OCR-Erkennung unter Linux, Part III

Kommerzielle und freie Texterkennung anhand ArchivistaBox 2014/XI

Contents

1 Einleitung

1.1 OCR-Erkennung unter Linux

Lange ist es her, seit es in den Jahren 2007 und 2008 zwei Vorträge zum Theme Texterkennung unter Linux seitens der ArchivistaBox gab.

In nachfolgendem Papier sei der aktuelle Stand quelloffener wie kommerzieller Texterkennung unter Linux im Jahre 2014 kurz zusammengefasst. Dabei werden neben einer kurzen Übersicht über die derzeitigen Projekte die drei aktuellen OCR-Engines der ArchivistaBox mit Stand 2014/XI miteinander vergleichen. Damit ein Vergleich zu den Jahren 2007 und 2008 möglich ist, werden die gleichen Test-Dateien wie bei den beiden ersten Tests verwendet.

2 Rückblick über die letzten 7 Jahre

2.1 Kommerzielle Texterkennung

Im Bereich der kommerziellen Texterkennung hat sich unter Linux wenig getan. Nach wie vor gibt es keine Desktop-Programme für Linux seitens der kommerziellen Anbieter.

Weder OmniPage, Abbyy und/oder ReadIris (um die wichigsten drei Player zu nennen) sind als Desktop-Applikationen für Linux erhältlich. Dagegen haben alle drei Anbieter SDKs (bei Abbyy heisst es – vereinfacht gesagt – CLI) nach Linux portiert.

Die Preise sind z.T. öffentlich, z.T. auch nicht. Derzeit (Stand Oktober) kostet z.B. ABBYY FineReader CLI 11 für jährlich 500'000 Seiten ca. 3500 EURO, bei 12'000 Seiten sind es ca. 150 Euro. Gemessen an der Anzahl Seiten sind dies mittlerweile moderatere Preise; allerdings stellt sich doch die Frage, wie flexibel eine Texterkennung daherkommt, wenn z.B. nach jährlich 12'000 Seiten Schluss ist bzw. ein Upgrade (das gut und gerne Kosten im vierstelligen Bereich verursacht) für weitere Seiten fällig wird.

2.2 Quelloffene Texterkennung

Über viele Jahre gab es keine vernünftigen Programme, um unter Linux mit Open Source Texterkennung (OCR) durchzuführen. Dies änderte sich in den Jahren 2006 bis 2008, weil gleichzeitig Tesseract (Google-Projekt) und Cuneiform erschienen.

Im Jahre 2007 setzte die Firma Archivista GmbH auf Cuneiform und liess bei ExactCode das Kommandozeilenprogramm hocr2pdf (welches aus Cuneiform PDF-Dateien erstellen kann) programmieren. Das Hilfsprogramm hocr2pdf wird heute in vielen verschiedenen Applikationen zum Erzeugen von durchsuchbaren PDF-Dateien verwendet, z.B. von pdfsandwitch, es hat aber seit einigen Jahren kaum mehr nennenswerte Änderungen erfahren.

An sich wäre es die Idee gewesen, dass der Verein freearchives.ch durch die entsprechenden Mitgliederbeitäge einen jährlichen Etat zusammentragen kann, um die Entwicklung von hocr2pdf bzw. cuneiform voranzutreiben. Leider konnten aber zu wenige Personen für eine Mitgliedschaft begeistert werden, sodass nicht genügend finanzielle Mittel zusammenkamen, um mehr als kleine Verbesserungen in Angriff zu nehmen.

Nebenbemerkung: Im Verein freearchives.ch sind mittlerweile nur noch die Gründungsmitglieder vertreten, und es wird sich in naher Zukunft weisen, od ein erneuter Anlauf für den Verein genommen wird, oder ob er aufgelöst wird.

Neben den fehlenden finanziellen Mitteln (weder in hocr2pdf noch cuneiform wollte die Community Zeit oder Geld investieren) bereitete auch das darunterliegendeb Cuneiform Probleme. Da die Zeichensätze alles andere als quelloffen vorliegen, konnte bis heute niemand das Format der Zeichensätze (wäre eine Grundbedingung, um neue Zeichen/Muster zu trainieren) entschlüsseln. Zudem stürzt Cuneiform dann und wann ganz einfach ab, der Code selber ist eher schlecht dokumentiert (Kommentare alle in Russisch) und der Versuch, Cuneiform auf eine andere Plattform (Stichwort ARM) zu portieren, scheiterte bisher ebenfalls. Kurz und gut, die gesmate Entwicklung rund um Cuneiform für Linux ist seit ca. drei Jahren 'eingeschlafen'.

Nebenbemerkung: Für die ArchivistaBox bedeutet dies, dass Cuneiform ab der Version 2014/XI nicht mehr enthalten ist. Keine Bange, durchsuchbare PDF-Dateien können nach wie vor mit (einer anderen) Open Source Texterkennung (siehe unten) erstellt werden.

Besser ergangen ist es dagegen Tesseract und (in bescheidenem Masse) Ocrad. Vorallem Tesseract erlebte mit der Version 3.x einen wahren Entwicklungsschub. Beide Programme lassen sich problemlos auf der ARM-Plattform übersetzen bzw. können dort auch produktiv eingesetzt werden.

Nachfolgend wird deshalb nur noch auf die kommerzielle Texterkennung der ArchivistaBox, Tesseract 3.04 und Ocrad (aktuelle Version 0.25pre2) eingegangen. Erstere Engine ist nicht Open Source, die beiden letzteren sind es; alle drei Produkte finden sich auf jeder ArchivistaBox (siehe z.B. ArchivistaCloud). Bevor die Demo-Dokumente aus dem Jahre 2007 erneut begutachtet werden sollen, einige allgemeine Hinweise zu den Neuerungen in Tesseract und Ocrad.

2.3 Tesseract 3.04

Tesseract wird laufend weiterentwickelt. Mit der Version 3.x wurde eine Layout-Erkennung für die Seiten eingeführt, seit der Version 3.03 kann Tesseract ebenfalls durchsuchbare PDF-Dateien erstellen. Ab Version 3.x können mehrere Sprachen (z.B. deu+fra+eng) für die Erkennung herangezogen werden. Kurz und gut, Tesseract 3.x bietet heute ziemlich alles, was ein OCR-Herz erfreuen lässt.

Hinweis: Fast alle Distributionen besitzen leider nicht aktuelle Tesseract-Versionen. War es früher eher schwierig, Tesseract zu kompilieren, so ist dies heute einfach(er) geworden. Im Prinzip wird zunächst die Bibliothek leptonica benötigt, danach kann Tesseract mit 'autogen.sh' und 'configure –prefix=/usr' bzw. 'make' und 'make install' übersetzt und installiert werden.

Wichtig zu wissen ist weiter, dass nach dem Erstellen der Binärdateien die Sprachenpakete zusätzlich nach **/usr/share/tessdata** zu kopieren sind. Der Aufruf von Tesseract ist einfach:

```
tesseract -l deu+eng eins.jpg eins pdf
```

Mit diesem Befehl wird aus der Datei eins.jpg eine Textdatei eins.txt (mit dem erkannten Text), und eine durchsuchbare PDF-Datei mit dem Namen eins.pdf erstellt. Sofern keine PDF-Datei gewünscht ist, kann der letzte Parameter 'pdf' ganz einfach weggelassen werden.

Tesseract benötigt relativ viel Speicherplatz auf der Platte, so benötigen nur schon die gängigen Sprachenpakete eng, deu, fra, ita sowie spa ca. 120 MByte an Speicher auf der Festplatte. Auch ist Tesseract nicht unbedingt sehr schnell, eine Seite benötigt auf einem üblichen 64-Bit-Rechner ca. 5 bis 10 Sekunden (bei komplexem Seitenlayout auch mal 20 bis 30 Sekunden). Auf einer ARM-CPU ist Tesseract äussert gemächlich, wird dort pro Seite ca. 1 Minute benötigt.

2.4 Ocrad 0.25

Ocrad kann in Punkto Qualität nicht mit Tesseract mithalten, dies gleich vorne weg. Allerdings ist Ocrad extrem klein (ca. 300 KByte) und bei einigermassen gut gedruckten Schriften ist die Erkennung durchaus brauchbar. Die Verarbeitungsgeschwindigkeit ist hoch, auf einer 64-Bit-CPU erfolgt die Erkennung meist unter 1 Sekunde, auf einer ARM-CPU sind es einige wenige Sekunden.

3 Demo-Seiten für OCR-Tests

3.1 Text, 1spaltig, 12 Punkt Helvetica

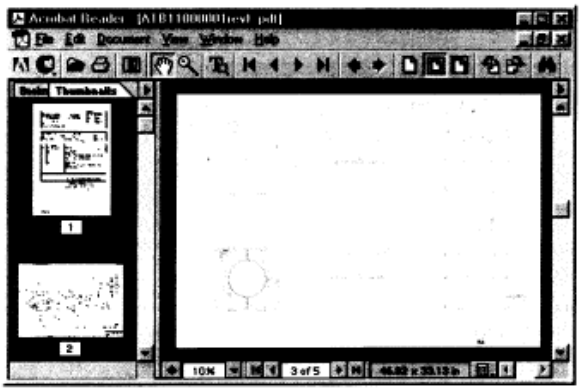
Fall 2: Engineering-Projekt mit PDF-Dokumentation

Kaum eine andere Technologie hat im letzten Jahr für mehr Furore gesorgt als das PDF-Format. PDF steht für Portable Document Format und wie der Name besagt, geht es darum, Inhalte digital so aufzubereiten, dass diese **auf unterschiedlichen Rechnern** (z.B. Mac und Windows) betrachtet und ausgedruckt werden können.

PDF-Dokumente sind unheimlich **flexibel.** Alles, was gedruckt werden kann, ist auch als PDF-Datei (in digitaler Form) publizierbar. Zudem sind PDF-Dateien mittlerweile **weit verbreitet;** der Viewer (Betrachter) für die Dateien ist kostenlos. Die Möglichkeiten dieses Formats machte sich eine internationale, im Bereich Engineering tätige Firma zunutze.

Problemstellung und Lösung

Ca. 75'000 auf der Vorder- und Rückseite bedruckte A4-Seiten sowie 200 zum Teil farbige Seiten im A3- bis A0-Format sollten gescannt und in ein **PDF-Format** überführt werden. Ziel waren selbsttragende CDs mit PDF-Dateien, welche die ursprüngliche Dokumentennummer aufweisen sollten und über die mit sogenannten Thumbnails ein Überblick besteht.



Archivista löste dieses Problem Schritt für Schritt:

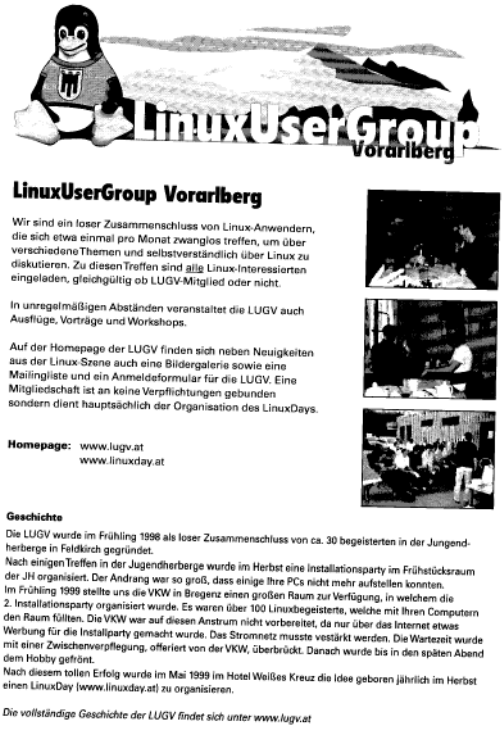
- **Duplex-Scanning** der 75'000 A4-Seiten mit Hochleistungsscanner mit 300dpi, Resultat: 150'000 Seiten im **TiffG4**-Format
- Individuelles Scanning der 200 **grösserformatigen** Belege, davon 50 **in Farbe**
- **Umwandlung** der TiffG4 und JPG-Dateien in **PDF**-Dateien
- Erstellen der **CDs mit Inhaltsverzeichnissen** (inkl. Link auf entsprechende Datei)

Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und dank des selber entwickelten **Produktmoduls Archivista TifToPDF** in der Lage, diese Wertschöpfungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendbaren PDF-File **kostengünstig** anzubieten. Das Engineering-Unternehmen verfügt heute über eine saubere, platzsparende Dokumentation des Projektes.

Scan-Dienstleistungen – Scanning und was dazu gehört, Fallbeispiele, Seite 2
Archivista GmbH, 8042 Zürich, Tel. 01 350 46 74, Fax 01 350 46 72, www.archivista.ch

Bei diesem Text wurde versucht, einen möglichst einfachen und gut gedruckten Text zu verwenden. Es gibt weder Spalten noch ist die Schriftgrösse klein. Das Beispiel entspricht in etwa dem, was bei normaler Geschäftskorrespendonz erwartet werden darf.

3.2 Flyer, 1spaltig, ca. 8 Punkt, Helvetica



LinuxUserGroup Vorarlberg

LinuxUserGroup Vorarlberg

Wir sind ein loser Zusammenschluss von Linux-Anwendern, die sich etwa einmal pro Monat zwanglos treffen, um über verschiedene Themen und selbstverständlich über Linux zu diskutieren. Zu diesen Treffen sind alle Linux-Interessierten eingeladen, gleichgültig ob LUGV-Mitglied oder nicht.

In unregelmäßigen Abständen veranstaltet die LUGV auch Ausflüge, Vorträge und Workshops.

Auf der Homepage der LUGV finden sich neben Neuigkeiten aus der Linux-Szene auch eine Bildergalerie sowie eine Mailingliste und ein Anmeldeformular für die LUGV. Eine Mitgliedschaft ist an keine Verpflichtungen gebunden sondern dient hauptsächlich der Organisation des LinuxDays.

Homepage: www.lugv.at
www.linuxday.at

Geschichte

Die LUGV wurde im Frühling 1998 als loser Zusammenschluss von ca. 30 begeisterten in der Jungendherberge in Feldkirch gegründet.
Nach einigen Treffen in der Jugendherberge wurde im Herbst eine Installationsparty im Frühstücksraum der JH organisiert. Der Andrang war so groß, dass einige Ihre PCs nicht mehr aufstellen konnten.
Im Frühling 1999 stellte uns die VKW in Bregenz einen großen Raum zur Verfügung, in welchem die 2. Installationsparty organisiert wurde. Es waren über 100 Linuxbegeisterte, welche mit Ihren Computern den Raum füllten. Die VKW war auf diesen Anstrum nicht vorbereitet, da nur über das Internet etwas Werbung für die Installparty gemacht wurde. Das Stromnetz musste verstärkt werden. Die Wartezeit wurde mit einer Zwischenverpflegung, offeriert von der VKW, überbrückt. Danach wurde bis in den späten Abend dem Hobby gefrönt.
Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel Weißes Kreuz die Idee geboren jährlich im Herbst einen LinuxDay (www.linuxday.at) zu organisieren.

Die vollständige Geschichte der LUGV findet sich unter www.lugv.at

Bei diesem Flyer ist die Text relativ klein gedruckt. Klein gedruckte Text stellen erhöhte Anforderungen an die Texterkennung. Der Aufbau der Seite ist nicht mehr ganz trivial, der Text selber ist aber noch immer einspaltig gesetzt.

3.3 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerkkabels. Jede Box ist fixfertig vorkonfiguriert; eine Installation ist nicht notwendig. Ob per FTP-Filetransfer (Kopiergeräte), angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich äusserst kontaktfreudig bei der Dokumentenannahme.

Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlagwortet (indexiert) und stehen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web- oder Rich-Client. Dokumente können aber auch als PDF oder per Mail weitergereicht werden. Ganz nach dem Motto: Connect your world.

ARCHIVISTA

Archivista GmbH
Postfach - CH-8042 Zürich
Tel: +41 (0)44 254 54 00
Fax: +41 (0)44 254 54 02
Web: www.archivista.ch
E-Mail: webmaster@archivista.ch

Etwas exotischere Schriften bereiten bei der Texterkennung oft Schwierigkeiten. Ansonsten ist der Textaufbau extrem einfach gehalten.

3.4 Prospekt, Seite, 10 Punkt, Eras

Für jeden Bedarf die Richtige

Sie heissen Rigi, Pilatus, Titlis, Eiger, Mythen und Rothorn – und sind punkto Unerschütterlichkeit und Stabilität genauso robust wie die gleichnamigen geografischen Erhebungen. Die Performance der Document-Server lässt sich von der relativen Höhe, die mit dem Bergnamen assoziiert ist, ableiten.

Rigi eignet sich primär für kleinere Umgebungen wie z.B. Rechtsanwaltspraxen oder PR-Büros. Aber auch grössere Unternehmen, die für Abteilungen PDF-Dokumentenserver suchen, sind mit der Rigi-Box gut bedient.

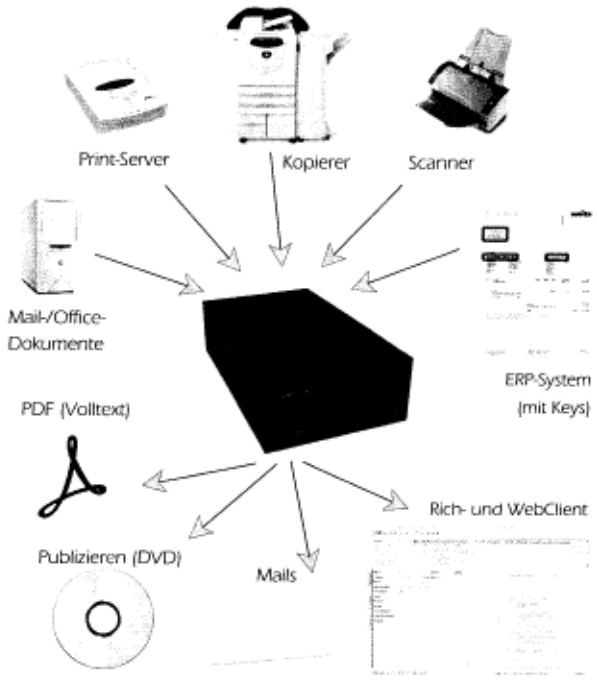
Pilatus ist für mittlere Firmengrössen, oder besser ausgedrückt, für das mittlere Datenvolumen gedacht. Die Titlis-Box ebenfalls, allerdings ergibt die redundante Hardware ein erhöhtes Sicherheitselement. Die Eiger-Box – mit entsprechender Festplatte, zweiter Box und Tape-Laufwerk – ist für Archive ab ca. 500'000 bis einige Millionen Seiten ausgelegt.

Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Titlis und Eiger gescannt werden kann. Ebenfalls führen sie z.B. eine Text- und/oder Barcode-Erkennung durch.

Und falls Sie nun einen Dokumenten-Cluster à la Mount Everest benötigen, auch kein Problem; wir stellen Ihnen diesen mit der entsprechenden Hardware gerne individuell zusammen.

Rigi	1797 m.ü.M.	Einzelplatz-Dokumenten-Server mit bis zu 20'000 Akten und 100'000 Seiten
Pilatus	2132 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten
Titlis	3238 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten, redundant (2 Boxen)
Eiger	3790 m.ü.M.	Dokumenten-Server für unlimitierte Anzahl Akten, bis ca. 2 Mio Seiten, redundant (2 Boxen) und mit Backup-Tape-Drive
Mythen	1899 m.ü.M.	Scan- und OCR-Box, welche Daten zum Pilatus und Titlis transportiert
Rothorn	2351 m.ü.M.	Scan- und OCR-Box, passend zum Eiger

ArchivistaBox: Connect your world



Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerkkabels. Jede Box ist fixfertig vorkonfiguriert; eine Installation ist nicht notwendig. Ob per FTP-Filetransfer (Kopiergeräte), angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich äusserst kontaktfreudig bei der Dokumentenannahme.

Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlagwortet (indexiert) und stehen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web- oder Rich-Client. Dokumente können aber auch als PDF oder per Mail weitergereicht werden. Ganz nach dem Motto: Connect your world.

ARCHIVISTA

Archivista GmbH
Postfach - CH-8042 Zürich
Tel: +41 (0)44 254 54 00
Fax: +41 (0)44 254 54 02
Web: www.archivista.ch
E-Mail: webmaster@archivista.ch

Die gesamte Seite des Prospektes enthält ein relativ komplexes Seitenlayout. Grundsätzlich dreispaltig gesetzt, mit einem einspaltig über zwei Seiten gesetzten Text in der Mitte sowie einem kleineren Text (siehe Ausschnitt) in der unteren Hälfte.

4 Kommerzielle Texterkennung

Informationen zur kommerziellen OCR-Erkennung der ArchivistaBox finden sich im Skript aus dem Jahre 2007. Nachfolgend die Ergebnisse dieser Engine.

4.1 Text, 1spaltig, 12 Punkt Helvetica

```
Fall 2: Engineering-Projekt mit PDF-Dokumentation
Kaum eine andere Technologie hat im letzten Jahr für mehr Furore gesorgt als das
PDF-Dokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auc
Problemstellung und Lösung
Ca. 75'000 auf der Vorder- und Rückseite bedruckte A4-Seiten sowie 200 zum Teil
S Ariob.it lUMcJer \!

fi^^^^j^J^M^Sm^^wm^B
Archivista löste dieses Problem Schritt für Schritt:
? Duplex-Scanning der 75'000 A4-Seiten mit Hochleistungsscanner mit 300dpi, Resu
? Individuelles Scanning der 200 grösserformatigen Belege, davon 50 in Farbe
? Umwandlung der TiffG4 und JPG-Dateien in PDF-Dateien
? Erstellen der CDs mit Inhaltsverzeichnissen (inkl. Link auf entsprechende Date
Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und d
Scan-Dienstleistungen - Scanning und was dazu gehört, Fallbeispiele, Seite 2 Arc
Jt
```

4.2 Flyer, 1spaltig, ca. 8 Punkt, Helvetica

```
iäcKsKiiüS;
Mii^Mk
LinuxUserGroup Vorarlberg
Wir sind ein loser Zusammenschluss von Linux-Anwendern, die sich etwa einmal pro
In unregelmäßigen Abständen veranstaltet die LUGV auch Ausflüge, Vorträge und Wo
Auf der Homepage der LUGV finden sich neben Neuigkeiten aus der Linux-Szene auch
Homepage: www.lugv.at
www.linuxday.at
Geschichte
Die LUGV wurde im Frühling 1998 als loser Zusammenschluss von ca. 30 begeisterte
Nach einigen Treffen in der Jugendherberge wurde im Herbst eine Installationspar
Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel Weißes Kreuz die Idee gebor
```

Die vollständige Geschichte der LUGV findet sich unter www.lugv.at

4.3 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig
Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlag
f Archivista GmbH
*t Postfach – CH–8042 Zürich –?s ’ Tel: +41 (0)44 254 54 00 Fax: +41 (0)44 254 5
Web: www.archivista.ch ARCHIVISTA E–Mail: webmaster@archivista.ch

4.4 Prospekt, Seite, 10 Punkt, Eras

Für jeden Bedarf die Richtige
Sie heissen Rigi, Pilatus, Titlis, Eiger, Mythen und Rothorn – und sind punkto U
Rigi eignet sich primär für kleinere Umgebungen wie z.B. Rechtsanwaltspraxen ode
Pilatus ist für mittlere Firmengrössen, oder besser ausgedrückt, für das mittler
platte, zweiter Box und Tape– Laufwerk – ist für Archive ab ca. SOO’OOO bis eini
Mythen und Rothorn sind Scan– und OCR–Cluster–Stationen, mit welchen auf die Box
führen sie z.B. eine Text– und/oder Barcode–Erkennung durch.
Und falls Sie nun einen Dokumenten– Cluster à la Mount Everest benötigen, auch k
Rigi 1797 m.ü.M. Einzelplatz–Dokumenten–Server mit bis zu 20’OOu Akten und 10O’O
Pilatus 2132 m.ü.M. Dokumenten–Server für bis zu 200’000 Akten und 1 Mio Seiten
Titlis 3238 m.ü.M. Dokumenten–Server für bis zu 200’uOO Akten und l Mio Seiten,
Eiger 3790 m.ü.M. Dokumenten–Server für uniimitierte Anzahl Akten, bis ca. 2 Mio
Mythen 1899 m.ü.M. Scan–und OCR–Box, welche Daten zum Pilatus und Titlis transpo
Rothorn 2351 m.ü.M. Scan–und OCR–Box, passend zum Eiger
ArchivistaBox: Connect your world
MailVOffice– Dokumente
ERP–System mit Keys)
Rieh– und WebClient
Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig
Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlag
Archivista GmbH
Postfach – CH–8042 Zürich Tel: +41 (0)44 254 54 00 Fax: +41 (0)44 254 54 02
Web: www.archivista.ch E–Mail: webmaster@archivista.ch
ARCHIVISTA

5 Ocrad 0.25)

Ocrad liefert bei gut gescannten Belegen durchaus anständige Resultate, bei kleinen Schriften oder schlechter Qualität bei den Scans sind die Ergebnisse mässig bis eher unbrauchbar.

5.1 Text, 1spaltig, 12 Punkt Helvetica

Fall z: Engineering-Projekt mit PDF-Dokumentation

Kaum eine andere Technologie hat im letzten Jahr für mehr Furore gesorgt als das Format. PDF steht für Portable Document Format und wie der Name besagt, geht es darum, lnhalte digital so aufzubereiten, dass diese auf unterschiedlichen Rechne Mac und Windows) betrachtet und ausgedruckt werden können.

PDF-Dokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auc PDF-Datei (in digitaler Form) publizierbar. Zudem sind PDF-Dateien mittlerweile breitet; der Viewer (Betrachter) für die Dateien ist kostenlos. Die Möglichkeite Formats machte sich eine internationale, im Bereich Engineering tätige Firma zun

Problemstellung und Lösung

Ca. 75'000 auf der Vorder- und Rückseite bedruckte A4-Seiten sowie 200 zum Teil Seiten im A3- bis AO-Format sollten gescannt und in ein PDF-Format überführt wer waren selbsttragende CDs mit PDF-Dateien, welche die ursprüngliche Dokumentennum aufweisen sollten und über die mit sogenannten Thumbnails ein Überblick besteht.

_ _

&_r__ _ . '__ ' ' '_ __'_ _ ___ , _,__,_, _@p ___ _ ..'_'.5 _ _ , . '_____>
. , . ,, , , , ,_ _, ____ __ _____ _ __rf , __ _ t ____, _.
***__e__ _ '_ ' , . _ _ _ '_'_ __'_,__ , ___ ,_,_,, _,_,_____ _. _!,,_.,.__, _/_**
r|_ __ _ , l_r,t"' _ ' __ _ ___.___' _ ' _ ,__ _ _ _
5
_ _ _ _ , _7 _ _ __. " __ _ _ y,_| . _ _, __ _ . . , |
_'_ _ ' _
' __ .

_..._ rçl __!, °

%_ ._

_ _

. . ! _._
..__:_._-._..._:_.._ _: .i : __ ?
. . Z . . - . ! __ ’
_i i_"_ -_ __-_ _,‘’__’_ ’____, _ __ ION a _ _ 3 ofS _ `_ _’

.’ . _ _

__ _

Archivista löste dieses Problem Schritt für Schritt:

. Duplex-Scanning der 75’000 A4-Seiten mit Hochleistungsscanner mit 300dpi,
Resultat: | 50’000 Seiten im TiffG4Format
. lndividuelles Scanning der 200 grösserformatigen Belege, davon 50 in Farbe
. Umwandlung der TiffG4 und JPG-Dateien in PDF-Dateien
. Erstellen der CDs mit lnhaltsverzeichnissen (inkl. Link auf entsprechende Date

Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und d
selber entwickelten Produktmoduls Archivista TifroPDF in der Lage, diese Wertsch
fungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendba
PDF-File kostengünstig anzubieten. Das Engineering-Unternehmen verfügt heute übe
saubere, platzsparende Dokumentation des Projektes.

Scan-Dienstleistungen - Scanning und was dazu gehört, Fallbeispiele, Seite 2
Archivista GmbH, 8042 Zürich, Tel. 01 350 46 74, Fax 01 350 46 72, www.archivist

’_!
_m_n

5.2 Flyer, 1spaltig, ca. 8 Punkt, Helvetica

_ , , a_ ,h % , , ,,_,__ _, , , , , _ __,, , ,,,_ _,_ _ , ,,,,, , ,,,, ,,, ,, ,

_ ’’’’’’ _i’mi__’_____’____i__’’’’ _ b __’’__m_’_d_,

_’’‘’"’’’"’’’’’___ ’ , _| _ _ _ _ _ _ ||||, ,,||| _ _ _ ’’’’’ _ __ _ _ V , _ ’
tr

_inuxU_ekroup Vor#rlberg

Wir sind ein loser Zusammenschluss von Linux-Anwendern,
die sich etwa einmal pro Monat zwanglos tre_en, um über
verschiedeneThemen und selbstverständlich über Linux zu
diskutieren. Zu diesenTrefren sind _ Linux-lnteressierten
eingeladen, gleichgültig ob LUGV-Mitglied oder nicht.

,|_||
,,,,M_|,|,,,¡|,_|
,lli’’|_’’’’_’’,,lh’’’|’_’’|’’’,||,_’’’__’’’’’..
_ _ , ’’

|
|

ln unregelmä_igen Abständen veranstaltet die LUGV auch
Ausflüge, Vo_räge und Workshops.

_’|_||

Auf der Homepage der LUGV finden sich neben Neuigkeiten
aus der Linux-Szene auch eine Bildergalerie sowie eine
Mailingliste und ein Anmeldeformular für die LUGV. Eine
MitgliedschaW ist an keine Verpflichtungen gebunden
sondern dient hauptsächlich der Organisation des LinuxDays.

Homepage: www.lugv.at
www.linuxday.at

_, ,|| _ _ , ’’’’ ’ %_ ’’ ’’’ ’ ’ _, ,

|’

_____ _, , ’_’’ _____ | _’||, | ’ | _ ’ , , |_| _’i’| ’ ’ ’_’’’’._’ _’ ’’
|_||||||,|,,,|,¡|,’’’’’,|’__’|’’j’h’’

_,_,_ ’ , ,_ , _, ,,,,,, ,, ’’ | |
_ ___ _, _, . ,

_._ ___, |

,,||’|_||,|,|’||’li_ll
_,||’,li’_’__’_’_’lill,’|’_’’||’|
II
III

Geschichte

Die LUGV wurde im Frühling I998 als loser Zusammenschluss von ca. 30 begeisterte
herberge in Feldkirch gegründet.
Nach einigenTre_en in der Jugendherberge wurde im Herbst eine lnstallatiDnspa_y
der JH organisie_. Der Andrang war so groˆ, dass einige lhre PCs nicht mehr aufs
lm Frühling I999 stellte uns dieVKW in Bregenz einen gro_en Raum zurVerfügung, i
2. lnstallationsparty organisie_ wurde. Es waren über loo Linuxbegeisterte, welc
den Raum füllten. Die VKW war auf diesen Anstrum nicht vorbereitet, da nur über
Werbung für die lnstallparty gemacht wurde. Das Stromnetz musste vestärkt werden
mit einer Zwischenverpflegung, o_erie_ von derVKW, überbrückt. Danach wurde bis
dem Hobby gefrönt.
Nach diesem tollen Erfolg wurde im Mai I999 im HotelWeiˆes Kreuz die ldee gebore
einen LinuxDay (www.linuxday.at) zu organisieren.

Die vollständige Geschichte der LUGV rindet sich unter www.lugv.at

5.3 Prospekt, Ausschnitt, 10 Punkt, Eras

D_mit eine Archivist_Box mit der
Aussenwelt kommunizieren k_nn,
bed_rf es einzig elnes Netzwerk-
k_bels. Jede Box ist fixrerIig vor-
konfiguriert; eine lnst_ll_tion ist nicht
nonNendig. Ob per FTF-Filetr_nsfer
IKopierger_tel, _ngeschlossenen
Sc_nnern, oder Druckvorg_ng, die
Archivist_Box zeigt sich _usserst

kont_kIfreudig bei der Dokumenten-
_nn_hme.

Alle Dokumente, dle in der Archivlst_-
Box eintre_en, werden _utom_tisch
beschl_gwortet lindexiertl und 5te-
hen unmittelb_r r_r eine kecherche
zur Verfügung. Der Zugrirr _ur dle
Box errolgt _ber den Web- oder kich-
Client. Dokumente können _ber
_uch _ls FDF oder per M_il wei-
tergereicht werden. G_nz n_ch dem
Motto: Connect your world.

_Archivisla GmbH

__ '= PosWach - CH-8042 Zurlch
,_a __| Tel: +41 (0)44 254 54 00
_Fax: +41 (0)44 254 5q 02
Web: www.archivista.ch
AR_HIVISTA E-Mail: wPbmaster_archivista.ch

5.4 Prospekt, Seite, 10 Punkt, Eras

rür je_en Bedarf die Áchtige

5ie helssen kigi, Fll_tus, Titlis, Eiger,
Mythen und kothorn - und sind
punkto UnerschütCerlichkeit und
5t_bilit_t gen_uso robust wie die
gleichn_migen geogr_ri5chen Erhe
bungen. Die Ferform_nce der
Document-5ewer |_sst sich von der
rel_tiven Höhe, die mit dem Berg-
n_men _ssoziiert ist, _bleiten.

pl_tte, zweiter Box und T_pe
L_uhNerk - ist rür Archive _b c_.
soo'ooo bis einige Millionen Seiten

_usgelegt.

Mythen und kothorn sind Sc_n- und OCk-Cluster-St_tionen, mit welchen _ur die Boxen Fil_tus, Titlis und Eiger gesc_nnt werden k_nn. Ebenr_lls

rühren sle z.B. eine Text- und/oder B_rcode-Erkennung durch.

Und f_lls Sie nun elnen Dokumenten-Cluster _ |_ MounI Everest be nötigen, _uch kein rroblemi wir stellen lhnen diesen mit der entsprechenden H_rdw_re gerne in-dividuell zus_mmen.

kigi elgnet sich prim_r für klelnere Umgebungen wie z.B. kechts_n-w_ltspr_xen oder Pk-Büros. Aber _uch grössere Unternehmen, die fur Abteilungen PDF-Dokumentensewer suchen, sind mic der kigi-Box gut bedient.

ril_tus ist fur miCtlere Firmengrössen, oder besser _usgedrückt, rür d_s mlttlere D_tenvolumen ged_cht. Die Titlis-Box ebenr_lls, _llerdings ergibt dle redund_nte H_rdw_re ein erhóhtes Slcherheitselement. Die Eiger-Box - mit entsprechender FesI-

nigi

ílhtu

Títli_

Eiger

Mythen

I797 m.ü.M.

2|_2 m.ü.M.

3238 m.ü.M.

3790 m.ü.M.

18_9 m.ü.M.

nothorn 23S1 m.ü.M.

Einzelpl_tz-Dokumenten-Server mit bis zu 20’000 Akten und loo’ooo Seiten

Dokumenten-Server für bis zu 200’000 Akten und l Mio Seiten

Dokumenten-Sewer für bis zu 200’000 Akten und l Mio Seiten, redund_nt I2 Boxenl Dokumenten-5erver für unlimitierte Anz_hl Akten, bis c_. 2 Mio Seiten, redund_nt I2 Boxenl und mit B_ckup-T_pe-Drive 5c_n- und OCk-Box, welche D_ten zum _il_tus und Titlis tr_nsportiert

Sc_n- und OCk-Box, p_ssend zum Eiger

ArchivistaBox: Connect your _orld

_’_ /
" ’_’ ;,_,_,,___i"

Frint-5ewer

_ __ ’_ ,
_ _

_ h
’_
M_il-/Orrice-
Dokumente

PDF IVolltext)

A,
`____ _’
_____rer 5c_nner
rf

rublizieren IDVDl

_ ’_,
’|
, O _
_ _"’

M_ils

_____ .
|
i

__. __
m__,_ _,,_’`

÷-

kich- und WebClient

|

EkP-System
lmiI Keysl

D_mit eine Archivist_Box mit der
Aussenwelt kommunizieren k_nn,
bed_rf es einzig eines Netzwerk-

k_bels. Jede Box lst rixrertig vor-
konriguriert; elne lnst_ll_tion lst nlcht
nonNendig. ob per FTF-Filetr_nsrer
IKopierger_te), _ngeschlossenen
5c_nnern, oder Druckvorg_ng, die
Archivist_Box zeigt sich _usserst
kont_ktrreudig bei der Dokumenten-
_nn_hme.

Alle Dokumence, die in der Archivist_-
Box eintrerren, werden _utom_tisch
beschl_gwortet lndexiertl und ste
hen unmlttelb_r rur eine kecherche
zur Verrügung. Der Zugritr _ur die
Box errolgt uber den Web- oder kich-
Client. Dokumente können _ber
_uch _ls rDF oder per M_il wei-
tergereicht werden. G_nz n_ch dem
Motto: Connect your world.

_Archivista GmbH

__ _ _ '= PosWach - CH-8042 Zúrlch
q

_OTel: +41 (0)44 254 54 00
L' Faw: +41 (0)44 254 54 02
_ Web: www.archlvista.ch
ARCHIVISTA E-Mail: webmasler_archivista.ch

6 Tesseract 3.04

Tesseract arbeitet mittlerweile sehr gut, die Unterschiede zur kommerziellen Texterkennung sind erst bei schlecht gescannten Vorlagen erheblich. Dabei darf beachtet werden, dass mit Tesseract exotischere Schriften wie die Eras trainiert werden können, womit mit etwas Initial-Aufwand auch bei exotischen Schriften sehr gute Ergebnisse erzielt werden können.

6.1 Text, 1spaltig, 12 Punkt Helvetica

Fall 2: Engineering-Projekt mit PDF-Dokumentation

Kaum eine andere Technologie hat im letzten Jahr für mehr Furore gesorgt als das Format. PDF steht für Portable Document Format und wie der Name besagt, geht es darum, Inhalte digital so aufzubereiten, dass diese auf unterschiedlichen Rechne Mac und Windows) betrachtet und ausgedruckt werden können.

PDF-Dokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auc PDF?Datei (in digitaler Form) publizierbar. Zudem sind PDF?Dateien mittlerweile breitet; der Viewer (Betrachter) für die Dateien ist kostenlos. Die Möglichkeite Formats machte sich eine internationale, im Bereich Engineering tätige Firma zun

Problemstellung und Lösung

Ca. 75?000 auf der Vorder- und Rückseite bedruckte A4-Seiten sowie 200 zum Teil Seiten im A3? bis A0-Format sollten gescannt und in ein PDF-Format überführt wer waren selbsttragende CDs mit PDF-Dateien, welche die ursprüngliche Dokumentennum aufweisen sollten und über die mit sogenannten Thumbnails ein Überblick besteht.

Archivista löste dieses Problem Schritt für Schritt:

o Duplex-Scanning der 75?000 A4-Seiten mit Hochleistungsscanner mit 300dpi, Resultat: 150?000 Seiten im Ti?G4wFormat

o Individuelles Scanning der 200 grösserformatigen Belege, davon 50 in Farbe

o Umwandlung der TiffG4 und JPG-Dateien in PDF?Dateien
o Erstellen der CDs mit Inhaltsverzeichnissen (inkl. Link auf entsprechende Date

Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und d
selber entwickelten Produktmoduls Archivista TifToPDF in der Lage, diese Wertsch
fungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendba
PDF-File kostengünstig anzubieten. Das Engineering-Unternehmen verfügt heute übe
saubere, platzsparende Dokumentation des Projektes.

Scan-Dienstleistungen ? Scanning und was dazu gehört, Fallbeispiele, Seite 2
Archivista GmbH, 8042 Zürich, Tel. 01 350 46 74, Fax 01 350 46 72, www.archivist

ÄA y)

WWII

6.2 Flyer, 1spaltig, ca. 8 Punkt, Helvetica

linuxllserGroup Vorarlberg

mum
Wir sind ein loser Zusammenschluss von Linux-Anwendern, A
die sich etwa einmal pro Monat zwanglos treffen, um über
verschiedeneThemen und selbstverständlich über Linux zu
diskutieren. Zu diesenTreffen sind ? Linux-lnteressierten
eingeladen, gleichgültig ob LUGV-Mitglied oder nicht.

In unregelmäßigen Abständen veranstaltet die LUGV auch
Ausflüge, Vorträge und Workshops.

Auf der Homepage der LUGV finden sich neben Neuigkeiten
aus der Linux-Szene auch eine Bildergalerie sowie eine
Mailingliste und ein Anmeldeformular für die LUGV. Eine
Mitgliedschaft ist an keine Verpflichtungen gebunden
sondern dient hauptsächlich der Organisation des LinuxDays.

Homepage: www.lugv.at
www.linuxday.at

Geschichte

Die LUGV wurde im Frühling 1998 als loser Zusammenschluss von ca. 30 begeisterte herberge in Feldkirch gegründet.

Nach einigenTreffen in der Jugendherberge wurde im Herbst eine lnstallationspart der JH organisiert. Der Andrang war so groß, dass einige Ihre PCs nicht mehr auf

Im Frühling 1999 stellte uns die VKW in Bregenz einen großen Raum zur Verfügung,

2. lnstallationsparty organisiert wurde. Es waren über 100 Linuxbegeisterte, wel den Raum füllten. Die VKW war auf diesen Anstrum nicht vorbereitet, da nur über Werbung für die Installparty gemacht wurde. Das Stromnetz musste vestärkt werden mit einer Zwischenverpflegung, offeriert von der VKW, überbrückt. Danach wurde b dem Hobby gefrönt.

Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel Weißes Kreuz die Idee gebor einen LinuxDay (www.linuxday.at) zu organisieren.

Die vollständige Geschichte der LUG V findet sich unter www.lugv.at

6.3 Prospekt, Ausschnitt, 10 Punkt, Eras

**Damit eine ArchivistaBox mit der
Aussenweit kommunizieren kann,
bedarf es einzig eines Netzwerk.
kabels. Jede Box ist ?xfertig vor?
konfiguriert; eine lnstaiiation ist nicht
notwendig. Ob per FTP-Filetransfer
{Kopiergeräte)? angeschlossenen
Scannern, oder Druckvorgang, die
ArchivistaBox zeigt sich äusserst
kontaktfreudig bei der Dokumenten-
annahme.**

**Alle Dokumente, die in der Archivista-
Box eintreffen, werden automatisch
beschlagwortet (indexiert) und ste-**

hen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web? oder Rich-Client. Dokumente können aber auch als PDF oder per Mail wei? tergereicht werden. Ganz nach dem Motto: Connect your world.

ä J, Archivista GmbH

ß .

4" Postfach – CH-8042 Zürich
? Tel: +41 (0)44 254 54 00
Fax: +41 (0)44 254 54 02

Web: www.archivista.ch

ABGHWIS?I?A E-Mail: webmaster@archivista.ch

6.4 Prospekt, Seite, 10 Punkt, Eras

Für jeden Bedarf die Richtige

Sie heissen Rigi, Pilatus, Titlis, Eiger, Mythen und Rothorn ? und sind punkto Unerschütterlichkeit und Stabilität genauso robust wie die gleichnamigen geografischen Erhe bungen. Die der Document?Server lässt sich von der relativen Höhe, die mit dem Berg-namen assoziiert ist, ableiten.

Performance

platte, zweiter Box und Tape Laufwerk ? ist für Archive ab ca. 500'000 bis einige Millionen Seiten

ausgelegt.

Mythen und Rothorn sind Scan? und
OCR-Cluster?Stationen, mit welchen
auf die Boxen Pilatus, Titlis und Eiger

gescannt werden kann. Ebenfalls

führen sie z.B. eine Text? und/oder
Barcode?Erkennung durch.

Und falls Sie nun einen Dokumenten?
Cluster a la Mount Everest be
nötigen, auch kein Problem; wir
stellen lhnen diesen mit der
entsprechenden Hardware gerne in?
dividuell zusammen.

Rigi eignet sich primär für kleinere

Umgebungen wie z.B. Rechtsan- Rigi 1797 m.ü.M.
waltspraxen oder PR?Büros. Aber

auch grössere Unternehmen, die für pnatus 2|32 mum
Abteilungen PDF-Dokumentenserver

suchen, sind mit der Rigi-Box gut Tims 3238 m.üM
bedient.

Pilatus ist für mittlere Firmengrössen, Eiger 3790 m.ü.M.
oder besser ausgedrückt, für das

mittlere Datenvolumen gedacht. Die

Titlis-Box ebenfalls, allerdings ergibt Mythen 1899 m_ü_M_
die redundante Hardware ein

erhöhtes Sicherheitselement. Die Rothorn 235] m.üM.

Eiger-Box ? mit entsprechender Fest-

Einzelplatz-Dokumenten-Server mit bis zu 20’000 Akten und lOO’OOO Seiten

Dokumenten?Server für bis zu 200’000 Akten und l Mio Seiten

Dokumenten-Server für bis zu 200’000 Akten und l Mio Seiten, redundant (2 Boxen)

Dokumenten-Server für unlimitierte Anzahl Akten, bis ca. 2 Mio Seiten, redundant l2 Boxen) und mit Backup-Tape?Drive

Scan- und OCR?Box, welche Daten zum Pilatus und Titlis transportiert

Scan? und OCR-Box, passend zum Eiger

ArehivistaBOX: Connect your world

l

WWWWM

Kopierer Scanner

4/ ?

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerk? kabels. Jede Box ist fixfertig vorkonfiguriert; eine Installation ist nicht notwendig. Ob per FTP-Filetransfer (Kopiergeräte), Scannern, oder Druckvorgang, die

angeschlossenen

ArchivistaBox zeigt sich ausserst
l ""h kontaktfreudig bei der Dokumenten-
annahme.

Alle Dokumente, die in der Archivista-
Box eintreffen, werden automatisch

Mam/Of?ce?? beschlagwortet (lndexiert) und Ste
DOkumente ? hen unmittelbar für eine Recherche
ERP?System zur Verfügung. Der Zugriff auf die

PDF (Volltext) (mit Keys) Box erfolgt über den Web? oder Rich-

Rich- und WebClient

Client. Dokumente können aber
auch als PDF oder per Mail wei-
tergereicht werden. Ganz nach dem
Motto: Connect your world.

Archivista GmbH

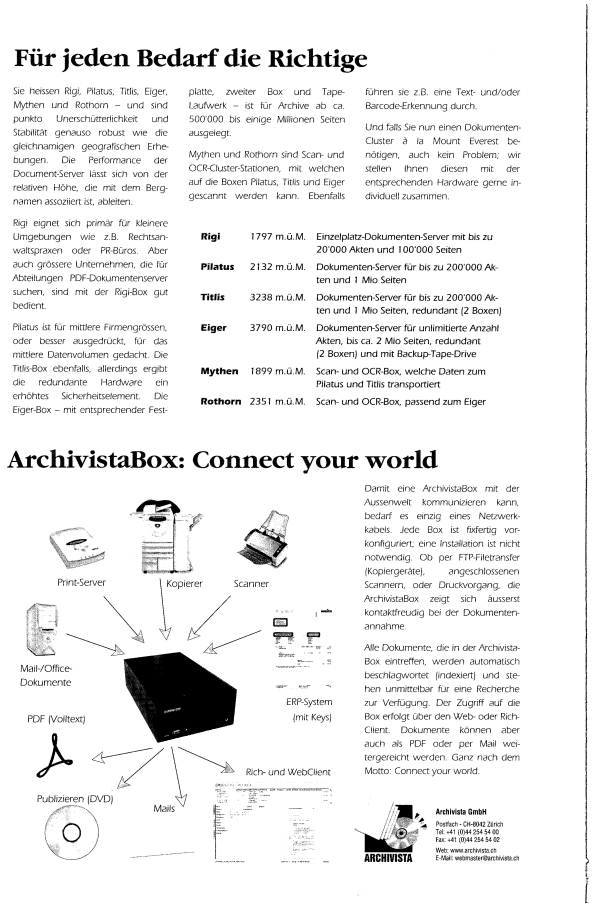
Postfach - CH-8042 Zürich
Tel: +41 (0)44 254 54 00
Fax: +41 (0)44 254 54 02

Web: www.archivista.ch
E-Mail: webmaster@archivistalch

ARCHIVISTA

7 Qualität beim Scannen

Das vierte Beispiel, die Seite des Prospektes weist eine schlechte Qualität des Scannens auf. Aus diesem Grunde wurde die gleiche Seite nochmals gescannt, diesmal mit den heutigen Default-Werten, mit denen die ArchivistaBox heute ausgeliefert wird. Zunächst eine Ansicht dieser Seite:



Für jeden Bedarf die Richtige

Sie heissen Rigi, Pilatus, Titlis, Eiger, Mythen und Rothorn – und sind punkto Unerschütterlichkeit und Stabilität genauso robust wie die gleichnamigen geografischen Erhebungen. Die Performance der Document-Server lässt sich von der relativen Höhe, die mit dem Bergnamen assoziiert ist, ableiten.

Rigi eignet sich primär für kleinere Umgebungen wie z.B. Rechtsanwaltspraxen oder PR-Büros. Aber auch grössere Unternehmen, die für Abteilungen PDF-Dokumentenserver suchen, sind mit der Rigi-Box gut bedient.

Pilatus ist für mittlere Firmengrössen, oder besser ausgedruckt, für das mittlere Datenvolumen gedacht. Die Titlis-Box ebenfalls, allerdings ergibt die redundante Hardware ein erhöhtes Sicherheitselement. Die Eiger-Box – mit entsprechender Festplatte, zweiter Box und Tape-Laufwerk – ist für Archive ab ca. 500'000 bis einige Millionen Seiten ausgelegt.

Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Titlis und Eiger gescannt werden kann. Ebenfalls führen sie z.B. eine Text- und/oder Barcode-Erkennung durch.

Und falls Sie nun einen Dokumenten-Cluster à la Mount Everest benötigen, auch kein Problem; wir stellen Ihnen diesen mit der entsprechenden Hardware gerne individuell zusammen.

Rigi	1797 m.ü.M.	Einzelplatz-Dokumenten-Server mit bis zu 20'000 Akten und 100'000 Seiten
Pilatus	2132 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten
Titlis	3238 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten, redundant (2 Boxen)
Eiger	3790 m.ü.M.	Dokumenten-Server für unlimitierte Anzahl Akten, bis ca. 2 Mio Seiten, redundant (2 Boxen) und mit Backup-Tape-Drive
Mythen	1899 m.ü.M.	Scan- und OCR-Box, welche Daten zum Pilatus und Titlis transportiert
Rothorn	2351 m.ü.M.	Scan- und OCR-Box, passend zum Eiger

ArchivistaBox: Connect your world

Print-Server
Kopierer
Scanner
Mail-/Office-Dokumente
PDF (Volltext)
ERP-System (mit Keys)
Rich- und WebClient
Publizieren (DVD)
Mails

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerkkabels. Jede Box ist fixfertig vorkonfiguriert; eine Installation ist nicht notwendig. Ob per FTP-Filetransfer (Kopiergeräte), angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich äusserst kontaktfreudig bei der Dokumentenannahme.

Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlagwortet (indexiert) und stehen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web- oder Rich-Client. Dokumente können aber auch als PDF oder per Mail weitergereicht werden. Ganz nach dem Motto: Connect your world.

ARCHIVISTA
Archivista GmbH
Postfach · CH-8042 Zürich
Tel. +41 (0)44 254 54 00
Fax. +41 (0)44 254 54 02
Web: www.archivista.ch
E-Mail: webmaster@archivista.ch

Nachfolgend den erkannten Text aus Tesseract 3.04:

```
Für jeden Bedarf die Richtige

Sie heissen Rigi, Pilatus, Titlis, Eiger,
Mythen und Rothorn und sind
punkto Unerschütterlichkeit und
Stabilität genauso robust wie die
gleichnamigen geografischen Erhe
bungen. Die Performance der
DocumentServer lässt sich von der
relativen Höhe, die mit dem Berg-
namen assoziiert ist, ableiten;
```

Rigi eignet sich primär für kleinere Umgebungen wie z.B. Rechtsan waltspraxen oder PR-Büros. Aber auch grössere Unternehmen, die für Abteilungen PDFDokumentenserver suchen, sind mit der ngiBox gut bedient.

Pilatus ist für mittlere Firmengrössen, oder besser ausgedrückt, für das mittlere Datenvolumen gedacht. Die TitlisBox ebenfalls, allerdings ergibt die redundante Hardware ein erhöhtes Sicherheitselement. Die EigerBox mit entsprechender Fest-

platte, zweiter Box und Tape-Laufvverk ist für Archive ab ca. SOO'OOO bis einige Millionen Seiten ausgelegt.

Mythen und Rothorn sind Scan- und OCRClusterStationen, mit welchen auf die Boxen Pilatus, Titlis und Eiger gescannt werden kann. Ebenfalls

führen sie z.B. eine Text und/oder Barcode-Erkennung durch.

Und falls Sie nun einen Dokumenten-Cluster a la Mount Everest be nötigen, auch kein Problem; wir stellen Ihnen diesen mit der entsprechenden Hardware gerne in-dividuell zusammen.

Rigi 1797 m.ü.M. Einzelplatz-DokumentenServer mit bis zu 20'000 Akten und IOO'OOO Seiten

Pilatus

2 i 32 m.ü.M. Dokumenten-Sewer für bis zu 200'000 Ak

ten und 1 Mio Seiten

Titlis 3238 m.ü.M. DokumentenServer für bis zu 200'000 Ak
ten und 1 Mio Seiten, redundant (2 Boxen)

Eiger 3790 m.ü.M. Dokumenten-Server für unlimitierte Anzahl
Akten, bis ca. 2 Mio Seiten, redundant
(2 Boxen) und mit Backup-TapeDrive

Mythen i899 m.ü.M. Scan- und OCRBox, welche Daten zum
Pilatus und Titlis transportiert

Rothorn 2351 m.ü.M. Scan und OCR-Box, passend zum Eiger

ArchivistaBox: Connect your world

Print-Server

Mail-/Office
Dokumente

PDF (Volltext)

Mails

Kopierer

Scanner

ERPSystem
(mit Keys)

Damit eine ArchivistaBox mit der
Aussenwelt kommunizieren kann,
bedarf es einzig eines Netzwerk-
kabels. Jede Box ist fixfertig vor

konguriert; eine lnstallation ist nicht notwendig. Ob per FTPFiletransfer {Kopiergerätei angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich ausserst kontaktfreudig bei der Dokumenten- annahme.

Alle Dokumente, die in der Archivista Box eintreffen, werden automatisch beschlagwortet (indexiert) und ste hen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web- oder Rich- Client. Dokumente können aber auch als PDF oder per Mail wei tergereicht werden. Ganz nach dem Motto: Connect your world.

Archivista GmbH

Postfach - CH-8042 Zürich
Tel: +41 (0)44 254 54 00
Fax: +41 (0)44 254 54 02

Weh: www.archivisia.ch
E-Mail: webmaster@archivista.ch

ABDHIWSTA

MM

Unschwer lässt sich erkennen, dass Tesseract 3.04 bei einigermassen gescannten Texten extrem gute Resultate liefert.

8 Frakturerkennung (Tesseract)

92

Wir werden zunächſt die erſtere zu betrachten haben, in den Nebenzonen aber weiter noch mehrere Specialgebiete, und zwar die Grauwackenzonen, die nördlichen Kalkalpen, die Wiener Sandſteinzone und die ſüdlichen Kalkalpen, denen ſich die auf unſer Staatsgebiet fallenden Gebirge des Balkanſyſtems unmittelbar anſchließen, abgeſondert behandeln.

1. Centralzone.

Die Centralalpen oder die kryſtalliniſche Mittelzone der Alpen beſteht durchwegs aus Geſteinen der archaiſchen Epoche, unter welchen allerorts die kryſtalliniſchen Schiefergeſteine über die kryſtalliniſchen Maſſengeſteine weitaus vorwalten. Die Grenzlinie übrigens, welche dieſelben von den Sedimentgeſteinen ſcheidet, ſtimmt nicht überall genau mit jener überein, welche man vom orographiſchen Standpunkte zwiſchen den Centralalpen und den Kalkalpen gezogen hat. So finden wir beiſpielsweiſe auf der Karte Seite 27 die Gruppen des Hochſchwab und der Veitſch, die aus meſozoiſchen Kalkſteinen beſtehen, noch der Centralzone zugezählt, anderſeits ſind die ganzen Ortler Alpen und die Adamello-Gruppe, ſowie im Oſten das Bachergebirge, obgleich ſie zum Theil oder ganz aus kryſtalliniſchen Geſteinen beſtehen, mit der ſüdlichen Nebenzone vereinigt, und analoge Abweichungen ergeben ſich auch an anderen Stellen. Auch mag hier gleich hervorgehoben werden, daß, wenngleich die Centralzone das eigentliche Herrſchfeld der archaiſchen und die Nebenzonen jenes der Sedimentgeſteine bilden, ſich doch einerſeits beträchtliche Maſſen der letzteren, an manchen Stellen der mittleren Kette, in iſolirten Schollen über den kryſtalliniſchen Geſteinen vorfinden, wie z. B. an der Landesgrenze in den Ortler Alpen oder am Brenner, oder endlich auf der zu den ſteiriſchen Alpen gehörigen Stangalpe, und daß anderſeits an manchen Stellen der ſüdlichen, nicht aber auch der nördlichen Nebenzonen Inſeln kryſtalliniſcher Geſteine aus den umgebenden Sedimentgeſteinen emportauchen. Die wichtigſten der letzteren auf unſerem Staatsgebiete ſind der gewaltige, von kryſtalliniſchen Schiefergeſteinen umgebene Granitſtock der Cima d'Aſta in Südtirol, der ſchmale Zug von Glimmerſchiefer, welcher der Einſenkung des Gailthales in Kärnten folgt, im Weſten aber mit der Centralzone doch in Verbindung ſteht, und ein ähnlicher langer und ſchmaler Zug von kryſtalliniſchen Schiefer- und Maſſengeſteinen, der ſüdlich von der Karavankenkette, den Längsthälern der Miß und Javoria entlang, fortſtreicht.

So wenig wie in der Bodenplaſtik, ebenſowenig zeigt ſich auch in der geologiſchen Zuſammenſetzung im Gebiete der Mittelzone eine regelmäßige, dem weſtöſtlichen Hauptſtreichen des ganzen Gebirges folgende Anordnung. Hier wie in anderen Gebieten hat man erkannt, daß von den drei Hauptarten der kryſtalliniſchen Schiefergeſteine der Gneiß das tiefſte und älteſte, der Glimmerſchiefer das nächſt jüngere und der Thonſchiefer das jüngſte Gebilde iſt. Keines dieſer Geſteine aber erſcheint, der ganzen Erſtreckung der Centralkette

Die Frakturerkennung liefert gute bis sehr gute Resultate, wenn wir einen Blick auf den erkannten Text werfen:

Wir werden zunächst die erstere zu betrachten haben, in den Nebenzonen aber wei
mehrere Specialgebiete, und zwar die Grauwackenzonen, die nördlichen Kalkalpen,
Wiener Sandsteinzone und die südlichen Kalkalpen, denen sich die auf unser Staat
fallenden Gebirge des Balkansystems unmittelbar anschließen, abgesondert behande

1. Centralzone.

Die Centr alalp en oder die krystallinische Mittelzone der Alpen besteht durchwe
aus Gesteinen der archaischen Epoche, unter welchen allerorts die krystallinisch
gesteine über die krystallinischenMassengesteine weitaus vorwalten. Die Grenzlin
welche dieselben Von den Sedimentgesteinen scheidet, stimmt nicht überall genau
überein, welche man vom orographischen Standpunkte zwischen den Centralalpen und
Kalkalpen gezogen hat. So finden wir beispielsweise auf der Karte Seite 27 die G
des Hochschwab und der Veitsch, die aus mesozoischen Kalksteinen bestehen, noch
Centralzone zugezählt, anderseits sind die ganzen Ortler Alpen und die Adamello-
sowie im Osten das Bachergebirge, obgleich sie zum Theil oder ganz aus krystalli
Gesteinen bestehen, mit der südlicheu Nebenzone vereinigt, und analoge Abweichun
ergeben sich auch an anderen Stellen. Auch inag hier gleich hervorgehoben werden
wenngleich die Cenwalzone das eigentliche Herrschfeld der archaischen und die Ne
jenes der Sedimentgesteine bilden, sich doch einerseits beträchtliche Massen der
an manchen Stellen der mittleren Kette, in isolirten Schollen iiber den krystall
Gesteinen vorfinden, wie z. B. an der Landesgrenze in den Ortler Alpen oder am B
oder endlich auf der zu den steirischen Alpen gehörigen Stangalpe, und daß ander
an manchen Stellen der südlichen, nicht aber auch der nördlichen Nebenzonen Jnse
krystallinischer Gesteine aus den umgebenden Sedimentgesteinen emportauchen. Die
wichtigsten der letzteren auf unserem Staatsgebiete sind der gewaltige, von krys
Schiefergesteinen umgebene Granitstock der Cima d?Asta in Südtirol, der schmale
Glimmerschiefer, welcher der Einsenkung des Gailthales in Kärnten folgt, im West
mit der Centralzone doch in Verbindung steht, und ein ähnlicher langer und schm
von krystallinischen Schiefer- und Massengesteinen, der südlich von der Karavank
den Längsthälern der Mifz und Javoria entlang, fortstreicht. .

So wenig wie in der Vodenplastik, ebensowenig zeigt sich auch in der geologische
Zusammensetzung im Gebiete der Mittelzone eine regelmäßige, dem westöstlichen Ha
streichen des ganzen Gebirges folgende Anordnung. Hier wie in anderen Gebieten h
erkannt, daß von den drei Hauptarten der krystallinischen Schiefergesteine der G
tiefste und älteste, der Glimmerschiefer das nächst jüngere und der Thonschiefer
Gebilde ist. Keines dieser Gesteine aber erscheint, der ganzen Erstreckung der C

Hinweis: Mit der ArchivistaBox kann die Fraktur-Erkennung direkt aufgerufen werden, indem in WebAdmin bei den OCR-Definition bei der Sprache 'DeutschNeu' bzw. 'German-NewSpelling' sowie als OCR-Engine 'Tesseract 3.04' gewählt wird.

9 Neue Zeichensätze

Im Unterschied zu allen übrigen Open Source Texterkennungsprogrammen lässt sich Tesseract 3.04 trainieren. Dies sei am Beispiel der Schweizer ESR-Einzahlungsscheine ausgeführt.

Die Schweizer ESR-Einzahlungsscheine besitzen eine sogenannte Referenzzeile (weisser Bereich unten rechts), der zwingend mit der Schrift OCRB zu drucken ist. Diese Schrift kann relativ einfach von vielen Lesegeräten erkannt werden. Wer schon einmal eine Einzahlung an einem Schweizer Postschalter gemacht hat, wird bemerkt haben, dass die Einzahlungsscheine von einem Lesegerät erfasst werden.

Derartige Geräte sind nicht ganz günstig. Ein Gerät, dass die Belege mit einem Einzug erfasst, ist ab ca. 1200 bis 1300 Franken erhältlich, die Verarbeitungsgeschwindigkeit liegt bei ca. 15 Belegen pro Minute. Ein klarer Fall für die ArchivistaBox. Ganz so einfach war es jedoch nicht. Nachfolgend wird der gewählte Weg in groben Zügen aufgezeigt.

9.1 Tesseract mit Layout-Modus

Die ersten Versuche, die OCRB-Schrift mit der Standardsprache (Deutsch oder Englisch) zu erkennen, war recht ermutigend. Mit den Optionen für das gewünschte Seitenlayout wie z.B. **`-psm 6`** konnten bereits recht gute Ergebnisse erzielt werden. Dennoch stellten sich dann und wann Fehler ein. Ebenso gelang es nicht, aus dem Einzahlungsschein den ESR-Block immer passgenau zu extrahieren. Die ESR-Zeile wurde gelegentlich in mehrere Blöcke geteilt, was das anschliessende Zusammentragen nicht wahnsinnig einfach machte. Mit einem Skript bei der Nachbearbeitung konnte bei den ESR-Scheinen eine Trefferquote von ca. 95 Prozent errreicht werden. Dies bedeutet, dass Tesseract 3.04 ohne trainierte Schriften bereits 19 von 20 ESR-Scheinen korrekt verarbeiten konnte.

Nebenbemerkung: Eine solche Quote war bei der kommerziellen Texterkennung nicht möglich (hier waren es ca. 80 Prozent). Zwar liesse sich unsere kommerzielle Engine ebenfalls trainieren, doch wären dazu zwingend Windows-Applikationen notwendig. Anzufügen gilt hier, dass längst nicht alle verfügbaren kommerziellen OCR-Engines für Linux trainiert werden

können. Im besten Falle sind die Linux-Produkte zu den Produkten unter Windows (die meistens trainiert werden können) kompatibel. Aufgrund dessen, dass für vertiefte Informationen zunächst meist ein Geheimhaltungsvertrag (NDA) zu unterzeichnen ist, ehe die Informationen erhältlich sind, wurde darauf verzeichtet. Dies auch deshalb, da besagte NDA-Verträge es meistens verbieten, dass Informationen daraus öffentlich gemacht werden dürften – und dann bringt die ganze 'Übung' ja nichts.

9.2 Trainieren OCRB-Schrift

In einem zweiten Schritt ging es darum, die Erkennungsrate der Zeichen zu erhöhen. Dazu gibt es mittlerweile viele Anleitungen, in denen beschrieben ist, wie Tesseract für bestimmte Schriften trainiert werden kann. Es würde den Rahmen dieses Skriptes sprengen, alle notwendigen Punkte dieses Prozesses zu beschreiben. Letztlich war es ein Mix vieler Anleitngen, um zu einem Resultat zu kommen. Im Grundsatz geht es darum, dass zunächst Bildmaterial der zu trainierenden Schrift vorhanden sein muss. Danach können (sofern Tesseract korrekt kompilliert ist) mit der Option **`batch.nochop makebox`** sogenannte Box-Dateien erstellt werden. Ein Beispiel findet sich untenstehend:

```
0 240 2999 264 3035 0
1 277 3000 290 3035 0
0 308 2999 332 3035 0
0 342 2999 366 3035 0
0 377 2999 401 3035 0
0 411 2999 435 3035 0
0 445 2999 469 3035 0
1 482 3000 495 3035 0
6 513 2999 537 3035 0
0 547 2999 571 3035 0
9 582 3000 606 3035 0
0 616 2999 640 3035 0
2 651 3000 672 3035 0
> 685 3002 708 3033 0
7 720 3000 743 3035 0
```

Darin werden Buchstaben und die entsprechenden Positionen in der Bilddatei gespeichert. Diese Dateien dienen als Muster für die spätere Zeichenerkennung. Wichtig ist hier, dass diese Datei keine Fehler enthalten sollte. Es gibt einige Programme, mit denen Box-Dateien nachbearbeitet werden können (z.B. jTessBoxEditor). Um eine neue Sprachdatei zu erstellen, wird am Schluss aus der Bild-Datei mit den Mustern und der Box-Datei eine Sprachen-

datei erstellt. Auch dazu gibt es entsprechende Programme (wiederum sei hier jTessBoxEditor genannt). Am Ende entsteht die Musterdatei **`name.traineddata`**, die in den Ordner **`/usr/share/tessdata`** zu legen ist. Anschliessend kann die Sprache mit **`-l name`** für die Erkennung aktiviert werden.

Ein Test mit der erstellten OCRB-Musterdatei brachte herrvorragende Werte, allerdings werden die übrigen Schriftmuster auf dem Einzahlungsschein mehr schlecht den recht erkannt – diest ist jedoch nicht weiter verwunderlich, letztlich wurde Tesseract nun ja ausschliesslich auf die OCRB-Schrift trainiert.

9.3 Mehr Speed mit Blockwahl

Die Erkennung mit Tesseract benötigt mitunter etwas Zeit. Für eine A4-Seite kann der Vorgang gut und gerne zwischen 5 und 20 Sekunden Zeit in Anspruch nehmen. Um diese Wartezeit zu eliminieren, wurde der ArchivistaBox der 'ESR-Modus' spendiert. Dabei wird bei gescannten Einzahlungsscheinen immer nur jener Bereich zugewiesen, auf dem die OCRB-Zeile steht (weisser Bereich unten rechts). Damit steigt die Geschwindigkeit markant, im Durchschnitt werden nun noch ca. 0.1 bis 0.2 Sekunden pro ESR-Schein benötigt.

9.4 Fehlerquote und Aktivierung

Mit den beiden obenbeschriebenen Optimierungen konnte bei etwas mehr als 350 ESR-Belegen kein Beleg mehr mit einer fehlerhaften Erkennung festgestellt werden. Das Feature 'ESR-Schein einlesen' steht daher ab sofort allen Kunden (inkl. ArchivistaCloud) zur Verfügung. Um den ESR-Modus zu aktivieren, genügt es, bei der gewählten OCR-Definition (die einer Scan-Definition zuzuweisen ist) bei Sprache die Option 'Zahlen' zu wählen. Ist diese Sprache aktiviert, wird automatisch auf 'ESR-Verarbeitung' umgeschaltet. Die erkannte ESR-Zeile erscheint danach im Feld 'Titel'. Abschliessend auch dazu ein Beispiel:

```
0100000031001>922352000000781278105892649+ 010026722>
```

10 Abschliessende Bemerkungen

Tesseract 3.04 bietet mittplerweile die gleiche Qualität, wie dies von kommerziellen Produkten her der Fall ist. Einzige Bedingung ist, dass die gescannten Seiten eine eingermassen gute Qualität aufweisen. Bei qualitativ gut gescannten Seiten ist faktisch kein Unterschied mehr zu erkennen. Mit der umfassenden Möglichkeiten, Zeichensätze zu trainieren (Beispiel siehe obenstehend) können noch bessere Resultate erzielt werden, siehe dazu 'Report on the comparison of Tesseract and ABBYY FineReader OCR Engine', wo mit trainierten Zeichensätzen gearbeitet wurde. Gerne daraus ein Zitat:

When comparing results of both engines in test, there is no single winner that would outperform the second engine in all test cases.

Damit wir uns richtig verstehen, bei Tesseract gibt es absolut keine Seitenbegrenzungen, der Code ist offengelegt, es fallen keine Lizenzgebühren an. Bei einer kommerziellen Engine (wie z.B. Abbyy FineReader) dagegen sind die Sourcen nicht offengelegt, es fallen Seitenbegrenzungen an und die Preise sind doch recht happig (eine SDK-Lizenz ist derzeit für ca. 5000 Euro zu haben).

Ocrad dagegen ist nur für gute Schriftstücke brauchbar, benötigt aber fast keinen Platz auf der Festplatte (ca. 300 KByte) und läuft selbst auf einer ARM-CPU recht flott. Aus diesem Grunde findet sich auf den ArchivistaBoxen Albis und Bachtel (beide mit ARM-Prozessor) derzeit nur Ocrad. Tesseract 3.04 läuft aber grundsätzlich auf jeder Plattform, bei einer kommerziellen Engine ist für jede Plattform wieder eine andere Lizenz erforderlich.

Was Open Source Produkten mitunter noch fehlt, ist eine einfache Installation. Zwar bieten viele Distributionen die gängigen Pakete (z.B. für Tesseract) an, doch sind diese oftmals veraltet. Aktuelle Versionen müssen dann 'mühsam' kompilliert (übersetzt) werden. Mühsam kann dies daher sein, weil dabei oft Abhängigkeiten nicht aufgelöst werden können. So erfordert der Layout-Modus von Tesseract 3.04 mittlerweile C++11, was bei Debian im 'stabilen' Bereich mit dem Standard-Kompiler schon mal nicht machbar ist. Dies sollte nicht sein!

Muss es auch nicht, denn wer die drei OCR-Engines testen möchte, kann dies bequem und kostenfrei mit der ArchivistaBox Cloud in Angriff nehmen. Pro Konto stehen hier 20'000 Dokumente, 100'000 Seiten bzw. 20 GByte zur freien Verfügung. Eizige Bedingung ist, dass die/der Kontoinhaber/in entweder eine Privatperson ist oder es sich um einen nicht kommerziellen Verein handelt.

Die Registrierung kann hier erfolgen:

archivista.ch/de/pages/support/community.php

Zum Abschluss noch dies: In den Jahren 2006 bis 2008 war keinesfalls absehbar, dass dereinst ein Open Source Projekt kommerzielle Produkte übertreffen wird. Umso erfreulicher ist es, dass dies im Jahre 2014 der Fall ist. In diesem Sinne war die Wahl der ArchivistaBox, bereits 2005 auf Open Source zu setzen, in allen Ecken und Kanten richtig. So kann die ArchivistaBox 2014 mal nebenbei schnell eine ESR-Erkennung zuschalten, wo dies früher teure Runtime-Lizenzen erfordert hätte. In diesem Sinne viel Spass mit quelloffener Texterkennung und/auf der ArchivistaBox.

Kontakt: Archivista GmbH, Stegstr. 14, CH-8132 Egg, www.archivista.ch