

The world’s first open source text recognition with searchable PDF files

Pfaffhausen, Switzerland, 19th September 2008: *With their launch of the ArchivistaBox 2008/IX, Archivista, a Swiss open source software company, has released the only open source text recognition software worldwide that can create searchable PDF files.*

The majority of current text recognition or OCR (optical character recognition) programs run only on Windows systems and can be purchased for prices from around 100 Euro upwards. When, however, thousands or millions of pages are to be processed, then expensive volume licenses, that are based on a price per scanned page, are required.

The ArchivistaBox is a web based DMS (document management system), that can be installed on every commercially available computer. Depending on the hardware used, the **page volume processed can vary between several thousand up to several million pages** per day.

Release of the 2008/IX marks **the launch of the first open source text recognition system that is able to generate searchable PDF files** directly from scanned pages. More than 20 languages are available and the recognition quality is comparable with that of commercial systems (>99 percent).

PDF files generated with the ArchivistaBox are stored in an Archivista database and automatically indexed, allowing the whole document stock can be researched. Documents scanned can be called up with a web-browser at any time. Sensitive data can be encrypted before being made available. If required, the ArchivistaBox can create complete DVD publications (‘self-supporting’ archives).

100 % of the source code used in the ArchivistaBox comes under the GPLv2 license. **Tesseract (including fracture / black-letter recognition)** and the **Linux port of Cuneiform** (BSD licence) OCR engines are used for text recognition. The hocr2pdf module (see www.exactcode.de) is used to generate the searchable PDF files.

The ArchivistaBox 2008/IX CD (700 MByte) can be downloaded from https://sourceforge.net/projects/archivista/ or www.archivista.ch.