: Universities that Require PDF/A for Deposit of ETDs

The following research institutions require students to submit their thesis or dissertation deposit in PDF/A format. All additional files included in their thesis should be submitted as separate files (e.g., embedded video, audio, images, etc.). Note then, that when the theses or dissertations are accessed in the patron-facing digital repository, not all content will be accessible because only PDF/A files are uploaded onto the access server.

Institution	**Creation Directions**	**Required**
Carleton University[24]	Yes[25]	Yes
Colorado College[26]	Yes[27]	Yes; PDF/A-1a specified in converting from MacOS directions
Concordia University	Yes[28]	Yes; PDF/A-1b
Dalhousie University[29]	Yes[30]	Yes
Harvard University	Yes[31]	Requirements vary across colleges within the university
Johns Hopkins[32]	Yes[33]	Yes; PDF/A-1a recommended
McGill University[34]	No	Yes (PDF format also requested); PDF/A-1b
Memorial University	Yes[35]	Yes
Rutgers University	Yes[36]	Preferred; From Acrobat Pro X and XI: PDF/A-1a; from Acrobat 7.0 and 9.0: PDF/A-1b; from Microsoft Word 2013: PDF/A; from Mac: PDF/A-1b

[24] https://gradstudents.carleton.ca/thesis-requirements/electronic/

[25] https://gradstudents.carleton.ca/thesis-requirements/pdfa-formatting/

[26] http://coloradocollege.libguides.com/c.php?g=286906

[27] http://spectrum.library.concordia.ca/HowtoPrepareYourThesisForDepositinSpectrum.pdf

[28] https://www.concordia.ca/students/graduate/thesis/ethesis.html

[29] https://www.dal.ca/faculty/gradstudies/currentstudents/thesesanddefences/submission.html

[30] https://www.dal.ca/faculty/gradstudies/currentstudents/thesesanddefences/submission/pdfa-acrobat.html

[31] https://www.gsd.harvard.edu/wp-content/uploads/2016/07/2015-FINAL-THESIS-SUBMISSION-TO-THE-FRANCES-LOEB-LIBRARY.pdf

[32] http://guides.library.jhu.edu/etd

[33] http://guides.library.jhu.edu/etd/pdfa

[34] https://www.mcgill.ca/gps/thesis/final-e-thesis/students

[35] https://www.mun.ca/sgs/go/guid_policies/Converting_Word_Latex_PDFA.pdf

[36] https://rucore.libraries.rutgers.edu/collab/ref/doc_sawg_pdfa_acrobat_tutorial.pdf

University of Alberta[37]	Yes[38]	Yes
University of Oulu	Yes[39]	Yes
Virginia Tech	No	Yes; PDF/A-1b

[37] https://www.ualberta.ca/graduate-studies/current-students/academic-requirements/thesis-requirement-and-preparation

[38] https://cloudfront.ualberta.ca/-/media/gradstudies/current-students/academicrequirements/thesisrequirementandpreparation/20160301instructions-how-to-save-a-thesis-as-a-pdfa-archiveeffectivemarch-12016eralinkchanged-only.pdf

[39] https://laturi.oulu.fi/instructions/index.php?uilang=en-US

Appendix 3: PDF/A Creation and Conformance Tools

Software	License	Compatibility	Conform/Create	Notes
* callas pdfaPilot Desktop	Proprietary	Windows, MacOS	Conform/Create	
eDocPrintPro PDF/A		Windows	Create	PDF/A-1, PDF/A-2, PDF/A-3
FileConverterPro	Proprietary	Windows		
* Intarsys PDF/A Live		Windows, Mac OS, Linux	Conform	PDF/A-1, PDF/A-2, PDF/A-3
* LibreOffice	GNU LGPLv3 MPLv2.0	Windows, Mac OS, Linux	Create	PDF/A-1a
OpenOffice.org	GNU LGPLv3	Windows, Mac OS, Linux	Create	
* PDF/A Manager	Proprietary	Windows, Mac OS, Linux	Conform	
PDF Converter	http://bit.ly/2t7RLGw	Windows	Conform	PDF/A-1b, PDF/A-2b
* PDFCreator	http://www.pdfforge.org/	Windows	Conform	PDF/A-1b, PDF/A-2b
PDF Creator Lite	http://www.simopdf.com/pdf-creator-lite.html	Windows		Converts Office Word, Excel PowerPoint, images, HTML, and more to PDF

PDF Creator Pro	http://www.simpopdf.com/pdf-creator.html	Windows		Converts Office Word, Excel PowerPoint, images, HTML, and more to PDF
* PDF Studio Pro	https://www.qoppa.com/pdfstudio/	Windows, Mac OS, Linux		PDF/A validation (Version: Qoppa jPDFPreflight v2017R1.04 - Demo Version) and conversion
Pdftex	https://www.tug.org/applications/pdftex/			For "Generating PDF/A compliant PDFs from pdftext"[40]

* Software used for testing

[40] http://support.river-valley.com/wiki/index.php?title=Generating_PDF/A_compliant_PDFs_from_pdftex

Appendix 4: Flowchart for Methodology

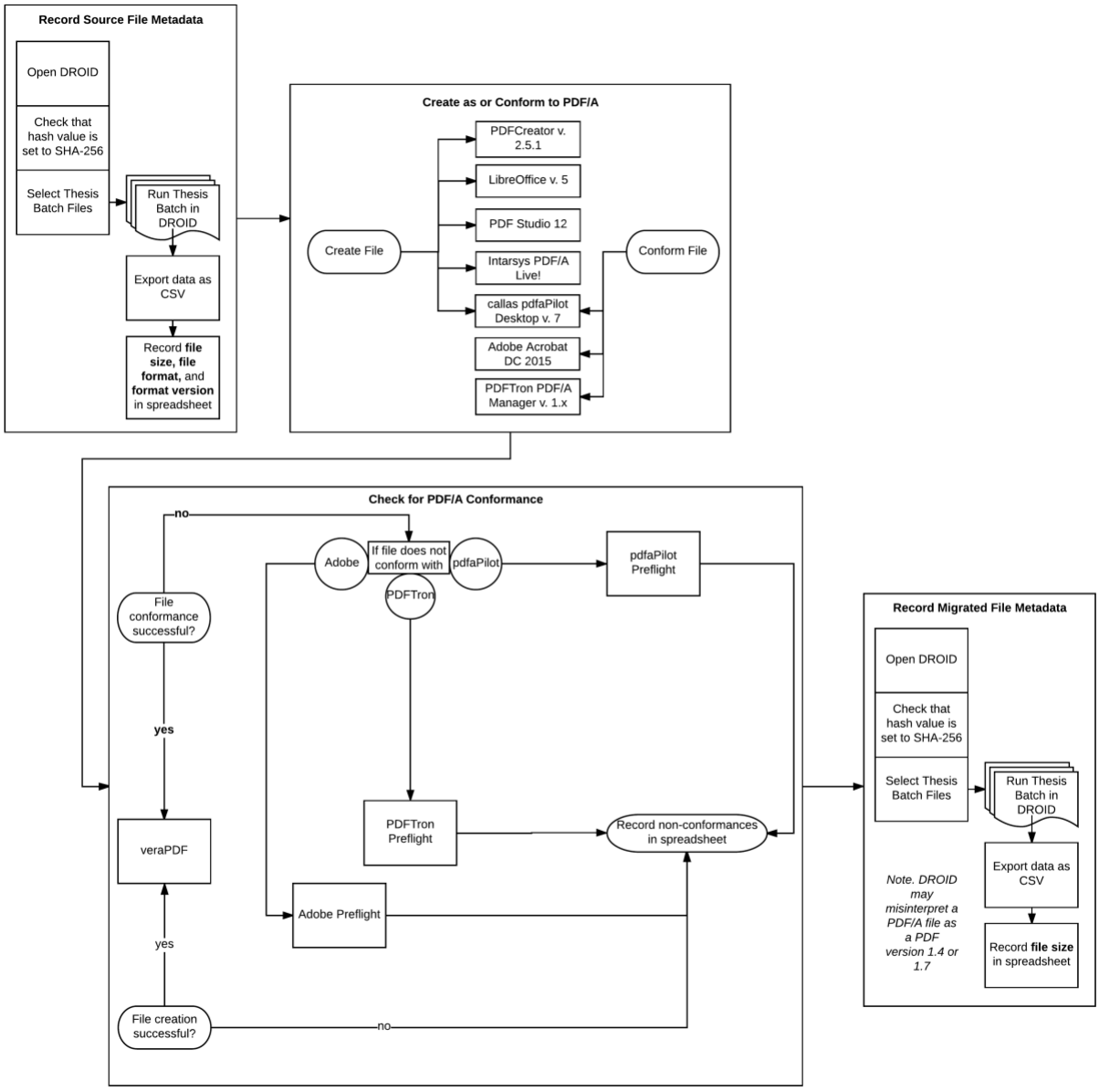


Figure 17. Complete flowchart of migration workflow.

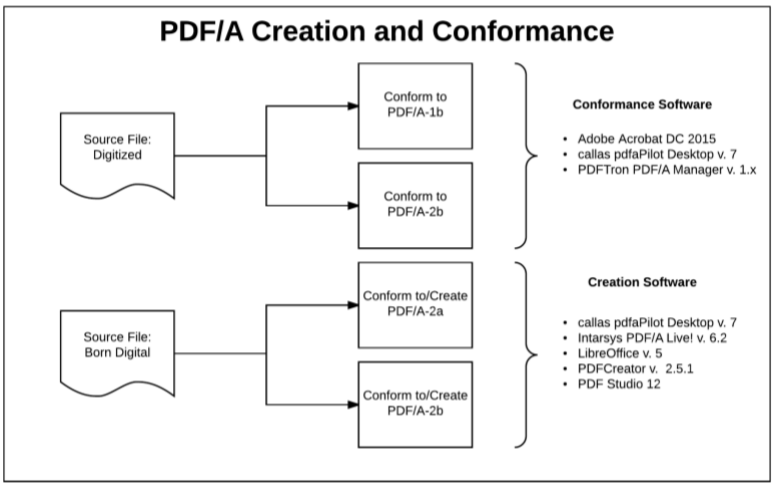


Figure 18. Decision tree for migration file destination and migration software.

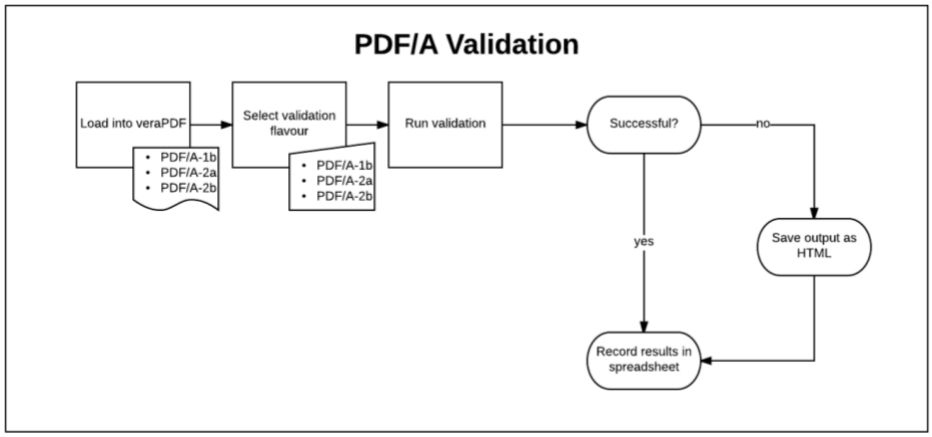


Figure 19. Flowchart of validation with veraPDF.

Appendix 5: IRB Exempt Approval

IRB EXEMPT APPROVAL

RPI Name: J Downie
Project Title: Oxford-Illinois Digital Library Placement Program 2017 – Bodleian Libraries
IRB #: 18056
Approval Date: August 1, 2017

Thank you for submitting the completed IRB application form and related materials. Your application was reviewed by the UIUC Office for the Protection of Research Subjects (OPRS). OPRS has determined that the research activities described in this application meet the criteria for exemption at 45CFR46.101(b)(2). This message serves to supply OPRS approval for your IRB application.

Please contact OPRS if you plan to modify your project (change procedures, populations, consent letters, etc.). Otherwise you may conduct the human subjects research as approved for a period of five years. Exempt protocols will be closed and archived at the time of expiration. Researchers will be required to contact our office if the study will continue beyond five years.

Copies of the approved consent form(s) (page(s) 14-15 in the attached, approved protocol) are to be used when obtaining informed consent.
We appreciate your conscientious adherence to the requirements of human subjects research. If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me at OPRS, or visit our website at http://oprs.research.illinois.edu
Sincerely,

Rebecca Miller, MSW
Human Subjects Research Specialist, Office for the Protection of Research Subjects

C: Ryan Dubnicek, Anna Oates

Office of the Vice Chancellor for Research | Office for the Protection of Research Subjects University of Illinois | Urbana-Champaign
805 West Pennsylvania Avenue, MC-095 | Urbana, IL 61801
Phone: (217) 333-2670 | Email: irb@illinois.edu
Website: http://oprs.research.illinois.edu

Appendix 6: Consent protocol

Consent Protocol – video conference, phone call, and email interviews
Oxford-Illinois Digital Library Placement Program 2017 – Bodleian Libraries
2017

You are invited to participate in this research being conducted by Professor J. Stephen Downie and Masters Student Anna Oates of the University of Illinois School of Information Sciences. The goal of this research project is to investigate the measures that research and cultural heritage institutions are taking to implement PDF/A (Portable Document Format—Archival) as a primary file format for long-term preservation of electronic documents. If your institution does not have policies or procedures for PDF/A, the research team is also interested in any implemented standards for working with standard PDF files. By participating in this study, you are helping the Bodleian Digital Library Systems & Services (BDLSS) of the University of Oxford to institute policies of best practice for their Oxford University Research Archive (ORA) digital collections of theses and dissertations. In the primary interview, you should express your interest in receiving a follow-up for the research team to collect institutional policies or workflows for working with PDF or PDF/A files.

You are free to discontinue participation in the study at any time. You are free to request that we cease recording at any time.

Activities

During *video conferences* and *phone calls*, you will be asked a series of questions, and your responses will be captured using audio recording.

For *email responses*, you will respond to a questionnaire, and your written answers will be saved.

Time commitment

Interviews conducted via *video conference* or *phone call* will last no more than 60 minutes.

Email responses to the written questionnaire should take no more than 45 minutes to compose.

Data capture and storage

Audio recordings will be transcribed and redacted, as possible (in other words, identifiable information will be deleted from transcripts wherever possible). Once transcribed, audio recordings will be deleted. Transcriptions and written responses to the questionnaire will be retained securely for 5 months after the project results have been disseminated.

Dissemination

The results of this study may be reported in papers, scholarly journals, or research conferences.

Benefits of participation

Your participation will help to create policies for the ORA theses collection and to guide the use of PDF/A in institutional repositories.

Risks of participation

Your participation in this research project is entirely voluntary, with no risks besides those of everyday life. Your participation, or your decision not to participate, will not affect your future relations with the University of Illinois at Urbana-Champaign, the University of Oxford, or any of the investigators.

Confidentiality and privacy

All answers will remain confidential and your name or identifying information will not be associated with your responses in reporting or dissemination of this research. Unless you give specific permission (see below) to link your name with your interview responses in research reports and presentations, you will not be attributed in any way. Contact information and identifiable responses will only be accessible to project personnel, and will be stored on a secure hard drive.

Laws and university rules might require us to tell certain people about you. For example, your records from this research may be seen or copied by the following people or groups: Representatives of the university committee and office that reviews and approves research studies, the Institutional Review Board (IRB) and Office for Protection of Research Subjects; University and state auditors, and departments of the university responsible for oversight of research.

Consent to participate in the interview

By continuing with this interview, you consent to participate and to the following summary points:

- You are 18 years of age or older.
- You consent for this interview to be recorded.
- You can request that the recording to be stopped at any time and the interview can proceed without being recorded.
- You can discontinue participation at any time, and you do not have to answer any questions you do not wish to answer.
- Your identity will be kept confidential

If you have questions, please contact Anna Oates at annaio2@illinois.edu. You may also contact Research Principal Investigator J. Stephen Downie at jdownie@illinois.edu.

If you feel you have not been treated according to the descriptions in this form, or if you have any questions about your rights as a research subject, including questions, concerns, complaints, or to offer input, you may call the Office for the Protection of Research Subjects (OPRS) at 217-333-2670 or e-mail OPRS at irb@illinois.edu.

Appendix 7: Call for participation

Version 1
Dear all,

Please consider participating in a 60-minute interview on your institutions' uptake of PDF/A. Optionally, consider participating simply by providing written responses to an email questionnaire.

This summer, the Bodleian Libraries at the University of Oxford hosted a placement student from the Illinois School of Information Sciences to work with the Polonsky Digital Preservation Fellows for a research project on the PDF/A (Portable Document Format—Archival) file format. This project seeks to identify which flavour(s) of PDF/A best suit the content and repository needs for theses ingested into the Oxford University Research Archive (ORA).

Our approach has been to conform PDF files to different flavours of PDF/A using Adobe Acrobat [2015] and pdfaPilot [v. 7]. Alternatively, we create files of other types (e.g., docx, doc) as PDF/A flavours using Adobe Acrobat [2015], pdfa Pilot [v. 7], LibreOffice [v. 5.2], PDF Studio [v. 12], and Intarsys PDF/A Live [v. 6.2]. For born digital documents, we create or conform to PDF/A-2a and PDF/A-2b. For digitized documents, we conform files to PDF/A-2b (in the first workflow, we conformed digitized documents to PDF/A-1b). Using veraPDF for validation, many of the PDF/A-2a documents fail due to the presence of non-Unicode (i.e., non-conforming) glyphs. Among the collection of theses containing these non-conforming glyphs are scientific papers with mathematical formulas and language papers with non-Latin glyphs.

Since we have encountered so many files that fail validation with veraPDF, my project team and I are considering investigating the possibility of disregarding some aspects of non-conformance if those factors present no preservation risks.

We are curious to learn other institutions' approaches to PDF/A or standard PDF validation, in addition to any issues encountered in the process or PDF/A creation and conformance.

I thank you in advance for your willingness to assist in my own research process.

Cheers,
Anna

Version 2

Dear all,

Please consider participating in a 60-minute interview on your institutions' uptake of PDF/A. Optionally, consider participating simply by providing written responses to an email questionnaire.

This summer, the Bodleian Libraries at the University of Oxford hosted a placement student from the Illinois School of Information Sciences to work with the Polonsky Digital Preservation Fellows for a research project on the PDF/A (Portable Document Format—Archival) file format. This project seeks to identify which flavour(s) of PDF/A best suit the content and repository needs for theses ingested into the Oxford University Research Archive (ORA).

Using veraPDF for validation, many of the theses fail due to the presence of non-Unicode (i.e., non-conforming) glyphs. Since we have encountered so many files that fail validation with veraPDF, my project team and I are considering investigating the possibility of disregarding some aspects of non-conformance if those factors present no preservation risks.

We are curious to learn other institutions' approaches to PDF/A or standard PDF validation, in addition to any issues encountered in the process or PDF/A creation and conformance.

I thank you in advance for your willingness to assist in my own research process.

Cheers,
Anna

Appendix 8: Interview questions

Oxford-Illinois Digital Libraries Placement Program – Bodleian Libraries: PDF/A Interview Questions

[online written questionnaire with introduction]

Dear [name of participant],

Attached to this email you will find a Written Consent Form, which you should read prior to continuing with the questionnaire below. If you prefer to change your method of participation to a phone interview or video conference, do not hesitate to ask.

If you would like to continue with a written response, see the questionnaire below:

A. Does your institutional repository use PDF/A? (If yes, please skip to question C.)
B. If your institution has not integrated PDF/A as an institutional repository standard file format, do you have other technical requirements for your PDF files (e.g., structural composition, visual rendering, semantic properties, embedded fonts, colour space, linearization, etc.)?
C. If your institution uses PDF/A, did your institution research the format before implementing its use into the repository?
D. Does your institutional repository have policies regarding the use of PDF/A for long-term preservation of electronic documents? (If available for public access, please attach in your response.)
E. If your institution uses PDF/A, what **Versions** (i.e., ISO 19005-1:2005 PDF/A-1, ISO 19005-2:2011 PDF/A-2, ISO 19005-3:2012 PDF/A-3) and **Conformance Levels** (i.e., Level A (Accessible), Level B (Basic) Level U (Unicode)) does your institution recommend?
F. Has your institution integrated veraPDF into the workflow for PDF/A validation?
G. If not, are there specific reasons for not using veraPDF as the primary mode of validation of PDF/A files?
H. If not, what is your institution's PDF/A validation process?
I. If a PDF/A does not validate, what non-conformances does your institution allow? (Please provide a list of non-conformance exceptions; e.g., glyph-related, image related.)
J. Speaking on behalf of your institution, would you recommend the use of PDF/A?
K. If so, what are some barriers that institutions might encounter when implementing PDF/A into their workflows?

As stated in the Written Consent Form, remember that you are free to discontinue participation in the study at any time.

Thank you for your time and participation.

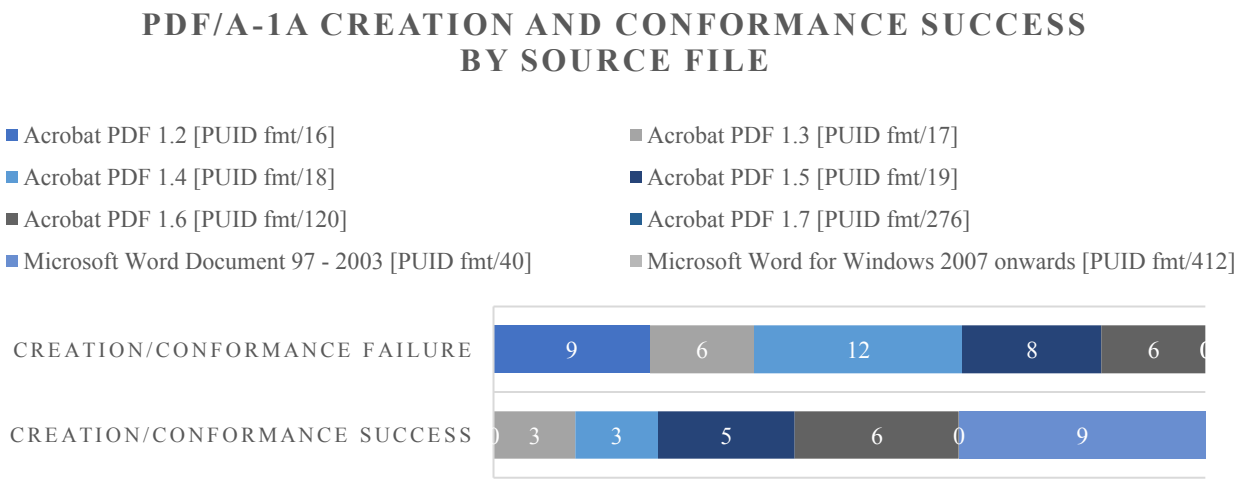
Cheers,
Anna Oates

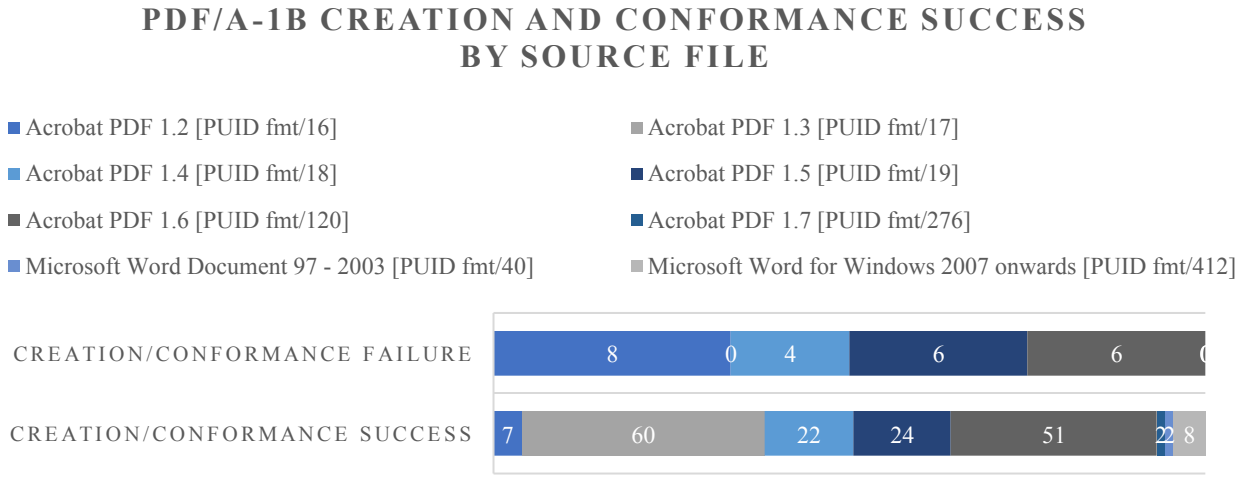
Oxford-Illinois Digital Libraries Placement Program – Bodleian Libraries: PDF/A Interview Questions

[oral interview schedule]

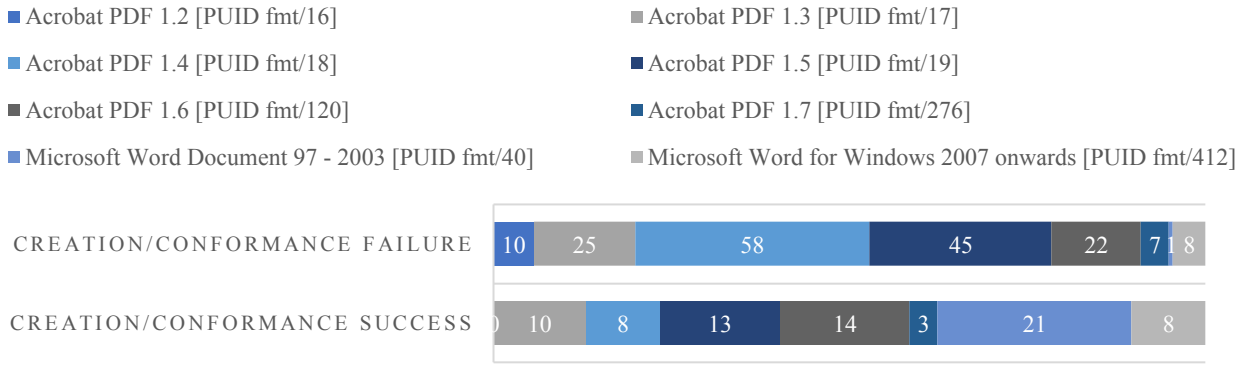
1. Does your institutional repository use PDF/A? (If yes, skip question 2.)
2. If your institution has not integrated PDF/A as an institutional repository standard file format, do you have other technical requirements for your PDF files (e.g., structural composition, visual rendering, semantic properties, embedded fonts, colour space, linearization, etc.)?
3. If your institution uses PDF/A, did your institution research the format before implementing its use into the repository?
4. Does your institutional repository have policies regarding the use of PDF/A for long-term preservation of electronic documents?
5. If so, are the policies available for public access?
6. If your institution uses PDF/A, what **Versions** (i.e., ISO 19005-1:2005 PDF/A-1, ISO 19005-2:2011 PDF/A-2, ISO 19005-3:2012 PDF/A-3) and **Conformance Levels** (i.e., Level A (Accessible), Level B (Basic) Level U (Unicode)) does your institution recommend?
7. Has your institution integrated veraPDF into the workflow for PDF/A validation?
8. If not, are there specific reasons for not using veraPDF as the primary mode of validation of PDF/A files?
9. If not, what is your institution's PDF/A validation process?
10. If a PDF/A does not validate, what non-conformances does your institution allow? (Please provide a list of non-conformance exceptions; e.g., glyph-related, image related.)
11. Speaking on behalf of your institution, would you recommend the use of PDF/A?
12. If so, what are some barriers that institutions might encounter when implementing PDF/A into their workflows?

Appendix 9: Comparison of creation/conformance success by source file.

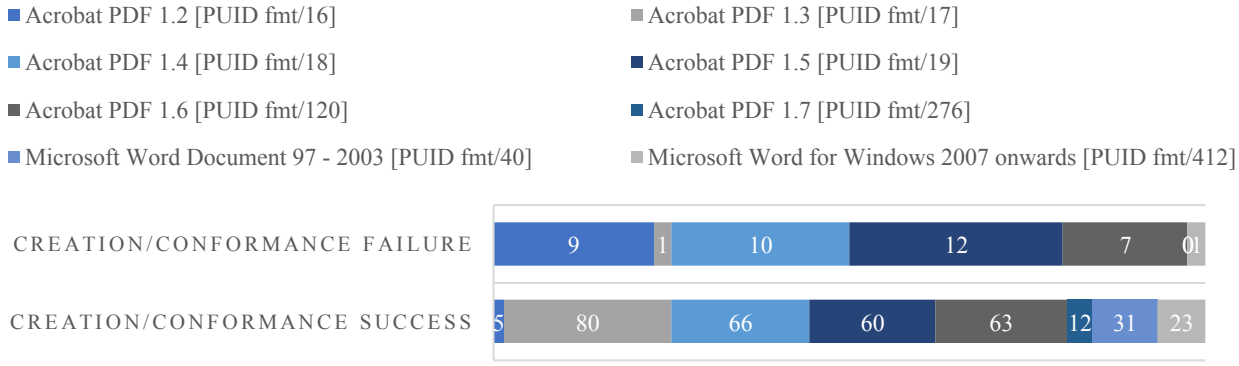




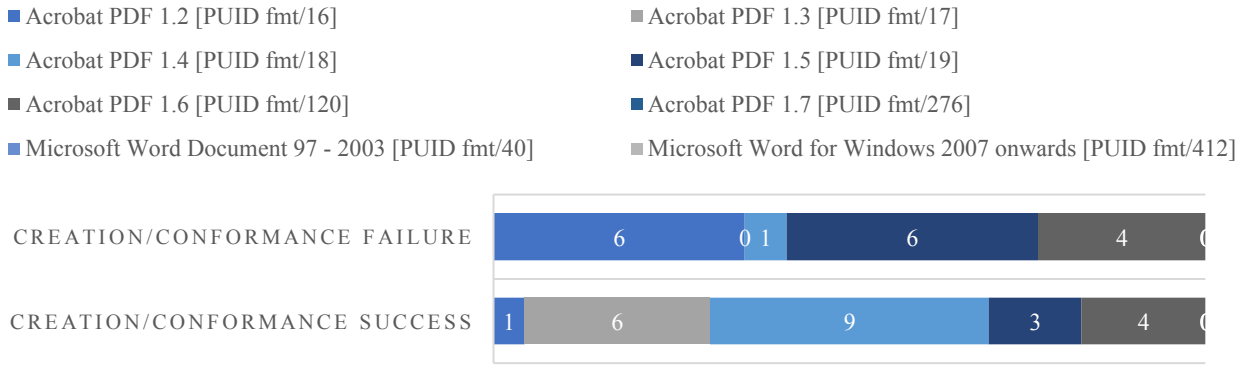
PDF/A-2A CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE



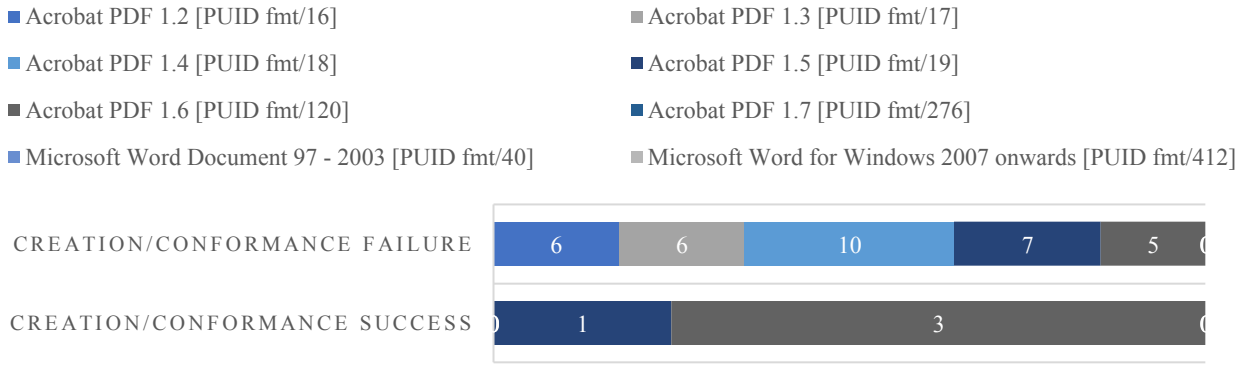
PDF/A-2B CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE



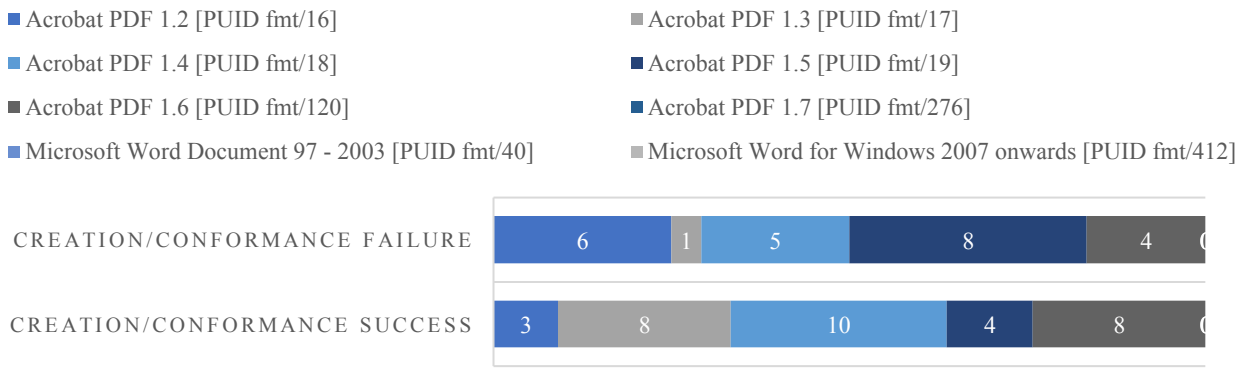
PDF/A-2U CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE



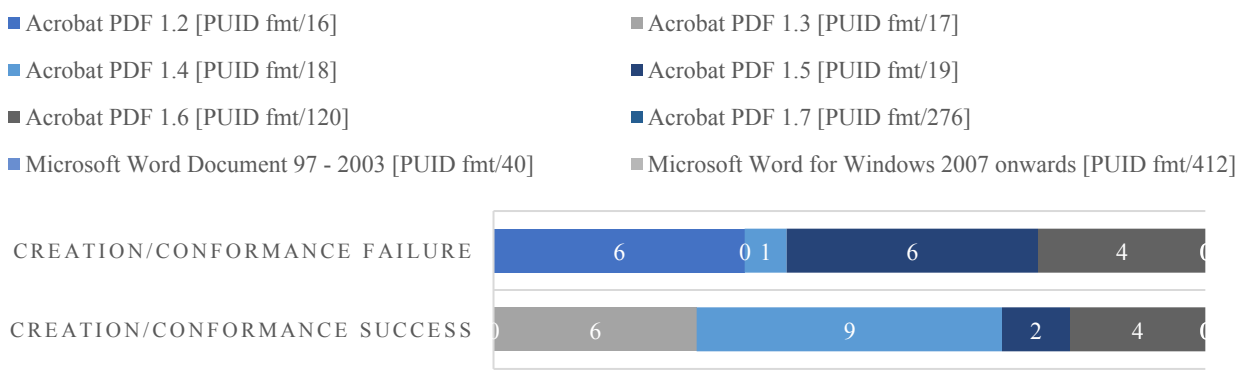
PDF/A-3A CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE



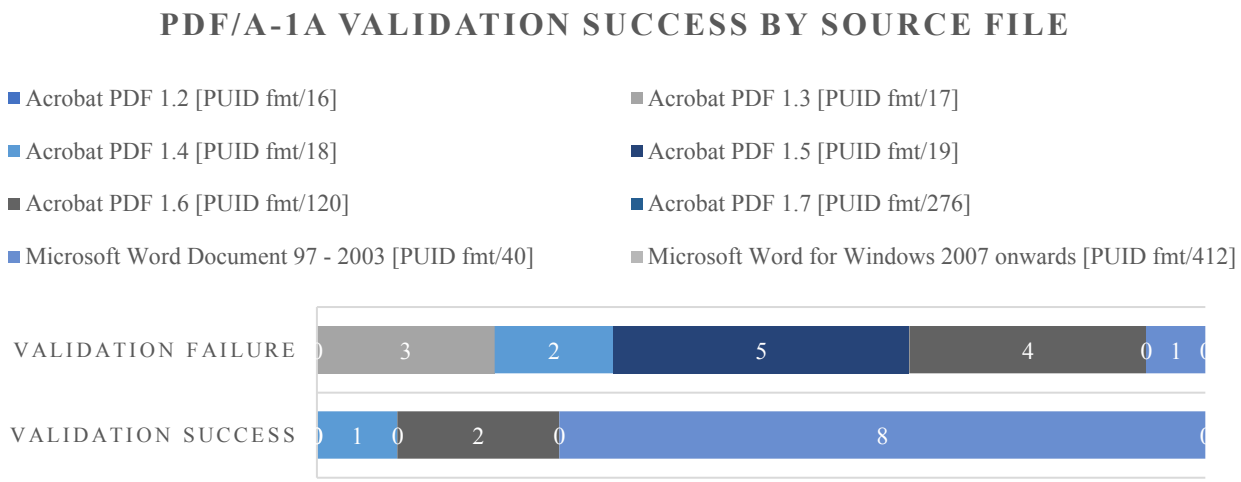
PDF/A-3B CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE

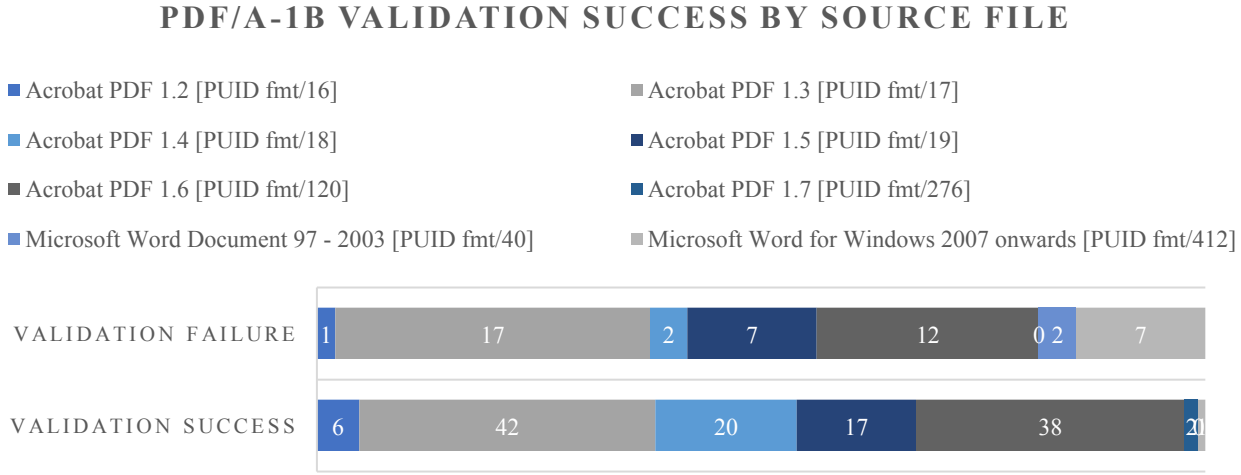


PDF/A-3U CREATION AND CONFORMANCE SUCCESS BY SOURCE FILE

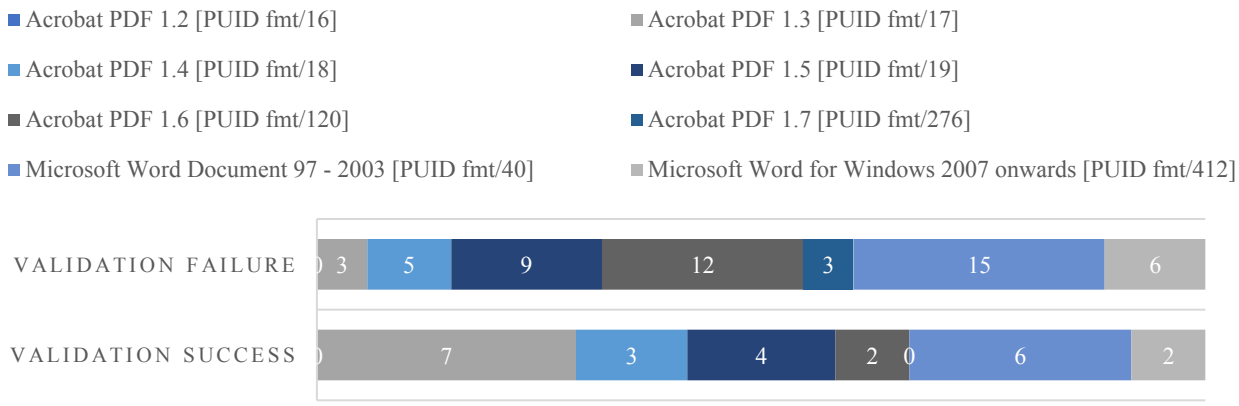


Appendix 10: Comparison of validation success by source file.

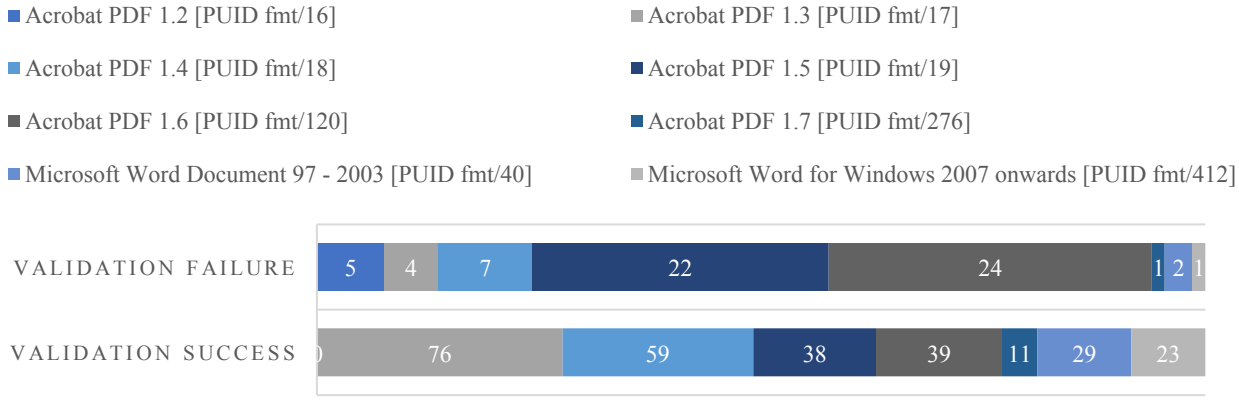




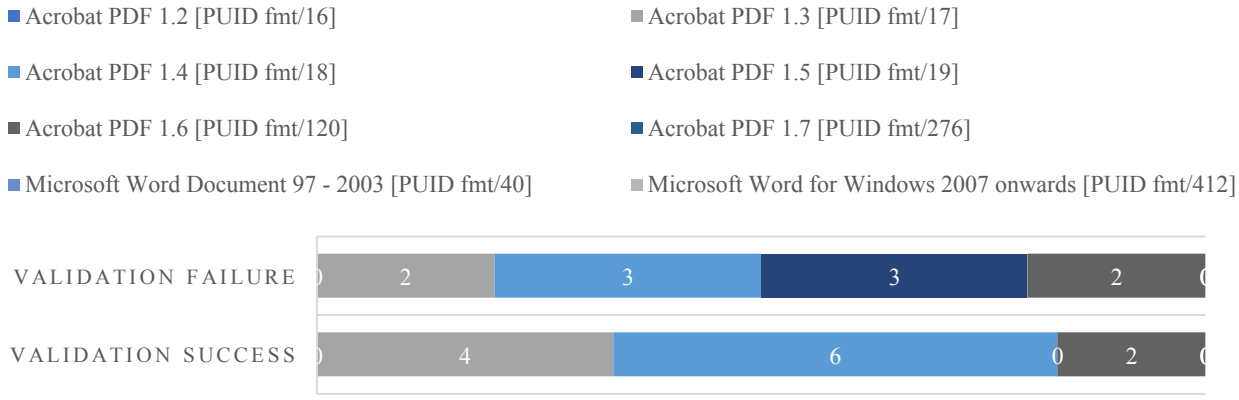
PDF/A-2A VALIDATION SUCCESS BY SOURCE FILE



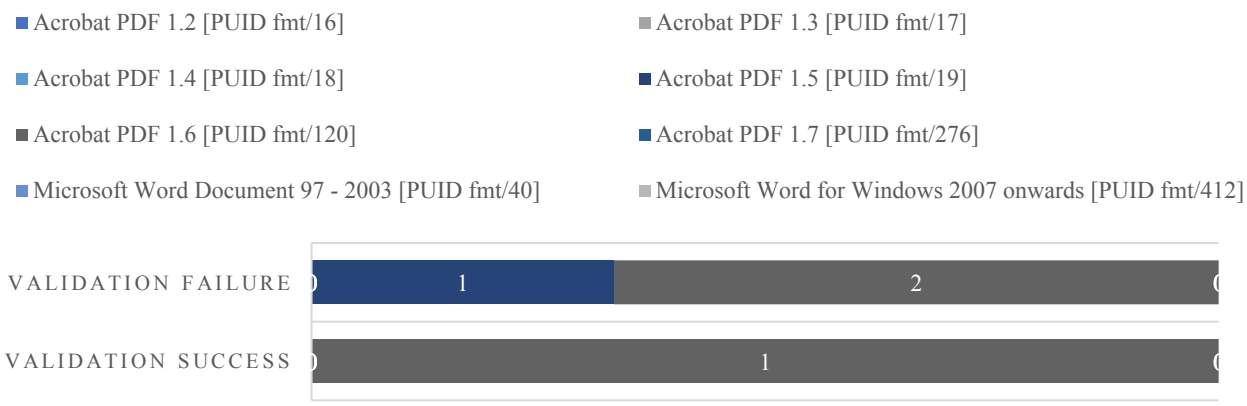
PDF/A-2B VALIDATION SUCCESS BY SOURCE FILE



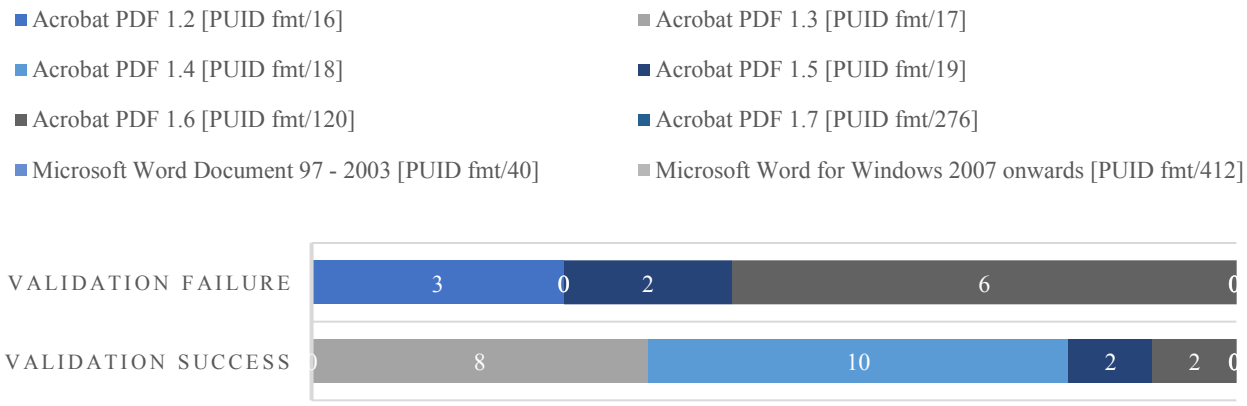
PDF/A-2U VALIDATION SUCCESS BY SOURCE FILE



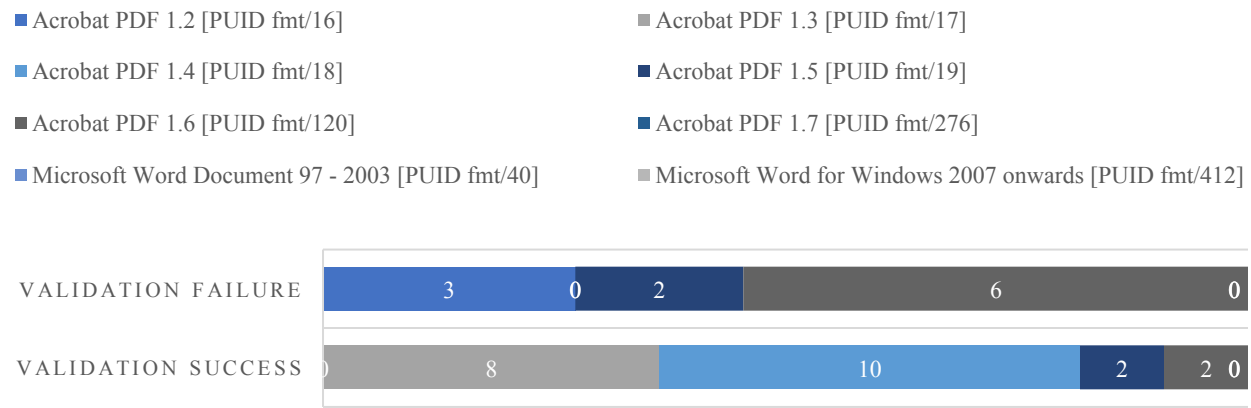
PDF/A-3A VALIDATION SUCCESS BY SOURCE FILE



PDF/A-3B VALIDATION SUCCESS BY SOURCE FILE



PDF/A-3B VALIDATION SUCCESS BY SOURCE FILE



Appendix 11: List of software used for testing and analysis.

DROID used to extract file metadata

Excel used to make charts

HexEd.it[41] used to view hex values

JHOVE used to review image metadata

OpenRefine used to combine and clean the datasets

[41] https://hexed.it