The Case For A Standard That's Old News

Recommendation of PDF/A for Digitized Newspaper Preservation

Anna Oates
Federal Reserve Bank of St. Louis
United States of America
anna.oates@stls.frb.org
0000-0002-1848-6382

William Schlaack
University of Illinois at Urbana-Champaign
United States of America
schlaac2@illinois.edu
0000-0002-8658-9885

Abstract – Since 2004, the Library of Congress, a beholden stakeholder in the risk assessment of and consideration for file formats, has supported the preservation of and access to digitized historic newspapers through the National Digital Newspaper Program (NDNP), a distributed, mass digitization program. This paper evaluates the implementation and validation of PDF as specified for NDNP, explores the benefits of PDF/A, and analyzes the adverse effects for digital preservation as realized in current digitization workflows.

Keywords – file formats, file validation, newspaper digitization, PDF, PDF/A

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The Portable Document Format (PDF), a prominent file format in the digitized newspaper ecosystem plays a vital role in the continued access for newspapers, particularly within the aegis of the National Digital Newspaper Program (NDNP), a collaboration between the National Endowment for the Humanities (NEH) and the Library of Congress (LC), which enables access to and preservation of digitized historic newspapers. NDNP specifications set requirements for PDF profiles, recommending PDF/A where no conflict exists with the NDNP PDF Profile [1].[1] While PDF/A is recommended as a preservation and access format, revision of NDNP's current validation of PDF is requisite to ensuring PDF conformance. This paper explores what features of existing PDFs in the NDNP collection do not conform to PDF/A, identifies challenges effacing PDF validation, and offers recommendations for investigating alternative validation workflows for NDNP PDF Profiles.

II. Background

A. National Digital Newspaper Program

In 2003, NDNP originated with a formal agreement between the NEH and the LC with the stated goal to provide permanent digital access to historic newspapers published in the United States. The project began in 2004 with NEH funding multiple state-level institutions ranging from public universities to state historical societies, and LC supporting awardees throughout the digitization process and hosting Chronicling America,[2] a website dedicated to providing access to freely available newspapers digitized through the NDNP. NDNP builds upon the success of the 29 year NEH and LC collaboration, the United States Newspaper Program (1982-2011), which saw all fifty U.S. states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands contribute bibliographic data for over 140,000 newspaper titles, and preserve approximately 65 million pages of newspaper on microfilm [2]. Through NDNP, grant awardees from 46 states, the District of Columbia, and one U.S. territory (Puerto Rico) have contributed over 14.8 million newspaper pages representing 24 languages published between 1789 and 1963. As the content steward for NDNP, LC works to establish and maintain imaging and bibliographic standards to manage the large-scale preservation of newspapers. When selecting titles, awardees must meet the standards specified in the program's "Technical Guidelines for Applicants".[3] These technical specifications describe an "extendable, scalable, and

[1] The most recent version of NDNP technical specifications removes specific PDF/A requirements, replacing them with general PDF/A recommendations, with a goal to minimize any conflicts with the current NDNP PDF specification (NDNP PDF Profile (Version 2.4) which can be found on the The National Digital Newspaper Program (NDNP)Technical Guidelines for Applicants 2019-2021 located here: http://www.loc.gov/ndnp/guidelines/NDNP_201921TechNotes.pdf, p. 44-45.

[2] Located here: https://chroniclingamerica.loc.gov/

[3] Both current and previous guidelines are located here: https://www.loc.gov/ndnp/guidelines/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

sustainable workflow" for awardee institutions [3]. As part of the project deliverables, awardees provide an information package that includes a TIFF, JPEG2000, ALTO XML, and PDF file for each newspaper page.

B. Portable Document Format—Archival

Since 2007, the NDNP technical specifications have asserted that, "Except where conflicting with any of the other requirements of [the NDNP PDF] profile, conforming to PDF/A (ISO 19005-1) is recommended" [4]. In 2005, the International Organization for Standardization introduced ISO 19005, a standard supporting the long-term preservation of electronic documents. Based upon PDF 1.4, ISO 19005-1:2005 specifies the use of PDF 1.4 for long-term preservation as the Portable Document Format—Archival. Since the first publication, two additional specifications have been released as ISO 19005-2:2011 and ISO 19005-3:2012, which specify how to implement PDF 1.7, standardized as ISO 32000, as PDF/A. Despite chronological ordering, the later releases of the standard do not indicate a "better or more advanced" format [5]. They are simply different approaches to creating a sustainable PDF. The implementation of ISO 19005 as PDF/A permits three levels of conformance: level a (accessible); level b (basic), and level u (unicode). Each conformance level enables the use or disuse of requirements in the specifications "to prevent the onerous requirements for full conformance presenting a barrier to software developers" [6].

TABLE I
ISO STANDARDS, LEVELS OF CONFORMANCE, AND THEIR RESPECTIVE PDF/A VARIATIONS

	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	NA	PDF/A-2u	PDF/A-3u

Among the features introduced with ISO 19005-2:2011, PDF/A-2 permits embedding of JPEG2000 images. Since its release, JPEG2000 has been subject to scrutiny for long-term preservation [7]. Despite its criticism, JPEG2000 is ubiquitous to digitization projects and is included as a required format for the NDNP file package. ISO 19005-3:2012 differs from ISO 19005-2:2011 in that the specification permits embedding of any file type, a provision which, in response to widespread criticism, resulted in a report from the National Digital Stewardship Alliance on the risks and benefits of the format [8].

Since positing the benefits of PDF/A in digitization workflows as a preservation master [5], Yan Han and Xueheng Wan tested the use of PDF/A-2b in a digitization workflow and found that a TIFF might be processed to a PDF/A-2 container and reverted to the TIFF image without altering the file fixity [9]. Han recommends PDF/A for its suitability as any information package throughout the preservation cycle—submission, archiving, dissemination—rendering the requirement of an archival TIFF redundant. Excluding TIFF from the archival information package enables less burdensome storage requirements, which is requisite to supporting long-term preservation of content produced at scale.

C. Validation

1. Digital Viewer and Validator

Before the LC accepts a submission package, informally referred to as a "batch," the content creator must validate the batch through the Digital Viewer and Validator (DVV). Developed by the Library of Congress for NDNP, the DVV "'wraps' JHOVE [(JSTOR/Harvard Object Validation Environment)] and extends JHOVE's existing TIFF, PDF, and JPEG2000 modules with the NDNP-specific validation rules…For example, while JHOVE will validate that a TIFF file conforms to the TIFF 5.0 specification, the NDNP extension validates that the TIFF file is uncompressed, 8-bit grayscale, and contains the microfilm reel number in tag 269" [10]. In addition to validating the file image and container formats against JHOVE and additional rules implemented in the NDNP specification, DVV validates METS and ALTO records using Schematron [10].

As part of validation, the DVV generates and embeds in the METS file a SHA-1 digital signature. Successful appendage of a digital signature documents fixity of the original object, indicates that a file is valid, and "proves that the technical metadata…was created by the NDNP Validation Library" [10]. With multiple stakeholders of the data, including the LC, the awardee, and in many cases, a digitization vendor, frequent fixity checks are essential. In an effort to ensure no changes occurred during file transfer, the DVV's verification function enables users to verify the file fixity by checking the digital signatures generated during validation [10].

2. Validation Challenges

With myriad PDF versions, substandards, and versions of substandards, PDF validation is challenging. As found in Lindlar, Tunnat, and Wilson's paper on "A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad and the Ugly," JHOVE fails to meet a ground truth requirement for PDF validation. Thus, while DVV includes an extension of the JHOVE PDF module, a

validation module specific to the validation of the NDNP PDF Profile may be better suited for accurate assessment and to avoid constraints imposed by the capacity of JHOVE's PDF module.

Given that the the DVV's validation profile is not testing for compliance with the ISO 19005 specification for PDF/A, awardees who wish to submit PDF files which conform to the NDNP PDF Profile in addition to ISO 19005 must rely upon workflows outside of the formal NDNP process to validate compliance.

In evaluating PDF/A converter software[4] for the Florida Virtual Campus digital repository workflow, Jamin Koo and Carol Chou identified that the pdfaPilot, 3-Heights, and PDF/A Manager occasionally failed to identify non-conformance, resulting in 90%, 74%, and 91.30% accuracy rates, respectively [11]. As with the PDFlib "Bavaria Report on PDF/A Validation Accuracy," Koo and Chou limited their evaluation of tools against the PDF/A-1b (ISO 19005, conformance level b) profile [12]. With an imperfect accuracy rate for PDF/A conversion across tools, PDF/A-specific validation is requisite.

3. veraPDF

As a result of the EU PREFORMA (PREservation FORMats), the veraPDF consortium released a validation tool dedicated to validation of PDF/A: veraPDF [13].[5] In addition to validating against each PDF/A 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), veraPDF allows users to create a profile unique to their institutional needs. For example, NDNP might extend the validation to validate the NDNP PDF Profile specification that, "The PDF will open to single page layout," a rule not required for a valid PDF/A [4].

III. Method

In this section, the authors introduce the approach for selecting a testing corpus, define testing processes, and outline a metric for analysis of results.

A. File Selection and Retrieval

69 newspaper issues were identified and downloaded from the Chronicling America website, totaling to 382 unique PDFs. In an effort to establish a diverse testing corpus, one issue was selected from each contributing awardee, totaling to 45 issues.[6] From the awardees' corporea, the authors selected the first issue from the most voluminous English-language newspaper.[7] Following the same selection process, one issue representing each non-English language was selected, totaling to 24 newspaper issues.

B. File Validation

The authors selected all PDF/A versions for a comprehensive testing scope and conformance level b for its simplicity. "Digitized documents [without embedded text], files must be stored as 'b' compliant documents that are minimally conforming" [6]. While NDNP PDFs contain embedded OCR, the logical structural requirements for PDF/A are difficult to achieve. Using the veraPDF [v. 1.12.1] GUI on a Windows 10 OS, the authors batch processed the 382 PDF files against the auto-detection module. Results were saved as XML for ease of data manipulation. The PDF files were re-validated against the PDF/A-2b and PDF/A-3b modules to investigate Han's recommendation for the use of PDF/A as a digitization container [5]. The authors then parsed the veraPDF XML output to a CSV for increased ease of analysis.

C. Validation Evaluation

Upon identifying the rules failed between the PDF/A-1, PDF/A-2, and PDF/A-3 validation profiles, rules were categorized by type. Four overarching types were identified as encompassing considerations fundamental to newspaper digitization: XMP Metadata, Embedded Images, Embedded Fonts, and Object Streams (in order of importance). The authors have identified and ranked categories in accordance to their salience for access and preservation of digitized newspapers within the NDNP framework.

IV. Discussion

As expected, all files conformed to the NDNP PDF specification, in that the files are PDF version 1.4, the PDF base required by both PDF/A-1 and the NDNP PDF specification. Also as anticipated, all files failed validation against the PDF/A-1b,[8] PDF/A-2b, and PDF/A-3b veraPDF modules.

A. XMP Metadata

[4] There are two methods of converting or generating a PDF/A: create or conform. To "create" a PDF/A file, a user begins with a source file (e.g., Microsoft Word Document or LaTeX). To "conform" to PDF/A, a user begins with another PDF.

[5] See also the homepage for VeraPDF for further information and resources: https://verapdf.org/

[6] As noted in the background section above, NDNP has 46 participating awardees, of which only 45 have contributed content to date.

[7] Puerto Rico has contributed only Spanish-language newspapers. An issue from the most voluminous newspaper was selected to represent their awardeeship; an issue from the second most voluminous newspaper was selected to as the overall most voluminous Spanish-language newspaper.

[8] PDF/A-1b profile defaulted for all files in the auto-detect module.

Perhaps the most ubiquitous element of access and preservation, metadata supports robust contextualization of a digital object. The XMP header metadata embedded in PDFs and other image and container files supports descriptive and administrative elements. ISO 19005-1 requires that, if metadata properties are contained in the document information dictionary, XMP "analogous properties…shall also be embedded…with equivalent values" [14]. The document information dictionary "contains the creation and modification dates of the file, together with some simple…metadata," which are not included in the XMP serialization but may be embedded in XMP extensions [15]. While non-conformance to this rule does not impact the NDNP workflow, it does suggest poor metadata practices. (See CosDocument, Table 2.) Poor metadata practice is further exemplified in the content objects' failure to meet the XMPPackage and XMPProperty requirements for ISO 19005, which require valid XMP serialization, stating that "XMP form shall use predefined schemas defined in XMP Specification, or extension schemas that comply with XMP Specification" [16, 17]. An essential feature included in the NDNP Technical Specification, best practice suggests that XMP metadata should conform to the XMP Specification.

B. Embedded Images

Validation failure for images included undefined OutputIntents of color profiles and use of interpolation, both which introduce potential harm of long-term sustainability. The NDNP Technical Specification requires that images be digitized at grayscale but does not impose requirements upon the colorspace. As seen in Table 2., results indicate that content creators fulfilled the grayscale requirement but did not define the colorspace (e.g., DeviceRGB, DeviceCYMK, DeviceGray) in the OutputIntent. The inclusion of image interpolation suggests that artifacts were generated through the image conversion process. The interpolation algorithm is specific to rendering software and thus cannot be embedded in a file. Use of interpolation impacts access by imposing an appearance which may not be rendered in every implementation.

C. Embedded Fonts

While font object streams render underneath the image object stream, fonts are captured during the OCR process and therein are embedded in the PDF. Thus, while it may seem that the visual appearance of a font is unnecessary, font information is captured in the OCR. All validation profiles indicated 137 failures to PDF/A-1 Rule 6.3.4-1, and PDF/A-2–PDF/A-3 Rule 6.2.11.4-1, which suggests that awardees are using Type 3 fonts. This further suggests that the DVV did not identify when PDF files contained Type 3 fonts, despite the requirements of the NDNP specification, which states that, "Only the 14 standard Type 1 fonts may be used. These fonts will not be embedded" [4]. This suggests that there are inconsistencies between the NDNP Specification and the digitization output. Increased validation not supported in the DVV's current validation profile may enable content creators and stewards to adhere to the NDNP Specification, as well as general recommendations for long-term preservation as identified in ISO 19005.

Perhaps the most notable difference between the PDF/A-1b and PDF/A-2b–PDF/A-3b validation profiles as seen in the results of this test is the relationship between Rule 6.3.5-2 (PDF/A-1b) and 6.2.11.4-4 (PDF/A-2b–PDF/A-3). The validation output for PDF/A-1b and PDF/A-2b–PDF/A-3b differed in that PDF/A-1b validation identified 28 failed instances of Rule 6.3.5-2, which requires that, "For all Type 1 font subsets referenced within a conforming file, the font descriptor dictionary shall include a CharSet string listing the character names defined in the font subset" [16]. PDF/A-3b validation identified 34 failed instances of Rule 6.2.11.4-4, which, requires that CID fonts be completely identified, whether or not all fonts are referenced. While no correlation has been identified between the rules, the fact that the PDF/A-1b validation profile did not detect failure for CID fonts in Rule 6.3.3-2 and vice versa such that the PDF/A-2b and PDF/A-3b validation profiles did not detect failure for CharSet with Rule 6.2.11.4-3 is of note.

Although evaluation of files which contained errors was not added to the methodology for this preliminary research, the authors are interested in the validation of fonts for non-English newspaper content. In skimming the results, it was found that Fraktur, a prominent glyph set used in many early 20th century German-language texts, did not return any font errors.

D. Object Streams

Object streams are necessary to long-term sustainability of content by exacting, for example, the beginning and end of an indirect object. However, the logical structure requirements for PDF have not been included in the NDNP PDF Profile, and as such has been considered out of scope for evaluation of PDF in the context of newspaper digitization under the current aegis of NDNP.

TABLE II
PDF/A Failed Validation Rules[9]

Object Type	PDF/A-1b Files Failed	PDF/A-2b Files Failed	PDF/A-3b Files Failed
CosDocument	235	-	-
CosIndirect	100	100	100
CosStream	40	40	40
CosStream	16	16	16
MainXMP Package	382	382	382
PDCIDFont	-	34	34
PDDeviceGray	344	344	344
PDDeviceRGB	19	19	19
PDExtGState	120	-	-
PDExtGState	120	-	-
PDFont	137	137	137
PDTrueType Font	16	16	16
PDType1Font	28	-	-
PDXImage	105	105	105
XMPPackage	12	12	12
XMPProperty	370	370	370

V. Conclusion

The authors posit that, for institutions digitizing newspapers using a framework that places emphasis on PDF, the validation tool employed for quality assurance must ensure that the requirements for PDF are actually met. The discussion demonstrates that, without adequate validation software, non-conforming files may be approved unnoticed. Additional research is necessary to provide recommendations for implementation of veraPDF validation within newspaper digitization workflows, to remediate existing errors in PDFs as found in the non-conformance of fonts to the NDNP PDF Profile, and finally, to investigate the use of PDF/A as an alternative to TIFF as a preservation format.

ACKNOWLEDGEMENT

The authors thank the Library of Congress for making newspaper content freely accessible for all scholarly purposes, including validation of PDFs.

REFERENCES

[1] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants 2019-2021," 24 Oct. 2018, www.loc.gov/ndnp/guidelines/NDNP_201921TechNotes.pdf.

[2] M. Sweeney (Contributor), "The National Digital Newspaper Program: Building on a Firm Foundation," *Serials Review*, vol. 33 no.3, pp. 188-189, https://doi.org/10.1080/00987913.2007.10765120.

[3] D. Thomas and M. Sweeney, "Sustainability in the United States National Digital Newspaper Program," *IFLA International Preservation News*, No. 56, May 2012 p.12-20, http://www.ifla.org/files/pac/ipn/IPN%2056.indd.def.pdf.

[4] Library of Congress, "NDNP PDF Profile Version 2.1" https://www.loc.gov/ndnp/guidelines/archive/PDFSpecs0708.pdf.

[5] Y. Han, "Beyond TIFF and JPEG2000: PDF/A as an OAIS submission information package container." *Library Hi Tech*, vol. 33, no. 3, 2015, pp. 409-423 https://doi.org/10.1108/LHT-06-2015-0068.

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

[7] C. Adams, "Is JPEG-2000 a preservation risk?" [Blog post], 28 January 2013, retrieved from http://blogs.loc.gov/thesignal/2013/01/is-jpeg-2000-a-preservation-risk/.

[8] C. Arms, et al,. "The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions: An NDSA Report," February 2014, Retrieved from http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014.pdf.

[9] Y. Han and X. Wan, "Digitization of Text Documents Using PDF/A," *Information Technology and Libraries*, vol. 37, no.1, 2018, pp. 52-64 https://doi.org/10.6017/ital.v37i1.9878.

[10] J. Littman, "A Technical Approach and Distributed Model for Validation of Digital Objects," *D-Lib Magazine*, vol. 12, no. 5, May 2006, retrieved from http://www.dlib.org/dlib/may06/littman/05littman.html.

[11] J. Koo and C.H.C. Chou, "PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow," N*ew Review of Information Networking* vol. 18, no. 1, 2013, pp. 1-15, https://doi.org/10.1080/13614576.2013.771989.

[12] PDFlib GmbH, "Bavaria Report on PDF/A Validation Accuracy," May 4, 2009, retrieved from https://web.archive.org/web/20170103204840/http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf.

[13] C. Wilson, R. McGuinness, and J. Jung, "veraPDF: building an open source, industry supported PDF/A validator for cultural heritage institutions," Digital Library Perspectives vol. 33, no. 2, 2017, pp. 156-165, https://doi.org/10.1108/DLP-08-2016-0031.

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

[15] J. Whitington, "Chapter 4. Document Structure," *PDF Explained*, retrieved from https://www.oreilly.com/library/view/pdf-explained/9781449321581/ch04.html.

[16] veraPDF, *PDF/A-1 validation rules*, retrieved from http://docs.verapdf.org/validation/pdfa-part1/.

[17] Adobe Systems Incorporated, *XMP Specification Part 1: Data Model, Serialization, and Core Properties*, April 2012, retrieved from https://wwwimages2.adobe.com/content/dam/acom/en/devnet/xmp/pdfs/XMP%20SDK%20Release%20cc-2016-08/XMPSpecificationPart1.pdf.

[9] Not all tests were equivalent between the PDF/-1 and PDF/A-2–PDF/A-3 validation profiles. In particular, the PDF/A-2–PDF/A-3 validation profile does not contain rules for testing the CosDocument and PDExGState objects. Furthermore, the PDCIDFont and PDType1Font objects, which failed against PDF/A-2–PDF/A-3 and PDF/A-1b, respectively, contain different rules in the validation profile, thus resulting in inconsistent validation output.