How to create a PDF/A file with LaTeX

Jarmo Niemelä (jarmo.niemela@tuni.fi)

27th September 2023

Contents

1 Introduction

PDF/A is a standardised version of the pdf file format, and it is intended for archiving and long-term preservation of electronic documents. A PDF/A file must contain all the information needed for displaying and printing the document. This includes text, graphics, fonts and colour information. Audio and video content and encryption are forbidden.

There are four versions of the PDF/A standard: PDF/A-1, PDF/A-2, PDF/A-3 and PDF/A-4. Versions 1–3 specify conformance levels a and b. Level a (accessible) meets all requirements for the standard. Level b (basic) conformance requires only that the document's visual appearance is preserved. PDF/A-2 and PDF/A-3 contain a third level u (Unicode), which expands conformance level b with an additional requirement that all text in the document has Unicode mapping. PDF/A-4 does not use conformance levels a, b and u but defines a new level f corresponding to version PDF/A-3. In general, PDF/A-4 does not have as strict requirements as the previous versions.

PDF/A format is required in Tampere University theses. Any PDF/A version can be used.

2   Creating a PDF/A file

Below is a LaTeX template for creating a PDF/A compliant file:

```
\pdfminorversion=7
\begin{filecontents*}[overwrite]{\jobname.xmpdata}
    \Title{Document's title}
    \Author{Author's name}
    \Language{en-GB}
    \Subject{The abstract or short description.}
    \Keywords{keyword1\sep keyword2\sep keyword3}
\end{filecontents*}
\documentclass[a4paper,12pt]{article}
\usepackage[utf8]{inputenc}
\usepackage[UKenglish]{babel}
\usepackage{colorprofiles}
\usepackage[a-2b,mathxmp]{pdfx}[2018/12/22]
\hypersetup{pdfstartview=}
\begin{document}
The text of the document goes here.
\end{document}
```

The command `\pdfminorversion=7` at the beginning of the file sets pdf version to 1.7, on which PDF/A-2 and PDF/A-3 are based.

The pdfx macro package

Use the pdfx package to create a PDF/A file from the LaTeX source:

```
\usepackage[a-2b,mathxmp]{pdfx}[2018/12/22]
```

The option `a-2b` selects PDF/A version and conformance level. The default value is `a-1b`. Always use the option `mathxmp` because it allows mathematical symbols in metadata, and it also corrects some errors in presenting metadata in pdf viewers. The option `[2018/12/22]` at the end of the command means that no version of the package older than the specified date is accepted. The pdfx macro package still has some shortcomings and problems, and there are more of these in the older versions.

The pdfx package loads hyperref and xcolor packages, among others, so you don't have to call these explicitly. You can set hyperref options using the `\hypersetup` command.

Colour profile

The colorprofiles package takes care of the colour profile required by the PDF/A format. The pdfx package loads colorprofiles automatically. However,

pdfx will not call colorprofiles if it is not installed. So it must be installed first. MiKTeX usually installs the missing packages automatically when needed, in which case it is sufficient to add the command

```
\usepackage{colorprofiles}
```

before calling the pdfx package. After the colorprofiles package is installed, this command is no longer needed.

Metadata

The PDF/A standard requires that document metadata, such as the document's title, author's name and keywords, are embedded in a specified format. The pdfx package reads the metadata from a text file `\jobname.xmpdata`, where the command `\jobname` contains the base name of the document's main LaTeX file, for example `pdfa-guide.xmpdata`. Usually, the most convenient way is to include the `\jobname.xmpdata` file at the beginning of the main file, within a `filecontents*` environment:

```
\begin{filecontents*}[overwrite]{\jobname.xmpdata}
    \Title{Document's title}
    \Author{Author's name}
    \Language{en-GB}
    \Subject{The abstract or short description.}
    \Keywords{keyword1\sep keyword2\sep keyword3}
\end{filecontents*}
```

If your version of LaTeX was released before October 2019, the `overwrite` option of the `filecontents*` environment does not work. In that case, replace the command

```
\begin{filecontents*}[overwrite]{\jobname.xmpdata}
```

with commands

```
\RequirePackage{filecontents}
\begin{filecontents*}{\jobname.xmpdata}
```

All the supported metadata fields are listed in the documentation of the pdfx package. None of the metadata fields are mandatory, but it is recommended to specify at least the fields `\Title`, `\Author` and `\Language`. Multiple authors, keywords and languages should be separated by the command `\sep`.

Within the metadata fields, you can type LaTeX's reserved characters `$`, `&`, `#`, `_`, `~`, `^`, `{` and `}` as themselves, but use the commands `\%` and `\textbackslash` for the characters `%` and `\`. This implies that, for example, superscripts and subscripts can only be represented with the corresponding Unicode characters.

If the metadata fields contain mathematical symbols, either as Unicode characters or represented with LaTeX's commands, you must use the `mathxmp` option of the pdfx package. Mathematical symbols must not be surrounded with dollar signs because these would show in the metadata as such.

3   Checking and validating a PDF/A file

After the PDF/A file is ready, check its text, images, bookmarks, links and metadata (*File → Properties → Description*) in Adobe Acrobat Reader or in PDF-XChange Editor. Lastly, you must ensure that the file actually complies with the PDF/A standard. Do this validation always before archiving PDF/A files. In particular, before submitting your thesis for evaluation.

You can validate PDF/A files with the following programs:

3-Heights PDF-validator online tool is a free online tool for validating PDF/A documents.

veraPDF is a free open source PDF/A validator. veraPDF requires the Java Runtime Environment. A quick start guide: docs.verapdf.org/gui.

Adobe Acrobat Pro's Preflight tool (*Tools → PDF Standards → Preflight*) can validate PDF/A documents and repair possible validation errors. Acrobat Pro is commercial software, but a 7-day trial version is available.

Callas pdfaPilot can validate and repair PDF/A files in a similar manner as Adobe Acrobat Pro. Callas pdfaPilot is commercial software, but a 14-day trial version is available.

Sometimes these validation programs may give different results. Usually, it is sufficient that the file validates with one of the programs listed above. Thus if a pdf file passes the validation tests of one program, you do not have to validate it with other programs. If a pdf file fails the validation tests of one program, you can test it with another program.

4   Recommendations and problems

Bookmarks

Check the document's section titles in Adobe Reader's or PDF-XChange Editor's bookmarks. Avoid using mathematical symbols in section titles because mathematical symbols and other special characters may not show up correctly in bookmarks. In many cases, you can correct these with the command `\texorpdfstring{}{}`, whose second argument is used in the bookmarks. For example,

```
\section{The integral \texorpdfstring
    {$\displaystyle\int\sqrt{x^2 - a^2}\,dx$}
    {\int\textsurd(x\texttwosuperior \textminus
        a\texttwosuperior)\unichar{"2009}dx}}
```

The second argument of `\texorpdfstring` can contain Unicode characters, either copied from somewhere or entered with the `\unichar{"XXXX}` command, where `XXXX` is the character's hexadecimal code in uppercase.

Included pdf files

Imported pdf graphics files and other included pdfs must have their fonts embedded. Otherwise, the pdf file produced from the LaTeX document does not contain those fonts and will not pass PDF/A validation. Embed the fonts with the same programs with which the included pdf files were made. If this is not possible, you can embed the fonts with the free Ghostscript program[1]. The required command in Windows command prompt is

```
"C:\Program Files\gs\gs10.02.0\bin\gswin64c.exe" -dBATCH
    -dNOPAUSE -sDEVICE=pdfwrite -dAutoRotatePages=/None
    -sOutputFile=output.pdf -c "<</NeverEmbed [ ]>>
     setdistillerparams" -f input.pdf
```

The file path at the beginning of the command depends on the version of Ghostscript. If you need anything other than Ghostscript's default fonts, you can give the list of font directories with the option `-sFONTPATH`. For example, `-sFONTPATH="C:/Windows/Fonts"`.

Alternatively, you can make an ordinary pdf file with LaTeX and convert it to PDF/A format with another program. The missing fonts will be embedded automatically during the conversion.

Limitations of the pdfx package

The current version (v1.6.3) of the pdfx package cannot generate level a (1a, 2a, 3a) conforming PDF/A files.

PdfLaTeX cannot produce a valid PDF/A file if one of the font-defining macro packages arev, kpfonts or mathdesign is used along with pdfx.

The pdfx package expects that XeLaTeX is used with the optional argument `-shell-escape`. Because using `-shell-escape` is potentially a security risk, it is better to create a new processing tool than to change XeLaTeX's default options. You can do this in the TeXworks editor in the following way:

1. Select *Edit → Preferences → Typesetting*.

---

[1]You will also need the Ghostscript fonts: https://sourceforge.net/projects/gs-fonts/files/latest/download. Copy the fonts folder to Ghostscript's installation directory.

2. Click the plus button next to *Processing tools.*

3. Create a new processing tool "XeLaTeX shell-escape" according to figure 1, when using MiKTeX, and according to figure 2, when using TeX Live.
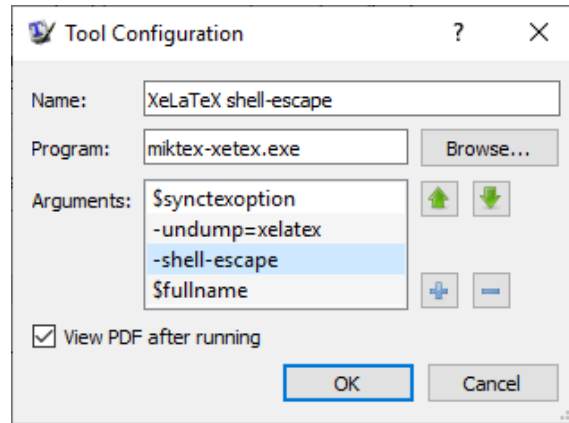

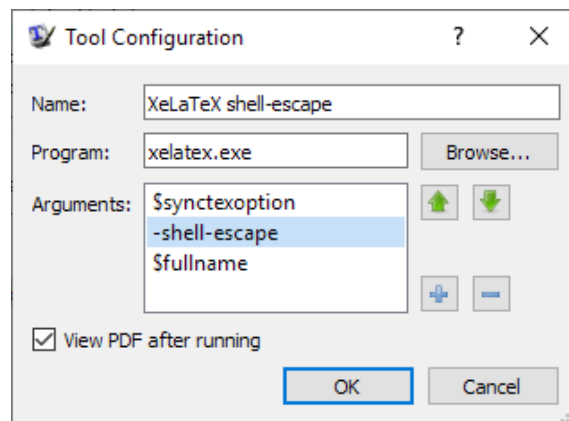
Figure 1: Adding the option `-shell-escape` in MiKTeX.

Figure 2: Adding the option `-shell-escape` in TeX Live.

The option `pdfstartview` of the hyperref package selects the pdf viewer's startup page view. See the hyperref manual for further details. The values `FitH`, `FitV`, `FitR`, `FitBH` and `FitBV` of this option must be followed by the distance from the bottom left corner of the page in 1/72 inch units. Otherwise this will cause a validation error in 3-Heights PDF validator. If the option is given without a value as `pdfstartview=`, then the pdf viewer's default setting is used as the startup page view.

5 PDF/A conversion with other programs

If you cannot create a valid PDF/A file with LaTeX, you must use another program for the conversion. In any case, write the document metadata with LaTeX. You don't necessarily need the pdfx package for this because you can save the metadata with the hyperxmp and hyperref packages too:

```
\pdfminorversion=7
\documentclass[a4paper,12pt]{article}
\usepackage[utf8]{inputenc}
\usepackage[UKenglish]{babel}
\usepackage{hyperref}
\usepackage{hyperxmp}
\hypersetup{%
  pdftitle={Dokument's title},
  pdfauthor={Author's name},
  pdflang={en-GB},
  pdfsubject={The abstract or short description.},
  pdfkeywords={keyword1, keyword2, keyword3},
  pdfstartview=}
\begin{document}
The text of the document goes here.
\end{document}
```

If the metadata contains LaTeX's reserved characters, you must write them with the commands \%, \$, \& etc. Include mathematical symbols as Unicode characters.

Muuntaja service

The students and staff of Tampere University can create PDF/A files with the Muuntaja service at muuntaja.tuni.fi.

Adobe Acrobat Pro

With Adobe Acrobat Pro you can convert pdf files to PDF/A format: *Tools → PDF Standards → Save as PDF/A*. You can repair validation errors with the Preflight tool (*Tools → PDF Standards → Preflight*). In Preflight, select the *Profiles* view, then choose the desired conversion profile from the *PDF/A* group, and finally click the *Analyze and fix* button in the lower right corner of the window.

Callas pdfaPilot

Callas pdfaPilot can create and repair PDF/A files in a similar manner as Adobe Acrobat Pro. Editing document metadata with Callas pdfaPilot is

laborious, so write the metadata with LaTeX instead.

Other programs

There are several online services for converting pdf-files to PDF/A format. At least PDFen seems to work well.