Weltweit erste OpenSource-Texterkennung mit durchsuchbaren PDF-Dateien

Pfaffhausen, 19. September 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 vonr ArchivistaDMS 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 aktuelle ArchivistaBox 2008/IX wird am am **24./25. September 2008 auf der**

OpenExpo in Winterthur (Archivista-Stand) präsentiert.