

## 3-Heights™ OCR Enterprise Add-On

The 3-Heights™ OCR Enterprise Add-On complements several products of PDF Tools AG with a high performance optical character recognition (OCR) function. This allows for converting images such as TIFF or JPEG to PDF or PDF/A, or converting PDF to PDF/A and applying OCR at the same time. The customer has a free choice of the OCR Engine he wants to use. At this time ABBYY FineReader 10 and Tesseract are available in different types of licensing models. Depending on the requirements for recognition rate, throughput and costs, an adequate model can be selected.




### Properties and Use

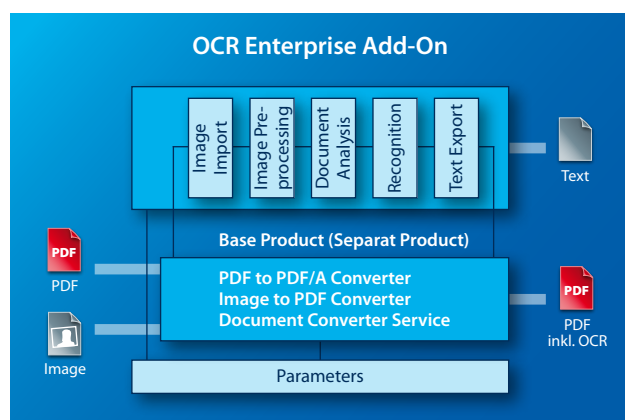
The 3-Heights™ OCR Enterprise Add-On is a module, which can be used in conjunction with other products of PDF Tools AG.

Based on the selected OCR engine, it recognizes textual content and embeds it as

Unicode text into the PDF and PDF/A document. Thus, created PDF documents are fully text searchable. Various options related to image manipulation, pre-processing and text recognition allow for an optimally tuned recognition process.

### Product Variants

API	Shell	Service
		



### Areas of Use

#### Inbox

Recognition of texts while scanning incoming mail. Usage of texts in the metadata of incoming documents and in the downstream business processes, for example ERP and Workflow Systems. Direct archiving of incoming documents with text recognition. Text recognition in scanned email attachments for easier processing.

#### Archiving

Apply text recognition when converting archives from TIFF or PDF to PDF/A. Convert proprietary formats to PDF/A and embed texts. Recognize information on index pages and transmit them to the metadata of the document or dossier.

#### Further Areas of Use

- Unpacking scanned email attachments
- Preparing for archiving
- Archive migration

# Technical Details

## Formats

### Input Formats

Defined by base product:

- 3-Heights™ PDF to PDF/A Converter
  - PDF
- 3-Heights™ Image to PDF Converter
  - TIFF (Tagged Image File Format)
  - JPEG (Joint Photographic Expert Group)
  - PNG (Portable Network Graphics)
  - GIF (Graphics Interchange Format)
  - BMP (Window Bitmap)
  - EPS (Encapsulated Post Script)
  - JB2 (JBIG2, Joint Bi-level Image Experts Group)
  - JP2 (JPEG2000)
  - JPX (Extended JPEG2000)
  - PBM (Portable Bitmap File Format)
  - JIF (GIF Flate)
- 3-Heights™ Document Converter Service
  - Microsoft Office 2003 and 2007 documents
  - Document of older Microsoft Office versions
  - Simple Text
  - WordPerfect
  - HTML
  - Outlook (MSG)
  - PDF
  - Internet Mail Message Format
  - Image formats (TIFF, JPEG, PNG, JBIG2, JPX, GIF, BMP, etc.)
  - ZIP und TAR Archive
  - Add-ins for customer specific formats

### Target Formats

- PDF, PDF/A

## Variants and Required Base Products

### Required Base Products

The 3-Heights™ OCR Enterprise Add-On can be used in conjunction with the following products:

- 3-Heights™ PDF to PDF/A Converter
- 3-Heights™ Image to PDF Converter
- 3-Heights™ Document Converter Service

### Platforms

#### Operating Systems

- ABBYY:
  - Windows 2000, XP, Vista, 7
  - Windows Server 2003, 2008, 2008 R2 – 32 and 64 Bit
- Tesseract: Windows, Linux, others on request

### Interfaces and Programming Languages

The OCR engine is installed as a separate product. However it does not come with a visible interface. Instead it is accessed directly from within the base product. E.g. the base product is an API: this API provides the OCR related functions and properties; the base product is a shell tool or service: there are corresponding switches available to control the OCR related features.

#### Interfaces

Defined by base product:

- API: C, Java, .NET, COM
- Shell Tool: Command line tool for batch processing
- Windows Service: Windows service with watched folders

## Performance Characteristics

- Choice of the OCR engine, available are:
  - ABBYY FineReader OCR Engine 10
  - Tesseract OCR Engine 2
- Depending on the OCR engine, up to 200 languages are supported and for many languages, there is additional support by means of dictionaries and morphologic tools
- Migration of license keys (ABBYY 10 only)

## Functions

### General Functions

- Add OCR text information to PDF documents
- Set the language of the OCR text to increase the recognition rate
- Direct access to the OCR engine or synchronized use via a service
- Recognition of multi language documents

### Tesseract Related Functions

- Support for 8 different languages by means of dictionaries
- Optimized for images with a resolution of 300 dpi
- Automatic detection of the base line

### ABBYY Related Functions

- Recognition of almost 200 languages with machine generated contents
- Extended support of almost 50 languages with dictionaries and morphological tools
- Recognition of typewriter scripts
- Recognition and decoding of barcodes (1D) Recognition of type of content (images vs. texts)
- Modules to support additional languages
  - Chinese, Japanese, Korean
  - Old European Languages
  - 2D Barcode
- Select normal, fast and balanced mode
- De-Skewing: Automatic image alignment
- Image clean-up: Unwanted artifacts are recognized and eliminated
- Filtering of non-relevant backgrounds
- Recognition and correction of page orientation
- Creation and use of profiles that summarize the above features