door Guido Socher (homepage)

Over de auteur:

Een tijdje terug hadden we gemeld dat LinuxFocus artikelen ook als pdf beschikbaar wilden maken. We kregen een hoop suggesties die zijn opgesomt in deze tip. Hartelijk dank voor alle suggesties.

Vertaald naar het Nederlands door:
Guus Snijders
<ghs(at)linuxfocus.org>

LF Tip: PDF genereren van HTML documenten

Kort:

Dit is een korte tip. Vanaf nu zal LinuxFocus minstens een tip per maand hebben. Als je ideëen hebt voor nieuwe tips, kun je sturen naar guido(apenstaartje)linuxfocus.org

Introductie

Je hebt wellicht al gemerkt dat we nu PDF bestanden hebben van alle artikelen voor de talen die de iso8859−1 karakter set gebruiken. Het was niet makkelijk te implementeren, vooral omdat we ze automatisch wilden genereren om verschillen tussen HTML tekst en PDF documenten te voorkomen.

Hier volgen onze ervaringen met een aantal opties voor het genereren van PDF in het algemeen.

Het idee

Alle Linux systemen komen met de ghostscript utility ps2pdf. ps2pdf werkt erg goed en de kwaliteit van de genereerde PDF is goed. Hiermee kun je altijd PDF bestanden genereren als je het document beheerd als een PostScript bestand.

Het volledige Linux afdruksysteem is gebaseerd op PostScript, dus zou het vrij eenvoudig moeten zijn!? Het probleem is om een manier te vinden om het via een script vanaf de opdrachtregel te doen. Je wilt niet met de muis door een paar duizend artikelen klikken.

Als je je geen zorgen maakt over tabellen, kleuren en afbeeldingen, dan zal een combinatie van "lynx

–dump... | nenscript" en ps2pdf werken. Als je echter tabellen en kleuren nodig hebt, lees dan verder.

De kandidaten

html2ps

Dit is een Perl script, de geteste versie is html2ps 1.0 beta3. De homepage is te vinden op http://user.it.uu.se/~jan/html2ps.html
Het programma werkt goed. Het vereist echter wel een aantal perl modules als dependancy en heeft problemen met pagina's die tabellen voor de structuur bevatten. Het is een goede oplossing als je een eenvoudige layout hebt.

latex

Er is een latex to pdf converter. Met behulp van xslt zou je HTML naar Latex kunnen transformeren. Hiervoor dien je wel syntactisch–correcte HTML te hebben. Dit kun je je oplossen met de utitility tidy:

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

Ik heb dit niet verder onderzocht omdat xslt en latex naar mijn mening te zwaar en te complex zijn.

web browser remote control

Als het op een of andere manier mogelijk zou zijn om een web browser op afstand te bedienen, zouden we het voordeel hebben dat de gegenereerde PDF identiek is aan wat je anders ziet in je browser. Het probleem is dat een X11 scherm nodig is. Daardoor is het niet mogelijk om een cronjob te gebruiken.

Het mozilla project heeft verbeterde printing en rendering, er zijn echter een aantal van de "remote control" features verwijderd, die de Netscape Communicator wel had. De volgende oplossing werkt dan ook alleen met 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
```

Een aantal lezers hadden gemeld dat ze dachten dat remote printing met Konqueror mogelijk was, maar ze konden geen werkende oplossing aanleveren.

htmldoc

HTMLdoc is een erg goed geschreven utility van http://www.htmldoc.org/. Het volgende commando doet precies wat we wilden:

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

We hebben versie 1.8.24 gebruikt en het werkt perfect. Het enige probleem is dat de resulterende pdf bestanden gemiddeld 10 keer zo groot zijn als de pdf bestanden die waren gegenereerd met de andere oplossingen, ongeacht welke compressie methode je in htmldoc gebruikt. Een groot probleem als je duizenden documenten hebt.

Conclusie

We gebruiken nu een combinatie van netscape remote control en htmldoc. Door de grootte van de gegenereerde bestanden konden we ons niet verlaten op htmldoc. Als je nog suggesties en/of ideëen hebt over dit onderwerp, schrijf ons dan.

2005−03−27, generated by lfparser_pdf version 2.51