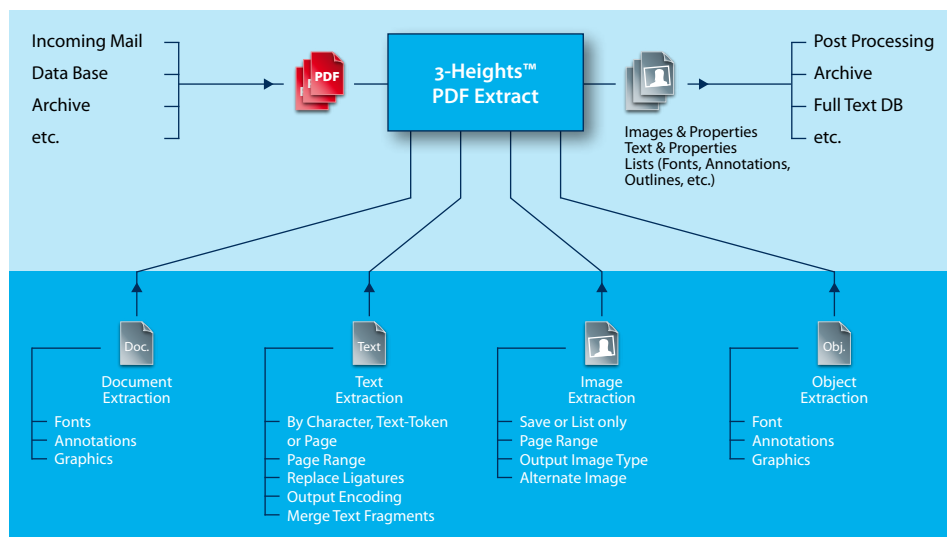
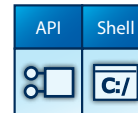


3-Heights™ PDF Extract

3-Heights™ PDF Extract is a component for reading out the contents and properties of PDF documents. PDF documents are used to store important information relating to products, customer data and corporate knowledge. Meta information such as the document's creator, date of creation or date of modification are further integral parts of a PDF document. PDF documents are often used as „containers“ to enable the transfer of text, images, videos and other data to other processes independently of the platforms in use. This component can extract information quickly and efficiently, regardless of whether document content or document properties. The results can be stored in a database, for instance, or used for evaluations and statistics or to secure internal corporate knowledge



Product Variants



Properties and Benefits

Texts extracted using the 3-Heights™ PDF Extract Tool can be used for indexing documents or in search engines, for instance. The component is generally used to extract data and resources from a PDF document for further processing. Highly detailed information is available for the purpose, which can also be transferred to document management systems (DMS) in various forms.

Areas of Use

Incoming Mail and Document Processing

Content from PDF files such as forms or scanned incoming invoices, for instance, is extracted and processed for characterization or indexing.

Outgoing Mail

PDF documents are restructured in preparation for use by other target groups. The process reads out processing information such as barcodes, address information or page formats that can then be used for controlling printing and packaging lines or sorting processes.

Archiving

Texts or their components are extracted for separate storage in metadata. This allows document indexing to be extended as required.

Other Areas of Use

- Convert PDF documents into text documents

- Extract information such as addresses, invoice data and report data from documents for process control purposes
- Extract information for document classification and document indexing
- Process data in forms
- Extract images for further processing (scans, photos, etc.)
- Analyze and evaluate the content of PDF documents in mass processing

Technical Details

Formats

Input Formats

- PDF

Compliance

- Standards: ISO 32000 (PDF 1.7)

Platforms

Operating Systems

- Windows 2000, XP, Vista, 7
- Windows Server 2003, 2008, 2008 R2 – 32 and 64 Bit
- HP-UX – 32 Bit and Itanium
- IBM AIX – 32 and 64 Bit
- Linux (SuSE and Red Hat on Intel)
- Mac OS X
- Sun Solaris

Interfaces and Languages

Interfaces

- API: C, Java, .NET, COM

Programming Languages

All program libraries are written in efficient and thread-safe C++. API offers a selection of the following connections to programming languages:

- C#, VB .NET, J# via .NET
- Java via JNI
- MS Visual Basic, Borland Delphi, MS Office products such as Access and C++ via COM
- C and C++ via native C

Variants and Options

Product Variants

- Shell tool (command line)
- API (programming interface)

Performance Characteristics

- Extract text by the character, word or page (including invisible text)
- Search for keywords and retrieve their position
- Extract images (including alternative images)
- Retrieve form fields
- Extract document information such as version, encryption, linearization and metadata
- List fonts and color spaces
- Extract page information and page descriptions (graphic objects, position and other attributes)
- Extract bookmarks

Functions

Information is extracted on the basis of the object type. The component supports the following objects and their respective properties:

- Document
- Page
- Page Content
- Text
- Font Type
- Color Space
- Image
- Graphics State
- Transformation Matrix
- Annotation
- Bookmarks
- Destination

The detailed list, see www.pdf-tools.com

