ArchivistaBox 2015/X: Texterkennung, durchsuchbare PDF-Dateien und Optimierungen um den Faktor 2.x

Egg, 7. Oktober 2015: *Mit dem Release 2015/X stehen Neuerungen zur Verfügung, welche die Verarbeitungszeiten für viele Jobs um den Faktor 1:2 oder noch höher verbessern. So kann die Verarbeitung in ArchivistaDMS über sämtliche CPU-Kerne nach Bedarf verteilt werden. Dies ergibt neben einer höheren Verarbeitungsgeschwindigkeit beim Einlesen von neuen Dokumenten eine deutlich höhere Geschwindikgeit bei der Texterkennung (OCR). Dabei erstellte PDF-Dokumente können neu direkt in einen externen Windows-Ordner erstellt werden, womit die ArchivistaBox zur vollautomatischen Erstellung von durchsuchbaren PDF-Dateien verwendet werden kann. Selbstragende Archivie können nun auch von den ARM-basierten ArchivistaBox-Systemen erstellt werden, die dazu notwendige ISO-Datei ist noch 80 MByte gross.*

Verarbeitung mit beliebig vielen CPU-Kernen

Seit einigen Jahren werden Computer vermehrt mit mehreren Prozessoren (CPUs) ausgeliefert. Davon profitieren Programme aber nur, wenn sie dafür optimiert wurden. Bei der ArchivistaBox war dies bislang einzig bei der Texterkennung der Fall, und auch da nur, wenn hintereinander viele Dokumente zur Bearbeitung anstanden. Mit dem aktuellen **Release 2015/X werden die Dokumente über die vorhandenen**

Prozessoren parallel verarbeitet. Bei acht Prozessoren kann z.B. ein 200 seitiges Dokument in einem Achtel der bisherigen Zeit verarbeitet werden, wenn sämtliche Prozessoren gleichzeitig angeworfen werden. Klingt einigermassen banal, ist es aber nicht. Denn wenn die gesamte Rechenzeit für einen Job freigegeben wird, kann an anderer Stelle ein Engpass entstehen.

Nun überwacht das Betriebssystem die laufenden Programme dahingehend, dass nicht ein Job sämtliche Ressourcen erhält. Vielmehr wird die verfügbare Kapazität geteilt. Trotzdem ist es nicht zwingend eine gute Idee, gleichzeitig zuviele Programme zu starten. Kleines Beispiel: Wenn gleichzeitig 1000 Jobs für die Texterkennung gestartet werden, werden sämtliche Dokumente zwar alle gleichzeitig abgearbeitet, dies freilich um den Preis, das kein Job bevorzugt beendet werden kann. Im dümmsten Fall steht für die 1000 Jobs zuwenig Speicher (RAM) zur Verfügung, womit zumindest ein Teil der Dokumente in der Vearbeitung "hängen" bleibt.

Als Grundregel gilt: Anzahl der verfügbaren CPU-Kerne gleich Anzahl der gleichzeitig laufenden Programme. Damit ist sichergestellt, dass die Jobs mit hoher Priorität abgearbeitet werden können. Genau hier setzt die neue **Version 2015/X an. Je nach Einsatzzweck können die CPU-Kerne pro Kunde individualisiert eingesetzt werden.** Beispiel A: Es greifen viele Benutzer gleichzeitig auf das Archiv zu, das Volumen der neu zu erfassenden Dokumente ist eher niedrig. Lösung A: Für die Verarbeitung werden nur 1 oder 2 CPU-Kerne reserviert. Beispiel B: Es gibt sehr viele gescannte Dokumente, die möglichst schnell in durchsuchbare PDF-Dateien konvertiert werden müssen. Lösung B: Für die Verarbeitung werden sämtliche CPU-Kerne freigegeben. Die Zugriffsgeschwindigkeit auf das Archiv leidet darunter ein wenig, dafür werden die Dokumente um den Faktox X schneller abgearbeitet.

Zum Abschluss ein zwei Messungen aus der Praxis: Das Einlesen des deutschen und englischen ArchivistaBox-Handbuches (PDF-Dateien mit 205 bzw. 204 Seiten) inklusive der Texterkennung mit Tesseract benötigt neu auf der ArchivistaBox Matterhorn rund 7 Minuten 30 Sekunden. Dies ergibt eine Leistung von knapp 1 Seite pro Sekunde, bzw.

eine Tagesleistung von ca. 80'000 Seiten. Geht es darum, bereits digital vorliegende Dokumente zu verarbeiten, so können die **409 Seiten in ca. 50 Sekunden abgearbeitet werden; dies entspricht einer Tagesleistung von über 700'000 Seiten.**

Im Vergleich dazu benötigt der 'alte' Code ca. 19 Minuten für den Job mit der Texterkennung und ca. 1 Minute 50 Sekunden für das reine Importieren der Handbücher. Dies bedeutet, dass mit dem aktuellen neuen Code eine Optimierung zwischen dem Faktor 2,2 und 2,5 erreicht werden kann. Selbstverständlich kann diese Leistung durch den Einsatz von schnelleren CPUs um den Faktor vier bis sechs weiter erhöht werden, ebenso könnte mit einem Cluster die Leistung fast beliebig nach oben skaliert werden. **Unter dem Strich bleibt jedoch, dass mit dem aktuellen Code die Hardware weniger als halb so schnell sein muss, um die gleiche Leistung zu erzielen.**

Erstellen von durchsuchbaren PDF-Dokumenten

Wie obenstehend ausgeführt wurde, werden mit den aktuellen ArchivistaBox-Systemen insbesondere bei der Texterkennung sehr gute Werte erreicht. Die ArchivistaBox eignet sich daher nicht nur als DMS-System, sondern auch als "Durchlauferhitzer" zum Erstellen von durchsuchbaren PDF-Dateien. Bislang mussten die erstellten Dateien mit einem Skript oder über das Application Programming Interface (API) weiterverarbeitet werden. Neu können die erstellten durchsuchbaren PDF-Dateien jederzeit direkt in ein anderes Netzlaufwerk kopiert werden.

Die dazu notwendigen **Einstellungen können in WebAdmin,** dort unter **'OCR-Definitionen'** sowie **'Optionen Texterkennung (OCR)',** festgelegt werden:

Einmal aktiviert werden nach der Texterkennung die generierten PDF-Dateien direkt in den zuvor in WebAdmin festgelegten Freigabepfad gespeichert. Sofern kein Netzwerklaufwerk zur Verfügung steht, werden die generierten PDF-Dateien in die archivista-Freigabe im Verzeichnis 'temp' gespeichert.

Selbsttragende Archive für Alle

Neu können selbsttragende Archive auf allen ArchivistaBoxen (bisher Intel/AMD) erstellt werden. Sowohl auf den ARM- als auch den Intel/AMD basierten Modellen steht neu eine 'kleine' 80 MByte grosse ISO-Datei zur Verfügung, mit der selbsttragende Archive erstellt werden können. Diese ISO-Datei 'archivista_cd1.iso' ist bei den ARM-basierten Modellen in den ftp/smb-Ordner 'temp' zu legen. Bei den Intel/AMD-Boxen kann Sie über den Home-Button von ArchivistaVM in den Ordner /var/lib/vz/template/iso hochgeladen werden. Die Datei findet sich im Download-Ordner unter dem Namen 'selfrun.zip'; das Passwort für das Entpacken bleibt für sämtliche OS-Dateien das gleiche.

Weiter können in WebAdmin Optionen festgelegt werden, um die Archive platzoptimiert auszulagern. Es können sowohl die Bilddateien 'geschrumpft' werden, ebenso ist eine höhere Komprimierung bei JPEG-Bilder möglich und weiter können die Quell- bzw. durchsuchbaren PDF-Dateien vom Export ausgeschlossen werden. Dadurch können Archivista-Archive bis zu 50 GByte (bis in den Millionenbereich an Seiten) in eine ISO-Datei geschrieben werden und bequem auf einem Intel/AMD-Rechner hochgefahren werden. Die selbsttragenden Archive laufen dabei komplett im Hauptspeicher (RAM), wobei gilt: Grösse der ISO-Datei plus minimal 600 MByte RAM werden für das Arbeiten benötigt.

Bereit für den produktiven Einsatz

Die **ArchivistaBox 2015/X steht ab sofort zum produktiven Einsatz bereit,** Kunden mit gültigem Wartungsvertrag können per Mail/Telefon jederzeit eine aktualisierte Version beauftragen, das Update lässt sich anschliessend bequem über WebConfig einspielen. Dank den **neuen ARM-basierten ArchivistaBoxen, die acht Kerne im Grundumfang beinhalten, steht z.B. mit der ArchivistaBox Dolder eine Lösung bereit, mit der ab sFr. 360.– beliebig oft und unlimitiert pro Tag mehrere zehntausend Seiten durchsuchbare PDF-Dateien erstellt werden können.**