

# Open Source Texterkennung unter Linux

Texterkennung auf Ubuntu 8.x und der ArchivistaBox 2008/XI

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>		
1.1	Freie OCR-Erkennung unter Linux: Part 2 . . . . .	2		
1.2	Was dieser Vortrag will . . . . .	3		
<b>2</b>	<b>Installation unter Ubuntu 8.x</b>	<b>4</b>		
2.1	Vorbemerkung . . . . .	4		
2.2	Installation von ExactImage . . . . .	4		
2.3	Installation von Cuneiform . . . . .	6		
2.4	Installation von Tesseract bzw. OCROpus . . . . .	6		
2.5	Zusammenfassung . . . . .	7		
<b>3</b>	<b>Qualität der OCR-Programme</b>	<b>8</b>		
3.1	Gleiche Vorlagen wie 2007 . . . . .	8		
3.2	Text, 1spaltig, 12 Punkt Helvetica . . . . .	8		
3.3	Flyer, 1spaltig, ca. 8 Punkt, Helvetica . . . . .	9		
3.4	Prospekt, Ausschnitt, 10 Punkt, Eras . . . . .	9		
3.5	Prospekt, Seite, 10 Punkt, Eras . . . . .	10		
<b>4</b>	<b>Ergebnisse Tesseract/OCROpus</b>	<b>12</b>		
4.1	Was Ocropus besser kann . . . . .	12		
4.2	Text, 1spaltig, 12 Punkt Helvetica . . . . .	12		
4.3	Flyer, 1spaltig, ca. 8 Punkt, Helvetica . . . . .	13		
4.4	Prospekt, Ausschnitt, 10 Punkt, Eras . . . . .	14		
4.5	Prospekt, Seite, 10 Punkt, Eras . . . . .	15		
4.6	Zusammenfassung . . . . .	16		
<b>5</b>	<b>Ergebnisse mit Linuxport-Cuneiform</b>	<b>17</b>		
5.1	Flyer, 1spaltig, ca. 8 Punkt, Helvetica . . . . .	17		
5.2	Flyer, 1spaltig, ca. 8 Punkt, Helvetica . . . . .	18		
5.3	Prospekt, Ausschnitt, 10 Punkt, Eras . . . . .	18		
5.4	Prospekt, Seite, 10 Punkt, Eras . . . . .	19		
5.5	Zusammenfassung Cuneiform . . . . .	21		
<b>6</b>	<b>Webbasierte OCR mit der ArchivistaBox</b>	<b>22</b>		
6.1	Was bleibt für die/den Anwender/in? . . . . .	22		
6.2	ArchivistaBox als Open Source Projekt . . . . .	22		
6.3	ArchivistaBox installieren . . . . .	23		
6.4	OCR-Erkennung mit der ArchivistaBox . . . . .	24		
6.5	OCR-Erkennung und PDF-Dateien erstellen überprüfen . . . . .	25		
<b>7</b>	<b>Was bringt die Zukunft</b>	<b>27</b>		
7.1	Es gibt noch viel zu tun . . . . .	27		

© 28.11.2008 by Urs Pfister, [www.archivista.ch](http://www.archivista.ch)

# 1 Einleitung

## 1.1 Freie OCR-Erkennung unter Linux: Part 2

Vor ziemlich genau einem Jahr, anlässlich des letztjährigen Linuxday.at 2007 hatte ich bereits das Vergnügen, einen Überblick über den damaligen Stand der Texterkennung unter Linux vorzutragen zu dürfen.

An sich finde ich es nicht so spannend, den gleichen Vortrag am gleichen Ort ein Jahr später nochmals zu halten. Wenn ich in diesem Jahr eine Ausnahme machen will, dann alleine deshalb, weil die Änderungen derart stark sind, dass der Vortrag vom letzten Jahr einfach nicht mehr aktuell genug ist.

Ich bin damals davon ausgegangen, dass es etwa noch ca. zwei bis drei Jahre dauern wird, bis eine einigermaßen vernünftige Alternative zu Closed Source Texterkennungsprogrammen unter Linux zur Verfügung stehen würde.

Auf dieser Annahme habe ich das Projekt [freearchives.ch](http://freearchives.ch) vorgestellt, um im nichtkommerziellen Umfeld auf eine gute Alternative hinarbeiten zu können. In der Zwischenzeit sind folgende Dinge passiert:

- OpenSource-Barcodeerkennung mit ExactImage (Teilfinanzierung [freearchives.ch](http://freearchives.ch))
- Linux-Port von Cuneiform erscheint.
- [freearchives.ch](http://freearchives.ch) wird abgemahnt.
- Mit [hocr2pdf](http://hocr2pdf) können durchsuchbare PDF-Dateien erstellt werden (Finanzierung durch Archivista GmbH)

Und um es auf einen Blick zu sagen. Drei Punkte sind hocheifrig, zum vierten Punkt nur soviel: Der Verein [freearchives.ch](http://freearchives.ch) ist im vergangenen Sommer abgemahnt worden. Aufgrund der geltenden Rechtslage hat der Verein eine strafbewehrte Unerlassungserklärung abgegeben. Zur Hauptsache ging es darum, dass ein Hersteller nicht wollte, dass Preise bez. Vertragsdetails öffentlich bekannt sind.

Und um allfällige Unsicherheiten gleich vorweg auszuräumen. Ich halte diesen Vortrag als Privatperson Urs Pfister und nehme für diesen Vortrag kein Entgelt und keine Reisekostenpauschale entgegen, war aber heute morgen nicht unfroh, dass mich das OK-Team am Bahnhof St. Margrethen abholte. Was damit gesagt werden soll, es ist kein kommerzieller Vortrag bzw.

es ist zumindest nicht beabsichtigt, dass dieser Vortrag unter alle möglichen wettbewerbsrechtlichen Gesetze in allen möglichen Ländern fällt. Sollte dies dennoch der Fall sein, so würde ich denn für mich das Recht auf Meinungsäußerungsfreiheit in Anspruch nehmen wollen. Soviel zum Kleingedruckten.

## 1.2 Was dieser Vortrag will

Im letzten Vortrag habe ich aufgezeigt, wie die Texterkennung sich historisch entwickelt hat. Am Schluss habe ich einige Open Source Texterkennungspakte mit einem kommerziellen Produkt verglichen. Heute möchte ich den Schwerpunkt auf die Open Source Produkte legen und eine kurze Einführung in diese Produkte geben.

Und im Unterschied zum letzten Jahr, wo keine fundierteren Linux-Kenntnisse gefragt waren, so wird es dieses Jahr so sein, dass sie noch immer nicht benötigt werden, dass das Skript aber auch dokumentiert, wie die Produkte aufgesetzt werden können und wie ein OCR-Vorgang ausgelöst werden kann. Und ja, dabei kann es dann und wann etwas technischer werden.

# 2 Installation unter Ubuntu 8.x

## 2.1 Vorbemerkung

Wir benötigen ein Linux-System, mit dem sich Quelltext (Programme) in ausführbare Pakete übersetzen lassen. Dies ist bei Ubuntu 8.x erst der Fall, wenn 'apt-get install build-essential' ausgeführt wird. Weiter benötigen wir Subversion.

## 2.2 Installation von ExactImage

### 2.2.1 Was ist ExactImage?

ExactImage ist eine Bildbearbeitungs-Suite für Linux. Mittlerweile kann ExactImage aber weit mehr als nur Bilder bearbeiten wie dies bei ImageMagick der Fall ist.

Historisch gesehen ist ExactImage entstanden, weil die ArchivistaBox eine schnelle Bibliothek zum Bearbeiten von Bildern benötigte. ImageMagick war leider zu langsam für die Jobs, die es auf den relativ bescheidenen Prozessoren, welche auf der ArchivistaBox z.T. im Einsatz stehen, zu lösen gilt.

Daneben beinhaltet ExactImage Optimierungsprogramme für Barcode- und Formular-Erkennung sowie ein Konvertierungsprogramm, um aus sogenannten hocr-Dateien PDF-Dateien zu erstellen. Hocr-Dateien werden von Open Source OCR-Programmen erstellt und enthalten neben dem erkannten Text auch Formatierungsauszeichnungen.

### 2.2.2 Installation von ExactImage

Beziehen können wir ExactImage über die folgende Adresse:

```
svn co http://svn.exactcode.de/exact-image/trunk ei
```

Nach dem Download wechseln wir nach `cd ei` und führen dort den folgenden Befehl aus:

```
./configure --prefix=/usr --without-python
```

Danach erhalten wir eine ganze Reihe von Meldungen:

```
./configure --prefix=/usr
checking whether the C++ compiler works ... yes
checking for C++ STL support ... yes
checking for C++ templates ... yes
checking for C++ template specialization ... yes
```

checking for C++ partial template specialization ... yes  
checking whether C++ supports templates ... yes  
checking for header iostream ... found  
checking for header string ... found  
checking for header iostream ... found  
checking for header sstream ... found  
checking for header fstream ... found  
checking for package x11 (atleast 11.0) ... yes (11.0)  
checking for package libagg (atleast 2.4) ... no

Anti-Grain Geometry was not found - an internal copy  
will be built as AGG is required to render vector  
graphics (such as even lines and text).  
AGG can be obtained from: [www.antigrain.com](http://www.antigrain.com)

checking for package freetype2 (atleast 9.5.0) ... yes (9.18.3)  
checking for package evas (atleast 0.9.9) ... yes (0.9.9.002)  
checking for header Evas\_Engine\_GL\_X11.h ... found  
checking for package libjpeg ... yes  
checking for package libtiff ... yes  
checking for package libpng (atleast 1.2) ... yes (1.2.8)  
checking for package libungif ... yes  
checking for package jasper ... no  
checking for package expat ... yes  
checking for package OpenEXR (atleast 1.2.0) ... no  
checking for package lcms (atleast 1.10) ... yes (1.14)  
checking for package bardecode ... no

For proprietary barcode recognition, place the commercial, binary-only blob  
library (from <http://www.bardecode.com/>) into the 'external/' directory.

checking for package swig (atleast 1.3.32) ... yes (1.3.33)  
checking for package lua (atleast 5.1) ... no  
checking for package perl (atleast 5.8.0) ... yes (5.8.7)  
checking for package php (atleast 5.2.0) ... yes (5.2.6)  
checking for package python (atleast 2.5.0) ... no (2.4.1)  
checking for package ruby (atleast 1.8.5) ... no

Nicht alle Abhängigkeiten müssen erfüllt sein. Evas z.B. benötigen wir nur, wenn wir die Bilder  
mit edisplay anzeigen möchten. Die Option `--without-python` war in meinem Fall unter

Ubuntu 8.x notwendig, weil ansonsten ab SStange ein python-config-Skript nicht gefunden wurde, womit die Kompilierung nicht erfolgte. Mit `--without-python` war das Erstellen der Dateien problemlos machbar.

Nun können wir mit `make` und anschliessend `make install` ExactImage übersetzen und installieren. Die Installation dauert einige Minuten.

Zum Test sollten wir nun prüfen, ob die Programme `econvert`, `bardecode` sowie `hocr2pdf` vorhanden sind. Ist dies der Fall, ist alles ok.

## 2.3 Installation von Cuneiform

Auch Cuneiform wollen wir im Quelltext übersetzen. Dazu benötigen wir Bazaar (unter Ubuntu das Paket 'bzd'). Sobald wir Bazaar installiert haben, können wir auf der Konsole den folgenden Befehl eingeben:

```
bzd branch lp:cuneiform-linux
```

Danach wechseln wir mit `cd cuneiform-linux` zu den Quellen. Die Datei `readme.txt` gibt Hilfe zur Installation an:

```
mkdir builddir
cd builddir
cmake -DCMAKE_BUILD_TYPE=debug -DCMAKE_INSTALL_PREFIX=/usr ..
make
make install
```

In meinem Fall sind dann zunächst einige Fehlermeldungen aufgetreten, die ich erst lösen konnte, nachdem ich `cmake` in der Version 2.6.2 im Quelltext installiert habe. Um zu testen, ob alles ok ist, sollte es möglich sein, Cuneiform mit `cuneiform` zu starten. Danach sollten wir in etwa die folgende Ausgabe erhalten:

```
Cuneiform for Linux 0.5.0
Usage: cuneiform[-l language -f format --dotmatrix --fax -o result_file] imagefile
```

## 2.4 Installation von Tesseract bzw. OCROpus

Die Installation von Ocropus war bisher nicht immer einfach. Beim erneuten Versuch Ende November gelingt auch dies relativ einfach. Zunächst die aktuelle Version von Tesseract.

```
svn checkout http://tesseract-ocr.googlecode.com/svn/trunk/ tesseract-ocr-read-only
```

Danach im heruntergeladenen Ordner mit `./configure`, `make` und `make install` die Installation durchführen.

Nun können wir die Bibliothek iulib installieren.

```
svn checkout http://iulib.googlecode.com/svn/trunk/ iulib
```

In meinem Fall unter Ubuntu 8.04 habe ich es nur zum Laufen gekriegt, als ich die Installation mit `scons` und `scons install` vorgenommen habe. Zum Schluss ging dann Ocropus relativ einfach:

```
svn checkout http://ocropus.googlecode.com/svn/trunk/ ocropus
```

Auch hier habe ich `scons` und `scons install` verwendet. Eine Seite konnte ich mit dem folgenden Befehl erkennen:

```
ocrosript recognize eins.png >eins.hocr
```

Dabei entsteht eine sogenannte hocr-Formatdatei. Diese kann in eine PDF-Datei bzw. Textdatei konvertiert werden mit:

```
hocr2pdf -i eins.png -o eins.pdf -t eins.txt < eins.hocr
```

Hinweis: `hocr2pdf` entstammt ExactImage und ist kein Bestandteil von Ocropus. Das entsprechende Lua-Script von Ocropus hat (zumindest für mich nicht ersichtlich) keine Resultate hervorgebracht.

## 2.5 Zusammenfassung

Die Installation der Open Source Texterkennungspakete klappt heute mit Konsolenkenntnissen und Ubuntu 8.x bereits relativ gut. Es ist aber nach wie vor so, dass die Installation der Pakete zeitaufwendig ist. In meinem Fall benötigte ich für die verschiedenen Pakete je etwa eine Stunde, es kann aber gut und gerne auch ein halber oder Tag werden.

# 3 Qualität der OCR-Programme

## 3.1 Gleiche Vorlagen wie 2007

Damit wir einen Vergleich haben, was im letzten Jahr passiert ist, wollen wir die Beispiele des letzten Jahres verwenden.

## 3.2 Text, 1spaltig, 12 Punkt Helvetica

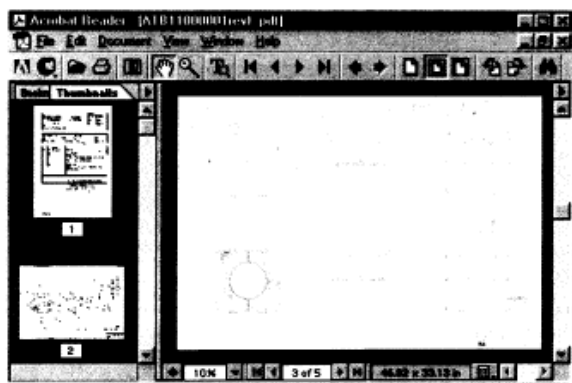
### Fall 2: Engineering-Projekt mit PDF-Dokumentation

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)

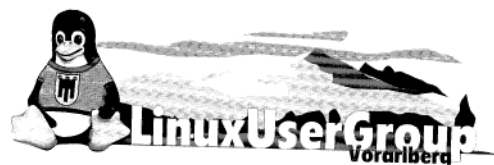
Archivista GmbH war aufgrund profunder Kenntnis der aktuellen Technologien und dank des selber entwickelten **Produktmoduls Archivista TifToPDF** in der Lage, diese Wertschöpfungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendbaren PDF-File **kostengünstig** anzubieten. Das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





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.

### 3.3 Flyer, 1spaltig, ca. 8 Punkt, Helvetica



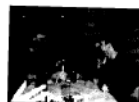
#### LinuxUserGroup Vorarlberg

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](http://www.lugv.at)  
[www.linuxday.at](http://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 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](http://www.linuxday.at)) zu organisieren.

Die vollständige Geschichte der LUGV findet sich unter [www.lugv.at](http://www.lugv.at)

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

### 3.4 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivistaBox mit der Auserwelt 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 GmbH  
Pfaffen - CH-8030 Zürich  
Tel. +41 (0)44 254 54 00  
Fax +41 (0)44 254 54 00  
Web: [www.archivista.ch](http://www.archivista.ch)  
E-Mail: [webmaster@archivista.ch](mailto:webmaster@archivista.ch)

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

### 3.5 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.

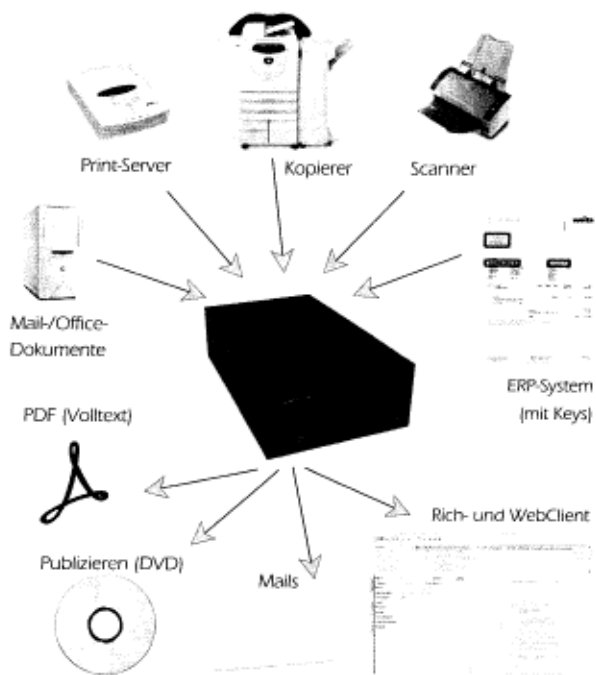
Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Titlis und Eiger gescannt werden kann. Ebenfalls

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.

<b>Rigi</b>	1797 m.ü.M.	Einzelplatz-Dokumenten-Server mit bis zu 20'000 Akten und 100'000 Seiten
<b>Pilatus</b>	2132 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten
<b>Titlis</b>	3238 m.ü.M.	Dokumenten-Server für bis zu 200'000 Akten und 1 Mio Seiten, redundant (2 Boxen)
<b>Eiger</b>	3790 m.ü.M.	Dokumenten-Server für unlimitierte Anzahl Akten, bis ca. 2 Mio Seiten, redundant (2 Boxen) und mit Backup-Tape-Drive
<b>Mythen</b>	1899 m.ü.M.	Scan- und OCR-Box, welche Daten zum Pilatus und Titlis transportiert
<b>Rothorn</b>	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 Netzkabels. 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 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)44 254 54 00  
 Fax +41 (0)44 254 54 02  
 Web: [www.archivista.ch](http://www.archivista.ch)  
 E-Mail: [webmaster@archivista.ch](mailto:webmaster@archivista.ch)

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

## 4 Ergebnisse Tesseract/OCROpus

Man möge mir verzeihen, wenn ich die beiden Programmpakete zusammen betrachte. Letztlich verwendet Ocropus im Moment Tesseract. Die Resultate unterscheiden sich daher kaum.

### 4.1 Was Ocropus besser kann

Ocropus beinhaltet in erster Linie eine Blockerkennung, d.h. Spalten und Trennungen werden sauber(er) erkannt. So erreichte ich im vierten Beispiel recht gute Resultate bei der Blockerkennung, hier leistet Ocropus wertvolle Dienste. Allerdings ist es mir innerhalb einer Stunde nicht gelungen, Ocropus die deutsche Sprache beizubringen. Im Prinzip müsste es so funktionieren:

```
ocrosript recognize --tesslanguage=deu --output-mode=test eins.png > eins.txt
```

Nun funktioniert OCROpus so, dass ein grosser Teil der Kommandos über die Skriptsprache Lua realisiert ist. Die derzeit verfügbaren Programme liegen unter:

```
/usr/local/share/ocropus/scripts/
```

Dort findet sich das Skript `recognize.lua`, welches für den OCR-Teil betr. Tesseract zuständig ist. Letztlich habe ich noch versucht, die Defaultsprache von 'eng' auf 'deu' zu setzen. Doch leider fruchtete auch das nichts.

Abschliessend möchte ich festhalten, dass Ocropus einen extrem interessanten (aber auch anspruchsvollen) Weg geht. Die Möglichkeit, die gesamte OCR-Engine mit Open Source Programmen zu verfeinern bzw. zu erweitern beinhaltet ein Potential, das schon als ausserordentlich mächtig erscheint.

Allerdings, im Moment ist Ocropus noch immer nicht wirklich für mich als anspruchsvolleren Anwender bzw. für die ArchivistaBox als professionelle Lösung genügend interessant. Nun aber zu den Testresultaten.

### 4.2 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. 1\Iac und Windows) betrachtet und ausgedruckt werden können.

PDFDokumente sind unheimlich flexibel. Alles, was gedruckt werden kann, ist auch als PDFDatei (in digitaler Form) publizierbar. Zudem sind PDFDateien mittlerweile weit ver-

breitet; 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 A0Format sollten gescannt und in ein PDF-Format überführt werden. Ziel waren selbsttragende CDs mit PDFDateien, 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 A4Seiten 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 PDFDateien

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 Tif'l'oPDF in der Lage, diese \\\/ertschöpfungskette vom Papierbeleg zum jederzeit elektronisch zugänglichen und versendbaren PDF-File kostengünstig anzubieten. Das EngineeringUnternehmen verfügt heute über eine saubere, platzsparende Dokumentation des Projektes.

Scan-Dienstleistungen Scanning und was dazu gehört, Fallbeispiele, Seite 2 Öl";

Archivista GmbH, 8042 Zürich, Tel. 01 350 46 74, Fax 01 350 46 72, www.archivista.ch

## 4.3 Flyer, 1spaltig, ca. 8 Punkt, Helvetica

lmuxllserüroup Vorarlberg

Mllllllliiri W

Wir sind ein loser Zusammenschluss von LinuxAnwendern, in = =

die sich etwa einmal pro Monat zwanglos treffen, um über l = i- lei

verschiedeneThemen und selbstverständlich über Linux zu l;r diskutieren. Zu diesenTreffen s

eingeladen, gleichgültig ob LUGVMitgIied oder nicht. ' il .

In unregelmäßigen Abständen veranstaltet die LUGV auch ,,,,,,,

Ausflüge, Vorträge und Workshops. ua W l , ,,g, ,wl;;l,lllllll

Auf der Homepage der LUGV finden sich neben Neuigkeiten llllllj ll

aus der LinuxSzene auch eine Bildergalerie sowie eine lll; U,,lll »

» »illllllllllllll i im

Mailingliste und ein Anmeldeformular für die LUGV. Eine ilr iiri «

Mitgliedschaft ist an keine Verpflichtungen gebunden " ° m". " lli

sondern dient hauptsächlich der Organisation des LinuxDays. . M lllllli " M

Homepage: www.lugv.at

Geschichte

Die LUGV wurde im Frühlin 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 Frühstü

der JH organisiert. Der Andrang war so groß, dass einige Ihre PCs nicht mehr aufstellen konn

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

2. Installationsparty organisiert wurde. Es waren über 100 Linuxbegeisterte, welche mit Ihre

den Raum füllten. Die VKW war auf diesen Anstrum nicht vorbereitet, da nur über das Interne

Werbung für die Installparty gemacht wurde. Das Stromnetz musste verstärkt werden. Die Wartezeit

mit einer Zwischenverpflegung, offeriert von der VKW, überbrückt. Danach wurde bis in den sp

dem Hobby gefrönt.

Nach diesem tollen Erfolg wurde im Mai 1999 im Hotel Weißes Kreuz die Idee geboren jährlich i

## 4.4 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine Archix/istaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerk-kabels. Jede Box ist üxfertig vor-konfiguriert; eine Installation ist nicht notwendig. Ob per FTPFiletransfer (Kopiergeräte), angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich äusserst kontaktfreudig bei der Dokumenten-annahme.

Alle Dokumente, die in der Archix/ista-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.

Q, Archivista GmbH  
i ,\*lr,. \_i-' Postfach-CH-8042 Zürich  
· E: gl L-je TeI:+41(0)442545400  
\_' i Fax: +41 (0)44 254 54 02  
Web: www.archivista.ch  
Aßc|·f|\||\$'|'A email: wgmásler@arcnivls1a.n

## 4.5 Prospekt, Seite, 10 Punkt, Eras

0 Q 0 I

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 500'000 bis einige Millionen Seiten \_ V

Stabilität genauso robust wie die ausgelegt. Erd fans ?||m'mMm Dgkumemgnl

leichenami en eo rafischen Erhe [MH 8 8 wm \gest G

Ehingen gme gpeäcormance der Mythen und Rothorn sind Scan- und nötigen, auch kein Problem; w

Documésén/er lässt sich von der OCR-Cluster-Stationen, mit welchen stellen ihnen diesen mit

relativen Höhe die mit dem Berg- auf die Boxen Pilatus, Titlis und Eiger entsprechenden Har

.. I . . gescannt werden kann. Ebenfalls dividuell zusammen.

namen assoziiert ist, ableiten.

Rigi eignet sich primär für kleinere

Umgebungen wie 2,B, Rechtsan- Rigi I797 m.ü.M. Einzelplatz-Dokumenten-Server mit bis zu  
waitspraxen oder PR-Büros. Aber 20000 Akten und 100'000 Seiten

allen größere Unternennien, Clie iir mmm; 2132 m.0.ivi. Dokumenten-semer für bis zu 200\*000

Abteilungen PDF-Dokumentenserver ten und 1 M50 Seiten

SUCÜBÜ sind mn dw RIQPBOX gut Titlis 3238 m.ü.M. Dokumenten-Sen/er für bis zu 200'000 Ak-  
bemem ten und 1 Mio Seiten, redundant (2 Boxen)

Pilatus ist für mittlere Firmen rossen, Ei er 3790 m.ü.M. Dokumenten-Server für unlimitierte

9 9

Qder besser ausgedrückt, für das Akten, bis Ca. 2 MIO Seiten, redundant

mittlere Datenvolumen gedacht. Die (2 BOXVll Und mit B@Cl<UP·T@P·DV1V

Titlis-Bex ebenfalls- allerdings ergibt Mythen 1899 m.0.ivi scan- und ociz-Box, Weiche Daten

die redundante Hardware ein Pilatus und Titlis transportiert

erhöhtes S'Chm'ts|mm DIG Rothorn 2351 m.ü.M. Scan- und OCR-Box, passend zum Eiger

Eiger-Box mit entsprechender Fest-

0 I

A1Ch1V1St21BOX: C01111Ct Y0111 W01°ld

Damit eine ArchivistaBox mit der

m Aussenwelt kommunizieren kann,  
 L" ie tr-. N bedarf es einzig eines Netzwerk-  
 ° kabels. Jede Box ist fixfertigt vor-  
 . ' " > g  
 -57 \_ «-aig ä i konfiguriert; eine Installation ist nicht  
 ' E I ' notwendig. Ob per FTP-Filetransfer  
 mw ( «i - Ä ( Kopiergeräte), angeschlossenen  
 l°rlnt·\$VVV Kopierer Scanner Scannern, oder Druckvorgang, die  
 ArchivistaBox zeigt sich ausserst  
 , T i ·\* kontaktfreudig bei der Dokumenten-  
 \_\_\_? Q annahme.  
 g. \ A Äj -= -g Alle Dokumente, die in der Archivista-  
 \_/ ff \_ T - Box eintreten, werden automatisch  
 Ma'- 0 ICQ- beschlagwortet (indexiert) und ste  
 Dokljmeme i hen unmittelbar für eine Recherche  
 ERP-System zur Verfügung. Der Zugriff auf die  
 PDF (Volltext) (mit Keys} Box erfolgt über den Web- oder Rich-  
 Client. Dokumente können aber  
 auch als PDF oder per Mail wei-  
 ter erreicht werden. Ganz nach dem  
 Q  
 Rrerq- und webgprenr Motto: Connect your world.  
 Publizieren (D\D) \_ A .  
 , ' iii)! . i , ig; Q Posnacn-cn-6042 Zürich  
 ß 0 §;£1. . rei; +41 (0)44 254 54 00  
 3 \_ Fax: +41 i0)44 254 54 02  
 \ 'W r K Z W6b;www.archivista.ch  
 E T rg V\_V { y M{ß|§|Y|\$'|'A EMaiI: wemasler@arnvrsra.cn

## 4.6 Zusammenfassung

Es gibt minime Fortschritte zum letzten Jahr. Es bleibt aber festzuhalten, dass die Texterkennung nur gute Ergebnisse liefert, wenn die Vorlage in extrem guter Qualität vorliegt und die Schriftgrösse 10 Punkt oder mehr beträgt. Mit gewissen Buchstaben kommt Tesseract schlecht zurecht (z.B. M oder W). Hier könnte allenfalls eine Trainingsdatei weiterhelfen.



# 5 Ergebnisse mit Linuxport-Cuneiform

Nun zu den Ergebnissen mit Cuneiform.

## 5.1 Flyer, 1spaltig, ca. 8 Punkt, Helvetica

Fall 2: Engineering-Projekt mit PDF-Dokumentation Kaum eine andere Technologie hat im letzten Jahr für mehr Furore gesorgt als das PDFFormat. 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,

## 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](http://www.lugv.at) [www.linuxday.at](http://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 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](http://www.linuxday.at)) zu organisieren. Die vollständige Geschichte der LUGV findet sich unter [www.lugv.at](http://www.lugv.at)

## 5.3 Prospekt, Ausschnitt, 10 Punkt, Eras

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerk- kables. Jede Box ist fixfertigvorkonfiguriert; eine Instaliation ist nicht notwendig. Ob

per FTP-Filetransfer (Kopiergeräte), angeschlossenen Scannern, oder Druckvorgang, die ArchivistaBox zeigt sich äusserst kontaktfreudig bei der Dokumenten- annahme.

beschlagwortet (indexiert) und stehen unmittelbar für eine Recherche zur Verfügung. Der Zugriff auf die Box erfolgt über den Web- oder RichClient. Dokumente können aber auch als PDF oder per Mail weitgereicht werden. Ganz nach dem Motto: Connect your world.

Q p

Arcblvlsta GmbH

Posffach - CH-8042 Zurich

Te): a41 (0)44 254 54 00

Pate +41 (0)44 254 54 02

Web: [www.archivista.ch](http://www.archivista.ch)

E-Mail: [webmastereaarchivista.ch](mailto:webmastereaarchivista.ch)

ARCHIVISTA

Alle Dokumente, die in der ArchivistaBox eintreffen, werden automatisch

## 5.4 Prospekt, Seite, 10 Punkt, Eras

Für jeden Bedarf die Richtige

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

führen sie z.B. eine Text- und/oder

Barcode-Erkennung durch.

Und falls Sie nun einen DokumentenCluster a la Mount Everest be-  
Mythen und Rothorn sind Scan- und OCR-Cluster-Stationen, mit welchen auf die Boxen Pilatus, Titfis und Eiger gescannt werden kann.

Ebenfalls

nötigen, auch kein Problem; wir stellen Ihnen diesen mit der entsprechenden Hardware gerne individuell zusammen.

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)  
Pilatus ist für mittlere Firmengrößen, oder besser ausgedrückt, für  
das mittlere Datenvolumen gedacht. Die Titlis-Box ebenfalls,  
allerdings ergibt die redundante Hardware ein

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  
erhöhtes Sicherheitselement. Die Eiger-Box -- mit entsprechender Fest-  
Archivistaßn : Cnnnect ynur wrld

Print-Server

Scanner

opierer

Alle Dokumente, die in der Archivista-

Mail-/Office- Dokumente

ERP-System (mit Keys)

PDF (Volltext)

Rich- und (/ebClient

Publizieren (DVD)

Mail

Archlvista GmbH

Pcsffach - CH-8042 Zur(eh

Tel: +41 (0)44 254 54 00

Pate e41 (0)44 254 54 02

Weh: www.archwista.ch

AIIIeIIIVISTA E-Mai(: webmasterebarchivista.ch

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 Berg-  
namen assoziiert ist, ableiten.

Rigi eignet sich primär für kleinere Umgehungen wie z.B.

Rechtsanwaltspraxen oder PR-Büros. Aber auch grössere Unternehmen,  
die für Abteilungen PDF-Dokumentenserver suchen, sind mit der  
Rigi-Box gut bedient.

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

Damit eine ArchivistaBox mit der Aussenwelt kommunizieren kann, bedarf es einzig eines Netzwerk- kables. 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 Dokumenten- annahme.

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

## 5.5 Zusammenfassung Cuneiform

Vergleichen wir die Ergebnisse mit dem letztjährigen Skript, dann erreicht der Cuneiform-Linuxport erstaunlich gute Resultate. Zwar gibt es im Vergleich zur kommerziellen Engine einige wenige Zeichen, die falsch erkannt werden, aber die Fehlerquote ist klein.

Die Ergebnisse stehen im übrigen im Einklang zu Cuneiform, das in meinem letzten Vortrag ja ebenfalls erwähnt wurde, weil Cuneiform um das Jahr 2000 herum bei den Closed Source Programmen die Nummer 2 darstellte.

Für mich persönlich ist/war der Linuxport von Cuneiform das absolute Highlight dieses Jahres. Und mit dem von der Firma Archivista GmbH gesponserten Programm 'hocr2pdf' können die Cuneiform-Resultate gleich auch in durchsuchbare PDF-Dateien umgewandelt werden.

Und damit hier keine Missverständnisse auftreten, ja die Firma Archivista GmbH ist in meinem Besitz, der 'hocr2pdf'-Code aber steht zu 100 Prozent unter der GPL-Lizenz und ist somit allen zugänglich.

# 6 Webbasierte OCR mit der ArchivistaBox

## 6.1 Was bleibt für die/den Anwender/in?

Wir haben gesehen, dass die OCR-Erkennung unter Linux massive Fortschritte gemacht hat. Es könnte nun der Eindruck entstehen, dass OCR-Erkennung unter Linux immer in Verbindung mit der Konsole gestartet werden müsste.


Falsch: Mit Linux ist es erstmalig machbar, ein OCR-Programm überhaupt über die Konsole zu starten. Die Automatisierung, die heute mit OCR-Programmen unter Linux möglich sind, waren mehr als ein Jahrzehnt unter Windows überhaupt nicht zugänglich.

Und dank der ArchivistaBox kann die OCR-Texterkennung komplett webbasiert und vollautomatisiert durchgeführt werden. Auch hier möchte ich betonen, dass ich gerne andere Projekte vorstellen würde, ich konnte aber auf den ersten Blick mit der Suchmaschine keine Alternativen finden. Daher werde ich nachfolgend die ArchivistaBox im Zusammenhang mit der OCR-Erkennung kurz vorstellen.

## 6.2 ArchivistaBox als Open Source Projekt



The screenshot shows the SourceForge search results for the project "ArchivistaBox". The page header includes the SourceForge logo and navigation links like "Log in", "Create account", "Community", "Jobs", and "Help". A search bar is visible in the top right. Below the header, the search results are displayed for "ArchivistaBox". A blue banner indicates "Exact matches found: ArchivistaBox". A table lists the project details:

Name	Relevance	Activity	Rank	Registered	Latest File	Downloads
<a href="#">ArchivistaBox</a>		99.79%	<a href="#">499</a>	2005-11-21	2008-11-28	14,275

Below the table, there is a description: "Archivista is an entirely webbased DMS and archiving application. The solution is independent of source formats: it turns all documents into image files. Additionally, it contains a PDF server. With it documents can be converted into searchable PDFs." There is a "Download" button with a green arrow icon and a "Members (1)" link. At the bottom, it shows "Page: 1" and "1 - 1 of 1 Results - Display 10".

Die ArchivistaBox wurde von Beginn weg als Open Source Projekt gestartet. Allerdings war es bis zu diesem Jahr leider nicht möglich, eine gute OCR-Erkennung mit auf die ArchivistaBox zu packen.

Mit dem Linuxport von Cuneiform ist das anders geworden. Nunmehr stehen (zusammen mit der Barcode-Erkennung von ExactImage) sämtliche Komponenten in guter Qualität unter der Open Source Lizenz zur Verfügung.

Dies hat sich auch in der Bewertung bei Sourceforge.net niedergeschlagen. Vor einem Jahr lag die ArchivistaBox noch etwa auf Position 3000, mittlerweile ist bereits eine Position um 500 erreicht. Dieses Jahr wurden mehr als 10'000 Downloads verzeichnet, deren ca. 8000 alleine in den letzten zwei Monaten (seit Cuneiform und hocr2pdf verfügbar sind).

Und noch etwas, die ArchivistaBox ist in diesem Monat erstmalig auf einer Titelseite eines Computermagazins gelandet:



Sorry, wenn ich das hier erwähne, ich hoffe, dass dies nicht als allzu eitel empfunden wird, aber etwas stolz darauf bin ich schon, das gebe ich gerne zu.

## 6.3 ArchivistaBox installieren

Zentrale Anlaufstelle ist die Homepage [www.archivista.ch](http://www.archivista.ch). Dort finden sich unter Support sämtliche Informationen, um eine ArchivistaBox selber unter der Open Source Lizenz aufzusetzen.

**ARCHIVISTA: OPENSOURCE DMS & ERP OUT OF THE BOX**

**Home**  
**Produkte**  
**Services**  
**Über uns**  
**Support**  
**Aktuell (Blog)**  
**Bezugsquellen**  
**English**

Suche

**ARCHIVISTA**  
 Archivista GmbH  
 Zürichstr. 80  
 CH-8118 Pfaffhausen  
 webmaster@archivista.ch  
 Tel: +41(0)44 2545400  
 Fax: +41(0)44 2545402

**ArchivistaBox: OpenSource-Lösungen für Ihr Business**

Ganze Berge von Dokumenten und Informationen auf einer kleinen Box verwaltet, eine Lösung, die zu 100 Prozent OpenSource ist und ein Daten-Konzept, bei dem ihre Dokumente auch in 30 Jahren ohne Plugins lesbar sind sind -- das und vieles mehr bietet die ArchivistaBox.



**Infos zur Box**  
 ArchivistaBox im Überblick  
 Produktpalette

**Online Testen**  
 ArchivistaBox testen  
 OpenSource-Version  
 Tutorial (Handbuch)

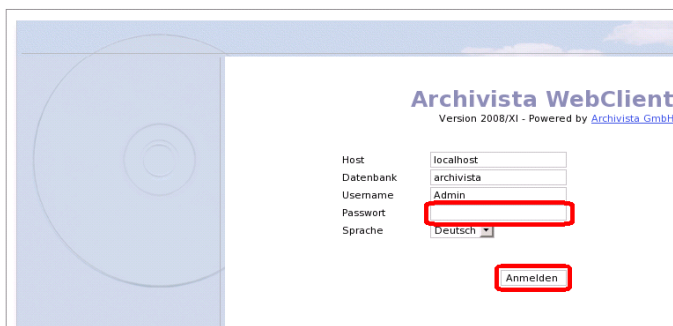
**Ja ich will eine**  
 Bezugsquellen  
 Preisliste (WebShop)

Ich möchte an dieser Stelle nochmals darauf hinweisen, dass die Firma Archivista GmbH ArchivistaBox-Systeme verkauft und wartet, dass die Open Source Version aber zu 100 Prozent mit derjenigen Version übereinstimmt, die auch produktiv bei Kunden eingesetzt wird. Es geht also explizit nicht darum, ein SZückerchenßu geben, um letztlich eine kommerzielle ArchivistaBox Bchmackhaftßu machen.

## 6.4 OCR-Erkennung mit der ArchivistaBox

Wenn die ArchivistaBox gestartet wird, erscheint das Anmeldefenster für den WebClient.

[WebClient](#) · [WebERP](#) · [WebAdmin](#) · [WebConfig](#) · [Manual](#) · [Handbuch](#)



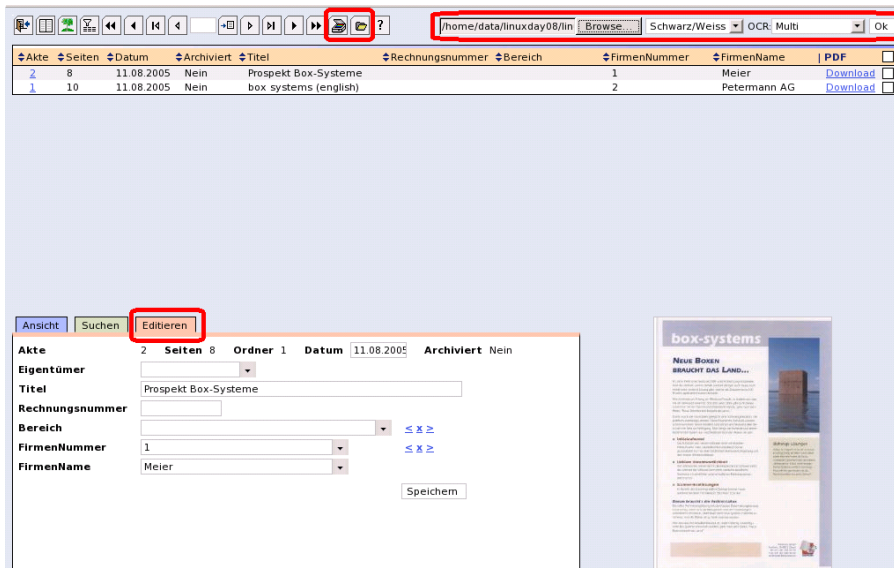
**Archivista WebClient**  
 Version 2008/XI - Powered by [Archivista GmbH](#)

Host: localhost  
 Datenbank: archivista  
 Username: Admin  
 Passwort:   
 Sprache: Deutsch

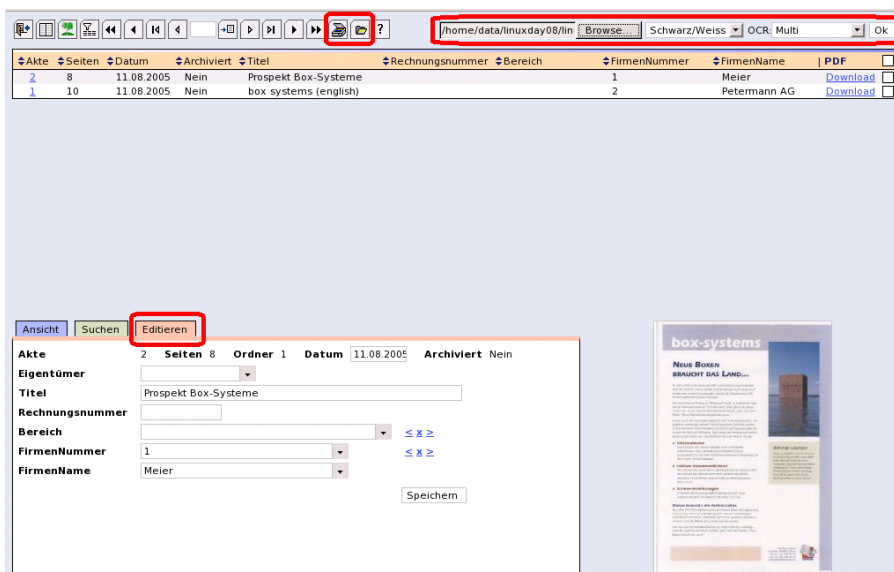
Anmelden

Die Anmeldung erfolgt mit dem Passwort 'archivista'. In der Hauptansicht müssen wir entweder das Symbol zum Scannen (Gerät muss angeschlossen sein) oder den Dateiupload wählen. Wählen wir den Dateiupload, weil ja nicht immer ein Scanner vorhanden ist.



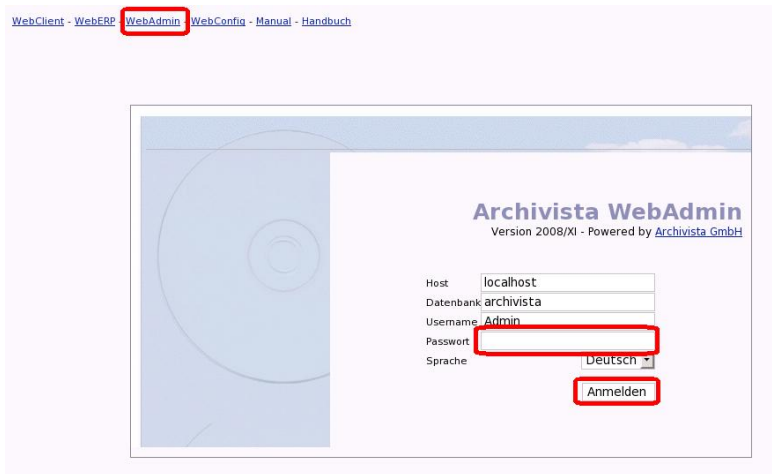


Sobald eine lokale Grafik- oder PDF-Datei ausgewählt ist, kann der Import-Vorgang gestartet werden. Die Datei wird umgehend bearbeitet und der OCR-Erkennung zugeführt. Im nächsten Beispiel sehen wir die importierte Seite samt OCR erkanntem Text.

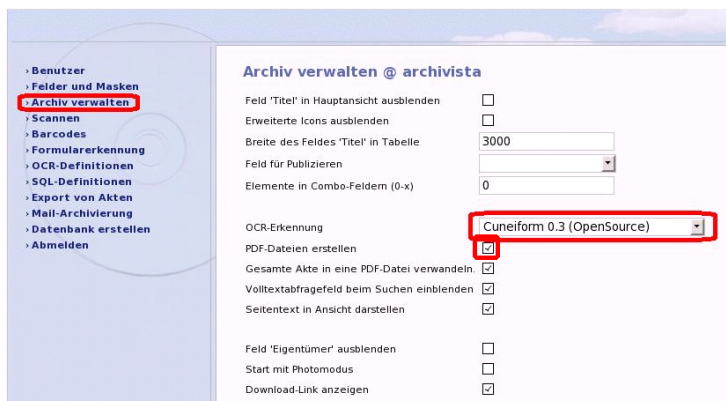


## 6.5 OCR-Erkennung und PDF-Dateien erstellen überprüfen

Sollte dies so nicht funktionieren, muss die OCR-Erkennung aktiviert werden. Dazu melden wir uns im WebAdmin-Modul an:



Nach dem Anmelden können wir unter 'Archiv verwalten' testen, ob die OCR-Erkennung aktiviert ist.



PDF-Dateien werden automatisch erstellt, sofern die Option 'PDF-Dateien erstellen' aktiviert ist.

# 7 Was bringt die Zukunft

## 7.1 Es gibt noch viel zu tun

Mit dem letzten Jahr bin ich an sich hochzufrieden. Ich hätte nicht gedacht, dass ich im Vergleich zum letzten Jahr derart tolle Produkte präsentieren darf.

Aber, es gibt noch immer viel zu tun. Folgende Wünsche würde ich an dieser Stelle gerne an den Weihnachtsmann senden wollen:

- Mehrsprachige OCR-Erkennung (z.B. Deutsch und Französisch auf einer Seite)
- Trainierbare Zeichensätze (z.T. bei OCROpus bzw. Tesseract vorhanden)
- Vollständige OCR-Applikation (inkl. Blöcke setzen/löschen im Rahmen einer Web-Applikation)

Ob mich der Weihnachtsmann dieses oder nächstes Jahr auch wieder derart generös beschenken wird, das werden wir sehen. Vielen Dank für Eure Aufmerksamkeit.