Hints & Tricks

Production notes: PDFs and urls

Karl Berry

Documents commonly include PDF, PNG, and JPEG files as images. It's straightforward to extract the latter two from a composed PDF, using, for instance, `pdfimages` from Poppler[1] (or Xpdf). But I wanted to extract an included PDF back out of a document PDF recently, and was surprised to find that there seemed to be no standard tool for this.

I asked Max Chernoff, fellow *TUGboat* TEXnician. He couldn't find anything either, so he wrote a LuaTEX script for the job. It's invoked like this:

```
./luatex-xobject-tub.tex somedoc.pdf
```

The output is written to `luatex-xobject-tub.pdf`, one page for each included PDF. It's available in the *TUGboat* repository.[2]

Another recurring job with PDFs is to check the live urls that are present. (We like to do this before each issue goes to press, so at least we know the urls are working at that time.) Max again came up with a solution, essentially using:

- `qpdf --qdf` (`github.com/qpdf`) for an ASCII transliteration of the PDF;
- `grep --only-matching` to extract the urls, and
- `wget --spider` to check them.

The full script is `check-pdf-urls-tub` in the same *TUGboat* repository directory. The license on these scripts is "do what you want to". Thanks Max!

That first step, converting a compressed PDF into a human-readable form, is generally needed from time to time — i.e., a `pdftype` program analogous to `dvitype` et al. Some other methods I know of:

- Use LATEX: `\DocumentMetadata{uncompress}`
- Use pdfTEX:
  ```
  \pdfcompresslevel=0 \pdfobjcompresslevel=0
  \immediate\pdfximage{in.pdf}%
  \pdfrefximage\pdflastximage \end
  ```
- `pdftk in.pdf output out.pdf uncompress`
- `mutool clean -d in.pdf`

Each has its own benefits and drawbacks, and there are surely yet more out there; I'd appreciate further information.

◇ Karl Berry
github.com/TeXUsersGroup

[1] poppler.freedesktop.org

[2] github.com/TeXUsersGroup/tugboat/blob/trunk/misc/

doi.org/10.47397/tb/45-3/tb141prod