von Guido Socher (homepage)

Über den Autor:

Wir haben vor einiger Zeit angekündigt, dass wir unsere Artikel auch als PDFs verfügbar machen wollen. Als Antwort darauf erhielten wir eine Reihe von Vorschlägen, die in diesem Tipp zusammengefasst sind. Vielen Dank für alle Vorschläge.

Übersetzt ins Deutsche von:
Sebastian Bauer
<sebastian.baua(at)t−online.de>

LF Tip: PDFs aus HTML Dokumenten erzeugen

Zusammenfassung:

Dies ist ein kurzer Tipp, wie es ihn von jetzt an mindestens einmal im Monat bei LinuxFocus geben wird. Wenn Sie eine Idee für einen neuen Tipp haben schicken Sie ihn bitte an guido("at" sign)linuxfocus.org

Introduction

Wahrscheinlich hast du bemerkt, dass wir für Sprachen, die den iso8859−1 Zeichensatz verwenden, jetzt auch PDF−Dokumente aller Artikel zur Verfügung stellen. Das war nicht leicht zu realisieren, besonders weil die PDFs automatisch erzeugt werden sollten, damit sich HTML− und PDF−Dokumente nicht unterscheiden.

Hier werden nun unsere Erfahrungen mit einigen Möglichkeiten zur Erzeugung von PDFs beschrieben.

Die Idee

Alle Linux−Systeme enthalten das Ghostscript−Tool `ps2pdf`. `ps2pdf` funktioniert sehr gut und die Qualität der erzeugten PDFs ist gut. Mit anderen Worten ist es kein Problem PDFs zu erstellen, wenn wir es schaffen die Dokumente in Postscript umzuwandeln.

Da das gesamte Linux Drucksystem auf Postscript basiert, sollte das also eigentlich kein Problem sein!? Das eigentliche Problem ist aber, das Ganze mit einem Script von der Kommandozeile aus zu erledigen. Wer will schon jedesmal mit der Maus klicken, wenn es gilt ein paar Tausend Artikel zu drucken.

Wenn du dir keine Gedanken um Farben, Tabellen oder Bilder machen musst, reicht eine Kombination von `lynx -dump .... | nenscript` und `ps2pdf` aus. Wenn du aber Tabellen oder Bilder brauchst, ließ weiter...

Die Kandidaten

html2ps

Hierbei handelt es sich um ein Perl Script, das wir in der Version html2ps 1.0 beta3 getestet haben. Es kann auf der Seite http://user.it.uu.se/~jan/html2ps.html heruntergeladen werden.

Das Programm funktioniert ziemlich gut. Es benötigt aber einige Perl Module als Abhängigkeiten und es hat Probleme mit Tabellen, die zur Strukturierung der Seite verwendet werden. Es ist eine sehr gute Wahl, wenn man ein sehr einfaches Layout hat.

latex

Es gibt ein Programm, das LaTeX nach PDF konvertiert. Mit xslt könnte man HTML in LaTeX umwandeln. Eine Voraussetzung für dieses Vorgehen ist syntaktisch korrektes HTML, das mit dem Tool `tidy` erzeugt werden kann:

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

Ich habe diese Möglichkeit nicht weiter untersucht, da ich xslt und LaTeX als zu schwer und komplex empfunden habe.

Browser Fernsteuerung

Wenn es möglich wäre, den Browser fernzusteuern, hätte das den Vorteil, dass das erzeugte PDF identisch mit der anzeigten Seite im Browser wäre. Das Problem ist, das ein X11 Server benötigt wird. Diese Möglichkeit kann also nicht durch einen Cronjob gestartet werden.

Das Mozilla Projekt hat das Drucksystem verbessert, aber auch einige Fähigkeiten zur Fernsteuerung des Netscape Communicator entfernt. Die folgende Lösung funktioniert deshalb nur mit dem 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
```

Einige Leser haben mir geschrieben, dass sie glauben Drucken per Fernsteuerung sei auch mit dem Konqueror möglich, aber leider konnte niemand eine funktionierende Lösung liefern.

htmldoc

`htmldoc` ist ein sehr gut geschriebenes Programm von http://www.htmldoc.org/. Das folgende Kommando macht genau das, was wir wollten:

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

Wir haben Version 1.8.24 verwendet und es funktioniert ausgezeichnet. Das einzige Problem ist die Größe der PDFs. Durchschnittlich sind sie zehn mal größer als die, die mit einer der anderen Möglichkeiten erzeugt wurden, egal welche Kompression bei `htmldoc` verwendet wurde. Ein großes Problem, wenn man tausende von Dokumenten hat.

Zusammenfassung

Wir verwenden jetzt eine Kombination von Netscape Fernsteuerung und `htmldoc`. Sich alleine auf `htmldoc` zu stützen war wegen der Größe der erzeugten PDFs leider nicht möglich. Wenn du weitere Vorschläge oder Ideen bezüglich dieses Themas hast schreib uns bitte.

Der LinuxFocus Redaktion schreiben © Guido Socher "some rights reserved" see linuxfocus.org/license/ http://www.LinuxFocus.org	Autoren und Übersetzer: en −−> −− : Guido Socher (homepage) en −−> de: Sebastian Bauer <sebastian.baua(at)t−online.de>

2005−03−04, generated by lfparser_pdf version 2.51