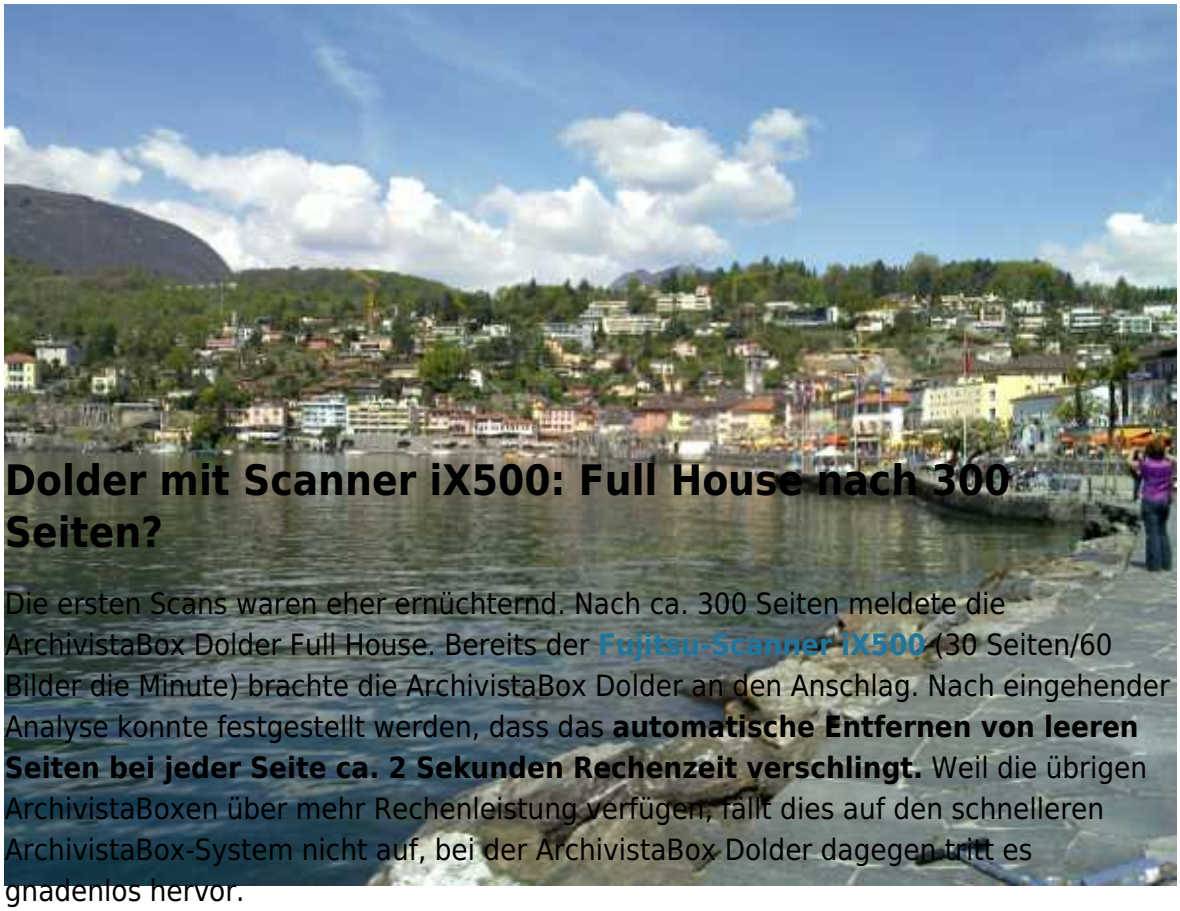
ArchivistaBox Dolder: 10'000 Seiten…

Egg, 5. April 2014: *Nachdem vor einigen Wochen an dieser Stelle die* ***ArchivistaBox Dolder vorgestellt werden konnte,*** *geht es in diesem Blog darum, wieviele Seiten mit der ArchivistaBox Dolder in einer Stunde gescannt werden können. An sich ist die kleinste ArchivisaBox Dolder pro Tag bis zu maximal 2000 Seiten freigegeben, diese Angabe ist bewusst konservativ gehalten. Der Titel dürfte nahelegen, dass die 2000 Seiten erreicht werden konnten, und doch geht es in diesem Blog um mehr als die ArchiistaBox Dolder, vielmehr wird aufgezeigt, welches Potential bei Optimierungen erreicht werden kann.*



Dolder mit Scanner iX500: Full House nach 300 Seiten?

Die ersten Scans waren eher ernüchternd. Nach ca. 300 Seiten meldete die ArchivistaBox Dolder Full House. Bereits der **Fujitsu-Scanner iX500** (30 Seiten/60 Bilder die Minute) brachte die ArchivistaBox Dolder an den Anschlag. Nach eingehender Analyse konnte festgestellt werden, dass das **automatische Entfernen von leeren Seiten bei jeder Seite ca. 2 Sekunden Rechenzeit verschlingt.** Weil die übrigen ArchivistaBoxen über mehr Rechenleistung verfügen, fällt dies auf den schnelleren ArchivistaBox-System nicht auf, bei der ArchivistaBox Dolder dagegen tritt es gnadenlos hervor.

Folglich galt es zu überlegen, wie das Erkennen der leeren Seiten optimiert werden kann. Mit den verschiedensten Bibliotheken wurden ausgiebige Messungen durchgeführt. Die nun **implementierte Lösung arbeitet um den Faktor 4 schneller, indem die Seiten vor dem Test verkleinert werden** (leere Seiten bleiben auch dann leer), sodass noch ca. 0.5 Sekunde pro Seite benötigt werden. So gelang es in einer Stunde ca. 3500 Seiten zu scannen; dies sind immerhin 7 Bundesordner.

Ein schnellerer Scanner muss her…

3500 Seiten pro Stunde sind grundsätzlich nicht schlecht, doch konnte beim Testen (nach der Optimierung) festgestellt werden, dass die beiden Prozessoren (CPUs) beim Scannen nur mässig ausgelastet waren. Aus diesem Grunde musste ein schnellerer Scanner her. Die Wahl fiel dabei auf den **neuen Fujitsu fi-7160,** der zu einem **Preis deutlich unter 1000 Euro immerhin 120 Bilder (60 Seiten) die Minute anliefert. Damit erbringt der fi-7160 eine Leistung, die bislang in dieser Preisklasse unerreicht sind.**

Beim ersten Scannen mit dem fi-7160 vermochte die ArchivistaBox nicht mitzuhalten, mehr als 4500 Seiten die Stunde sollten es nicht werden. Eine weitere vertiefte Analyse brachte zu Tage, dass unsere Bibliothek für das Verarbeiten der Bilder beim Setzen der DPI-Werte (Auflösung im Bild) sich viel Zeit lässt, deutlich über eine halbe Sekunde benötigt sie, um zweimal zwei Bytes zu erfassen. Dies deshalb, weil das gesamte Bild nochmals durchgerechnet wird. Dies ist sinnlos, und wurde daher deaktiviert. Nun konnten 7200 Seiten in einer Stunde realisiert werden.



Schnellste Scanner und weitere Optimierungen

Beim Versuch, mit dem **fi-6670 (80 Seiten bzw. 160 Bilder)** eine noch höhere Leistung zu erzielen, zeigte sich die Problematik, dass die beim A3-Scanner querformatig einzulegenden Seiten gedreht werden müssen, ehe sie in der Datebank abgelegt werden können. Würden die Seiten hochformatig eingelegt, sinkt die Geschwindikeit um ca. 20 Prozent, womit der fi-6670 kaum mehr schneller als der fi-7160 arbeitet.

Die Bibliothek libjpeg (Linux-Standard) arbeitet zwar recht flink. Auf den **schnelleren ArchivistaBoxen benötigt das Rotieren einer JPEG-Seite plus/minus ca. 0,3 Sekunden, bei der ArchivistaBox Dolder wird etwas mehr als 1 Sekunde** benötigt. Dies bedeutet nicht, dass die ArchivistaBox Dolder generell langsamer arbeitet, sondern einzig, dass CPU-intensive Dinge aufgrund der tiefen Leistungsaufnahme (6 Watt pro CPU unter Last) mehr Zeit erfordern, weil ansonsten die CPU stromfressend hochgefahren werden müsste, womit am Ende ein surrender Lüfter weitere Watt verwinden würde…

An dieser Stellte gerne die Messung von jpegtran. Mit diesem Programm können verlustfrei Bilder im JPEG-Format gedreht werden:

```
time jpegtran -rotate 90 job0085.img > job0085.jpg
real 0m1.035s
user 0m0.900s
sys 0m0.132s
```

Nach erheblichen Recherchen konnte die Bibliothek **libjpeg-turbo** gefunden werden. Hier die Messung einer A4-Seite mit 300dpi, um sie um 90 Grad zu drehen:

```
time /opt/libjpeg-turbo/bin/jpegtran -rotate 90 job0085.img >
job0085.jpg
real 0m0.507s
user 0m0.372s
sys 0m0.128s
```

1.035/0.507 ergibt 2.04, **um etwas mehr als den Faktor 2 ist der "Turbo" schneller** — und hat sich damit die Aufnahme in alle ArchivistaBoxen reichlich verdient. Allerdings ergeben 160*0.5 Sekunden auf der ArchivistaBox Dolder in der Minute noch immer 80 Sekunden, womit der Scanvorgang (selbst unter libjpeg-turbo) nicht mit vollem Speed erfolgen kann. Kurz und gut, die Kombination ArchivistaBox Dolder und fi-6670 ergibt kaum einen Sinn, für viel weniger "Moneten" erbringt der fi-7160 fast die gleiche Leistung.

ArchivistaBox Dolder scannt 10'000 Seiten die Stunde

Der neue **Fujitsu fi-7180 passt da gut,** weil die Seiten hochformatig eingelegt werden, das Gerät 80 Seiten bzw. 160 Bilder pro Minute scannt und es deutlich kostengünstiger als der A3-Scanner fi-6670 ist. Alle ArchivistaBox-Systeme enthalten sämtliche Scanner-Treiber bereits auf der Box. Scanner einfach an die Box anhängen und danach den Scanvorgang über das Keypad auslösen, und schon scannt und **verarbeitet die ArchivistaBox Dolder über 160 Seiten die Minute. Dies ergibt plus/minus 10'000 Seiten pro Stunde.**

Zur Feier der Stunde gibt es in diesen Blogs eine Premiere, ein Video, welches das Scannen mit der ArchivistaBox Dolder und dem fi-7180 veranschaulicht:

Das Video hat eine Dauer von 1:30, wobei 30 Sekunden für das Vorstellen der Komponenten und 60 Sekunden gescannt werden. Aufgrund nicht optimaler Lichtverähltnisse und einer sehr mässigen Handy-Kamera resultierte leider keine bessere Qualität, doch zeigt das Video sehr schön auf, mit welchem Speed die ArchivistaBox Dolder doppelseitig gescannte Farbblder (mit 300dpi) ohne jegliche Verzögerung verarbeiten kann.

Optimierungen sind immer sinnvoll — auf allen ArchivistaBoxen

Die hohe Scan-Leistung der ArchivistaBox-Dolder bedeutet im Prinzip, dass nach 10 Stunden (d.h. nach einem Tag, ob der fi-7180 dies am Stück überstehen würde, bleibe dahingestellt) das derzeitige Gesamtvolumen von 100'000 Seiten erschöpft wäre. Ergibt dies einen Sinn? Um es klar zu sagen, **wer täglich 100'000 Seiten verarbeiten möchte, sollte nicht unbedingt zur ArchivistaBox Dolder greifen.** Die Frage kann aber andersherum gestellt werden. Wenn täglich einige Dutzend bis einige Hundert Seiten deutlich schneller verarbeitet werden können, dann
ergibt dies immer einen Sinn.

Dabei bedeutet **Optimierung zunächst einen Mehraufwand.** Richtig umgesetzt resultieren daraus später aber (wie vorliegend) **enorme "Sparpotentiale",** die selbstverständlich nicht nur der ArchivistaBox Dolder zur Verfügung gestellt werden, sondern **all unseren ArchivistaBox-Systemen.** Dank der Optimierung arbeitet die ArchivistaBox Dolder nun beinahe so rank und schnell, wie die übrigen ArchivistaBox-Systeme vor der Optimierung. In diesem Sinne viel Spass beim Arbeiten mit unseren ArchivistaBox-Systemen.

P.S: Dieser Blog könnte den Eindruck erwecken, dass wir die Fujitsu-Scanner "gnadenlos" gut finden. Dieser Eindruck täuscht nicht, weil wir mit diesen Scannern sehr gute Erfahrungen gemacht haben. Neben den Fujitsu-Scannern gibt es mittlerweile (im Bereich um 25 Seiten) viele preisgünstige Duplex-Scanner, die mit der ArchivistaBox laufen (Stichwort SANE). Aber, uneingeschränkt empfehlen können wir nur Geräte, welche wir ausgiebig getestet haben. ***Ohne Vorbehalte und Tests empfehlen können wir dagegen Geräte, welche die Daten per Netzwerk anliefern. Solche Scanner (meist sind es multifunktionale Geräte) sind bereits in der Preisklasse ab 100 Euro erhältlich****. Achten Sie einfach darauf, dass die Daten per SMB (Windows-Ordner) oder FTP und im Format PDF angeliefert werden.*

P.S II: Die ArchivistaBox-Systeme erstellen von sämtlichen Dokumenten ***automatisch einen Volltext-Suchindex und durchsuchbare PDF-Dateien.*** *Bei der ArchivistaBox* ***Dolder können pro Tag 10'000 Seiten in Farbe (300dpi) bzw. ca. 20'000 Seiten in Schwarz/Weiss*** *verarbeitet werden. Bei der ArchivistaBox Matterhon können um den Faktor 10 mehr Dokumente texterkannt werden, wobei mit* ***weiteren Scan-Stationen das Tagesvolumen auf über 1 Mio. Dokumente für die Texterkennung*** *erweitert werden kann.*