

AVMultimedia und ArchivistaBox als PDF-Station

Egg, 25. März 2024: Eine neue Generation der ArchivistaBox erfordert aktualisierte Scanner-Treiber. Ab sofort stellen wir diese auch für AVMultimedia zur Verfügung. So gibt es neu die beiden Tools Simple-Scan und 'scantopdf', um direkt auf dem Desktop PDF-Dateien zu erstellen. Weiter wird die Frage erörtert, warum die JPEG-Komprimierung für die Fujitsu-Scanner auch 15 Jahre nach der Finanzierung durch die Firma Archivista GmbH nicht auf dem Linux-Desktop ankommt. Und, zum Abschluss wird die Thematik behandelt, warum Dokumenten-Scanner plötzlich um den Faktor 4.x teurer sind bzw. was dies für die ArchivistaBox bedeutet.



Scanning, Texterkennung und PDFs neu auf dem Desktop

Fairerweise darf die Frage gestellt werden, warum dies bisher nicht bereits der Fall war. Historisch gesehen war es recht schwierig, mit Scannern unter Linux zu arbeiten. Als die ArchivistaBox 2006 das Licht der Welt erblickte, konnten nur einige wenige Modelle unterstützt werden. Sowohl für die AVision-, Canon- als auch Fujitsu-Geräte musste ein mittlerer fünfstelliger Betrag aufgewendet werden, damit diese liefen.

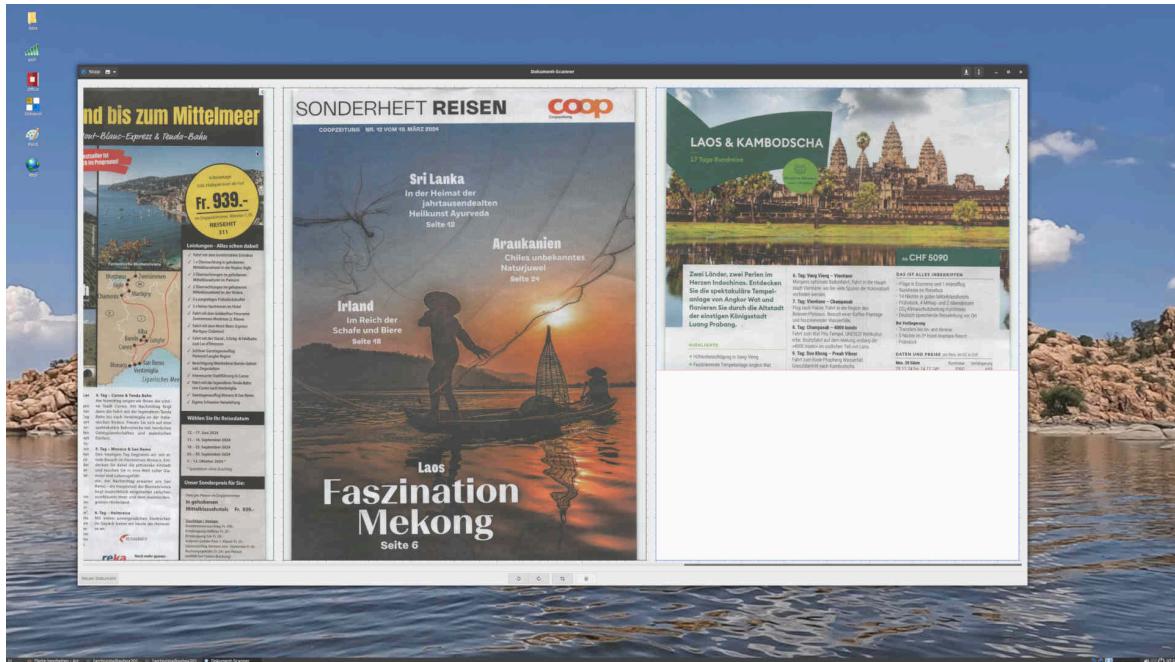
Die durch die Distributionen ausgelieferten Treiber konnten dabei nicht verwendet werden. Schlimmer noch, die Treiber der ArchivistaBox waren inkompatibel zu den Treibern, welche bei Debian/Devuan ausgeliefert wurden. Aus diesem Grund wurde darauf verzichtet, die ArchivistaBox-Treiber bei AVMultimedia auszuliefern.

In diesen Tagen wurden bzw. mussten die Treiber aktualisiert werden. Dabei konnte festgestellt werden, dass die bei sane-project.org ausgelieferten Treiber ohne Anpassungen übernommen werden können. Einzig die JPEG-Komprimierung erfordert noch einen Patch, dazu später mehr.

Dadurch, dass auf der neuen ArchivistaBox die Standard-Treiber zum Einsatz kommen, ist es vom Aufwand her recht einfach, diese auch für den AVMultimedia-Desktop zur Verfügung zu stellen. Ebenso ist die Texterkennung (Tesseract) mittlerweile derart ausgereift, dass auch hier die Standard-Pakete zum Einsatz kommen können.

Simple-Scan und scantopdf als ideale Helferlein

Etwas schwieriger präsentiert sich die Ausgangslage, wenn es darum geht, Tools für das Scannen auszuwählen. Entweder erfordern diese Pakete viel Platz auf dem Desktop oder dann zeigte sich beim Testen, dass die Tools nicht stabil liefen. Simple-Scan ist ein kleines Tool, um recht einfach Dokumente zu scannen und daraus PDF-Dateien zu erstellen. Als einziger Nachteil resultiert, dass Simple-Scan von Haus aus keine durchsuchbaren PDF-Dateien erstellen kann.



Mit dem neu erstellten kleinen Skript 'scantopdf' können Dokumente gescannt und als PDF gespeichert werden, mit dem optionalen Flag ,+' auch in der durchsuchbaren Form. Um z.B. sämtliche Seiten eines Dokumentenscanners zu erfassen, kann das Programm auf der Konsole wie folgt aufgerufen werden:

scantopdf test+

Der Fortschritt wird entsprechend protokolliert:

```
0=>scanimage --batch=/home/archivista/data/test%04d.jpg --
mode 'Color' --format=pnm --source 'ADF Duplex' --resolution
'300' -x 210 -y 297 -l 0 -t 0 --page-height 297 --page-width
210 --buffermode=On --df-action Continue
0=>tesseract -l deu /home/archivista/data/test0001.jpg
/home/archivista/data/test0001 pdf
0=>tesseract -l deu /home/archivista/data/test0002.jpg
/home/archivista/data/test0002 pdf
0=>tesseract -l deu /home/archivista/data/test0003.jpg
/home/archivista/data/test0003 pdf
0=>tesseract -l deu /home/archivista/data/test0004.jpg
/home/archivista/data/test0004 pdf
0=>tesseract -l deu /home/archivista/data/test0005.jpg
/home/archivista/data/test0005 pdf
0=>tesseract -l deu /home/archivista/data/test0006.jpg
/home/archivista/data/test0006 pdf
0=>pdftk /home/archivista/data/test0001.pdf
/home/archivista/data/test0002.pdf
/home/archivista/data/test0003.pdf
/home/archivista/data/test0004.pdf
/home/archivista/data/test0005.pdf
/home/archivista/data/test0006.pdf output
/home/archivista/data/test.pdf
```

Am Ende des Prozesses steht die durchsuchbare PDF-Datei 'test.pdf' zur Verfügung. Um die Optionen von 'scantopdf' zu erhalten, kann das Skript ohne Angabe des Namens aufgerufen werden, nachfolgend die Ausgabe:

```
scantopdf file+[deu|eng] Front|Back|Duplex|Flat
Lineart|Color dpi bright
file+ means create searchable pdf or file+deu with german
language
```

Front+/Back+/Duplex+/Flatbed+[0-7] activate jpeg compression) at /usr/bin/scantopdf line 17.

Soll z.B. eine durchsuchbare PDF-Datei mit italienischer Sprache mit Schwarz/Weiss erstellt werden, kann dies wie folgt erreicht werden:

scantopdf datei+ita Duplex Lineart

Angemerkt sei, dass es neben Color und Lineart die Optionen Gray und Halftone gibt. Dabei entspricht ‚Gray‘ einem Scan in Graustufen sowie Halftone führt zu einer anderen Art eines Scans in Schwarz/Weiss. Ob die Option ‚Lineart‘ oder ‚Halftone‘ zu besseren Ergebnissen beim Scannen führt, kann nicht pauschal festgehalten werden, vielmehr hängt dies von der Vorlage ab und ist entsprechend zu testen.



Wenn Open Source (fast) nicht zum Desktop kommt

Als Anbieter eines Dokumenten-Management-Systems (DMS) wünschen sich viele unserer Kunden, dass bei der Auslieferung der ArchivistaBox gleich auch ein Dokumenten-Scanner mitgeliefert wird. Ein Blick in unser **Web-Museum zeigt, dass zu Beginn unserer Firmengesichte 1998 die AVision-Scanner** zum Zuge kamen, weil diese damals recht günstig waren. Zwar kosteten diese auch vierstellige Beträge, im Vergleich zu den Mitbewerbern, die schnell über 10'000 Franken/Euro kosteten, waren die AVision-Geräte preiswert und solide. Im Jahre **2008 wurde der Wechsel zu Fujitsu** vollzogen. Primär, weil diese Geräte beim Scannen die Bilder im JPEG-Format (d.h. komprimiert) übermitteln konnten und dadurch die Seiten über UBS um Faktoren schneller übermittelt werden konnten. Damit dies unter Linux möglich wurde, investierte die Firma Archivista GmbH einen mittleren fünfstelligen Betrag, um diese Treiber mit Open Source-Lizenz (**SANE-Projekt, siehe sane-project.org**) entwickeln zu lassen. Allerdings ist es bis heute (aus unerklärlichen Gründen) so geblieben, dass diese **JPEG-Komprimierung beim publizierten Code deaktiviert ist.** Die

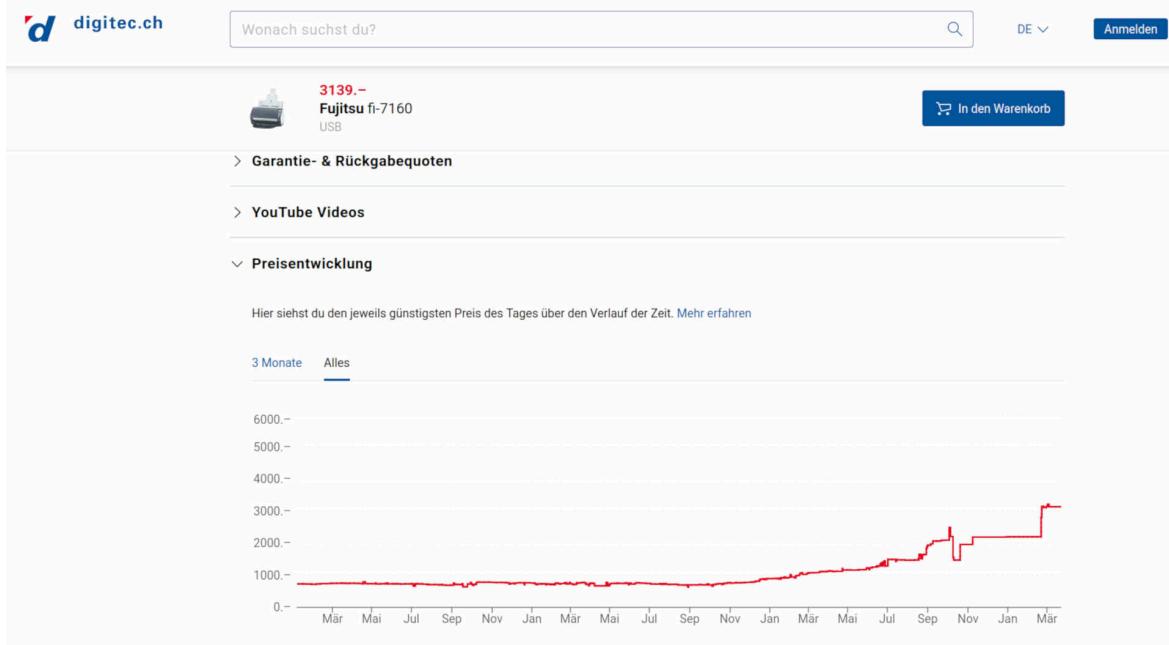
Begründung lautet immer, die JPEG-Komprimierung entspreche nicht den Konventionen des Sane-Standards. Folglich ist es so, dass jener Code, der durch unsere Firma finanziert wurde, bis heute nicht für die Allgemeinheit verfügbar ist.

Die notwendigen **Anpassungen, um die JPEG-Komprimierung einzuschalten, sind minim, siehe hier**, womit auch klar sein dürfte, dass auf der ArchivistaBox bzw. AVMultimedia selbstverständlich die JPEG-Komprimierung aktiviert ist.

Ricoh übernimmt Fujitsu und erhöht Preise um Faktor 4.x

Still und leise hat sich Ricoh die Dokumenten-Scanner-Sparte von Fujitsu

«geschnappt». Dies betrifft alle Geräte der fi- und ix-Serie. Der neue Eigner scheint die ix-Geräte neu als Consumer und die fi-Geräte im Business-Bereich anzupreisen, mit Preissteigerungen bis zum Faktor 4.x. Kostete der fi-7160 vor 2 Jahren noch ca. 750 Franken, so fallen derzeit mehr als 3100 Franken/Euro an (gemäss Homepage digitec).



Opps, da kostet also das gleiche Gerät plötzlich um Faktoren mehr und warum eigentlich wird der Marktleader Fujitsu bei den Dokumenten-Scannern übernommen? Zunächst einmal, auch wenn die Branche immer wieder Prognosen publiziert, der Markt für das Scannen würde weiter anwachsen, so kann dies nicht beobachtet werden. Wurde vor 25 Jahren fast zu 100% gescannt, so darf bei den Kunden der ArchivistaBox festgestellt werden, der Anteil der gescannten Belege liegt deutlich unter 50%, zuweilen auch unter 25%.

Fakt ist und dürfte bleiben, das Scanvolumen nimmt von Jahr zu Jahr ab, im privaten Umfeld sind die Scanner fast komplett verschwunden, eine jede Smartphone-Kamera liefert (mit etwas Optimierung) innert Sekunden genügend gute Resultate. Der Scanner-Markt ist seit vielen Jahren heiß umkämpft. Fujitsu hat bzw. hatte sehr gute Geräte, jedoch waren diese stets ein paar Franken oberhalb der Konkurrenz angesiedelt.

Die Gründe, warum Ricoh 80% der Fujitsu-Tochter (die restlichen 20% verbleiben bei Fujitsu) übernommen hat, sind nicht bekannt. **Öffentlich gemacht wurde jedoch der Übernahmepreis.** Mit ca. 625 Millionen USD erscheint der Preis recht bescheiden, denn wie gesagt Fujitsu war/ist globaler Marktleader. Die Übernahme erfolgte in den Jahren 2022/2023 in gelinde gesagt sehr leisen Tönen.

Hinzu kommt, die Übernahme wird ausdrücklich als Rebranding (eine Art Pinselanstrich) bezeichnet. Zwar findet sich auf den neueren Verpackungen neuerdings der Name Ricoh (wenn auch sehr dezent). Die Geräte selber enthalten jedoch nur die Modellbezeichnung (z.B. ScanSnap) und dies derart



marginal, dass die Geräte fast schon «namenlos» daherkommen. Oder, um es passend zu untenstehendem Bild auszudrücken, ganz sicher klösterlich bescheiden.



Folgen für die ArchivistaBox

Für die ArchivistaBox bzw. die Lauffähigkeit der Scanner im Zusammenspiel mit unseren Scan-Boxen darf angemerkt werden, dass keine Anpassungen notwendig sind. Ob das Gerät noch aus der Zeit von Fujitsu oder neu aus dem Hause Ricoh stammt, die internen Modellnummern bleiben gleich. Aktuell werden die Geräte beim Sane-Projekt nach wie vor unter Fujitsu gelistet und es darf angenommen werden, dass dies für eine gute Weile noch so bleibt.

Etwas schwieriger gestaltet sich die Sachlage bei den fi-Modellen, rein von der neuen Preispolitik her betrachtet. Der (neu) unterstützte IX-1600 kostet keine 400 Franken. Damit können 40 Blätter (80 Seiten) die Minute erfasst werden. Das Modell fi-7160 (bisher am meisten ausgeliefert) erreicht 60 Blätter (120 Seiten) die Minute. Bisher kostete das Gerät ca. 750 Franken, neu wären es ca. 3200 Franken. Diese happy Preisdifferenz von ca. 2500 Franken dürfte in vielen Kundenumgebungen keinen Sinn ergeben.

Generell darf bezweifelt werden, ob die neue Preisstrategie von Ricoh erfolgreich sein wird, denn die Preise der Mitbewerber (z.B. bei 60 Blätter die Minute) liegen deutlich unter 1000 Franken, wobei z.B. der Brother ADS-4900W über ein Display und einen Netzwerkanschluss verfügt, während der Fujitsu bzw. Ricoh fi-7160 einzig über einen USB-Anschluss besitzt. Natürlich reicht dieser, um effizient zu scannen, aber ob Firmen bereit sein werden für gleichartige Geräte weit über 2000 Franken Aufpreis zu zahlen, dies darf bezweifelt werden. Weil nun aber die Fujitsu/Ricoh-Preise derart volatil sind, können wir diese nicht mehr in unserem Shop anbieten. Die gute Ausnahme bildet die **Scan-Box Albis**, welche ab sofort perfekt mit dem IX-1600 zusammenarbeitet und auch entsprechend so ausgeliefert wird. Wird ein anderes Gerät gewünscht, wird die Lieferfähigkeit und der Preis gerne abgeklärt. Überdies besteht für unsere Kunden und Interessenten neu die Möglichkeit die Lauffähigkeit eines gewünschten Scanners neu mit AV Multimedia zu testen. Läuft das Gerät dort, steht einem Einsatz im Zusammenspiel mit der ArchivistaBox nichts im Wege.