

OCR- und Formularerkennung mit Linux

bersicht ber Projekte und Vergleich zu ClosedSource-Produkten

Inhaltsverzeichnis

| | | | | | |
|----------|---|-----------|-----------|--|-----------|
| 1 | Einleitung | 3 | | | |
| 1.1 | Freie OCR-Erkennung unter Linux | 3 | 6.2 | Flyer, 1spaltig, ca. 8 Punkt, Helvetica 12 | |
| 2 | Rckblick ber die letzten 20 Jahre | 4 | 6.3 | Prospekt, Ausschnitt, 10 Punkt, Eras 13 | |
| 2.1 | OmniPage um 1990 | 4 | 6.4 | Prospekt, Seite, 10 Punkt, Eras 13 | |
| 2.2 | FineReader in den spen 90er-Jahren | 4 | 7 | Ocrad 0.17) | 14 |
| 2.3 | FineReader ab 2000 Markt-leader | 5 | 7.1 | Text, 1spaltig, 12 Punkt Hel-vetica | 14 |
| 3 | Wo bleibt Linux? | 6 | 7.2 | Flyer, 1spaltig, ca. 8 Punkt, Helvetica | 16 |
| 3.1 | Die ersten Produkte | 6 | 7.3 | Prospekt, Ausschnitt, 10 Punkt, Eras | 19 |
| 3.2 | Tesseract alias Google-Projekt | 6 | 7.4 | Prospekt, Seite, 10 Punkt, Eras | 20 |
| 3.3 | OCR-Opus | 7 | 8 | Tesseract 2.01 | 24 |
| 3.4 | Und sonst? | 7 | 8.1 | Text, 1spaltig, 12 Punkt Hel-vetica | 24 |
| 4 | Verein freearchives.ch | 8 | 8.2 | Flyer, 1spaltig, ca. 8 Punkt, Helvetica | 25 |
| 5 | Demo-Seiten fr OCR-Tests | 9 | 8.3 | Prospekt, Ausschnitt, 10 Punkt, Eras | 26 |
| 5.1 | Text, 1spaltig, 12 Punkt Hel-vetica | 9 | 8.4 | Prospekt, Seite, 10 Punkt, Eras | 27 |
| 5.2 | Flyer, 1spaltig, ca. 8 Punkt, Helvetica | 10 | 9 | Frakturerkennung (Tesseract) | 29 |
| 5.3 | Prospekt, Ausschnitt, 10 Punkt, Eras | 10 | 10 | Formular- und Barcodeerken-nung | 31 |
| 5.4 | Prospekt, Seite, 10 Punkt, Eras | 11 | 10.1 | Formulare mit ExactImage . . | 31 |
| 6 | Archivista OCR-Option | 12 | 10.2 | Barcode-Erkennung | 31 |
| 6.1 | Text, 1spaltig, 12 Punkt Hel-vetica | 12 | 11 | Abschliessende Bemerkungen | 32 |

© 23.11.2007 by freearchives.ch, Homepage: www.freearchives.ch

1 Einleitung

1.1 Freie OCR-Erkennung unter Linux

Beim nachfolgenden Vortrag wird der Versuch gewagt aufzuzeigen, welche OpenSource-Produkte derzeit unter Linux verfügbar sind. Der Vortrag wurde für 'Fortgeschrittene' angekündigt. Es wird nun beruhigen, wenn ich vorweg anfragen darf, dass er allgemein verständlich gehalten ist und das Programmierkenntnisse an keiner Stelle gecheckt werden müssen. Dies ganz einfach deshalb, weil mittlerweile die ArchivistaBox die 'Kommandozeilen'-Eingriffe kapselt. So gesehen sind wir im fortgeschrittenen Stadium und können uns nun auch an Anfänger wenden.

Als Geschäftsführer der Archivista GmbH bin ich tagtäglich mit hunderten ja tausenden von Belegen konfrontiert, die zumindest meistens über minimal 10 Jahre aufzubewahren sind.

Meistens verwenden wir dabei für unsere Kunden in irgendeiner Form Texterkennung, sei es um den gesamten Text zu extrahieren oder sei es auch nur darum, dass wir im Falle der Formularerkennung aus einem bestimmten Beleg genau einzelne bestimmte Informationen extrahieren möchten.

Oft verwenden wir auch Barcodes, weil diese im Gegensatz zur Texterkennung weit näher bei 100 Prozent liegen als die Texterkennung. Zwar sind 99,9 Prozent Trefferquote bei erkanntem Text gut, bei 4000 Zeichen (soviel umfasst eine durchschnittliche A4-Seite in der Regel), sind es aber dennoch 4 Zeichen pro Seite.

2 Rückblick über die letzten 20 Jahre

2.1 OmniPage um 1990

Vor ca. 20 Jahren sind die ersten Texterkennungsprodukte auf den Markt gekommen. Ich selber durfte die ersten Erfahrungen mit Texterkennung bei einem staatlichen Betrieb (EMPA Döbendorf) sammeln, wo ich in den Jahren 1991 bis 1993 als DTP-Gestalter nebenher auch noch für das Scannen bzw. die Texterkennung zuständig war. Die damalige Analoge kostete ca. 30'000 Euro, alleine der Scanner kostete weit mehr als 5000 Euro, die OCR-Software gab es für ca. 1500 Euro, der Rechner (ein Apple-Gert) selber nochmals ca. 15'000 Euro.

Das Produkt, das damals in aller Munde war, hiess OmniPage, ein weiteres sehr gutes Produkt Recognita. Umgekehrt in den Jahren 1995 bis 1997 kam FineReader auf dem Markt, das russische Produkt brachte mit jeder Version eine bessere Erkennungsgenauigkeit als die übrigen beiden und weitere Konkurrenten.

Im Jahre 1998 gründete ich die Firma Archivista und baute damals für unser Produkt Xerox TextBridge ein. Die Qualität von TextBridge war deutlich besser als jene von OmniPage im Jahre 1993, der Preis für eine Lizenz belief sich auf ca. 100 Euro pro Lizenz, wobei keinerlei Seitenbeschränkungen vorhanden waren.

2.2 FineReader in den späten 90er-Jahren

Ein angehender Kunde, ein Pfarrer, der für den damaligen Bischof von St. Gallen, eine Archivierung suchte, hat an einer Messe unser Produkt gesehen. Am Produkt fand er Gefallen, bei der OCR-Texterkennung machte er mir klar, dass FineReader (siehe www.abbyy.com) das weit bessere Produkt sei.

So kam es, dass die Firma Archivista auf FineReader aufmerksam wurde. Im Jahre 1999 bauten wir schliesslich FineReader in Archivista ein. Die Lizenz (die Preise waren damals öffentlich) kostete im Einkauf ca. 50 USD, womit wir bei Endverkaufspreisen von ca. 200 DM bzw. 160 sFr. lagen.

In der Zwischenzeit wurde Recognita von der Firma aufgekauft, welche OmniPage entwickelte. Das Produkt OmniPage schliesslich wechselte in der Folge mehrmals den Eigentümer. Aufmerksam wurde ich während dieser Zeit auf genau drei weitere Produkte:

- Cuneiform (siehe www.ocr.com): Ebenfalls ein russisches Produkt, mit guter Erkennungsqualität, wobei die Stabilität der Windows-Anwendung nicht immer zu überzeugen vermochte.

- ReadIris (siehe www.readiris.com: Belgisches Produkt, dass von der Qualitt her nicht ganz an die Konkurrenten herankam.
- Wocar (siehe www.simpleocr.com: Ein franzsicher Programmierer entwickelte eine Freeware-Applikation, welche Texte in Franzsich und Englisch in ganz akzeptabler Qualitt erkennen konnte.

Allen Applikationen gemeinsam war, dass sie unter Windows liefen (im besten Fall fr den Mac verfgrbar waren), aber weder OpenSource waren noch unter Linux liefen.

2.3 FineReader ab 2000 Marktleader

Ganz klar Marktleader wurde FineReader, ganz einfach weil die Erkennungsgenauigkeit gegenber den brigen Produkten weit besser war. Mit dem Aufstieg an die 'Weltspitze' nderte die Firma Abbyy, welche FineReader entwickelt, aber radikal das sogenannte Business-Modell. Nicht primr im Consumer-Markt, sondern bei den Entwickler-Produkten.

Etwa im Jahre 2000 wurden die Preise fr die sogenannten Runtime-Lizenzen (das sind Lizenzen, welche die Entwickler, die mit FineReader selber ein Produkt entwickeln, an die Endkunden weitergeben) von ca. 50 auf einen Schlag auf ein hheres Mehrfaches angehoben.

Ursprnglich wurden die Preise auf der Homepage von ABBYY verffentlich. Leider werden die aktuellen Preise seit Jahren nicht mehr verffentlicht. Es kann aber gesagt werden, dass es stolze Summen (Stand 2007) sind.

Wenn ich mit diesen Zahlen spiele, dann deshalb, um aufzuzeigen, dass es im Bereich der OCR-Texterkennung ziemlich schnell ziemlich teuer werden kann. Natrlich hat nicht jeder Millionen von Seiten, die er scannen mchte. Wenn es aber darum geht, alte Buchbestnde zu scannen, dann fallen schnell erhebliche Summen an.

3 Wo bleibt Linux?

3.1 Die ersten Produkte

Ich selber arbeite nun seit ca. 10 Jahren mit Linux. Lange Zeit war es praktisch unmöglich, mit Linux und vernünftigem Aufwand eine OCR-Software zu betreiben. Die ersten Produkte, die mir unter die Augen gekommen sind, waren GOOCR (siehe ocr.sourceforge.net) sowie Clara (siehe www.geocities.com/claraocr). Beide Projekte werden meines Wissens von einem Entwickler betrieben.

Eine OCR-Erkennung zu entwickeln, ist leider keine einfache Angelegenheit. Die Zeichenerkennung an sich stellt dabei nur einen Teil der Aufgabe dar. Mindestens so wichtig ist das Erkennen der Textblöcke bzw. die Seitenanalyse. Bis vor einigen Monaten gab es unter Linux praktisch kein OpenSource-Produkt, das kommerziellen Produkten das Wasser hätte reichen können.

Am besten schnitt bei uns intern Ocrad (siehe www.gnu.org/software/ocrad) ab. Einmal ist Ocrad von Anfang an darauf ausgelegt, mit Sonderzeichen (Umlauten) umgehen zu können, weiter aber kann bei Ocrad relativ einfach ein Zeichensatz (insbesondere auch nur Zahlen) festgelegt werden.

3.2 Tesseract alias Google-Projekt

Ziemlich genau vor einem Jahr kündigte Google an, eine Software von HP zu übernehmen und diese unter der Apache-Lizenz zu veröffentlichen. Die Homepage von Tesseract findet sich unter code.google.com/p/tesseract-ocr. Wenn auch über das vergangene Jahr nicht immer klar ersichtlich wurde, ob und wie stark an der Software gearbeitet wird, so machte Tesseract im letzten Sommer einen gewaltigen Schritt nach vorne.

Mit der Version 2.0 ist es nun erstmalig möglich sprachspezifische Texte erkennen zu können. Mit der Version 2.0.1 können sogar Fraktur-Texte erkannt werden. War es am Anfang eher schwierig Tesseract unter Linux zum Laufen zu bringen, ist das heute relativ einfach machbar. Bei einigermaßen gut gescannten Seiten darf sich die Erkennungsgenauigkeit durchaus sehen lassen. Sie liegt etwa dort, wo kommerzielle Lösungen um 1998 bis 2000 lagen. Wobei hier angefügt werden darf, dass sich in Punkto Qualität in den letzten fünf Jahren nicht mehr sonderlich viel geändert hat.

3.3 OCR-Opus

Etwas nach Tesseract ist das Projekt OCR-Opus, siehe code.google.com/p/ocropus, entstanden. Im Unterschied zu Tesseract legt OCR-Opus das Augenmerk auf die Seitenanalyse und die Erweiterbarkeit (skriptföhrig) der Texterkennung. Im Moment arbeitet OCR-Opus nur sinnvoll mit Tesseract zusammen. Im brigen sei darauf hingewiesen, dass sich die Software im Alpha-Stadium befindet, eine stabile Version wird binnen eines Jahres folgen, sofern die Zeitplne eingehalten werden knnen.

3.4 Und sonst?

Die Auswahl an OCR-Produkten unter Linux ist noch immer karg. Ohne Tesseract bzw. OCR-Opus wrde sie gar kmmerlich aussehen. Dazu ein Beispiel: Einer unserer Kunde verlangte nach dem Einsatz einer Formularerkennung. Als wir ihm die Priesse kommerzieller Formularerkennungssoftware zeigte, fragte er uns, ob wir denn nicht eine solche Lsung implementieren knnten.

Offen gestanden muss ich eingestehen, dass mir die mathematischen Fertigkeiten fehlen, um eine solche Formularerkennungssoftware zu entwickeln. Einer unserer Partner, die Firma ExactCode in Berlin, siehe www.exactoce.de dagegen entwickelte fr uns eine Formularerkennungssoftware, welche in der Lage ist, auf einer Seite anhand eines Firmenlogos bzw. eindeutigen Merkmales eine Seite so vorzubereiten (Geradestellen), dass wir die gewünschten Blcke ,ot der ArchivistaBox extrahieren knnen und die einzelnen Blcke einer Texterkennung zufhren knnen. Der gesamte Sourcecode untersteht der GPL-Lizenz und ist pffanfertigauf der ArchivistaBox-CD enthalten.

Weiter gibt es derzeit keine schlanke Barcode-Erkennung. Vorhandene OpenSource-Lsungen bauen entweder auf .NET- oder Java-Frameworks auf. Das von uns verwendete Produkt zur Barcode-Erkennung (siehe www.softeksoftware.co.uk ist zwar keineswegs frei. Die Lsung kann aber insofern frei betrieben werden, als dass keine Runtime-Lizenzen auf der ArchivistaBox-CD anfallen. Sie darf daher von allen Interessierten auf der ArchivistaBox frei eingesetzt werden.

Natrlch wre uns eine reine OpenSource-Lsung wichtig, wir haben das Projekt aber auf das Jahr 2008 verschoben, ganz einfach deshalb, weil derzeit noch einige Erweiterungsanfragen fr die ArchivistaBox bestehen und wir unsere Krfte bndeln mssen. Ich hoffe nun einfach, dass wir zum 10. linuxday.at dann eine reinrassigefreie Barcode-Lsung in unser Produkt integriert haben werden.

4 Verein freearchives.ch

Es ist mir wichtig, dass dieser Vortrag nicht als Werbeveranstaltung für unsere ArchivistaBox verstanden wird. Trotzdem, wir benötigen für unsere ArchivistaBox eine OCR-Erkennung, die aktuellen Entwickler-Preise von FineReader scheinen uns jenseits von Gut und Böse und so haben wir uns Ende des letzten Jahres überlegt, ob und wie wir die Entwicklung einer OpenSource-OCR-Erkennung bzw. entsprechender Tools unterstützen können.

Dank einem Vertrag, den wir vor langer Zeit (genau genommen im Jahre 2000) mit der Firma Abby abgeschlossen haben, haben sich bei der Archivista GmbH über die Jahre hinweg eine stattliche Anzahl von OCR-Lizenzen für unsere Archivista-Produkte angesammelt. Weil wir in der Zwischenzeit nicht mehr mit Closed-Source-Software arbeiten möchten, d.h. unsere ArchivistaBox und sämtliche unsere Software, die wir entwickeln, unter der GPL-Lizenz stehen, haben wir Ende des letzten Jahres die 'übergebliebenen' Lizenzen dem damals gegründeten Verein freearchives.ch (siehe www.freearchives.ch) vermacht.

Der Verein hat mit Stichtag 22.12.2006 10000 Lizenzen der Archivista OCR-Option erhalten. Diese Lizenzen verteilt der Verein an seine Mitglieder. Weil der Verein zum Zwecke der Entwicklung von OpenSource-Produkten im Bereich der Dokumentenarchivierung gegründet wurde, kann dieses Ziel am besten erreicht werden, wenn die Mitgliederbeiträge zu 100 Prozent in die Entwicklung entsprechender Software fließen.

Im Moment stehen dem Verein 1200 Euro zur Verfügung, je nach Interesse bzw. Engagement der Vereinsmitglieder könnte es aber wesentlich mehr werden. Bei einem Vereinsbeitrag von 10 Euro pro Jahr (z.B. für minimal drei Jahre) würden dem Verein z.B. bereits 30'000 Euro zur Verfügung stehen, wenn 1000 Interessierte mitmachen würden. Immerhin erhalten die Vereinsmitglieder eine Software, mit der sich unter Linux unlimitiert durchsuchbare PDF-Dateien erstellen lassen. Die genauen Bedingungen betr. einer Mitgliedschaft können der Homepage von freearchives.ch entnommen werden.

5 Demo-Seiten fr OCR-Tests

5.1 Text, 1spaltig, 12 Punkt Helvetica

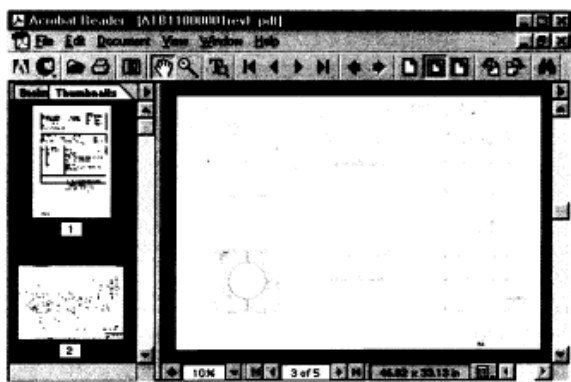
Fall 2: Engineering-Projekt mit PDF-Dokumentation

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)

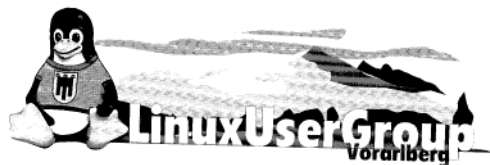
Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und dank des selber entwickelten **Produktmoduls Archivista TifToPDF** in der Lage, diese Wertschöpfungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendbaren PDF-File **kostengünstig** anzubieten. Das 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 Geschäftskorrespondenz erwartet werden darf.

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



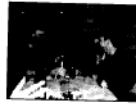
LinuxUserGroup Vorarlberg

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 Jugendherberge 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 Ansturm 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 geföhnt.

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.

5.3 Prospekt, Ausschnitt, 10 Punkt, Eras

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

Alle Dokumente, die in der Archivista-Box 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 GmbH
Postfach - CH-8042 Zürich
Tel: +41 (0)044 254 54 00
Fax: +41 (0)044 254 54 02
Web: www.archivista.ch
E-Mail: werner@archivista.ch

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

5.4 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 Fest-

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

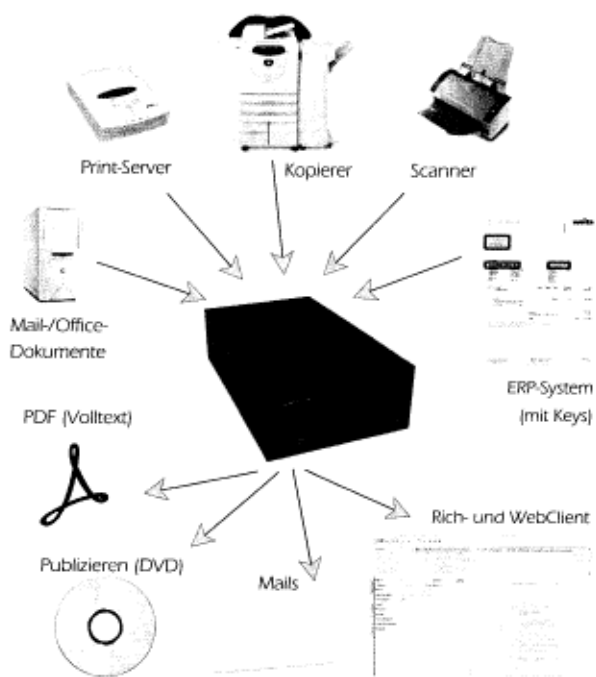
Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Titlis und Eiger gescannt werden kann. Ebenfalls

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

Und falls Sie nun einen Dokumenten-Cluster 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 Netzwerk-kabels. Jede Box ist fixfertig vor-konfiguriert; eine Installation 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 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.



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.

6 Archivista OCR-Option

Die kommerzielle OCR-Option wird seit Ende 2006 nicht mehr weiterentwickelt. Die Lizenzen wurden dem Verein freearchives.ch freundlicherweise von der Firma Archivista GmbH gespendet. Technologisch enthält die OCR-Option FineReader.

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 PDF-Format. PDF-Dokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auch als PDF-Dokument darstellbar. Problemstellung und Lösung

Ca. 75'000 auf der Vorder- und Rückseite bedruckte A4-Seiten sowie 200 zum Teil farbige Seiten. 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, Resultat: 150'000 Dateien

? Individuelles Scanning der 200 grsserformatigen 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 selbst entwickelten Scan-Dienstleistungen - Scanning und was dazu gehört, Fallbeispiele, Seite 2 Archivista GmbH
Jt

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

icKsKiiS;

Mii^Mk

LinuxUserGroup Vorarlberg

Wir sind ein loser Zusammenschluss von Linux-Anwendern, die sich etwa einmal pro Monat zwangsweise treffen.

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 Bilder-Galerie.

Homepage: www.lugv.at

www.linuxday.at

Geschichte

Die LUGV wurde im Frühling 1998 als loser Zusammenschluss von ca. 30 begeisterten in der Jugendherberge gegründet.

Nach einigen Treffen in der Jugendherberge wurde im Herbst eine Installationsparty im Frühstadium durchgeführt.

Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel Weies Kreuz die Idee geboren jährlich im
Die vollständige Geschichte der LUGV findet sich unter www.lugv.at

6.3 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerkes.
Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlagwortet (indexiert).
f Archivista GmbH

*t Postfach - CH-8042 Zrich -?s ' Tel: +41 (0)44 254 54 00 Fax: +41 (0)44 254 54 02
Web: www.archivista.ch ARCHIVISTA 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.
Rigi eignet sich primär für kleinere Umgebungen wie z.B. Rechtsanwaltspraxen oder PR-Bros. Aber auch für
Pilatus ist für mittlere Firmengrößen, oder besser ausgedrückt, für das mittlere Datenvolumen geeignet.
platte, zweiter Box und Tape- Laufwerk - ist für Archive ab ca. 500'000 bis einige Millionen Dokumente.
Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Eiger und Titlis
führen sie z.B. eine Text- und/oder Barcode-Erkennung durch.

Und falls Sie nun einen Dokumenten- Cluster für Mount Everest benötigen, auch kein Problem; wir haben
Rigi 1797 m..M. Einzelplatz-Dokumenten-Server mit bis zu 20'000 Akten und 100'000 Seiten
Pilatus 2132 m..M. Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten
Titlis 3238 m..M. Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten, redundant (2 Boxen)
Eiger 3790 m..M. Dokumenten-Server für unlimitierte Anzahl Akten, bis ca. 2 Mio Seiten, redundant (2 Boxen)
Mythen 1899 m..M. Scan-und OCR-Box, welche Daten zum Pilatus und Titlis transportiert
Rothorn 2351 m..M. Scan-und OCR-Box, passend zum Eiger

ArchivistaBox: Connect your world

MailOffice- Dokumente

ERP-System mit Keys)

Rieh- und WebClient

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerkes.
Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch beschlagwortet (indexiert).
Archivista GmbH

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

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

ARCHIVISTA


```

-- -
, -
_l .
..n rrl __!,
% .-
-,
..' .. '
- . - .\ .
t

_ i .
. - ...i ,
. . . . .
. . . . . _ .
_ . . . _ _ . . | | _ _
. . . . .
. .
_ . . _ - . . _ . . . : . _ .
\ . . . ' . 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
- . Individuelles Scanning der 200 grsserformatigen 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 TinoPDF in der Lage, diese Wertschöpfungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendbaren

_. U k ' r

nux _e roup Vor#rlberg

W.r sind ein loser Zusammenschluss von Linux-Anwendern, |' | _ __

die sich etwa einmal pro Monat zwanglos tre_en, um ber '''

verschiedeneThemen und selbstverstdlich ber Linux zu ,|_||

diskutieren. Zu diesenTrefren sind _ Linux-Interessierten lll,_|||'_ '|||', ''| ,|, IM_I

eingeladen, gleichgltig ob LUGV-Mitglied oder nicht llll , ,lh II , , ,_|'_ _ '||'' .. |

. - - ' , - ,

In unregelm_igen Abstnden veranstaltet die LUGV auch _ ' | _||

Ausflge, Vortrge und Workshops. '_ _' | ,

Auf der Homepage der LUGV finden sich neben Neuigkeiten ' , ,

aus der Linux-Szene auch eine Bildergalerie sowie eine

Mailingliste und ein Anmeldeformular fr die LUGV. Eine __ ,|||'''''''_

MitgliedschaW ist an keine Verpflichtungen gebunden _| '''''''.b_ , ,

sondern dient hauptsächlich der Organisation des LinuxDays. , _ __, , , , | |,

, | __, _ _ , | i_ | _ |' '|_i_ , ' | ' |/' ' ._, _'

,_||,,,_|,_!|_|||||,,'_|||_|_','|_''''''__

Homepage. www lugv at __|,|,|,, ,,, , ' |, _-----'' __ | ' ' , , , , - - -

www:linuxday.at _,_ ,lm '. i''''''''| ''''

|_,_||_i_|_ i_i , ' ' _ _

__||| ' , , , , , , , |__||_lli,_|,' 'i_|_ | , , , | _ _ _ | , | , _||| , |

Geschichte

Die LUGV wurde im Frhling |998 als loser Zusammenschluss von ca. 30 begeisterten in der Jugendherberge in Feldkirch gegrndet.

Nach einigenTre_en in der Jugendherberge wurde im Herbst eine Installationsparty im Frhstck der JH organisie_. Der Andrang war so gro^, dass einige ihre PCs nicht mehr aufstellen konnten. Im Frhling |999 stellte uns dieVKW in Bregenz einen gro_en Raum zurVerfgung, in welchem die 2. Installationsparty organisie_ wurde. Es waren ber 100 Linuxbegeisterte, welche mit ihren PCs den Raum fllten. Die VKW war auf diesen Anstrum nicht vorbereitet, da nur ber das Internet Werbung fr die Installparty gemacht wurde. Das Stromnetz musste vestrkt werden. Die Warteze

mit einer Zwischenverpflegung, o_erie_ von derVKW, berbrckt. Danach wurde bis in den spten A
dem Hobby gefrnt.

Nach diesem tollen Erfolg wurde im Mai 1999 im HotelWei^es Kreuz die Idee geboren jhrlich in
einen LinuxDay (www.linuxday.at) zu organisieren.

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

7.3 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 fixrerlig vor-
konfiguriert; eine lnst_ll_tion ist nicht
nonNendig. Ob per FTF-Filetr_nsfer
IKopierger_tel, _nges<hlossenene
Sc_nnern, oder Druckvorg_ng, die
Archivist_Box zeigt sich _usserst
kont_kIfreudig bei der Dokumenten-
_nn_hme.

Alle Dokumente, die in der Archivliste
Box eingetragen, werden automatisch
beschäftigt und indexiert und stehen
den unmittelbaren Mitarbeitern zur Verfügung.
Der Zugriff über die Web- oder mobile
Client. Dokumente können über
das FDF oder per Mail weitergeleitet
werden. Ganz nach dem
Motto: Connect your world.

Archivista GmbH

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

Web: www.archivista.ch
AR_HIVISfA E-Mail: wPbmaster_archivista.ch

7.4 Prospekt, Seite, 10 Punkt, Eras

Je nach Bedarf die richtige

Sie helfen kigi, Flüt, Titlis, Eiger, platte, zweite Box und Tpe rhen sie z.B. eine T

Mythen und Kothorn - und sind L_uhNerk - ist rr Archive _b c_. B_rcode-Erkennung durch.

punkto Unerschlichkeit und soo'ooo bis einige Millionen Seiten

Stabilität gen_uso robust wie die _usgelegt. Und falls Sie nun einen Dokumenten-

gleichmigen geogr_rischen Erhe Cluster _ |_ Mount Everest be

bungen. Die Performance der Mythen und Kothorn sind Sc_n- und ntigen, _uch kein Problem wird

Document-Server _sst sich von der O_k-Illustration, mit welchen Stellen Ihnen diesen mit

relativen Höhe, die mit dem Berg- _ur die Boxen Fil_tus, Titlis und Eiger entsprechenden H_r

nomen _ssoziiert ist, _bleiten. gesc_nnt werden k_nn. Ebenfalls individuell zusammen.

kigi eignet sich prim_r für kleinere

Umgebungen wie z.B. rechts_n- nigi | 797 m..M. Einzelplatz-Dokumenten-Server mit bis zu

Wirtspr_xen oder Pk-Bros. Aber 20'000 Akten und 100'000 Seiten

uch größere Unternehmen, die für Flhtu 2|_2 m..M. Dokumenten-Server für bis zu 200'000 Ak-
Abteilungen PDF-Dokumentenserver ten und 1 Mio Seiten

suchen, sind mit der kigi-Box gut Ttli_ 3238 m..M. Dokumenten-Server für bis zu 200'000 Ak-

bedient. ten und 1 Mio Seiten, redund_nt 12 Boxen

fil_tus ist für mittlere Firmengrößen, Eiger 3790 m..M. Dokumenten-Server für unlimitierte An

oder besser _usgedrckt, rr d_s Akten, bis c_. 2 Mio Seiten, redund_nt

mittlere Datenvolumen ged_cht. Die 12 Boxen und mit B_<kup-Tape-Drive

Titlis-Box ebenfalls, _llerdings ergibt Mythen 18_9 m..M. Sc_n- und Ock-Box, welche Daten zum

die redund_nte Hardware ein Fil_tus und Titlis transportiert

erhhtes Slcherheitselement. Die nothorn 23S1 m..M. Sc_n-und Ock-Box, p_ssend zum Eiger Eiger-Box - mit entsprechender FesI-

ArchivistaBox: Connect your _orld

D_mit eine Archivist_Box mit der

AAussenwelt kommunizieren k_nn,

/_ ,. ,\

' _ _ , _ _ ' , _ _ , bed_rf es einzig eines Netzwerk-Sm_ ' , , _ _ _ _ _ ' k_bels. Jede Box lst rixrertig vor-

___/____!__konfiguriert; elne lnst_ll_tion lst nlcht

"'_ , ; _ , _ , _ _ i" .'_ '_ " _ .' __ nonNendig. Ob per FTF-Filetr_nsrer

'_ "' __lh__ , __ , _ , IKopierger_te), _ngeschlossenen

Frint-5ewer Kopierer 5<_nner 5c_nnern, oder Druckvorg_ng, die

__Archivist_Box zeigt sich _usserst

_ __'_ _ | m kont_ktrreudig bei der Dokumenten-

_ ' _ _nn_hme.

_ha

_ - =-_-_- _ _! - Alle DokumenIe, die in der Archivist_-

M_il-/Orrrice- _ -. _ . : Box eintrerren, werden _utom_tisch

Dokumente bes<hl_gwortet lindexiertl und ste

-.. . hen unmlttelb_r rur eine kecher<he

EkP-System zur Verrgung. Der Zugritr _ur die

PDF IVolltext) lmiI Keysl Box errolgt uber den Web- oder kich-

<lient. Dokumente knnen _ber

__uch _ls rDF oder per M_il wei-

__tergereicht werden. G_nz n_ch dem

kich- und WebClient Motto: Connect your world.

rublizieren IDVDl _ .. _ ''..: _ ' _ . ' . ' .

M_ils _____ ' _ Archivista GmbH

0'''' _: _.. q'__ _ _ '= PosWach - CH-8042 Zrlch

, _ __!'_: . '0 Tel: +41 (0)44 254 54 00

___L' Faw: +41 (0)44 254 54 02

_, ' _ _ '

_ . Web: www.archlvista.ch

_ --_ . . _ _ ARCHIVISTA E-Mail: webmasler_archivista.ch

8 Tesseract 2.01

Tesseract kommt mittlerweile mit Umlauten relativ gut zurecht. Einzig bei schlechten Scans bzw. nicht ausgerichteten Seiten kann die Erkennungsgenauigkeit leiden.

8.1 Text, 1spaltig, 12 Punkt Helvetica

Fall 2: Engineering-Projekt mit PDF-Dokumentation

Keine 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. Linux 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.

```
'hnrmlml lleutlcr [AIU] Ifllllllllllllurvl pull]
```

```
i f s ,
```

```
W ' V .
```

```
[F E3,,...:' _ _ _
```

```
...4. ~l~< 'i< 3**r
```

Archivista löste dieses Problem Schritt für Schritt:

Duplex-Scanning der 75'000 A4-Seiten mit Hochleistungsscanner mit 300dpi,

Ergebnis: 150'000 Seiten im TiffG4-Format

Individuelles Scanning der 200 großformatigen 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 ihrer tiefen Kenntnis der aktuellen Technologien und dank des selbst 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 JQ;

Box erfolgt über den Web- oder Rich
Client. Dokumente können aber
auch als PDF oder per Mail wei-
tergereicht werden. Ganz nach dem
l\lotto: Connect your world.
Q, Archlvista GmbH
i, "i<, . _il Postlach-CH-8042 Zurich
5: _ {yp ' T&|:+41(0)442545400
_ Fa><;+41 (0)44 254 54 02
Web; www.archivista.ch
A\(\}|\|\|\\$'|'A E-Mau; wmasmr@arcnavisia.n

8.4 Prospekt, Seite, 10 Punkt, Eras

Für jeden Bedarf die Richtige

Sie heissen Rigi, Pilatus, Titlis, Eiger, platte, zweiter Box und Tape führen sie z.B. eine
Mythen und Rothorn und sind Laufwerk ist für Archive ab ca. Barcode-Erkennung durch.
punkto Unerschütterlichkeit und 500000 bis einige Millionen Seiten _
Stabilität genauso robust wie die ausgelegt. Und falls man nun Gmex DOKUMM
gleichnamigen geografischen Erde _ Ousmr 3 Ia Mmm E'/Nest DEL
mmgen Die Pmfmmang dg Mythen und Rothorn sind Scan- und ntigen, auch kein Problem; wir
DOWmm Sn/Gr mm Sim VOM dg OCR-Cluster-Stationen, mit welchen stellen Ihnen diesen mit der
relamen Hhg me mit dem BU? auf die Boxen Pilatus, Titlis und Eiger entsprechenden Hardware g
namen EBOZMH M ablmn gescannt werden kann. Ebenfalls individuell zusammen.

Rigi eignet sich primär für kleinere

Umgebungen wie z.B. Rechtsan- Rigi I797 m.d.M. Einzelplatz-Dokumenten-Server mit bis zu
waizspraxen oder PRBdros. Aber 20'000 Akten und 100'000 Seiten

300* 9V6\$\$1 Unternehmen Clie tilt params 2132 m.0.ivi. ookumenien-sewer rm bis zu 200000 Ak-
Abteilungen PDF-Dokumentenserver tgp Und 1 M50 ggitgn

SUCHn sind mit GU MQFBOX gut Titlis 3238 m.d.M. Dokumenten-Sewer für bis zu 200'000 Ak-
Dedien ten und 1 Mio Seiten, redundant (2 Boxen)

Pilatus ist für mittlere Firmengrossen, Eiger 3790 m.d.M. Dokumenten-Sen/er für unlimitierte
oder besser ausgedrückt, für das Akten, bis ca. 2 Mio Seiten, redundant

mittlere Datenvolumen gedacht. Die (Z B><el Und mit B@l<UPT6PDV ive

Titlis-B<>>< ebenfalls si allerclinqs erqibt Mythen iam m.0.ivi. scan- tm oci2-BX, Weime umn z
die redundante Hardware ein Pilatus und Titus transporrien

Gmms Simemenselemem DE Rcthorn 2351 m.d.M. Scan- und OCR-Box, passend zum Eiger
EigerBox mit entsprechender Fest-

ArchivistaBox: Connect your world

9 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, abgefordert 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 Weitsch, 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 emporragen. Die wichtigsten der letzteren auf unserem Staatsgebiete sind der gewaltige, von krystallinischen Schiefergesteinen umgebene Granitstock der Cima d'Alta 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:

P Wir werden zunchst die erstere zu betrachten haben, in den Nebenzonen aber weiter noch mehrere Specialgebiete, und zwar die Grauwackenzone, die nrdlichen Kalkalpen, die Wiener Sandsteinzone und die sdlichen Kalkalpen, denen sich die aus unser Staatsgebiet fallenden Gebirge des Baltansystems unmittelbar anschlieen, 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 ber die krystallinischen Massengesteine weitaus vorwalten. Die Grenzlinie brigens, welche dieselben von den Sedimentgesteinen scheidet, stimmt nicht berall genau mit jener s berein, welche man vom orographischen Standpunkte zwischen den Centralalpen und den s Kalkalpen gezogen hat. So finden wir beispielsweise aus der Karte Seite 27 die Gruppen des Hochschwab und der Veitsch, die aus mesozoischen Kalksteinen bestehen, noch der Centralzone zugezht, andererseits 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 sdlichen Nebenzone vereinigt, und analoge Abweichungen ergeben sich auch an anderen Stellen. Auch inag hier gleich hervorgehoben werden, da, wengleich die Centralzone das eigentliche Herrschfeld der archaischen und die Nebenzonen jenes der Sedimentgesteine bilden, sich doch einerseits betrchtliche Massen der letzteren, ein manchen Stellen der mittleren Kette, in isolirten Schollen ber den krystallinischen Gesteinen vorfinden, wie z. B. an der Landesgrenze in den Ortler Alpen oder am Brenner, oder endlich aus der zu den steirischen Alpen gehrigen Stangalpe, und da andererseits an manchen Stellen der sdlichen, nicht aber auch der nrdlichen 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 dAsta in Sdtirol, der schmale Zug von Glimmerschiefer, welcher der Einsenkung des Gailthales in Krnten folgt, im Westen aber mit der Centralzone doch in Verbindung steht, und ein hnlicher langer und schmaler Zug von krystallinischen Schiefer- und Massengesteinen, der stidlich von der Karavankenkette, Q den Lngsthlern der Mi und Javoria entlang, sortstreichet.

So wenig wie in der Bodenplastit, ebensowenig zeigt sich auch in der geologischen Zusammensetzung im Gebiete der Mittelzone eine regelmige, dem westtlichen Hauptstreichen des ganzen Gebirges folgende Anordnung. Hier wie in anderen Gebieten hat man erkannt, da von den drei Hauptarten der krystallinischen Schiefergesteine der Gnei das tiefste und ltteste, der Glimmerschiefer das nchst jngere und der Thonschiefer das jngste Gebilde ist. Keines dieser Gesteine aber erscheint, der ganzen Erstreckung der Centralkette

Hinweis: Mit der ArchivistaBox kann die Fraktur-Erkennung direkt aufgerufen werden, indem in WebAdmin bei den OCR-Definition bei der Sprache 'DeutschNeu' bzw. 'GermanNewSpelling' sowie als OCR-Engine 'Tesseract 2.0' gewhlt wird.

10 Formular- und Barcodeerkennung

10.1 Formulare mit ExactImage

Bei der Formularerkennung mit der ArchivistaBox muss zunächst ein eindeutiges Merkmal auf der Seite gescannt und aufbereitet (primär geradegestellt) werden. Dieses Bild wird anschließend als Logo hinterlegt. Beim Scannen findet die Software anschließend das Logo auf der Seite und kann aufgrund des Rotationswinkels zum geradegestellten Logo berechnen, um wieviele Grad die Seite geradegestellt werden muss und wo auf der Seite sich der Nullpunkt (Logoposition) befindet.



The image shows a scanned invoice header for Hommel Hercules Werkzeughandel. The logo 'HHW' is in the top right corner. The company name 'Hommel Hercules Werkzeughandel' is on the left, and 'Werkzeuge und Werkzeugmaschinen' is on the right. Contact information includes phone and fax numbers. Tax information includes USt.-Id.Nr. HHW. (DE 146279926) and USt.-Id.Nr. Kunde. (STEUERFREI DRITTLAND). The invoice number is 531352 and the customer number is 80311. A note at the bottom says 'Bei Zahlung und Rückfragen bitte stets angeben'.

| | | | |
|--|---------------------|--|--|
| Hommel Hercules Werkzeughandel | | HHW | |
| | | Werkzeuge und Werkzeugmaschinen | |
| HOMMEL HERCULES | Telefon | (06204) 739-0 | |
| HEIDELBERGER STR. 52 DE - 68519 VIERNHEIM | Telefax | 06204/739222 | |
| Firma | USt.-Id.Nr. HHW.: | DE 146279926 | |
| HHW SCHWEIZ AG | USt.-Id.Nr. Kunde.: | STEUERFREI DRITTLAND | |
| RECHNUNG NR. | | 531352 | |
| | | Kd.Nr.: 80311 | |
| Bei Zahlung und Rückfragen bitte stets angeben | | | |

10.2 Barcode-Erkennung

Die Barcode-Erkennung auf der ArchivistaBox ist zwar nicht OpenSource, kann aber auf der ArchivistaBox frei eingesetzt werden. Die Barcode-Erkennung wird direkt beim Scannen durchgeführt, d.h. die Erkennungszeiten liegen im Schnitt ca. bei 0.1 bis 0.2 Sekunden.

Dies ist auch der Grund, weshalb entsprechende Alternativen, welche es unter OpenSource gibt, bisher nicht implementiert wurden.

11 Abschliessende Bemerkungen

Wer vor einem Jahr mit einer OpenSource-Texterkennung arbeiten wollte, der scheiterte klglich. Heute kann die Antwort differenziert gegeben werden. Wer schwierige Vorlagen hat, der wird nach wie vor erhebliche Schwierigkeiten haben, um auch nur in irgendeiner Form einen halbwegs brauchbaren Text zu erhalten. Bei einfacheren Vorlagen dagegen kann Tesseract 2.01 durchaus mithalten. Ocrad dagegen kann allenfalls verwendet, wenn es darum geht, nur Zahlen zu erkennen, da Ocrad einen entsprechenden Filter besitzt.

Wer derzeit keine Sorgen mit der OCR-Erkennung haben möchte und damit leben kann, dass die Sourcen nicht offengelegt sind, dem steht für einen bescheidenen Betrag von Euro 30.– bzw. sFr. 50.– eine Mitgliedschaft bei freearchives.ch offen.

Ich gebe zu, dass hier die 'Werbeabteilung' des Vereins spricht, auf der anderen Seite fließt der gesamte Betrag in die Entwicklung einer guten OpenSource-OCR-Texterkennung. Bei 10'000 verfügbaren Lizenzen ergäbe das immerhin den Betrag von Euro 300000. Ob das reicht, eine zu kommerziellen Texterkennungspaketen gleichwertige OCR-Erkennung zu entwickeln, diese Frage kann hier nicht beantwortet werden. Einen guten Beitrag leistet die Summer aber ganz sicher und wer will darf auch mehr spenden.

Und Deine/Ihre Spende ist ja nicht ganz unntz, denn nochmals, wo gibt es für Euro 30.– bzw. sFr. 50.– eine Software, mit der unbeschränkt OCR-Erkennung im professionellen Umfeld möglich ist. Professionell heisst z.B. auch, dass ca. 100 Sprachen zur Verfügung stehen oder dass eine unlimitierte Anzahl von durchsuchbaren PDF-Dateien erstellt werden kann.

Offen gestanden, dem Autor ist keine Software (auch nicht unter Windows) bekannt, die das zu diesem Preis könnte. Im brigen wird die Zukunft zeigen, ob mit Tesseract und OCROpus zwei neue Sterne am Texterkennungshorizont entstehen. Die Chance ist heute (2007) noch nie so gross gewesen. Und das wiederum heisst, dass ein Nebeneffekt auch sein könnte, dass die kommerziellen Vertreter gezwungen sein könnten, ihre Preismodelle in nicht allzu ferner Zukunft massiv anpassen zu müssen.