Achieving Conformance to Document Standards:
Can PDF Files Conform to the PDF/A-1b Specification?

Thomas Fischer, Software Systems Research Group, University of Skövde, Sweden*

iD https://orcid.org/0000-0003-0272-7433

Björn Lundell, Software Systems Research Group, University of Skövde, Sweden

Jonas Gamalielsson, Software Systems Research Group, University of Skövde, Sweden

ABSTRACT

In the context of long-term archival of digital assets, file formats that are standardized and designed for longevity such as PDF/A are preferred. However, due to the complexity of and ambiguities in PDF standards, it is far from trivial to either create standard-conformant files or check the conformance of any given file. This study investigates the challenges when checking real-world PDF files from public sector organizations meant for long-term archival for PDF/A conformance. Results show that only a small set of PDF files claims to conform to the PDF/A-1b specification variant and even fewer files pass conformance checks by various conformance checking tools. Challenges for conformance checking tools include both ambiguities in the standards' technical specifications and limitations in the implementation.

KEYWORDS

Conformance Checking Tools, ISO 19005:1, Normative References, Open-Source Software, Portable Document Format, Public Sector Organizations, Technical Specifications, Validators, Vera PDF

INTRODUCTION

The process of long-term maintenance of digital assets for use and re-use imposes a number of challenges, including the limitations of storage technologies and the choice of future-proof file formats. In context of the latter challenge, digital archives, for example, must be able to handle a number of different media formats such as audio or video recordings or textual documents. One variant of digital assets are page-oriented, text-centric documents as, for example, generated in office productivity software. The native format in which those documents were originally created is often not suitable for long-term archival (Anderson, 2005). Dryden (2008) stresses the need for digital file formats designed for long-term archival stating 'it is not an exaggeration to say that long-term preservation of digital objects is the biggest challenge facing not just the archival profession but society as a whole.'

A common choice (Library of Congress, 2019) is, therefore, to convert those documents to PDF which has properties attractive for archival such as being 'read-only' and the ability to reproduce the

original document across different devices (even web browsers can display PDF files, see Mozilla Labs, 2020).

In the context of long-term archival, how can it be guaranteed that PDF files can be read in a future without today's computer systems? Here, 'reading' is not limited to the extraction of text and images, but includes as well the visual appearance, logical structure, and metadata of a document. Various ISO standards (ISO, 2005, 2011, 2012a) specify subsets of 'normal' PDF variants under the name 'PDF/A' in order to address those requirements, i.e. it should be possible to read a standard-conformant PDF/A file just by implementing the ISO standards.

Further, the importance of transitioning from PDF to PDF/A is elaborated by an analogy as follows:

Pressure from the preservation community provided the catalyst for many publishers to change over from acidic to acid-neutral paper in the production of published works. Introducing more stable materials at the beginning of the information production process represents in a significant victory for preservation interests which in the long run will reduce the need for salvage efforts. (Hedstrom, 1998)

Whereas there is a broad agreement on PDF/A standards are the preferred choice when archiving PDF files (Bundesarchiv, 2010; LAC, 2015; Riksarkivet, 2009; Rog, 2007; Swiss Federal Archives, 2020), adopting PDF/A standards in a PDF workflow has multiple challenges. A central aspect here is how to determine if a given PDF file actually conforms to a PDF/A standard, usually at least to the most basic specification, PDF/A-1b. Especially public sector organizations such as universities, which have a legal obligation to archive important documents (SFS, 1993, 2012), are motivated to adopt PDF/A in order to save costs (less physical storage required) and general 'modernization'.

This study investigates the following research questions specifically related to the long-term archival of PDF/A files by public sector organizations:

RQ 1: What characterizes PDF files provided by public sector organizations?
RQ 2: How successful are public sector organizations at providing PDF/A-1b-conformant files?
RQ 3: How and why does the outcome of assessments of PDF/A-1b conformance for files differ between conformance checking tools?

Through an investigation of research question 1, the study establishes properties of analyzed files which contributes an overarching characterization of state-of-practice concerning how PDF files are being generated and used in public sector organizations. Addressing the requirements for long-term archival, research question 2 allows for a quantitative assessment of the conformance to the PDF/A standard within the same set of files as well as documents the uncertainty of such an assessment due to the varying criteria applied by conformance checking tools. Finally, research question 3 investigates in greater detail the differences between conformance checking tools and challenges for determining conformance which is of relevance for archival processes that need to know the conformance properties of archived documents.

In acknowledging that there are a number of legal and licensing issues which impinge on implementation and use of PDF and PDF/A (Koo & Chou, 2013; Lundell et al., 2015, 2019), it should be noted that issues concerning standard-essential patents and copyright are outside the scope of this study.

30 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/achieving-conformance-to-document-standards/288523

Related Content

Key Challenges in the Design of Learning Technology Standards: Observations and Proposals
Adam R. Cooper (2010). *International Journal of IT Standards and Standardization Research (pp. 20-28).*
www.irma-international.org/article/key-challenges-design-learning-technology/46110

Global Perspectives on the Information Society
William J. McIver Jr. (2004). *Social, Ethical and Policy Implications of Information Technology (pp. 1-28).*
www.irma-international.org/chapter/global-perspectives-information-society/29303

Standardising the Internet of Things: What the Experts Think
Kai Jakobs, Thomas Wagnerand Kai Reimers (2011). *International Journal of IT Standards and Standardization Research (pp. 63-67).*
www.irma-international.org/article/standardising-internet-things/50575

Security Management in Heterogeneous Distributed Sensor Networks
Al-Sakib Khan Pathan (2013). *IT Policy and Ethics: Concepts, Methodologies, Tools, and Applications (pp. 773-793).*
www.irma-international.org/chapter/security-management-heterogeneous-distributed-sensor/75056

A Standards-Based Common Operational Environment
Jaroslav Blaha (2000). *Information Technology Standards and Standardization: A Global Perspective (pp. 152-167).*
www.irma-international.org/chapter/standards-based-common-operational-environment/23733