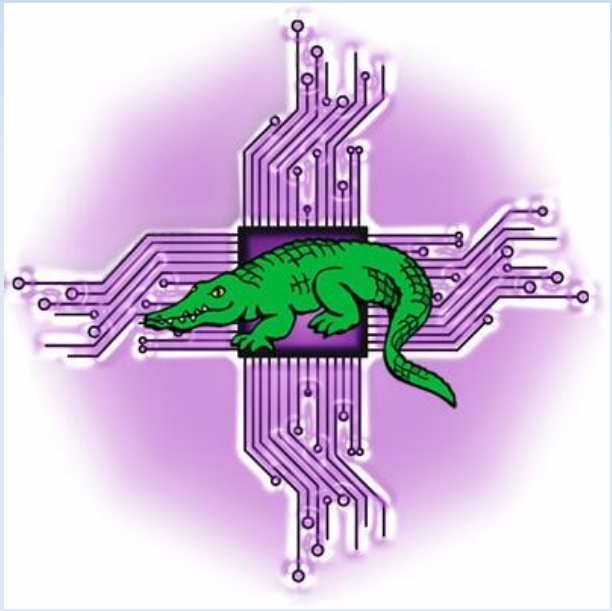
Being An Acrobat: Linux and PDFs

Presentation to Linux Users of Victoria



Melbourne, June 16, 2018

http://levlafayette.com

Objective and About the PDF File Format

- Objective of this presentation to impart understanding about the PDF file format and numerous ways it can be efficiently manipulated in Linux and other free software that may not be easy in proprietary operating systems or applications.

- Portable Document Format (PDF) is a file format first specified by Adobe Systems in 1993. It was a proprietary format until it was released as an open standard on July 1, 2008, and published by the International Organization for Standardization as ISO 32000-1:2008.

- PDF combines three technologies (1) a subset of the PostScript page description language, for layout and graphics (2) a font-embedding/replacement system, and (3) a structured storage system to bundle elements and content into a single file.

PDF Readers for Linux

- Support for Adobe Acrobat Reader was discontinued a while ago, however there are still sites which maintain a repository (e.g., `sudo add-apt-repository "deb http://archive.canonical.com/ precise partner"`, `sudo apt-get update`, `sudo apt-get install acroread`)

- Okular; universal document viewer developed by KDE. Includes embedded 3D model, table selection tool, copy images to clipboard, annotation, etc.

- Evince; simple document viewer, default on Gnome desktop environment, annoying scrolling issues.

- XPDF; a venerable PDF viewer that include text extractor and PDF-to-PostScript converter.

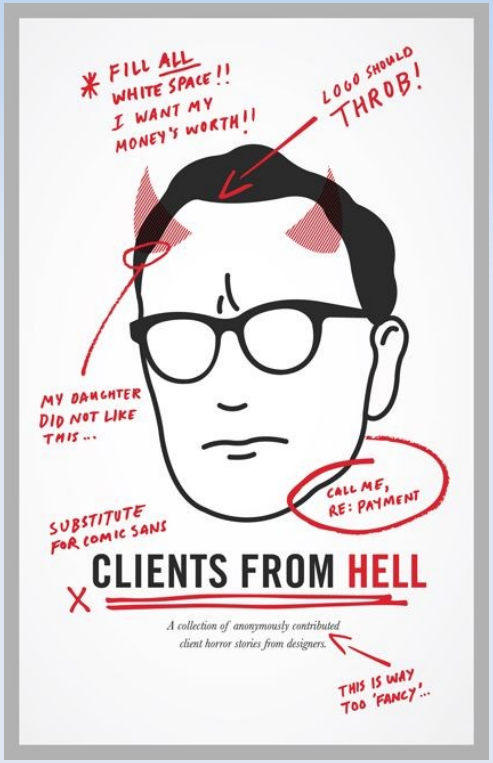
- Calibre eBook Reader; more specifically designed for formats like mobi and epub, handles PDFs just fine. Great for tablets!

Office Documents to PDF

- LibreOffice has built-in functionality for exporting word processing documents, presentations, spreadsheets, and drawings as PDF files. More powerfully, it can do it from the command-line e.g., A supposedly “client from hell” request: "Please do me a small favor and convert all 357 word documents into pdf format."

```
 /usr/bin/soffice --headless --convert-to-pdf
*.doc
```

- A simple method to edit PDFs is to open them in LibreOffice Draw. The GIMP can also be used to crop, copy, remove, and add text. PDFedit provided a powerful set of editing tools under the QT3.x framework.



Images, Postscript, and PDF

- Images to PDF (and vice-versa). Use `convert` command from ImageMagick. Many options available; consider using `-trim` (remove edge pixels), `-density` (dpi for output, declare before input), -`flatten` (flatten a sequence of images), to PDF (and vice-versa), `quality` (quality level for jpg, png).

```
convert -density 300 -flatten -trim foo.pdf foo.png
```

- Use `tesseract-ocr` when stuck with non-OCR pdf documents and the text is needed. Convert to an image, then run `tesseract imageinfile.png outputfile.txt`

- PostScript to PDF, use ps2pdf. Popular options for optimisation (send one page at a time, `-dOptimize`), embed fonts, and compression.

```
ps2pdf -dOptimize=true -dEmbedAllFonts=true
-dUseFlateCompression=true bar.ps bar.pdf
```

Text Documents and PDF

- The ps2pdf command can be used for rendering of man pages, redirct the man page to ps first (`-Tps`). e.g.,

```
man -Tps rsync | ps2pdf - rsync.pdf
```

- restructuredText documents can be easily converted to PDFs using rst2pdf.py; install with `sudo pip install rst2pdf`. Examples:

```
rst2pdf.py foo.rst -o foo.pdf
```

```
curl http://docutils.sourceforge.net/docs/user/rst/quickstart.txt |
rst2pdf > quickstart.pdf
```

- Markdown to PDF requires node, `sudo apt-get install npm sudo yum install npm` for installation `sudo npm install -g markdown-pdf, markdown-pdf -options markdownfile.md -o markdown.pdf .` Nota bene: Can be used in scripting logic.

CUPS, Pandoc, PDFBox for PDF Creation

- CUPS can render any document to a PDF file, thus anything that can be printed in Linux can become a PDF file.

- Pandoc converts from one markup format to another. Input is specificied with `-r/--read` or `-f/--from` and output using `-w/--write` or `-t/--to` options. Output is stout by default. PDFs are created by the LaTeX engine by default. Note that Pandoc can only output as PDF and the -o is required to print out to a file rather than to stdout.

```
pandoc -r markdown_mmd -t latex test.md -o test.tex
pandoc test.tex -o test.pdf
```

- A useful Java library for creation, manipulation, and extraction is Apache PDFBox.

Searching and Converting PDF Files

• PDF files can be simply searched with various reader applications. From the command-line, `pdfgrep` is a powerful choice that uses many of the expected options from grep, such as `-i` for case-insensitive, `-P` for a Perl compatible regular expression, `-r` for recursive, `-R` for recursive with symlinks. An additional bonus is `-n` which prints the page number where the expression occurs. As with other command-line options, the great advantage of `pdfgrep` is the ability to loop and include in scripts. e.g.,

```
pdfgrep -r -i regularexpression /path/to/files/
```

• PDFs can be convert to textfiles, as well as possible, with `pdftotext`. Options include page number range (`-f`, `-l`), output of simple html (-htmlmeta), and a valiant attempt to replicate formatting (`-layout`). Often combined with pipes and loops, and is part of the broader popper-utils.

```
pdftotext test.pdf - | wc
```

Poppler Utilities

- Derived from xpdf, Poppler is a software library for rendering PDF documents and is used in a variety of PDF readers (including Evince, KPDF, LibreOffice, Inkscape, Okular, Zathura etc).

- A collection of tools, poppler-utils, is built on Poppler's API provides a variety of useful functions e.g.,

pdffonts - lists the fonts used in a PDF (e.g., pdfonts filename.pdf)
pdfimages - extract images from a PDF (e.g., pdfimages -png filename.pdf images/)
pdfseparate - extract single pages from a PDF (e.g., pdfseparate sample.pdf sample-%d.pdf)
pdftohtml - convert PDF to HTML format retaining formatting
pdftops - convert PDF to printable PS format
pdftotext - extract text from a PDF
pdfunite - merges PDFs (pdfunite page{01..13}.pdf combined.pdf ; another option is convert page{0..13}.pdf combined.pdf)

Create PDF Presentations from Text

- The tool pinpoint for GNOME provides the options for exciting presentations with a simple markup syntax, e.g., pinpoint introduction.pin

- Can export files as PDF, `pinpoint introduction -o introduction.pdf`

- This presentation was not created with pinpoint (but maybe it should be!)

Fillable PDF Forms

- Scribus can be used to create PDF documents. When creating such documents the PDF Field selection allows for the inclusion of interactive elements (such as text-fields, check boxes, combo boxes, list boxes). Select the Field and select the area on the document desired. Properties for the field can be configured as desired. Select PDF Options and Field Properties, also consider the Options tab.

- A label should also be created so the user knows the purpose of the field or checkboxes. To create the label, use the standard Text Frame and then use the Story Editor to create the label.

Add and Recover Passwords etc

- **The most simple method to add passwords in a PDF in Linux is to use the security tab in LibreOffice.**

- **A password can be added or removed with the qpdf toolkit, a very useful set of tools!;**

```
qpdf --encrypt [readpass] [ownerpass] 256 -- infile.pdf
outfile.pdf
qpdf --password=userpassword --decrypt infile.pdf outfile.pdf
```

- **The qpdf toolkit can also be used to recover (at least in part) damaged and corrupted PDF files.**

```
qpdf infile.pdf outfile.pdf
```

- **To recover a lost password, use pdfcrack. Use options to specify range etc.**

```
pdfcrack filename.pdf
```

- **See also `pdf-parser.py` and `pdfid.py` for checking documents.**


THANKS FOR WATCHING
& LISTENING PATIENTLY
makeameme.org