



von Iznogood

<iznogood/at/iznogood-factory.org>

Über den Autor:

Schon seit einiger Zeit mit GNU/Linux befasst, benutze ich nun ein Debian-System. Trotz elektronischer Studien habe ich überwiegend Übersetzungsarbeiten für die GNU/Linux-Gemeinschaft gemacht.

Werkzeuge für die Umwandlung von Papier nach HTML



Zusammenfassung:

Hier geht es um eine Werkzeugkette zur Umwandlung eines herkömmlichen Papiermagazins in HTML. Ich werde den Prozess vom Scannen bis zur HTMLifizierung erläutern.

Einführung

Ich habe gelesen, dass einige US-Universitäten es Google erlauben und dabei helfen, ihre Bibliothek in numerischer Form zu digitalisieren. Ich bin nicht Google und ich verfüge nicht über eine Universitätsbibliothek, aber ich besitze einige alte Papiermagazine über Elektronik. Die Papierqualität war nicht die beste: Seiten lösen sich, das Papier graut ...

Daher habe ich mich entschlossen, es zu digitalisieren, denn obwohl die Ausgaben vor 10 Jahren stoppten, sind einige Artikel immer noch aktuell!

Hardware

Am Anfang musste ich die Daten in den Computer bringen. Ein Scanner ermöglicht mir dies: nach einigen Kompatibilitätsprüfungen kaufte ich einen alten gebrauchten, aber billigen ScanJet 4300C, und nach einiger Internetnavigation fand ich die erforderlichen Einstellungen zur Konfiguration.

Unter Debian installierte ich sane, xsane, gocr und gtk-ocr ganz normal mit:

```
apt-get install sane xsane gocr gtk-ocr
```

```
als root.
```

Sane und xsane sind die Scanner-Werkzeuge, die mein HP zum Arbeiten benötigt.

Gocr und gtk-ocr sind Werkzeuge, um ein Bild in einen Text zu wandeln.

Der Scanner ist ein USB-Scanner:

```
sane-find-scanner
```

dann wechselte ich nach /etc/sane.d/, um einige Dateien zu editieren:
in dll.conf aktivierte ich

```
hp  
niash
```

und alles andere wurde auskommentiert.

In hp.conf und niash.conf trug ich folgendes ein:

```
/dev/usb/scanner0  
option connect-device
```

und alles andere wurde auskommentiert.

Ich veränderte die Gruppenzugehörigkeit der Gerätedatei /dev/usb/scanner mit

```
chgrp scanner scanner0
```

und fügte iznogood als Anwender hinzu, um mir die Benutzung des Scanners zu ermöglichen, ohne root zu sein:

```
adduser iznogood scanner
```

Nach einem Reboot war alles erledigt!

Zum Speichern von Bildern sind DVD-Brenner billig genug, z. B. ein NEC 3520. Ich benutze einen alten Kernel (2.4.18), daher benutzte der IDE-Brenner die SCSI-Schnittstelle:

Mittels modconf lade ich ide-scsi

und erweiterte /etc/lilo.conf um:

```
append="hdb=ide-scsi ignore hdb"
```

dann ein Aufruf von

```
lilo
```

um es zu aktualisieren.

In /etc/fstab fügte ich

```
/dev/sdc0    /dvdrom    iso9660    user, noauto    0 0
```

hinzu. Dann änderte ich die Gruppe sdc0 auf cdrom

```
chgrp cdrom sdc0
```

Recht einfach.

Software

Zur Fortsetzung des Prozesses benötige ich einige Software:

sane, xsane, gimp, gocr, gtk-ocr, einen Text-Editor, einen HTML-Editor und etwas Plattenplatz.

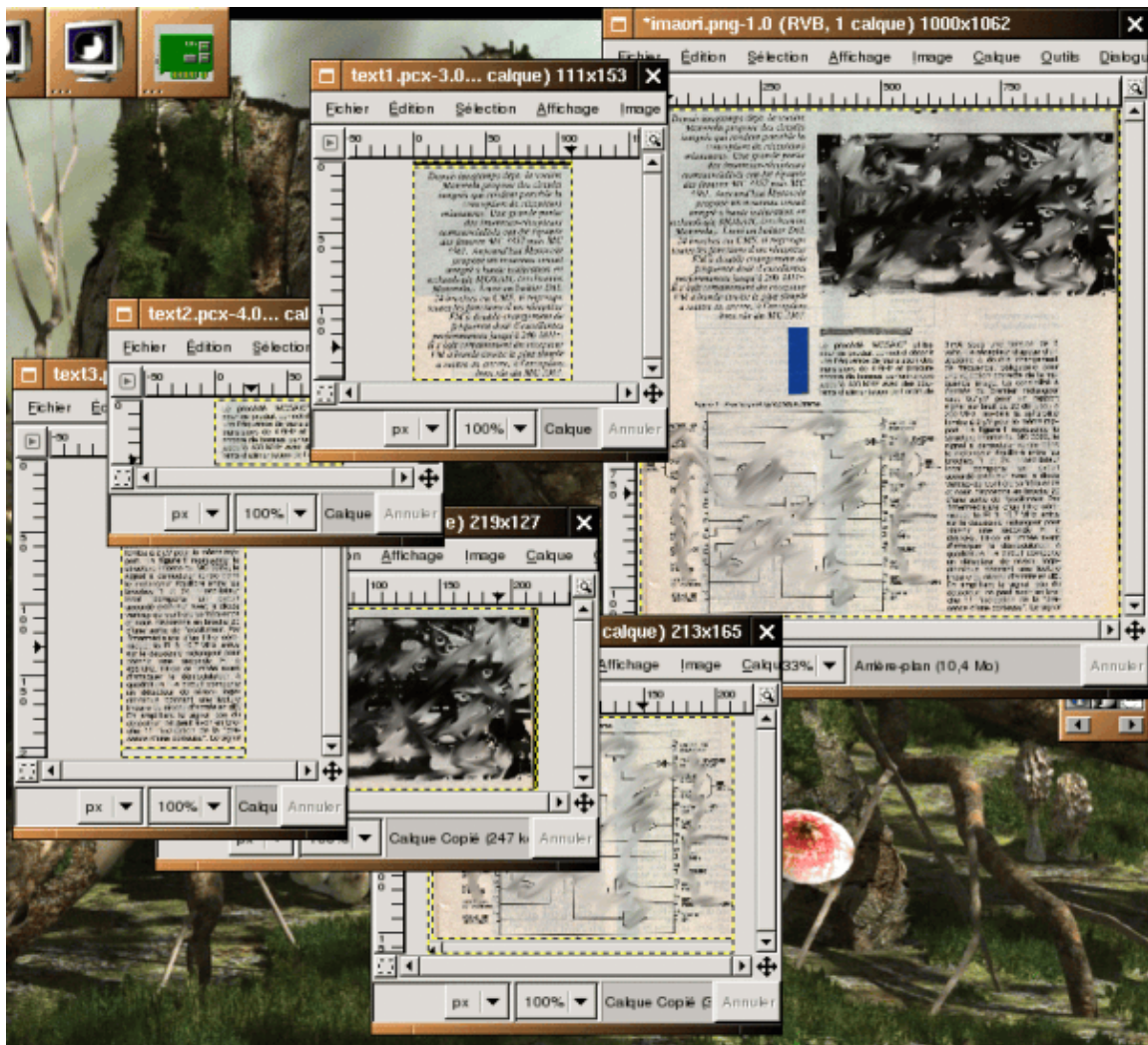
Sane ist das Scan-Programm und xsane ist die grafische Oberfläche.

Meine Vorstellung war, die maximale Auflösung beizubehalten und damit eine 50 MB-Datei für eine Seite zu erhalten, sie zur weiteren Verarbeitung auf Platte zu speichern und nach der Fertigstellung auf eine DVD-ROM zu brennen.

Ich setzte die Auflösung auf 600 dpi, etwas mehr Helligkeit und startete die Umwandlung. Da dies auf einem sehr alten Rechner (PII 350 MHz) lief, dauerte es etwas, aber ich erhielt ein gutes und präzises Bild. Ich speicherte es im png-Format.

Warum solch eine Auflösung und eine 50 MB-Datei? Ich wollte eine maximale Auflösung für das Archiv und für weitere digitale Verarbeitung.

Mittels Gimp schnitt ich die Seite in grafische Bilder und Bilder, die nur den eingescannten Text enthielten. Die Grafiken wurden mit einer reduzierten Größe in png gespeichert, damit sie auf eine HTML-Seite passen und die Textabbilder wurden nicht reduziert, aber von Farbe auf Grauwerte geändert (Werkzeuge, Farbwerkzeuge, Schwellwert und OK) und für die weitere Verarbeitung mit der OCR-Software unter der .pcx-Erweiterung gespeichert.

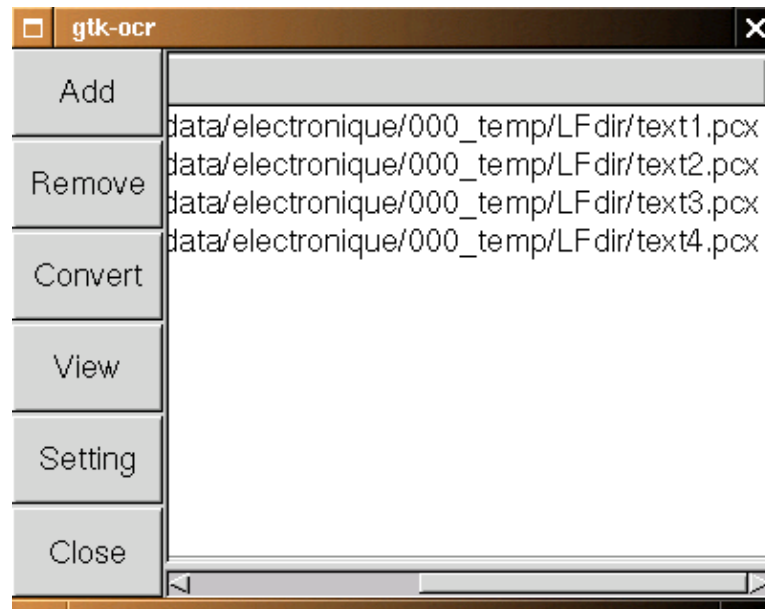


Sie können das vollständig gescannte Bild oben rechts und die ausgeschnittenen Teile auf der linken Seite sehen.

Wenn Sie die Bilder ausschneiden, können Sie Titel entfernen, da sie zuviel Platz wegnehmen und von gocr nicht erkannt werden.

Ich erstellte ein Unterverzeichnis ima für die Bilder und trennte es von den .pcx-Dateien.

Nun kommt gtk-ocr ins Spiel, die Oberfläche zu gocr. gocr ist eine Software für optische Zeichenerkennung. Es ist recht einfach zu benutzen: Ich musste nur die Dateien auswählen und gtk-ocr verwaltet alles. Ich erhielt eine .txt-Datei für jede bearbeitete .pcx-Datei.



Mit einem einfachen


```
cat *.txt > test.txt
```

erhalte ich eine Datei test.txt und kann mit einem Texteditor einige Anpassungen vornehmen (nicht französische Zeichen entfernt, Worte korrigiert ...).

Kopieren/Einfügen in den HTML-Editor (für mich Mozilla Composer) und ich konnte mit der HTML-Erstellung beginnen (achten Sie darauf, nur relative Links zu benutzen, wenn Sie einige Bilder hinzufügen).


L'lectronique, c'est bien

Les diodes sont de nos jours peu utilisées isolément. Il est commun de les voir dans des circuits ou de voir leur principe de fonctionnement étendu aux transistors. Néanmoins, on peut encore en trouver tel quel dans des circuits là où il est nécessaire d'installer une voie à sens unique pour le courant. Par exemple, dans un circuit où la polarité est vitale au bon fonctionnement, on peut installer une diode entre les bornes positives et négatives de l'alimentation, qui est passante quand la polarité est mauvaise, créant un court-circuit, détruisant le fusible de protection et, « sauvant » ainsi le reste du montage.



Fabrication

Les diodes sont fabriquées à partir de semiconducteurs et son principe physique de fonctionnement est à la base de tous les composants actifs en électronique.



Bash-Skript

Ich erinnere mich an einen Mathe-Lehrer, der mir, als ich jung war, folgende Maxime erzählte:

"Um faul zu sein, muss man intelligent sein".

Ok, ich wurde faul!!!! ;-)

Es gibt einige manuelle Aufgaben, die nicht leicht zu automatisieren sind (Verzeichnis-Erstellung, Scannen, Gimp-Ausschnitte und Dateierstellung). Der Rest kann automatisiert werden.

Es gibt ein fabelhaftes englisches Tutorial über Bash-Skripting, ABS (Advanced Bash Scripting Guide), und ich fand eine französische Übersetzung.

Sie finden die englische Version unter www.tldp.org.

Dieses Handbuch ermöglichte mir das Schreiben eines kleinen Programmes. Hier ist das Skript:

```
#!/bin/bash

REPERTOIRE=$(pwd)
cd $REPERTOIRE
mkdir ../ima
mv *.png ../ima/
for i in `ls *`
do
  gocr -f UTF8 -i $i -o $i.txt
done
cd ..
mv ima/ $REPERTOIRE
cd $REPERTOIRE
cat *.txt | sed -e 's/_//g' -e 's/(PICTURE)//g' -e 's/î/i/g' \
-e 's/í/i/g' -e 's/F/r/g' -e 's/î/i/g' > test.txt
```

Die Datei wurde ausführbar gemacht und unter root-Berechtigung als ocr-rp nach /usr/local/bin kopiert.

Damit es funktioniert, müssen wir uns in dem Verzeichnis befinden, das verarbeitet werden soll und folgendes eingeben:

```
ocr-rp
```

pwd übergibt den Verzeichnispfad an das Skript, dann wird ima ausserhalb des Verzeichnisses angelegt und alle .png-Dateien dorthin verschoben. Alle Textdateien werden aufgelistet, mit gocr bearbeitet, in test.txt zusammengefasst und zur Anpassung französischer Zeichen bearbeitet.

Und wir machen mit dem gleichen Prozess wie vorher weiter: Kopieren/Einfügen in Mozilla Composer. Die faulste Lösung würde es sein, dass das Skript der Textdatei einige Kopf- und Fusszeilen hinzufügt, es speichert und Mozilla Composer direkt öffnet, aber ich bin zu faul. Das werde ich morgen machen!!!! ;-)

Schlussfolgerung

Dies war nur ein Überblick über Digitalisierungswerkzeuge und es gibt offensichtlich mehr als einen Weg und sicherlich auch bessere. Aber es gibt eine Konstante in der GNU/Linux-Welt: die Hardware-Werkzeuge werden von Jahr zu Jahr besser unterstützt und sind leichter zu benutzen.

Z. B. benutzte ich einen DVD-Brenner zum Speichern meiner 50 MB-Bilder. Die Installation dauerte 10 Minuten und er funktionierte ohne Probleme mit k3b (Ich musste nur "apt-get install dvd+rwtools dvd+rwtools" aufrufen).

Aber mit einem alten PII 350, 192MB RAM, einem billigen Scanner, DVD-Brenner und etwas Plattenplatz haben Sie ein Digitalisierungs-Werkzeug, das gut genug ist, um einem alten Elektronik-Papiermagazin "Unsterblichkeit" zu verleihen. Hier sind die Webseiten der Hilfsmittel, die ich für die Digitalisierung benutzte:

- Scanner ist ein HP ScanJet 4300C
sane, www.sane-project.org
xsane, www.xsane.org
- gimp, www.gimp.org
- gocr, gtk-ocr jocr.sourceforge.net
- ABS findet sich unter www.tldp.org
- DVD-Brenner: NEC 3520
- k3b www.k3b.org

<p><u>Der LinuxFocus Redaktion schreiben</u> © Iznogood "some rights reserved" see linuxfocus.org/license/ http://www.LinuxFocus.org</p>	<p>Autoren und Übersetzer: en --> -- : Iznogood <iznogood@iznogood-factory.org> en --> fr: Iznogood <iznogood@iznogood-factory.org></p>
---	---