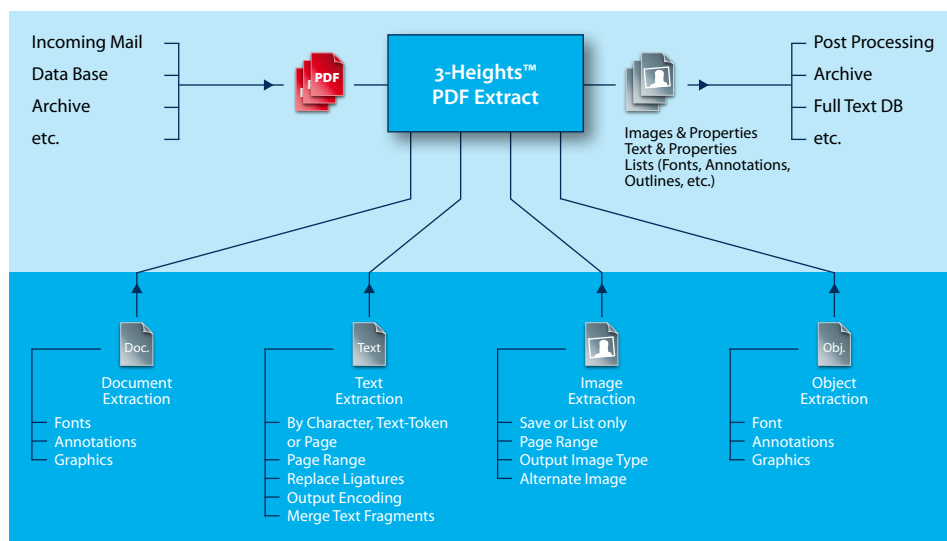


## 3-Heights™ PDF Extract

3-Heights™ PDF Extract ist eine Komponente zum Auslesen von Inhalten und Eigenschaften von PDF Dokumenten. Wichtige Informationen wie Produkte Informationen, Kundendaten oder Firmen Wissen werden in PDF Dokumenten abgelegt. Metainformationen, wie der Ersteller des Dokumentes, das Erstellungsdatum oder Änderungsdatum sind Bestandteil eines PDF Dokumentes. Oft werden PDF Dokumente als „Container“ verwendet, so dass Text, Bilder, Videos und andere Daten Plattform unabhängig an andere Arbeitsprozesse übermittelt werden können.

Die Komponente kann diese Informationen, sei es Inhalt oder Dokument Eigenschaften, schnell und effizient auslesen. Die Resultate können z. B. in Datenbanken gespeichert, für Auswertungen und Statistiken angewendet oder zur Sicherstellung von Firmen internem Wissen abgelegt werden.



### Eigenschaften und Nutzen

Die mit 3-Heights™ PDF Extract extrahierten Texte können beispielsweise für die Indexierung von Dokumenten oder für Suchmaschinen verwendet werden. Die Komponente dient generell zur Suche und Extraktion von Daten und Ressourcen aus einem PDF Dokument, um diese weiter verarbeiten zu können. Dazu stehen äußerst detaillierte Informationen zur Verfügung, die in verschiedenen Formen z. B. an DMS Systeme übergeben werden können.

### Einsatzgebiete

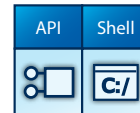
#### Posteingang und Dokumentverarbeitung

Inhaltsteile von PDF Dateien, z. B. von Formularen oder gescannten Eingangsrechnungen, werden extrahiert und für die Charakterisierung oder Indexierung aufbereitet.

#### Postausgang

PDF Dokumente werden umstrukturiert, um diese für andere Zielgruppen optimal aufzubereiten. Verarbeitungsinformationen wie Barcodes, Adressinformationen oder Seitenformate können ausgelesen und für die Steuerung von Druck- und Ver-

### Produktvarianten



packungsstrassen oder Sortierungsprozesse verwendet werden.

### Archivierung

Texte oder deren Bestandteile werden für eine separate Speicherung in den Metadaten extrahiert. Damit lässt sich die Indexierung der Dokumente bedürfnisgerecht erweitern.

### Weitere Einsatzgebiete

- Umwandlung von PDF in Textdateien

- Auszug von Informationen wie Adressen, Rechnungsdaten, Berichtsdaten aus Dokumenten für die Prozesssteuerung
- Auszug von Informationen für die Dokumentenklassifikation und Dokumentenindexierung
- Verarbeitung von Formulardaten
- Auszug von Bildern für die Weiterverarbeitung (Scans, Fotos usw.)
- Analyse und Auswertung von Inhalten in PDF Dokumenten in der Massenverarbeitung

## Technische Daten

### Formate

#### Eingangsformate

- PDF

#### Compliance

- Standards: ISO 32000 (PDF 1.7)

### Plattformen

#### Betriebssysteme

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

### Schnittstellen und Sprachen

#### Schnittstellen

- API: C, Java, .NET, COM

#### Programmiersprachen

Alle Programmbibliotheken sind in effizientem und Thread sicherem C++ geschrieben. In der API wird eine Auswahl der folgenden Anbindungen an Programmiersprachen angeboten:

- C#, VB .NET, J# via .NET
- Java via JNI
- MS Visual Basic, Borland Delphi, MS Office Produkte wie Access und C++ via COM
- C und C++ via native C

### Varianten und Optionen

#### Produktvarianten

- Shell Tool (Befehlszeile)
- API (Programmierschnittstelle)

### Leistungsmerkmale

- Text Zeichen-, Wort- und Seiten weise extrahieren (auch wenn nicht sichtbar)
- Nach Schlüsselwörter suchen und deren Position auslesen
- Bilder extrahieren (auch alternative Bilder)
- Formularfelder auslesen
- Dokumenteninformationen wie Version, Verschlüsselung, Linearisierung und Metadaten extrahieren
- Schriften und Farbräume auflisten
- Seiteninformationen und Seitenbeschreibung (Grafikobjekte, Position und weitere Attribute) extrahieren
- Lesezeichen extrahieren

### Funktionen

Informationen werden ja nach Objekt Typ extrahiert. Folgende Objekte und deren Eigenschaften sind unterstützt:

- Dokument
- Seite
- Seiteninhalt
- Text
- Schrift
- Farbraum
- Bild
- Grafikstatus
- Transformationsmatrix
- Annotation
- Lesezeichen
- Destination

Die ausführliche Liste finden Sie unter [www.pdf-tools.com](http://www.pdf-tools.com)

