PDF libraries and TeX

Martin Schröder

ConTeXt 2008, 20th August—24th August 2008, Bohinj

(BY:)   (=)

Introduction

TEX engines and the PDF libraries

Some PDF libraries

Conclusio

Why this talk?

- Over the last years a number of new PDF libraries have appeared

- And two new TEX engines with PDF output have been created

- So the question is: Should these projects switch to one of the new libraries?

What is in a PDF library?

- PDF is a relatively complex file format with a lot of different object types
- Most PDF libraries are designed for writing PDF
- Only a handfull of PDF libraries support reading PDF
- Very few PDF libraries are designed for *modifying* PDFs

What to look for in a PDF library

- Free (BSD or GPL)
- Actively maintained
- High level of abstraction
- Reading and writing
- Incremental writing (modifying)
- PDF 1.5
- Fonts and Colours

TEX engines and the PDF libraries

- We now have three free TEX engines that can read and write PDFs

- Ideally these engines would use one well designed cleanly written library for reading and writing PDF

pdfTEX

- pdfTEX uses XPDF for PDF inclusion
- XPDF is written in C++ and used only in one source file (`pdftoepdf.cc`) of pdfTEX (which is Pascal and C otherwise)
- There may be an additional layer of abstraction between pdfTEX and XPDF in pdfTEX 1.50
- XPDF is statically compiled into pdfTEX
- Writing PDF is done without an abstract concept of PDF objects by pdfTEX itself
- There's a patch by the debian guys for using poppler instead of XPDF

luaTEX

- luaTEX is the same as pdfTEX: It uses XPDF, and the PDF inclusion code is mostly unchanged. So is the PDF writing code.

- There is currently no layer of abstraction between luaTEX and XPDF

- XPDF is statically compiled into luaTEX

XƎTEX

- XƎTEX uses XPDF to find the bounding box and orientation of included PDFs

- XPDF is statically compiled into XƎTEX

- xdvipdfmx has its own PDF parser written in C used for reading *and* writing

XPDF

- XPDF is a PDF viewer (and some command line tools) started in 1996 and written in C++

- It's not designed as a library

- It's dual-licensed: © Glyph & Cog, GPLv2 and commercial licenses are available

- XPDF has a history of security problems

poppler

- poppler is a fork of XPDF started in 2005 aimed at creating a free (GPLv2) PDF rendering library which is API-compatible to XPDF

- poppler's core can be easily substituted for XPF's code; indeed the XPDF viewer can be compiled with poppler as a backend

- poppler is part of cairo

- Lately some work has been done on giving poppler PDF writing ability

podofo

- podofo is a PDF library (with reading and writing) started in 2006, written in C++ and licensed at GPLv2
- podofobrowser is a PDF object browser (using podofo and Qt) which can also modify PDFs

GNU PDF

- "The GNU PDF Library provides functions to read and write PDF documents conforming to the PDF 1.7 specification. This includes visualization (retrieving of bitmaps with rasterized page contents) and interactive features such as annotations and interactive forms. The library also support the generation of specific subsets of PDF conforming to the ISO standards PDF/A, PDF/X and ISO 32000.
  Right now the library is under heavy development and we have not released a version yet."

- GNU PDF is an FSF high-priority project

- It's written in C and (of course) licensed at GPLv3

- There will also be a full-fledged PDF viewer and editor called GNU Juggler

iText

- iText is a PDF library written in Java initially aimed at writing (lately reading and modifying has been added) licensed at MPL or LGPLv2; commercial licenses are available

- pdftk is a command line tool written in C using iText (thanks to gcj) which allows some manipulations of PDFs

MuPDF

- MuPDF is a PDF library with reading and writing started at GhostScript written in C and licensed at GPLv2

- MuPDF is part of Fitz (a graphics library) which also includes Samus (a Metro [XPS] parser) and FzView (a PDF and Metro viewer)

multivalent

- multivalent is a viewer written in Java for HTML, PDF, *DVI*, man pages, and other document formats written in Java licensed at GPLv2

- the PDF library supports reading, writing and modifying up to PDF 1.5

- the latest release is of 2006

Others

- PDFlib is commercial C library aimed at creating PDFs from web services; lately some PDF import functions have been added. There's also a free variant of which pdfTeX borrowed some code

- PJX is a Java library supporting reading, writing and modifying licensed at GPLv2

- PDFBox is a Java library with reading and writing written licensed at BSD

- Apache FOP has a Java library for writing PDFs licensed at Apache License 2

- Adobe and Global Graphics sell commercial PDF libraries

- There are many abandoned or unfinished free PDF libraries

Conclusion

- There is no ideal free PDF library yet
- XPDF is showing its age
- poppler is a ready substitute
- podofo, MuPDF and GNU PDF are the future