Thomas Fischer's Weblog

Life, Linux, LaTeX

Searchable PDFs with Linux

with 7 comments

Recently, I came across a news posting that there is an open source document management software called ArchivistaBox 2008/IX that can create searchable PDFs from scanned documents. Core components of this software package are Cuneiform (an OCR system) and hocr2pdf (a special PDF generator from ExactCODE).

Using these two programs (both are GPL-2), everyone can generated searchable PDFs which I will demonstrate in the following example.

Lacking a scanned document, I created a LaTeX document using a sample text from Project Gutenberg and generated a TIFF file using GhostScript:

```
pdflatex mammalia.tex
gs -r320 -dBATCH -sOutputFile=mammalia.tiff -sDEVICE=tiffgray mammalia.pdf
```

Tip: When scanning or generating TIFF images, try different image resolutions where the recognization rate is sufficient and the image size is still acceptable small.

Generating a searchable PDFs is a two-step process. First, cuneiform is used to generate a special HTML document which contains information where letters and words are located on the TIFF image.
This HTML document uses the suffix .hocr:

```
cuneiform -f hocr -o mammalia.hocr mammalia.tiff
```

Tip: You can use cuneiform to write its output in different other formats such as normal HTML or plain text. Use `cuneiform -f` to get a list of formats.
Tip: Linked against ImageMagick, cuneiform can read a large number of image formats, not only TIFF.

Once `mammalia.hocr` has been generated, the searchable PDF document is generated using hocr2pdf:

```
hocr2pdf -i mammalia.tiff -o mammalia-ocr.pdf <mammalia.hocr
```

Here, the TIFF image is used for the PDF's visual content, but when you search for text, the meta information from the .hocr file is used to find and highlight the search hits in the document.

Above example is rather artificial, as the used TIFF image has a much better quality compared to a scanned document. If scan results degenerate (not all letters are recognized and some word boundaries are detected wrong), you may want to try the optional switch `-s` for hocr2pdf to use a more sloppy approach on detecting words.

Now you can use above tools to run your own document management system at home e.g. to scan incomming letters. Happy OCRing… 😃

Note: Gentoo Linux users can use ebuilds from bug reports for cuneiform and exactimage.

Like Be the first to like this post.

Written by Thomas Fischer

November 26, 2008 at 22:23

Posted in Linux

Tagged with ocr, pdf

« Videos on using KBibTeX
Verteidigung »

7 Responses

Subscribe to comments with RSS.

1. This seems to work as you describe but only does the first page of my document. I can't see any way to specify to cuneiform which pages to process. Am I missing something?

   Jonathan

   July 22, 2009 at 2:57

2. For more than one page you'll need batch processing (shell scripts).

   I wrote an article about that, you'll find it with a search engine with the keywords 'linux ocr and pdf problem solved' (it seems I'm not allowed to post links here).

   Konrad Voelkel

   March 6, 2010 at 12:30

3. There is a script for processing multipage PDFs.

   http://superuser.com/questions/28426/how-to-extract-text-with-ocr-from-a-pdf-on-linux/33203%2333203

   Rodrigo Torres

   September 1, 2010 at 18:56

4. Ubuntu – from Konsole -

   $ cuneiform -f hocr -o scan-0001.hocr scan-0001.tiff
   Cuneiform for Linux 0.7.0
   PUMA_XFinalrecognition failed.

   Any idea what the interesting error is saying?

   Thank you

   Barry Smith

   September 7, 2011 at 20:45

5. Answered my own question… in a trial-and-error way.
   scan-0001.tiff was made at 600x600DPI.
   Created scan-0002.tiff at 150x150DPI, and it worked.

   QED.

   Barry Smith

   September 7, 2011 at 22:37

6. New Question:
   Mr. Torres responded about extracting text from a PDF on a multi-page document. That text file can't be used in your cuneiform & hocr2pdf process, can it?

   Yet _differently_, I want to work on a multi-page document to create a single searchable PDF.

An example — My resume is 23 pages long in .doc format for all of the silly recruiters out there.
If I create a 150dpi image file for each page, and run each file through your cuneiform & hocr2pdf process, I'm left with 23 PDFs that cannot be merged… am I not? The .hocr file would refer to a single page document.

Another wrinkle – I'm still working on KDE, but I have installed GNOME tools after switching to kdm-KDE,and they are working. xsane is working for scanning… I was able to scan a single-sheet to tiff, and use your process above. yet, how do I convert the multi-page .doc to searchable PDF, and then chain the single-sheet PDF – xsane-cuneiform-hocr2pdf – followed by the 23 individual PDFs of my resume?
While waiting for this complex answer, I'll continue to ponder. as I did above… but if you have the answer, please share. 🙂

Thank you again,
Barry

Barry Smith

September 8, 2011 at 11:50

7. If you have the .doc file, you can easily create PDF files using LibreOffice or OpenOffice. If you have a multipage PDF file which basically consists of scanned images and is not searchable, you can use the following Bash script:

```
TEMPDIR=$(mktemp -d)
INPUTPDF="$1"
OUTPUTPDF="${INPUTPDF/.pdf/-index.pdf}"

gs -r320 -dBATCH -dNOPAUSE -sOutputFile=${TEMPDIR}/page%05d.tiff -sDEVICE=tiffgray "${INPUTPDF}" || e
for tiff in ${TEMPDIR}/page*.tiff ; do
  hocr=${tiff/.tiff/.hocr}
  pdf=${tiff/.tiff/.pdf}
  cuneiform -f hocr -o ${hocr} ${tiff} && \
  hocr2pdf -i ${tiff} -o ${pdf} <${hocr} || \
  exit 2
done
pdftk ${TEMPDIR}/page*.pdf output "${OUTPUTPDF}"

rm -rf ${TEMPDIR}
```

Thomas Fischer

November 27, 2011 at 18:00

Leave a Reply

Enter your comment here...

Fill in your details below or click an icon to log in:

Email (required) (Not published)

Name (required)

Website

Notify me of follow-up comments via email. Post Comment

Visit my homepage at **www.t-fischer.net**

Learn more about **KBibTeX**, the BibTeX editor for KDE

Search

Recent Posts

- More on LaTeX Beamer: Linking images to an enlarged version
- Some new LaTeX Beamer Tricks
- KParts Browser Plugin
- Recording VoIP Phone Calls with ALSA
- Schnee zu Neujahr

Archives

- September 2010
- August 2010
- May 2010
- March 2010
- January 2010
- November 2009
- October 2009
- September 2009
- August 2009
- June 2009
- May 2009
- March 2009
- February 2009
- December 2008
- November 2008
- October 2008
- September 2008
- August 2008
- July 2008
- June 2008
- May 2008
- April 2008
- December 2007
- November 2007
- September 2007
- May 2007
- April 2007

Advertising

License

November 2008

M	T	W	T	F	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

« Oct Dec »

Meta

- Register
- Log in
- Entries RSS
- Comments RSS
- WordPress.com

Blog at WordPress.com. Theme: The Journalist v1.9 by Lucian E. Marin.