KB

PDF Guidelines

Recommendations for the creation of PDF files for long-term preservation and access

Author	Judith Rog (judith.rog@kb.nl) Koninklijke Bibliotheek (http://www.kb.nl) Digitale Duurzaamheid (digitalpreservation@kb.nl)
Current version	1.7
Change history	1.7 Updated current Adobe Acrobat version, minor textual changes 1.6 Correction on XMP metadata 1.5, October 2006, Added notes onPDF/A-1 conformance levels and some info on uncalibrated colour spaces. 1.4, July 2006, Minor textual changes 1.3, June 2006, Minor textual changes 1.2, June 2006, Changes in lay-out 1.1, May 2006, Minor textual changes 1.0, May 2006, First version
Date	31-05-2007

Introduction

To safeguard the long-term storage and access of digital publications in the e-Depot the National Library of the Netherlands/Koninklijke Bibliotheek (KB) needs to know all ins and outs of the supplied files. Although every file format is accepted, the choice of file format and the chosen settings within a file format can affect the degree to which long term preservation and access can be guaranteed. As 88 per cent of the files in the e-Depot are PDF files, the KB has chosen to publish recommendations for PDF first. Recommendations for other file formats will follow later.

"PDF" stands for "Portable Document Format". It was developed as a follow-up to Adobe's Postscript language. Adobe Systems invented PDF technology in the early 1990s to smooth the process of moving text and graphics from publishers to printing-presses and has been in use since 1993. PDF was originally envisioned as a way to communicate and view printed information electronically across a wide variety of machine configurations, operating systems and communication networks in a reliable manner. PDF relies on the same imaging model as the PostScript page description language to render complex text, images and graphics in a device and resolution-independent manner, bringing this feature to the screen as well as the printer. To improve performance for interactive viewing, PDF defines a more structured format than that used by most PostScript language programs. PDF also includes objects, such as hypertext links and annotations that are not part of the page itself but are useful for building collections of related documents and for reviewing and commenting on documents [1].

PDF files may be created natively in PDF form, converted from other electronic formats or digitized from paper, micro-film or other hard-copy format. When creating PDF files the application offers a choice of several settings. The specific settings can affect preservation of and access to the file in the short and long term. Certain choices of settings can change the appearance of the PDF file on different environments.

The ISO 19005-1 standard for PDF/A-1 was published in 2005 and geared towards long-term preservation. It provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files. The PDF/A-1 specifications are based on the specifications of PDF 1.4 and describe which aspects of a PDF are compulsory, optional or prohibited in a PDF/A-1 compliant file.

The KB prefers PDF files that are compliant with this PDF/A-1 standard. PDF/A-1 defines two conformance levels. A 'Level A' conformant file (PDF/A-1a) shall adhere to all requirements of the ISO standard. A 'Level B' conformant file (PDF/A-1b) does not have to adhere to two requirements. These requirements refer to the use of tagging to add structure to preserve the logical structure and reading order and the use of Unicode character maps that map character codes to Unicode values. As preserving the natural reading order and using Unicode are crucial for the preservation of the document as it was intended by the creator, the KB advises to create PDF/A-1a conformant files.

The PDF/A Competence Center [2] gives an overview of PDF/A compliant products for creating and validating PDF/A documents in accordance with ISO 19005-1. With regard to Adobe Acrobat products, as from version 8.0 when creating PDF files from Microsoft Office applications using PDFmaker, it is possible to indicate that the resulting PDF document must comply with the PDF/A-1a conformance level of the ISO standard. When creating PDF files using Adobe Acrobat Distiller 8.0, only PDF/A-1b settings are available. Adobe Acrobat 8.0 Professional can validate PDF files to see whether they comply to the

PDF/A profile. Furthermore, Adobe Acrobat 8.0 Professional can do an attempt at converting an existing PDF file to PDF/A. The exact procedures to do so in PDFmaker, Acrobat Distiller and Acrobat Professional are described in "PDF/A in Action: Creating and Conforming" [3].

By choosing for PDF/A, preferably conformance level A, you will have the most certainty that your PDF file is portable across systems and platforms without changing the content or look-and-feel of the document now and in the future.

However, it might not always be possible to meet with all the PDF/A-1 requirements. For this reason, and because the PDF/A-1 standard is quite hard to read, the KB decided to publish a set of recommendations to which PDF files that are submitted to the e-Depot should comply to as much as possible.

Guidelines

Accessibility and structure

1. Avoid the use of security measures such as password protection or any other form of encryption or Digital Rights Management that put restrictions on the management of or access to the file. Examples are restrictions on opening or printing the file or copying text from the file. These security settings can be set during the creation of the PDF file.

Note
From PDF version 1.1 it is possible to use password protection and encryption. However, any form of restriction on the access to and management of an object can create problems for the long-term preservation of that object. Without knowing the password of a password protected file it cannot be opened. Even when the key is known, the algorithm to decrypt the data can be lost, which would make it impossible to decrypt the data in the future. Security measures like these can hinder the access to the file. As a result, strategies for long-term preservation and access such as migration and emulation cannot be applied.

2. The KB advises to make sure that tags are added to the PDF using Adobe Acrobat. Tags provide the file with structure and preserve the natural reading order.

Note
PDF tags are loosely based on HTML tags. The tags give structure to the content of the PDF file and increase accessibility. The tags have no influence on the appearance of the file. For example, tags can define the direction of the text flow and the connection between images and captions. When migration is applied to a non-tagged PDF file, the text may get scrambled or the connection between an image and its caption might get lost. Tagged PDF files also work better with the screen-reader devices used by blind and visually impaired users or when reflowing the file to fit different page or screen widths [4]. Tags can be added to a PDF file as from PDF version 1.4.

Making a PDF file accessible through tagging also implies that 'alt' (alternative) descriptions must be added to images, formulas, tables and any other object that has no standard textual alternative. A screen-reader will use these alternative descriptions when reading the content out loud. [1]

[1] For further information on tagged PDF Files see Chapter 20 "Accessibility and Tagged PDF Files" in Ted Padova's PDF Bible [5]. 'Alt' descriptions can be added to a PDF file as from PDF version 1.3. A good explanation of how tags work, how they can be added and modified can be found on the WebAIM website [6] "Understanding PDF Tags" [7]. More information on making a PDF file accessible can be found on the Adobe website "Accessibility, Improving access to electronic information for people with disabilities" [8].

3. The KB recommends writing out the URL of a hyperlink in full (e.g. "Koninklijke Bibliotheek (http://www.kb.nl)" and not "Koninklijke Bibliotheek"). This can be done in either the running text or at any other location in the document (e.g. in the footnotes).

Note
When for some reason the underlying URL can no longer be accessed, a hyperlink such as "Koninklijke Bibliotheek" becomes useless. When writing out the URL in the text or at another location as in "Koninklijke Bibliotheek (http://www.kb.nl)" the destination address of the URL will always be human readable [9].
The KB does not take responsibility for maintaining the accessibility to the external resources that can be referred to within the PDF.

Fonts

4. The KB advises to embed as well as subset all fonts. Only use fonts that can be legally embedded and that are encoded using either WinAnsiEncoding or MacRomanEncoding.

Note
By embedding as well as subsetting all fonts the original look can be maintained as much as possible in a different environment.
Although it may seem that embedding alone is sufficient to maintain the original look, this involves a risk. In theory it is possible that a user has an installed font with the exact same name as an embedded font. In that case, the PDF reader could then use the installed font, assuming it is the same font as the embedded one, while it is in fact a different font that only shares its name with the embedded font. By not only embedding but also subsetting the font it will get an internal, unique name which will not occur as an installed font on the machine.

Please note that the Base 14 fonts (Courier, Courier Bold, Courier Bold Italic, Courier Italic, Helvetica, Helvetica Bold. Helvetica Bold Italic, Helvetica Italic, Times, Times Bold, Times Bold Italic, Times Italic, Symbol and ZapfDingbats) must also be embedded and subsetted. When using standard Adobe Acrobat Distiller settings with the option 'optimize for fast web view' checked, the Base 14 fonts will not be embedded, because it is presumed that a PDF reader includes these standard fonts [10]. However, because it is not certain that this will always be the case, the KB strongly advises to embed and subset the Base 14 fonts as well [11].

Adobe Acrobat Distiller does not allow the embedding of fonts with proprietary constraints. Although other applications might allow this, the KB still advises against doing so as it is legally not permitted [11].

The application of non-standard encoding algorithms poses a risk to long-term preservation because the algorithm to decode the data might get lost for future users, making the content inaccessible [11].

Compression

5. The KB advices against using compression. If it cannot be avoided, use only lossless compression algorithms such as ZIP. Do not choose algorithms that are subject to proprietary constraints.

Note
The use of lossy compression algorithms will cause quality loss. Although this may not be a problem for long-term preservation and access as such, the KB does strive for high quality images so that any further processing or other functionality (such as zooming) will not be hindered by poor quality of the image [11].

Using compression algorithms that are subject to proprietary constraints poses a risk to the long-term accessibility of the object [11].

Images

6. The KB recommends avoiding the down sampling of images.

Note
The down sampling of images will cause quality loss. Although this may not be a problem for long-term preservation and access as such, the KB does strive for high quality images so that any further processing or other functionality (such as zooming) will not be hindered by poor quality of the image [11].

Using compression algorithms that are subject to proprietary constraints poses a risk to the long-term accessibility of the object [11].

7. The KB advices against using transparency in images.

Note
As from PDF version 1.4, transparency in images can be used. However, the use of transparency poses a risk to the long-term accessibility of the object as the implementation of transparency has not yet been fully defined by Adobe. In the PDF specifications for version 1.4 the generic model on which transparency is based has been described, but not the actual implementation [11].

Executable actions

8. The KB advises to avoid the use of scripting or any other form of executable actions. Likewise, the use of form fields or the use of optional content or alternative views of images that could change the appearance, should be avoided.

Note
The use of scripting such as JavaScript, possible since PDF version 1.3, can create a dependency on external factors. Furthermore, it makes the file more complex and affects the look and content. These factors all threaten the long-term preservation and the authenticity of the object [11]. The use of interactive elements such as radio buttons and checkboxes is possible as from PDF version 1.2.

9. The KB strongly advises to include the information that is required to render all aspects of a publication in the PDF file *itself*. As a result, no data stream from an external object or other application than the PDF reader itself is needed to render the entire content of the publication. Embedding multimedia files is therefore dissuaded.

Note
When external applications or external data streams are used, rendering an object becomes dependent on external factors. Besides the PDF file itself, the application or data stream needs to be preserved and maintained accessible for future users as well, making the long-term preservation of the PDF file a complex matter [11]. If possible it is recommended to avoid dependencies such as these. The use of external data streams is supported as from PDF version 1.2.

Colour

10. The KB recommends using one of the four device-independent colour spaces: CalGray, CalRGB, Lab or an ICCbased colour space. When using ICC based colour spaces it is advised to embed the colour space as an ICC profile. Furthermore the uncalibrated colour spaces DeviceRGB or DeviceCMYK can be used. Only one of them should be used but not both in the same document and they should always be combined with a conforming PDF/A-1 OutputIntent that defines the conforming RGB or CMYK colour characteristics for the output device [11].

Note
Using a device-independent colour space is a good way to maintain the original colour display, independent of the type of monitor, printer or other devices that are used [11]. The device-independent colour spaces CalGray, CalRGB and Lab can be used as from PDF version 1.1. ICC based colour spaces can be used as from PDF version 1.3.

References:

All URLs were functional as of May 2007.

[1] "Request for Comments: 3778 ", Network Working Group, online available: http://www.rfc-editor.org/rfc/rfc3778.txt
[3] "PDF/A in Action: Creating and Conforming", online available at "Acrobat for Legal Professionals": http://blogs.adobe.com/acrolaw/2007/01/pdfa_in_action.html
[4] "What is Tagged PDF?", Duff Johnson, online available: http://www.planetpdf.com/enterprise/article.asp?ContentID=6067∓
[5] "Adobe Acrobat PDF Bible", Ted Padova, Wiley Publishing Inc, 2005, ISBN 0-7645-8378-6;
[6] "WebAIM, Web Accessibility in Mind", http://www.webaim.org/
[7] "Understanding PDF Tags", online available at WebAIM: http://www.webaim.org/techniques/acrobat/understandingtags
[8] "Accessibility, Improving access to electronic information for people with disabilities", online available: http://www.adobe.com/enterprise/accessibility/main.html
[9] "Empfehlungen zum Erzeugen archivierbarer Dateien im Format PDF", Michael Horvath, online available: http://www.onb.ac.at/about/lza/pdf/ONB_PDF-Empfehlungen_1-4.pdf
[10] "No 1. Font Issues", Rich Sprague, online available: http://www.planetpdf.com/mainpage.asp?WebPageID=362
[11] "Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)", ISO 19905-1, online available at charge: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=38920&ICS1=37&ICS2=100&ICS3=99
[12] "PDF Reference-third edition. Adobe Portable Document Format version 1.4", Adobe Systems Incorporated, Published by Addison-Wesley, ISBN 0-201-75839-3; online available: http://partners.adobe.com/public/developer/en/pdf/PDFReference.pdf