by Guido Socher (homepage)

About the author:

We posted a while ago that linuxfocus wanted to make articles as pdf files available. As a response we received a number of suggestions which are summarized in this tip. Thanks very much for all the suggestions.

LF Tip: generating PDF from html documents

Abstract:

This is a small tip. From now on LinuxFocus will have at least one new tip every month. If you have some ideas for a new tip then send them to guido("at" sign)linuxfocus.org

Introduction

You have probably noticed that we have now PDF files for all the articles from languages which use the iso8859−1 character set. It was not easy to implement especially since we wanted to have it generated automatically to avoid that html text and PDF documents differ.

Here is now our experience with a list of options how to generate PDF in general.

The idea

All linux systems come with the ghostscript utility ps2pdf. ps2pdf works very well and the quality of the generated PDF is good. In other words we can always generate PDF files if we manage to the document as postscript file.

The entire linux printing system is based on postscript so it should be easy!? The problem is really to find a way to do it with a script from the command line. You don't want to click with the mouse when you need to print a few thousand articles.

If you are not concerned about tables colors and images then a combination of "lynx −dump .... | nenscript" and ps2pdf will work. If you need however tables and images then read on.

The candidates

html2ps

This is a perl script and the version tested here was html2ps 1.0 beta3. The homepage is http://user.it.uu.se/~jan/html2ps.html
The program works quite well. It requires however an number of perl modules as dependency and it has problems with pages containing tables to structure the page. It is a good solution if you have a very simple layout.

latex

There is a latex to pdf converter. Using xslt you could transform html to Latex. A pre−requisite for this is to have syntactically correct html. This can be done with the tidy utility:

```
HTML --(tidy)--> XHTML --(XSLT)--> Latex --(pdflatex)--> PDF
```

I did not investigate this further because I find xslt and latex to heavy and complex.

web browser remote control

If it would somehow be possible to remote control a web browser then we would have the advantage that the generated PDF is identical to what you normally see in your web−browser. The problem is that a X11 display is needed. It is therefore not possible to run this from a cron−job.

The mozilla project has improved printing and rendering it did however remove some of the remote control features that netscape communicator has. The following solution will therefore only work with communicator 4.X

```
netscape -noraise -remote "openurl(http://somepage)"
sleep(10) # there is no way to know if the page is completely loaded
          # so we just wait a bit
netscape -noraise -remote saveas(somepage.ps,PostScript)
sleep(10)
ps2pdf somepage.ps
```

Some readers told me that they think that remote printing would also be possible with konqueror but nobody could provide a working solution.

htmldoc

Htmldoc is a very well written utility from http://www.htmldoc.org/. The following command will do exactly what we wanted:

```
htmldoc -t pdf --webpage -f file.pdf file.html
```

We used version 1.8.24 and it works perfectly. The only problem is that the resulting pdf files are in average 10 times bigger than any of the pdf files generated by the other solutions no mater what compression option you use in htmldoc. A big problem if you have thousands of documents.

Conclusion

We use now a combination of netscape remote control and htmldoc. We could not rely only on htmldoc due to the size of the generated files. If you have any further suggestions an ideas regarding this subject then write us.

2005−02−09, generated by lfparser_pdf version 2.51