OCR- und Formularerkennung mit Linux

Übersicht über Projekte und Vergleich zu ClosedSource-Produkten

Contents

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 anfügen darf, dass er allgemein verständlich gehalten ist und das Programmierkenntnisse an keiner Stelle gezückt 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 Erfahren mit Texterkennung bei einem staatlichen Betrieb (EMPA Dübendorf) 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 Analge 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-Gerät) selber nochmals ca. 15'000 Euro.

Das Produkt, das damals in aller Munde war, hiess OmniPage, ein weiteres sehr gutes Produkt Recognita. Umgefähr 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 Archiv-Lösung 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 Qualität her nicht ganz an die Konkurrenten herankam.
- Wocar (siehe www.simpleocr.com: Ein französischer Programmierer entwickelte eine Freeware-Applikation, welche Texte in Französich und Englisch in ganz akzeptabler Qualität erkennen konnte.

Allen Applikationen gemeinsam war, dass sie unter Windows liefen (im besten Fall für den Mac verfügbar 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 gegenüber den übrigen Produkten weit besser war. Mit dem Aufstieg an die 'Weltspitze' änderte die Firma Abbyy, welche FineReader entwickelt, aber radikal das sogenannte Business-Modell. Nicht primär im Consumer-Markt, sondern bei den Entwickler-Produkten.

Etwa im Jahre 2000 wurden die Preise für 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 höheres Mehrfaches angehoben.

Ursprünglich wurden die Preise auf der Homepage von ABBYY veröffentlich. Leider werden die aktuellen Preise seit Jahren nicht mehr veröffentlicht. 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. Natürlich hat nicht jeder Millionen von Seiten, die er scannen möchte. Wenn es aber darum geht, alte Buchbestände 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 GOCR (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 Textbläcke 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 schnittt 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 2.01 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 (skriptfähig) der Texterkennung. Im Moment arbeitet OCR-Opus nur sinnvoll mit Tesseract zusammen. Im übrigen sei darauf hingewiesen, dass sich die Software im Alpha-Stadium befindet, eine stabile Version wird binnen eines Jahres folgen, sofern die Zeitpläne eingehalten werden können.

3.4 Und sonst?

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

Offen gestanden muss ich eingestehen, dass mir die mathematischen Fertigkeiten fehlen, um eine solche Formularerkennungssoftware zu entwicklen. Einer unserer Partner, die Firma ExactCode in Berlin, siehe www.exactoce.de dagegen entwickelte für 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 Blöcke ,ot der ArchivistaBox extrahieren können und die einzelnen Blöcke einer Texterkennung zuführen können. Der gesamte Sourcecode untersteht der GPL-Lizenz und ist "pfannenfertig" auf der ArchivistaBox-CD enthalten.

Weiter gibt es derzeit keine schlanke Barcode-Erkennung. Vorhandene OpenSource-Lösungen 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 Lösung 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.

Natürlich wäre uns eine reine OpenSource-Lösung wichtig, wir haben das Projekt aber auf das Jahr 2008 verschoben, ganz einfach deshalb, weil derzeit noch einige Erweiterungsanfragen für die ArchivistaBox bestehen und wir unsere Kräfte bündeln müssen. Ich hoffe nun einfach, dass wir zum 10. linuxday.at dann eine "reinrassige" freie Barcode-Lösung 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 Abbyy 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 'übriggebliebenen' Lizenzen dem damals gegründeten Verein freearchives.ch (siehe www.freearchives.ch vermacht.

Der Verein hat mit Stichdatum 22.12.2006 1000 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 Bereiche der Dokumentenarchivierung gegründet wurde, kann dieses Ziel am besten erreicht werden, wenn die Mitgliederbeitrage 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 (fällig 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 für OCR-Tests

5.1 Text, 1spaltig, 12 Punkt Helvetica

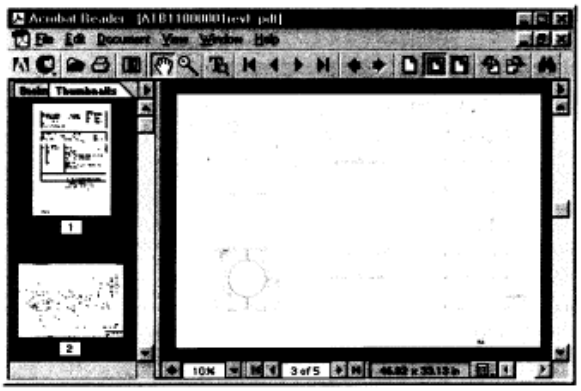
Fall 2: Engineering-Projekt mit PDF-Dokumentation

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:

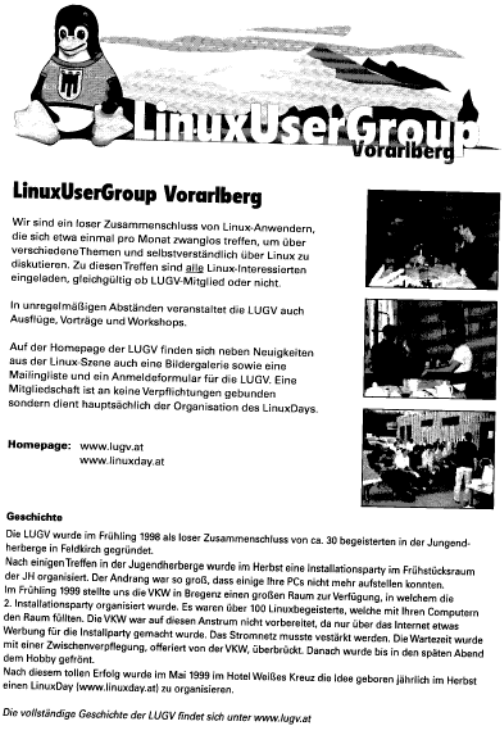
- **Duplex-Scanning** der 75'000 A4-Seiten mit Hochleistungsscanner mit 300dpi, Resultat: 150'000 Seiten im **TiffG4**-Format
- Individuelles Scanning der 200 **grösserformatigen** Belege, davon 50 **in Farbe**
- **Umwandlung** der TiffG4 und JPG-Dateien in **PDF**-Dateien
- Erstellen der **CDs mit Inhaltsverzeichnissen** (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 Geschäftskorrespondonz erwartet werden darf.

5.2 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.

5.3 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 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

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.

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.

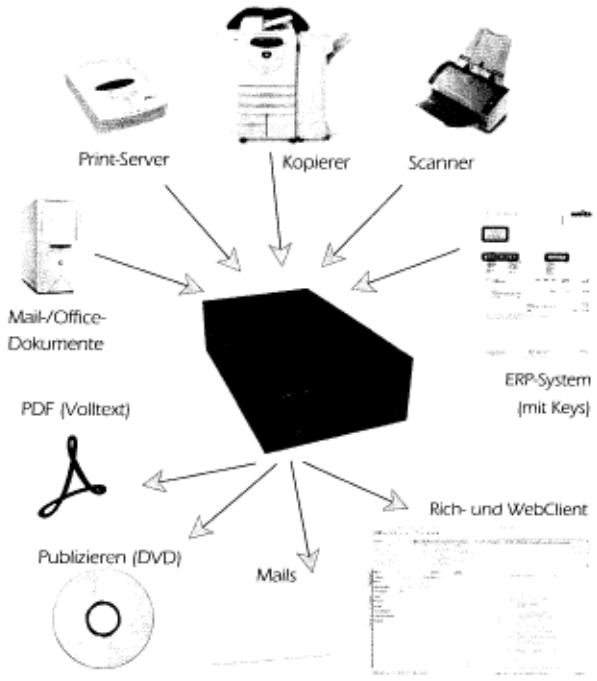
Pilatus ist für mittlere Firmengrössen, oder besser ausgedrückt, für das mittlere Datenvolumen gedacht. Die Titlis-Box ebenfalls, allerdings ergibt die redundante Hardware ein erhöhtes Sicherheitselement. Die Eiger-Box – mit entsprechender Festplatte, zweiter Box und Tape-Laufwerk – ist für Archive ab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

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-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

6.2 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

6.3 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

6.4 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

7 Ocrad 0.17)

Ocrad liefert bei der Formularerkennung mitunter gute Dienste. Dagegen hat Ocrad bei einigen Zeichen gewisse Probleme.

7.1 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 FormT 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\___ _ ..'_'.' _ ^ . '?_ _________,,`
__ _ '___ ' '_ __w , rf ` _ ___ _. ./__
_ _ _
, . _, , , , ,_ _, _ ___ _w_ __ __ ' __ , '_ _ F ____/ _.
, . , ,, ,_ . _ _ _ __ ___,__, y__ , _ _ ,_____ _ _,,,_ , __, _/_ ,_' ' _ _ !_ _
r|_e__|_____/__,_ ___^,,_ _ l., , | . .__ _ _ ., , h , _ ,,
__ _ _ _ , "' _ ' __ _ _, _5`_ __'`' _ ,__ _ _
_ _ , _ , , _ _ _ _ __rt _, ,_ _ _ y,_, . _ _, ,_ , . . , ,
, _ _ _
```

__ _
' _
_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

. 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 Produktmodul5 Archivista TinoPDF in der Lage, diese Wertschö 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

--- -
-

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

a_'h__,|||||,,|,,,,,,,,,||||,|',|||,||||,_.|_|'|||,|''|'',',,,,,,,,,,_,,,_,,,,,,
_ , _ |_ , , , | | | , , _ . ,, _, | , | , | , ,,,,
'_''||'|||,,___'_|'||||||¡||',|,|,'|''_'.lr'__'''''''''''_''|'||¡,,||||''||||||||
,|__'||'|||||||,|'|||,,,,.._.___|||||||,,__,
' ' ' ' '' ' ' ' ||||| _ '_'w_,',,, '._, , _'_ ' __` '' '_ ' .' _ ' ' i'' _ _' _

' '¡i__ _ _ _|'',_ __7,___ ___ __ _ _'__ ,,,
, _ _ _,i,il_'_ _ , _ ,w, , _|, |
|_|_|____¡|'____¡,li_llí___,_,,_||||__,
, | _,___ , _i_______'_ | _|_'
_|||||_|¡,____||llil___¡,¡|_í_'_'____i_'
'i_w'i'i_i_\|'|'|__||'||||,||||||
,,,,, | ' " ' ' | | | / '

_ 'r _ ''
|||___

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

_. 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 über '''

verschiedeneThemen und selbstverständlich über Linux zu ,|_||

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

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

ln unregelmä_igen 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 __ ,|||||| '''''' '_

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 Frühling |998 als loser Zusammenschluss von ca. 30 begeisterte

herberge in Feldkirch gegründet.

Nach einigenTre_en in der Jugendherberge wurde im Herbst eine lnstallationsparty

der JH organisie_. Der Andrang war so gro^, dass einige lhre PCs nicht mehr aufs

lm Frühling |999 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 |999 im HotelWei^es Kreuz die ldee gebore

einen LinuxDay (www.linuxday.at) zu organisieren.

Die vollständige 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 fixrerIig vor-

konfiguriert; eine lnst_ll_tion ist nicht

nonNendig. Ob per FTF-Filetr_nsfer

IKopierger_tel, _nges<hlossenen

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 rúr eine kecherche

zur Verfügung. Der Zugrirr _ur dle

Box errolgt úber den Web- oder kich-

Ilient. 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 54 02

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

7.4 Prospekt, Seite, 10 Punkt, Eras

rür je_en Bedarf die Áchtige

5ie helssen kigi, Fll_tus, Titlis, Eiger, pl_tte, zweiter Box und T_pe rühren sl

Mythen und kothorn – und sind L_uhNerk – ist rür Archive _b c_. B_rcode-Erkennun

punkto UnerschütIerlichkeit und soo’ooo bis einige Millionen Seiten
5t_bilit_t gen_uso robust wie die _usgelegt. Und f_lls 5ie nun elnen Dokumenten-

gleichn_migen geogr_ri5<hen Erhe Cluster _ |_ MounI Everest be

bungen. Die Ferform_nce der Mythen und kothorn sind Sc_n- und nötigen, _uch kein

Document-5ewer |_sst sich von der O<k-Iluster-St_tionen, mit welchen stellen lhn

rel_tiven Höhe, die mit dem Berg- _ur die Boxen Fil_tus, Titlis und Eiger entspr

n_men _ssoziiert ist, _bleiten. gesc_nnt werden k_nn. Ebenr_lls dividuell zus_mm

kigi elgnet sich prim_r für klelnere

Umgebungen wie z.B. kechts_n- nigi | 797 m.ü.M. Einzelpl_tz-Ookumenten-5erver mi

w_ltspr_xen oder Pk-Büros. Aber 20’000 Akten und TOO’OOO Seiten

uch grössere Unternehmen, die fur Fílhtu 2|_2 m.ü.M. Dokumenten-Server für bis
Abteilungen PDF-DokumentenseNer ten und l Mio Seiten

suchen, sind mil der kigi-Box gut Títli_ 3238 m.ü.M. Dokumenten-Sewer für bis zu

bedient. ten und l Mio Seiten, redund_nt l2 Boxenl

ril_tus ist fur mi_tlere Firmengrössen, Eiger 3790 m.ü.M. Dokumenten-5erver für

oder besser _usgedrückt, rür d_s Akten, bis c_. 2 Mio Seiten, redund_nt

mlttlere D_tenvolumen ged_cht. Die l2 Boxenl und mit B_<kup-T_pe-Drive

Titlis-Box ebenr_lls, _llerdings ergibt Mythen 18_9 m.ü.M. 5c_n-und OCk-Box, wel

dle redund_nte H_rdw_re ein Fil_tus und Titlis tr_nsportiert

erhóhtes Slcherheitselement. Die nothorn 23S1 m.ü.M. Sc_n-und OCk-Box, p_ssend z
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-

___/_____!__konriguriert; 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.
¡_h÷*a*

_ - =-_.-_- ._÷! ± −± *Alle DokumenIe, die in der Archivist_−*

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

Dokumente bes<hl_gwortet lindexiertl und ste
-.. . hen unmlttelb_r rur eine kecher<he

EkP-System zur Verrügung. Der Zugritr _ur die

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

<lient. Dokumente können _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

O’‘"’ _: _._ q‘__ _ _ ’= PosWach - CH-8042 Zúrlch
, _ __!’_: . ’O 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
I<aum eine andere Technologie hat im letzten Jahr fur mehr Furore gesorgt als da
Format. PDF steht fur Portable Document Format und wie der Name besagt, geht es
darum, lnhalte digital so aufzubereiten, dass diese auf unterschiedlichen Rechne
l\/lac und Windows) betrachtet und ausgedruckt werden konnen.
PDFDokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auch
PDF-Datei (in digitaler Form) publizierbar. Zudem sind PDFDateien mittlerweile w
breitet; der Viewer (Betrachter) fur die Dateien ist kostenlos. Die l\/loglichke
Formats machte sich eine internationale, im Bereich Engineering tatige Firma zun
Problemstellung und Lésung
Ca. 75'000 auf der Vorder- und Puckseite bedruckte A4-Seiten sowie 200 zum Tell
Seiten im A3- bis A0-Format sollten gescannt und in ein PDF-Format uberfuhrt wer
waren selbsttragende CDs mit PDF-Dateien, welche die ursprungliche Dokumentennum
aufweisen sollten und uber die mit sogenannten Thumbnails ein Uberblick besteht.
`hnrmlml lleutlcr [AIU] Ifllllllllllurvl pull]
i f s ,
W ' V .
[F E3,,..:.' _ _ _
...4. ~l~< `i< 3**r
Archivista Ioste dieses Problem Schritt fur Schritt:
Duplex-Scanning der 75'000 A4Seiten mit Hochleistungsscanner mit 300dpi,
Flesultat: 150'000 Seiten im TiffG4rFormat
lndividuelles Scanning der 200 grésserformatigen Belege, davon 50 in Farbe
Umwandlung der TiffG4 und JPGDateien in PDFDateien
Erstellen der CDs mit Inhaltsverzeichnissen (inkl. Link auf entsprechende Datei)
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 zuganglichen und versendba
PDF-File kostengunstig anzubieten. Das Engineeringllnternehmen verfugt heute ube
saubere, platzsparende Dokumentation des Projektes.
Scan-Dienstleistungen Scanning und was dazu gehort, Fallbeispiele, Seite 2 JQ;
```

Archivista GmbH, 8042 Zurich, Tel. 01 350 46 74, Fax 01 350 46 72, www.archivist

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

V V V y$yy$$$]$y]$]$$$yy]]]yVvvyyvvyyyy VV V V y]$$Vj$Vjy $y5$y $yy$$yy$$yyyyy
Q V ~·i =~ ~~*i iii;~1~ii=1:2~i~1i122~ii5;MV1%s:V:;ViPs4s$;?~~i;4JweFai;111iizsss
VVV. V VV `iii~·ii 1 ~·· 1 *~~`i*i·t=·· · * iiiiiiiiiiiii1iiiiililiiiiiiiiwiiiiiiiV
M VVVVVVVV VVV VVVVV VVVVVVVVVVVVVVVVW i iiiiiiiiiiiiiiiiiiii V ,VV,
~`iiiis"·i ;i~P1iiiiiiFiiiii=~ ~ i"i`iiiiiiiiiiiiiiiiiii1111111111iii`
iiiiiii i~i 1 V 1* 1 imiiiiiiiiiiiii iitt~t~r~ti
Vqiiiiiiiii Vi1iiiiiiii?aiiiiiigiiiii i V V
ii V V V V
VVViii11111111111it VVV V 1 V VV V V * t 1 1 V V VV
VVVV V VVVVVVVV ~ iiiiii`` i ii VV "w Hit V V V
F ~VV V ` ` M M M M m ""lm iiiiiii, iii ;ri;31;j1~~ W ii ·
~~ ·V~~VVVV iVVVVVVVVV_ V VVV V sm VVVVV.VVVVV V VVVV.VVVVVV i VVVVV Vi VVVVVVV
Mw i W · · I · – i*· t
O V
lmuxl|serGrup Vorurlherg ~ ~
Miiiiiiiiii
Wir sind ein loser Zusammenschluss von LinuxAnwendern, V """*%pV in V
die sich etwa einmal pro Monat zwanglos treffen, um Uber VV ~ i– 34* "
verschiedeneThemen und selbstverstandlich Uber Linux zu iVVVVi &iiM1~V"ii·_` W
· · · – – · V =· 1**9i;V;`ViiV. ii;i ~ in
diskutieren. Zu diesenTreffen sind & LinuxInteressierten VVVViVii==~~~ V_V ~ii ~
eingeladen, gleichgiiitig ob LUGVMitgIied oder nicht. ` **5 ~
V V V;Vi;y3?iI`!l;iVis1"~i~1 V W
In unregeImaBigen Abstanden veranstaltet die LUGV auch HF ViVV~VVV,i;· ii"i"""`"
.. .. V ii V ~`~~.~· ii ~~ii i V ii iii
Ausfluge, Vortrage und Workshops. ii i Q " ` V ,,V;Vwii, Vi. `iii**1iii1*i
iii 1 1 ° *1 iiii ii i*i wi ~Viiiiiiiiimiiiiiiiiii
. . . . V· i wasrFi=i:ii==~ ~i`````` VVVVVVVVVVVVV
Auf der Homepage der LUGV finden sich neben Neuigkeiten wi; T V W it i~` iii
aus der LinuxSzene auch eine Bildergalerie sowie eine Vim VViii~~1~1~iiiiiiii1iii
Nlaiiingliste und ein Anmeldeformular fLir die LUGV. Eine Vi ii= ~
Mitgliedschaft ist an keine Verpflichtungen gebunden i " ° m`··` " i"`i
sondern dient hauptséchlich der Organisation des LinuxDays. V V ;VV.,VVnVV_·i_V,V
iii iii iiiw ii
we iiiV V=Vi:Vi1iiii1VViV1 l= ii

Wl1i˜˜ ˜VViiiV VV Vi, i;1F4‘i#iV :!‘1 i˜ = It
Homepage: www.|ugv.at Ei wij VVVV VVV r ˜˜˜˜ in ˜˜i i f i W ,_V_
www.Iinuxday.at V V ˜ in M H it. ‘‘VV
ii ViVV 1 VVVVV . *1- t 9 *:* iii ˜ii˜ 1
11 Vwiiiiiiiii‘iiiil"i"iiI 4. ww ‘!ii i § Vm iiiim J"
iii˜1G˜1iiiii1i1iiiiiiii1iiiiiiiii1i1i1iiiii11iiiiiUiiUiiiiiiiiiuiiiiiiiiiiiiimw
Geschichte
Die LUGV wurde im Friihling 1998 als loser Zusammenschiuss von ca. 30 begeistert
herberge in Feidkirch gegriindet.
Nach einigenTreffen in der Jugendherberge wurde im Herbst eine Installationspart
der JH organisiert. Der Andrang war so gr0B, dess einige Ihre PCs nicht mehr auf
Im Friihling 1999 stellte uns die VKW in Bregenz einen gr0i$en Raum zur Verfiigu
2. Installationsparty organisiert wurde. Es weren Liber 100 Linuxbegeisterte, we
den Raum fiillten. Die VK\N war auf diesen Anstrum nicht vorbereitet, da nur fib
Werbung fiir die Instaliparty gemacht wurde. Das Stromnetz musste vestérkt werde
mit einer Zwischenverpflegung, offeriert von der VKW, iiberbriickt. Danach wurde
dem Hobby gefront.
Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel WeiBes Kreuz die Idee gebor
einen LinuxDay (vvvvw.|inuxday.at) zu organisieren.
Die vo/lsténdige Geschichte der LUG V finder sich unter www/ugmat

8.3 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivlstaBo>< mit der
Aussenvvelt kommunlzieren kann,
bedarf es einzlg eines l\letzwerl<˜
kabels. Jede Box ist Hxfertig vor-
konfiguriert; eine Installation ist nicht
notvvendig. Ob per FTPFiletransfer
(Kopiergerate), angeschlossenen
Scannern, oder Druckvorgang, die
ArchlvlstaBox zeigt sich ausserst
kontaktfreudig bei der Dokumenten-
annahme.
Alle Dokumente, die in der Archix/ista-
Box eintreffen, werden automatlsch
beschlagvvortet (indexlert) und ste-
hen unmittelbar fur eine Recherche
zur Verfugung. Der Zugriff auf die

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

8.4 Prospekt, Seite, 10 Punkt, Eras

Fiir jeden Bedarf die Richtige
Sie heissen Rigi, Pilatus, Titlis, Eiger, platte, zweiter Box und Tape fdhren si
Mythen und Rothorn und sind Laufwerk ist fdr Archive ab ca. Barcode-Erkennung du
punkto Unerschdtterlichkeit und 500000 bis einige Millionen Seiten _
Stabilitat genauso robust wie die ausgelegt. Und fans sm nun Gmen DOkUmmn
gleichnamigen geografischen Erhe _ Ousmr 3 Ia Mmm E`/Nest DEL
mmgen Die Pmfmmancg dg Mythen und Rothorn sind Scan- und ndtigen, auch kein Prob
DOwmm Sn/Gr mm Sim VOM dg OCR-Cluster-Stationen, mit welchen stellen lhnen diese
relamen Héhg me mit dem BU? auf die Boxen Pilatus, Titlis und Eiger entsprechend
namen EBOZMH M ablmén gescannt werden kann. Ebenfalls dividuell zusammen.
Rigi eignet sich primar fdr kleinere
Umgebungen wie z.B. Rechtsan- Rigi I797 m.d.M. Einzelplatz-Dokumenten-Server mit
waizspraxen oder PRBdros. Aber 20'000 Akten und 100'000 Seiten
300* 9V6$$l Unternehmen Clie tilt params 2132 m.0.ivi. ookumenien-sewer rm bis z
Abteilungen PDF-Dokumentenserver tgp Und 1 M50 ggitgn
SUChn sind mit GU MQFBOX gut Titlis 3238 m.d.M. Dokumenten-Sewer fdr bis zu ZOO'
Dediem ten und l Mio Seiten, redundant (2 Boxen)
Pilatus ist fdr mittlere Firmengrossen, Eiger 3790 m.d.M. Dokumenten-Sen/er fdr
oder besser ausgedrdckt, fdr das Akten, bis ca. 2 Mio Seiten, redundant
mittlere Datenvolumen gedacht. Die (Z B><el Und mit B@l<UP·T6P·DV ive
Titlis-B<>>< ebenfallsi allerclinqs erqibt Mythen iam m.0.ivi. scan- tm oci2-BX,
die redundante Hardware ein Pilatus und Titus transporrien
Gmémés Simemenselemem DE Rcthorn 2351 m.d.M. Scan- und OCR-Box, passend zum Eige
EigerBox mit entsprechender Fest-
ArchivistaBox: Connect your world

Damit eine ArchivistaBox mit der
m Aussenwelt kommunizieren kann,
*** { ter , Dedarf es einzig eines Netzwerk-**
" kabels. Jede Box ist fixfertig vor-
·'T , konfiguriert; eine Installation ist nicht
'°`` ` ` notwendig. Ob per FTPFiletransfer
````i`` ` ~i~i~ 4* (Kopiergerate), angeschlossenen
Print-Server Koblerer Scanner Scannern, oder Druckvorgang, die
Archix/istaBox zeigt sich ausserst
, 4 l ··*° kontaktfreudig bei der Dokumenten-
___? Q annahme.
y \ sj = ;j Alle Dokumente, die in der Arcnivista-
i OHi - Box eintreh°en, werden automatisch
beschlagwortet (indexiert) und ste
DOkUmm " hen unmittelbar fdr eine Recherche
ERPSystem zur Verfdgung. Der Zugriff auf die
PDF (VOMEXO (mit key;) Box erfolgt dber den Web- oder Rich-
Client. Dokumente konnen aber
auch als PDF oder per Mail wei-
M & tergereicht werden. Ganz nach dem
Rim- Und wgbgigm Motto: Connect your world.
Publizieren (DVD) A .
s---...__ lVl·?lllS l· 'A A n ArchivislaGmbH**
g O l r l§.?Ii2"i5Ei'l222i5$l°
_ Q { Fax: +41 (U)44 254 54 02
\ // · ~ léV3: iivww.archivista.ch
.._,_, R I ail.webmaster@archivista.ch

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, 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:

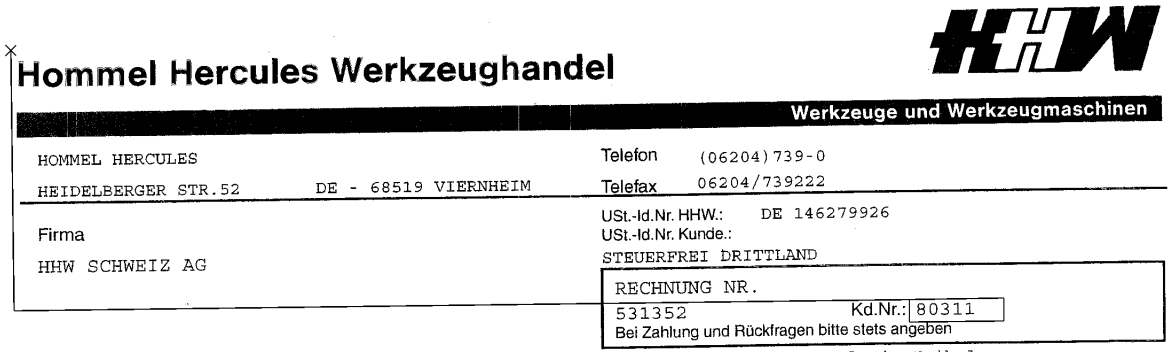
92
P Wir werden zunächst die erstere zu betrachten haben, in den Nebenzonen aber we
mehrere Specialgebiete, und zwar die Grauwaekenzonen, die nördlichen Kalkalpen,
Wiener Sandsteinzone und die südlichen Kalkalpen, denen sich die aus unser Staat
fallenden Gebirge des Baltansystems unmittelbar anschließen, abgesondert behande
1. Centralzone.
Die Centr alalpen oder die krystallinische Mittelzone der Alpen besteht durchweg
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
s überein, welche man vom orographischen Standpunkte zwischen den Centralalpen u
s Kalkalpen gezogen hat. So finden wir beispielsweise aus der Karte Seite 27 die
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 südlichen Nebenzone vereinigt, und analoge Abweichun
ergeben sich auch an anderen Stellen. Auch inag hier gleich hervorgehoben werden
wenngleich die Centralzone das eigentliche Herrschfeld der archaischen und die N
jenes der Sedimentgesteine bilden, sich doch einerseits beträchtliche Massen der
ein manchen Stellen der mittleren Kette, in isolirten Schollen über den krystall
Gesteinen vorfinden, wie z. B. an der Landesgrenze in den Ortler Alpen oder am B
oder endlich aus 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 dAsta in Südtirol, der schmale Z
Glimmerschiefer, welcher der Einsenkung des Gailthales in Kärnten folgt, im West
mit der Centralzone doch in Verbindung steht, und ein ähnlicher langer und schma
von krystallinischen Schiefer- und Massengesteinen, der stidlich von der Karavan
Q den Längsthälern der Miß und Javoria entlang, sortstreicht.
So wenig wie in der Bodenplastit, 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 Schiefergefteine der G
tiefste und älteste, der Glimmerschiefer das nächst jüngere und der Thonschieser
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 2.0' gewählt 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 Loge berechnen, um wieviele Grad die Seite geradegestellt werden muss und wo auf der Seite sich der Nullpunkt (Logoposition) befindet.



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 kläglich. Heute kann die Antwort differenziertet gegeben werden. Wer schwierigie 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 300°000. Ob das reicht, eine zu kommerziellen Texterkennungspakten 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 unnütz, 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 übrigen 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.