SEALSYSTEMS
THE DIGITAL PAPER FACTORY

FACTSHEET
Text Recognition - OCR

Optical Character Recognition (OCR) of Text in Scanned Documents

What is OCR Text Recognition?

Optical character recognition allows text in scanned documents to be digitized.The OCR process from SEALSystems functions for rasterized and vector data and can be integrated into automated processes.

What does OCR Text Recognition?

OCR techniques can make text machine-readable, which makes them automatically searchable. Large numbers of files are then pre-examined by search engines, so that searching all the data can be executed very quickly.

Who needs OCR Text Recognition?

Users who want to process documents that have no character codes.

Your benefits

- **Information in files can be found more quickly if you are searching not only keywords in the DMS but also searching directly in the files for relevant terms. Therefore, the visible text must be searchable.**
- **Data exchange in supply chains requires that documents not be managed solely by a DMS. The usability of files is greatly increased if you can enter relevant keywords for ranking the files directly from the file.**
- **PDF/A is increasingly replacing the raster format TIFF as an archive format. Inventory files in TIFF and scanned documents are especially easy to convert to PDF format. But without additional OCR processing, this conversion does not add value. The resulting PDF contains no other useful data other than a raster image. Only the enrichment with text elements adds value.**

OCR - Text Recognition

Applications

There are several ways in which text in documents is not recognized as such by search engines:

- Scanned documents usually contain only raster data (pixels).
- Engineeringtools (CAD) and layout applications display letters as lines or planes.
- **Images of text are treated like photos**.

To make text readable to computers, it must be rendered as characters with a unique character code from a font. If these character codes are not correct, the text may be visible, but not searchable.

The process

As a first step, the files are examined for possible existing text. All graphic elements that contain or represent text, such as zigzags, images of texts or rasterized texts are converted into a raster format. These uniform data serve as the basis for OCR recognition. The found text can now be stored as an additional layer in the output file can bemade available as a separate text file in the subsequent process.

Integration

SEAL Systems OCR comes as a Working Unit for the DPF (Digital Process Factory®), so this functionality can be quickly integrated into all processes:

- Conversion processes
- Release processes
- Checking documents in DMS
- Converting old files

Input formats

BMP, PCX, DCX, JPEG, **TIFF**, PNG, GIF, DjVu, **PDF** (through 1.6)

Output formats

Text in separate text file or XML files or as an additional layer inPDF.

System environment

See all details for reccommended computer equipment under www.sealsystems.com/service-support/computer-equipment/

Product code

OCR-25PM, OCR-75PM, OCR-200PM, OCR-500PM, OCR-500T

Uwe Wächter and Debra Garls are specialist for your questions concerning:

Conversion and PDF processing

Europe/Asia/Australia
Dr. Uwe Wächter
Tel +49 6154 637 372
uwe.waechter@sealsystems.de

USA/Canada/Americas
Debra Garls
Tel +1 774 200 0933
debra.garls@sealsystems.com

SEALSYSTEMS
THE DIGITAL PAPER FACTORY

E-Mail: info@sealsystems.com
Web: www.sealsystems.com

OUTPUT MANAGEMENT
CORPORATE SOLUTIONS BY SEAL SYSTEMS

We would be happy to answer all of your questions around conversion and integration in your company.

Status: 21. January 2019 V512-120327-0-de