

ARCHIVISTABOX 2008/IX: WELTWEIT ERSTE OPENSOURCE- TEXTERKENNUNG MIT DURCHSUCH- BAREN PDF-DATEIEN

Pfaffhausen, 19.09.2008: Die Schweizer OpenSource-Firma Archivista veröffentlicht mit der ArchivistaBox 2008/IX die weltweit erste OpenSource-Texterkennung, welche durchsuchbare PDF-Dateien erstellen kann.

Gängige Texterkennungsprogramme (OCR) laufen derzeit fast ausschliesslich unter Windows und sind ab Preisen von ca. 100 Euro an aufwärts käuflich zu erwerben. Geht es darum Tausende oder Millionen von Seiten zu verarbeiten, so fallen kostspielige Volumenlizenzen an, d.h. bezahlt wird pro erkannte Seite.

Die ArchivistaBox ist ein webbasiertes DMS-System (Dokumenten-Management), das auf jedem handelsüblichen Rechner installiert werden kann. Je nach Hardware können dabei Seitenvolumen von einigen Tausend Seiten bis in den Millionenbereich pro Tag verarbeitet werden.

Das neue Release 2008/IX beinhaltet die weltweit erste OpenSource-Texterkennung, welche direkt aus gescannten Seiten durchsuchbare PDF-Dateien erstellen kann. Dabei stehen mehr als 20 Sprachen zur Verfügung. Die Erkennungsqualität ist mit kommerziellen OCR-Programmen gut und gerne vergleichbar (>99 Prozent).

Mit der ArchivistaBox erstellte PDF-Dateien werden direkt in einer Archivista-Datenbank abgelegt und automatisch beschlagwortet, d.h. es kann über den gesamten Dokumentenbestand recherchiert werden. Einmal erfasste Dokumente sind jederzeit mit einem Webbrowser abrufbar. Sensitive Daten können verschlüsselt zur Verfügung gestellt werden. Bei Bedarf erstellt die ArchivistaBox fertige DVD-Publikationen (selbsttragende Archive).

Die Quellen der ArchivistaBox liegen zu 100 Prozent in der GPLv2-Lizenz vor. Für die Texterkennung stehen Tesseract (inkl. Frakturerkennung) und der Linux-Port von Cuneiform (BSD-Lizenz) zur Verfügung. Die durchsuchbaren PDF-Dateien werden mit dem Hilfsprogramm hocr2pdf erstellt (siehe www.exactcode.de).

Die ArchivistaBox kann als ISO-CD direkt bei sourceforge.net bezogen werden. Ebenfalls erhältlich ist die Lösung auf einer 1 kg schweren Embedded-ArchivistaBox direkt beim Hersteller oder einem Vertriebspartner. Die ArchivistaBoxen laufen mit ca. 2000 Scannern (Liste siehe www.archivista.ch, Stichwort Scanner-Parade) sowie sämtlichen FTP-fähigen Digitalkopierern. Dabei müssen keinerlei Treiber installiert werden, da diese bereits auf der CD bzw. der Box enthalten sind.

Die aktuelle ArchivistaBox 2008/IX wird am 24. September 2008 auf der MediaMit in Kaiserslauten (KaMUX-Stand) sowie am 24./25. September 2008 auf der OpenExpo in Winterthur (Archivista-Stand) präsentiert. Die ca. 700 MByte grosse Installations-CD steht zum Download unter <https://sourceforge.net/projects/archivista/> bereit.



Über Archivista GmbH

Archivista GmbH vertreibt seit 1998 DMS-Produkte. Seit dem Jahre 2005 liegt der Fokus bei der Entwicklung quelloffener Applikationen wie der ArchivistaBox. Die Firma wird seit 1998 vom Geschäftsführer Urs Pfister geleitet und befindet sich zu 100 Prozent in privatem Besitz.

Archivista GmbH
CH-8118 Pfaffhausen
Tel: +41 (0)1 254 54 00
Fax: +41 (0)1 254 54 02
webmaster@archivista.ch

