International Journal of Standardization Research
Volume 20 • Issue 1

On PDF/A Conformance and Font Usage in PDF Documents Provided by Public Sector Organizations

Thomas Fischer, University of Skövde, Sweden*
https://orcid.org/0000-0003-0272-7433
Björn Lundell, University of Skövde, Sweden
Jonas Gamalielsson, University of Skövde, Sweden

ABSTRACT

The use of appropriate fonts and file formats for long-term maintenance of digital assets is a challenge for organizations in the public sector. The article reports from a study which investigated the PDF/A conformance and font usage in PDF files provided by Swedish public sector organizations (PSOs). This article presents an analysis of the PDF files' properties and font usage including a categorization of fonts' licenses. This study is motivated by the PDF/A-1 standard's requirement that 'only fonts that are legally embeddable in a file for unlimited, universal rendering shall be used.' Analyzing PDF sets from three PSOs, the finding shows that the proportion of files that claim or succeed at conforming to PDF/A greatly varies among the sets despite similar backgrounds. Although the most popular way to make use of fonts is by embedding a subset of the font data, for some fonts expected to be 'always available,' a considerable proportion of PDF files does not include any font data. This puts the onus of locating this data on the PDF reader which is problematic for long-term archival.

KEYWORDS

Archival, Embedding, Font, ISO, License, Long-Term, Open Source, Standardization, Subset, Typeface

PDF/A CONFORMANCE AND FONT USAGE IN PDF DOCUMENTS PROVIDED BY PUBLIC SECTOR ORGANIZATIONS

Long-term maintenance and archiving of digital assets such as electronic office documents requires the consideration of how to prepare those digital assets for future use. Multiple challenges exist such as the choice of storage technology and file format. Development and use of file formats impose a number of technical and legal challenges (Lundell et al., 2019), and in particular when formats are to be implemented in software (Egyedi, 2007). To allow for a future use of digital assets, file formats that are clearly specified and provided under terms that allow for implementation and use by software

DOI: 10.4018/IJSR.329605 *Corresponding Author

projects should be used (Lundell et al., 2023). This includes many formats which are made available as standards by standard-setting organizations such as the International Organization of Standardization (ISO[1]) and the Organization for the Advancement of Structured Information Standards (OASIS[2]).

An example of where an existing file format was standardized is the Portable Document Format (PDF), which, somewhat simplified, allows to describe the content of pages that can be printed. The PDF file format has properties relevant for archiving, such as the read-only option. But the format has drawbacks: documents may refer to external data, which may not be available when reading a PDF file, and the format's specifications may rely on normative references which may be unavailable when implementing tools for those file formats. To address those limitations, a standard—commonly known as PDF/A— was specified by ISO (2005, 2011, 2012, 2020a) with the intent to define a self-sufficient subset of the PDF file format.

Adherence to the PDF/A standard is required by several national archives and national libraries (Bundesarchiv, 2010; Library and Archives Canada, 2015; Rog, 2007). For example, the Swedish National Archives mandate the use of PDF/A-1 if PDF files are to be archived (Riksarkivet, 2009). Determining the conformance to the PDF/A standard faces challenges both due to deficits in the technical specifications of the PDF/A standard and its normative references, and due to implementation deficits in the tools that may get employed to assess PDF/A conformance (Fischer et al., 2021).

To display a PDF file on screen or to print it, the text contained in such a file must be rendered (i.e., put into a graphical representation by using a so-called font program, commonly referred to as "font"). The font program must be available both to the system where the PDF file was originally created as well as to every PDF reader where the file is to be displayed. The technical specifications of each specific version of PDF outline various alternatives of how to make the font available to the PDF reader: embedding parts[3] of the font program into the PDF file, relying on standard fonts that are expected to be generally available, and putting the onus on the PDF reader to locate or synthesize a suitable font when displaying the file.

Only embedding the font data into the PDF file allows to recreate the text's shapes and thus guarantees a faithful visual reproduction of the file on the PDF reader's side irrespectively whether the font is locally available to the reader or not. This, however, introduces the question whether the font can be for legal reasons (Groves, 1992) included into the PDF file (i.e., if the PDF file's author has the right to use and include copyrighted font data into the PDF file and then distributes the file). In the context of long-term archival, it is difficult to determine for a PDF file whether the author has legal right. It's difficult for an archiving organization to determine the legal status of used fonts, and if necessary, acquire the permission to use the font.

Addressing those legal challenges, the PDF/A-1 specification imposes the requirement on conforming PDF files that "only fonts that are legally embeddable in a file for unlimited, universal rendering shall be used." (ISO, 2005, p. 10). The aspects of embedding fonts are further elaborated: "This part of ISO 19005 precludes the embedding of fonts whose legality depends upon special agreement with the font copyright holder. Such an allowance places unacceptable burdens on an archive to verify the existence, validity and longevity of such claims." (ISO, 2005, p. 11).

Based on these arguments, we address the following research questions:

RQ 1: To what extent do PDF files from public sector organizations (PSOs) conform to the PDF/A standard and what characterizes those files?
RQ 2: How are different fonts used in the collected PDF files?

We investigate how well PSOs perform in providing PDF files that conform to the PDF/A standard and which fonts are used and into which licensing category those fonts can be put. The investigation is focused on PSOs as transparency laws apply to this type of organizations which further motivated our investigation: only file formats that are suitable for long-term archivalprovide citizens the best possible access to those files.