Typing a thesis that meets the PDF/A-1b ISO-19005-1:2005 specification with LaTeX, PDFLaTeX or XeLaTeX

by

J Concepción Loredo-Osti[1]

1 Introduction

Many institutions (ours included) require that the thesis are formatted/produced using the PDF/A-1b ISO-19005-1:2005 standard. The PDF/A-1b specification is an standard for long term digital archiving. It defines a profile for PDF documents which ensures preservation of their content over extended periods of time as well as the capability of be retrieved and rendered with a consistent and predictable result each time they are viewed [3]. With this, we expect the visual reproduction of the PDF documents in the future to be faithful to what the author had seen when they were composed. To achieve this goal, it is necessary that the PDF/A-1b documents are totally self contained; this means that embedded in the PDF file is all the information needed to display the document every time that is accessed. Obviously, a PDF/A-1b document shall not contain external references linked.

2 Main features of the PDF/A-1b specification

The PDF/A protocol attempts to maximise: device independence, self-containment and self-documentation [1, 2]. Its main features are

- The PDF/A identifier and the level of compliance shall be stated.
- Use of standards-based metadata is mandated, in particular, specifications such as author, document title, creation data, and source program must be XMP-compliant.
- Encrypted security settings are prohibited; it must be possible to open/process the PDF file in question without requiring a password.

[1] Department of Mathematics and Statistics, Memorial University; St. John's, NL, Canada; 2015.

- Colorspaces specified in a device-independent manner and identified by output intent.

- Transparency layers in images and other forms PDF layers of are not permitted.

- Both LZW and JPEG2000 compression are not permitted.

- All fonts used must be embedded (at least as subsets) and also must be legally embeddable for unlimited, universal rendering. Mapping of character codes to glyphs must be unambiguous and each letter must have a unicode equivalent.

- Audio and video content are forbidden.

- Referenced (non-embedded) images or page content are not permitted.

- Alternate images (for lower-resolution screen display) are not permitted.

- Embedded JavaScript and executable file launchers are prohibited.

- Certain actions, such as opening movies or sound files or sending or resetting forms, are not permitted.

- Forms are permitted, but with restrictions.

There is a lot of documentation on the web describing in detail each of the points above.

3 Typing PDF/A-1b documents with LaTeX, PDFLaTeX or XeLaTeX

There is one general PDFLaTeX package (`pdfx.sty`) that claims to produce PDF/A-1b documents as one of its options. However, it contains bugs and, frequently, the documents produced are not PDF/A-1b compliant. For a discussion on this issue check Peter Selinger's PDF/A web page [4].

Elsewhere, I have introduced the PDFLaTeX package *thesispdfa* whose scope is limited to process LaTeX files to produce PDF/A-1b ISO-19005-1:2005 documents specifically oriented for typesetting of theses. However, *thesispdfa* calls the package *hyperref* and there are well documented features and issues with *hyperref* that may prevent PDFLaTeX from successfully producing a PDF/A-1b document. Another issue with the use of PDFLaTeX with or without *thesispdfa* is that some times the metrics of the fonts are not correctly embedded. It is known that PDFLaTeX takes the metric from the TEX-font metric files '.tfm' but not always data in these files corresponds to the metrics of the fonts embedded (for example many of the .tfm files are based on the adobe fonts metrics, but the actual fonts used are not the adobe proprietary fonts, but some free replacement of them with some slight variation in the metrics of some glyphs). So, it is desirable to have the option of producing PDF/A-1b files without the use of *hyperref* or through other LATEX engines like, 'plain' LATEX or XeLATEX. This is the objective of *pdfathesis* and accomplishes it by post-processing the file with *ghostscript*, i.e., by using the program *gs*.

4 Installation of *pdfathesis*

The package contains eight files

```
pdfathesis.sty
pdfathesis.gs
pdfathesis.sh
8bit.def
sRGB_IEC61966-2-1_black_scaled.icc
sRGB_IEC61966-2-1_black_scaled.icc.terms_of_use.txt
pdfathesis.cls
MUNLogoRGB.png
MUNLogoRGB.eps
```

The file 'sRGB_IEC61966-2-1_black_scaled.icc' contains the colour scheme and can be replaced with your favourite one. Of course, if you do not know what this file does, you should better leave it alone.

When called as package, *pdfathesis* only uses the first five files. The last files (MUNLogoRGB) of this list contain the logo of the university and it is only used to set up the front page when *pdfathesis* is called as a class. You can replace this file with whatever mylogo.png or mylogo.eps (depending upon the program used to process your document) by inserting the macro '\UniversityLogo{mylogo}' in the preamble of your LATEX file.

The files are packed in 'pdfathesis.tar.gz' and 'pdfathesis.tds.zip' to be used in unix-like and windows systems, respectively. These files should be unpacked in a place where your main LATEX processing program can find them.

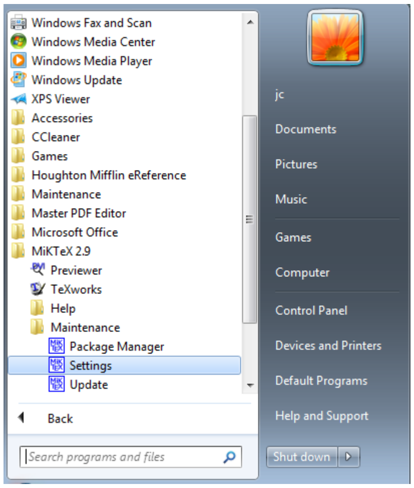
4.1 Linux

If the target is a linux system with tex-live installed, depending upon the kind of installation, the file `pdfathesis.tar.gz` should be unpacked in the directory '`/usr/share/lexlive/texmf-local`' or in `/usr/local/texlive/texmf-local` for an overall system installation that all users can access. If you do not have administrator rights, unpack the file inside '`$HOME/texmf`' for a user installation. Notice that Linux terminals in the Department of Mathematics and Statistics of MUN have already installed *pdfathesis*.

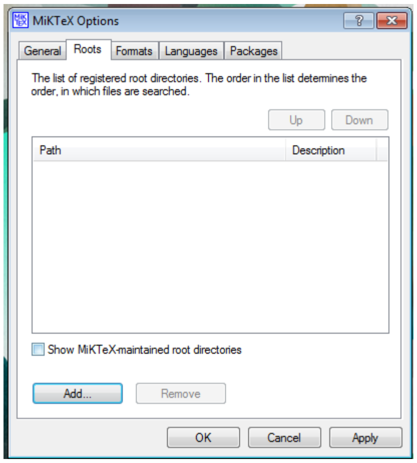
4.2 Windows

If the target is a windows system with MikTeX installed, then the following instructions will do a personal installation of *pdfathesis*.

1. If it does not exist, create the folder '`C:\Users\<UserName>\texmf`' and unpack '`pdfathesis.tds.zip`' within this folder.
2. Navigate through the MikTeX programme
   `Start > Programs > MiKTeX > Maintenance > Settings`



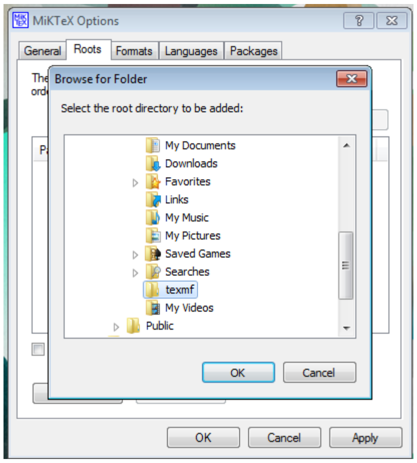
3. In the `<MikTeX Options>` window select `<Roots>` and press the `<Add>` button.



MiKTeX Options

The list of registered root directories. The order in the list determines the order, in which files are searched.

□ Show MiKTeX-maintained root directories



Browse for Folder

Select the root directory to be added:

4. Now select the user '`texmf`' folder and press the `<Apply>` button.

4.3 OSX

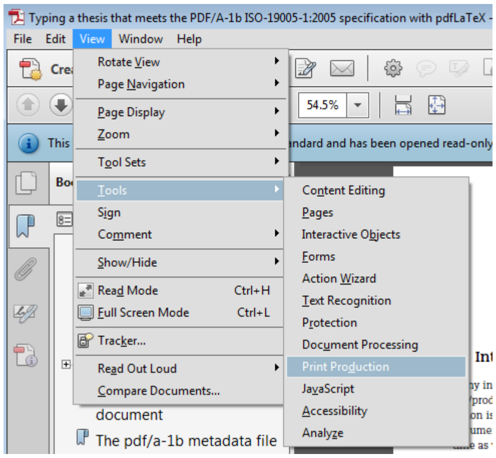
If the target is a OSX system with MacTeX, then the installation is similar to the linux one with '`$HOME/texmf`' replaced with '`~/Library/texmf`'.

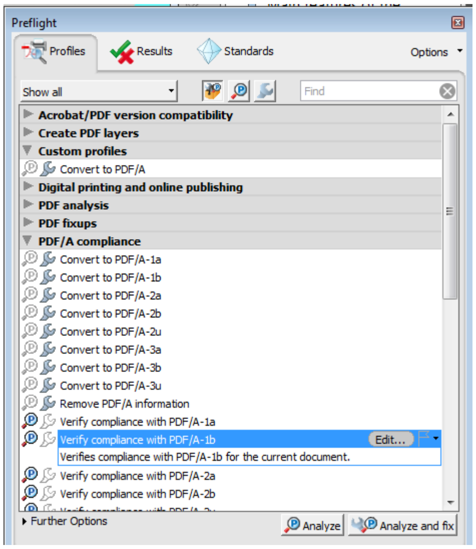
If you are not using one of the above TEX distributions, you shall to consult your distribution's documentation.

5 Validation the PDF/A-1b document

Before submitting the document, it is important to carry out some checking regarding its compliance with the PDF/A-1b specification. Although there is not an infallible validator, there are few with acceptable ratings

1. The Adobe Acrobat® plug-in application `preflight`.
   Navigate through `Acrobat > View > Tools > Print production`
   then select the button `<Preflight>`





In the `<Preflight>` window select `<Verify compliance with PDF/A-1b >`.

2. The Apache PDFBox command-line application `preflight`. Note that versions earlier than `v2.0.0` do not handle properly some time zones like the Newfoundland time zone and, wrongly, report an XMP discrepancy error. The `preflight` that comes with the PDFBox `v2.0.0` is one of the best validators.

3. The 3-Heights™ PDF Validator Online Tool

   `http://www.pdf-tools.com/pdf/validate-pdfa-online.aspx`

   and `https://www.pdf-online.com/osa/validate.aspx`.

For a comprehensive report about validation and validators, check the the PDFLib website (`http://www.pdflib.com/knowledge-base/pdfa/validation-report`).

6 The PDF/A-1b metadata file

There is a file with the name of your LaTeX file and extension '.xmpi' generated by *theispdfa*. This file holds the XMP-metadata content to be included in the PDF/A-1b file. The information used to compose this file must be defined the preamble through the following self-explanatory macros:

```
%%% Metadata macros shall be placed in the document's preamble
%%
\Title{The title of the document}
```

```
\Author{Author’s name}
\Degree{Whatever Degree Applies}
     %% ‘\DocumentType’ options: thesis, practicum, project, dissertation
\DocumentType{dissertation}
\DateSubmitted{Month Year}        % default: current Month and Year
\ConvocationYear{Year}            % default: the year of next convocation
     %% to fill ‘\Subject’ check http://www.ams.org/msc/msc2010.html
     %% or the subject classification system used in your discipline
\Subject{MSC-code The subject of this document}
     %% the keyword separator is ‘\sep’
\Keywords{keyword1\sep keyword2\sep keyword3}
\University{University’s name}
\AcademicUnit{Department/School/Institute’s name}
     %% don’t know what a ‘\ColorspaceProfile’ is, then leave this alone
\ColorspaceProfile{file name}{profile identifier}{number of components}
     %% The last two entries are optional and by default undefined
%\Supervisor{Supervisor’s name}
%\CoSupervisor{Co-supervisor’s name}
%%
%%%  End of metadata
```

All first eleven metadata entries are mandatory, but notice that only three of them (`\DateSubmitted`, `\ConvocationYear` and `\ColorspaceProfile`) have reasonable default values. All the other macros get default the values that must be replaced in the preamble.

7 Usage of the *pdfathesis* package

If you are already using a LaTeX class/package to format your thesis and you are happy with it, you only need to insert in the preamble of your document the calling for the *pdfathesis* package, i.e., place at the beginning of your preamble

```
%
\usepackage{amsmath}  % this must go before because of ‘hyperref’
             % although, you may be better off without ‘hyperref’
\usepackage{pdfathesis}
%
```

followed by the metadata information (see previous section) and you are done. Next time that you run LaTeX or pdfLaTeX or XeLaTeX you will get your document (dvi or pdf) and two additional files: one with extension `.gs` that holds the *ghostscript* commands, another with a shell script (the one with extension `.sh`) to prepare and execute *ghostscript* which is the program that actually produces the PDF/A-Ib document. Once you run this shell script, for example, in a linux terminal window, by typing

```
    sh mythesis.sh
```

you will have a PDF/A-Ib document, unless you are using *hyperref* or some other package that requires it, like *bookmarks*, and some of the options collide and/or are incompatible with PDF/A. Then, you check that the file produces is PDF/A-Ib compliant. This can be done with the command

```
    preflight mythesis-PDFA.pdf
```

(The PDFBox program *preflight* is a free PDF/A-Ib validator and it is also installed in our linux terminals). If you were unable to obtain the PDF/A-Ib file the first time, remove the offending packages/options and try again. Remember that many of the options of *hyperref* are not compatible with the production of a PDF/A-Ib document. Sometimes the problem is with the fonts, change the font package and try again.

8 Usage of the *pdfathesis* class

Another option is to load *pdfathesis* as a class, i.e., the firs line of your main LaTeX file is

```
    \documentclass{pdfathesis}
```

If you exercise this option, the class will load the packages *amsmath*, *pdfathesis*, *setspace* and *graphicx* in that order. In addition to the two auxiliary files mentioned in the previous section (`.gs` and `.sh`), after running your LaTeXengine you will have another file with extension `.bmks` with postscript commands to produce first level bookmarks. Also, it will set the document margins, the inter-line spacing and define or renew the following macros:

```
\frontpage   % generates the front page,
\dedication  % marks the start of the dedication page
```

The following macros mark the beginning of their respective page and put the headings in the table of contents

```
\abstract
\aknowledgements or  \aknowledgement
\contributions or \contribution
```

```
\laysummary
\tableofcontents
\listoftables
\listoffigures
\listofsymbols
\listofabbreviations
\thebibliography
```

also, it defines the environments `prefatory` (sets up the prefatory section), `symbols` (a single-spaced two-column table for symbols) and `abbreviations` (a single-spaced two-column table for abbreviations and acronyms). Finally, it defines the command `\thesispagestyle` to make the heading page number only. It can be overwritten with the command

```
    \renewcomman{\thesispagestyle}{ ...  }
```

if you like something fancier. For example, you can put the following code in the preamble somethig like

```
\usepackage{fancyhdr}
\renewcommand{\thesispagestyle}{%
  \pagestyle{fancy}
  \renewcommand{\chaptermark}[1]{\markboth{##1}{##1}}
  \renewcommand{\sectionmark}[1]{\markright{\thesection\ {##1}}}
  \fancyhf{}
  \fancyhead[L]{\sc \rightmark}\fancyhead[R]{\thepage}
  \renewcommand{\headrulewidth}{0.5pt}
  \renewcommand{\footrulewidth}{0pt}
  \addtolength{\headheight}{0.5pt}
  \fancypagestyle{plain}{\fancyhead{}\renewcommand{\headrulewidth}{0pt}}
  \setlength{\headheight}{15pt}\setlength{\topmargin}{0pt}
  \setlength{\textheight}{210mm}
}
```

and type

```
\thesispagestyle
```

after the prefatory section. If you ask me, I like a simple page style so I would go for the default.

The *pdfathesis* requires that ‘`\GradSchool`’ and ’`\UniversityAddress`’ are defined in the preamble. Optionaly, if ‘`\UniversityLogo`’ is also defined and point out to a valid PNG or JPEG, the front page will show the logo. Of course, the logo file must contain a PDF/A-1b compatible image.

9 Post-processing

Once that the document has been processed through LaTeX, PDFLaTeX or XeLaTeX, you post-process the output to obtain the final PDF/A-1b document. As you might know LaTeXproduced `.dvi` files while PDFLaTeX and XeLaTeX produce `.pdf` documents.

If you have used *pdfathesis* as LaTeX package, after running your favourite LaTeX processor you will have your document plus two additional files, one with extension '*.gs*' containing the postscript commands that ghostcript needs to produce a PDF/A-1b file with your metadata and a shell script (the one with extension '*.sh*') to prepare and execute *ghostscript* to post-process the document.

If you have used *pdfathesis* as LaTeX class, in addition to the `.gs` and `.sh` auxiliary files, you will have also a file with first level bookmarks postscript commands (the file with extension `.bmks`).

To carry on the post-processing, move to the directory where the document was generated and run the shell script. For example, in linux, open a terminal window and move to the directory were the document lives. Then, execute the script as

```
sh myfilename.sh
```

and if everything is in place you will get a PDF/A-1b document. The resulting file will have extension '*-PDFA.pdf*'. Validate the produced file with

```
preflight myfilename-PDFA.pdf
```

If it passes the validation, rename the file according to regulations and you may be ready to submit. If you have chance and means, you should use other PDF/A-1b validators.

If you have processed the file in windows and you do not have a shell program installed, you can take the `.pdf` or `.dvi` file with the other auxiliary files to a linux terminal. Copy the files in a directory, open a linux terminal window and before executing the shell script as indicated before, remove the '\r' from the files with extension `.sh`, `.gs` and `.bmks` (this last only if you used *pdfathesis* as class) by using the command

```
dos2unix myfilename.sh myfilename.gs myfilename.bmks
```

Of course, you can install a windows shell program, like *win-bash* or the *Cygnus* environment amongst many other options. If you do this, you may also need to edit the `.sh` file to reflect how the ghostscript callings are done in your windows installation. Needless to say that this route is not supported and if you decide to go this way this manual won't help and you are by your own. Anyways, if this is too much for you, just go to a linux terminal and do the post-processing there.

10 Image transparency and PDF/A-1b

The PDF/A-1b specification forbids the use of image transparency. You can remove a transparency 'alpha channel' of your images with `gimp` as follows:

1. Start `gimp`.
2. Open the file with the image whose alpha channel has to be removed.
3. Navigate through the menu `Gimp > Image > Flatten image`.
4. Save the file.

References

[1] Library of Congress Collections. PDF/A-1, PDF for long-term preservation. Use of PDF 1.4. http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

[2] Library of Congress Collections. PDF/A, PDF for long-term preservation. http://www.digitalpreservation.gov/formats/fdd/fdd000318.shtml.

[3] A. Oettler. *PDF/A in a nutshell* 2.0. Association for Digital Document Standards, Berlin, 2013. http://www.pdfa.org/publication/pdfa-in-a-nutshell-2-0.

[4] P. Selinger. Creating high-quality PDF/A documents using LaTeX, 2015. http://www.mathstat.dal.ca/ selinger/pdfa.