

## Datenblatt

# PDFlib TET PDF IFilter 4

## Enterprise PDF Search für Windows

### Was ist PDFlib TET PDF IFilter?

TET PDF IFilter extrahiert Text und Metadaten aus PDF-Dokumenten, um sie Retrieval-Produkten unter Windows zugänglich zu machen. Damit haben Sie die Möglichkeit, die PDF-Dokumente auf Ihrem Desktop-Computer, dem Enterprise-Server oder im Web zu durchsuchen. TET PDF IFilter basiert auf dem patentierten Entwicklungswerkzeug PDFlib Text Extraction Toolkit (TET), mit dem sich Text zuverlässig aus PDF-Dokumenten extrahieren lässt. TET PDF IFilter ist eine stabile Implementierung der Microsoft IFilter-Schnittstelle zur Volltextindizierung und arbeitet mit allen Produkten zur Textabfrage zusammen, die die IFilter-Schnittstelle unterstützen, z.B. SharePoint oder SQL Server. Diese Produkte verwenden für jedes Dateiformat, z.B. HTML, ein anderes format-spezifisches Filterprogramm, das IFilter genannt wird. TET PDF IFilter ist ein solches Filterprogramm für PDF-Dokumente. Die Benutzerschnittstelle zum Durchsuchen der Dokumente kann Windows Explorer, ein Web- oder Datenbank-Frontend, ein Abfrageskript oder eine selbst entwickelte Anwendung sein. Alternativ zur interaktiven Suche mittels Benutzeroberfläche lassen sich Anfragen über eine Programmierschnittstelle absetzen.

### Patenterte TET-Technologie

PDFlib TET, das die Grundlage von TET PDF IFilter bildet, wurde erstmals 2002 veröffentlicht und hat sich weltweit in Server- und Desktop-Systemen bewährt. TET extrahiert nicht nur PDF-Seiteninhalte und Metadaten als Rohtext, sondern liefert den Dokumentinhalt alternativ auch im XML-Format. TET ist auch als kostenloses Plugin für Adobe Acrobat verfügbar; damit können Sie die hervorragende Text- und Bildextraktion von TET interaktiv testen und evaluieren.

### Besondere Vorteile

TET PDF IFilter bietet folgende Vorteile:

- ▶ Unterstützt westlichen, chinesischen, japanischen und koreanischen Text, sowie von rechts nach links laufende Sprachen wie Arabisch und Hebräisch.
- ▶ Indiziert auch geschützte Dokumente und extrahiert Text sogar aus PDFs, bei denen Acrobat scheitert
- ▶ Unterstützt Unicode-Nachbearbeitung durch Folding, Decomposition und Normalisierung
- ▶ Leistung: thread-sicher, schnell und stabil, 32- und 64-Bit
- ▶ Automatische Erkennung von Sprache und Schriftsystem

### Unternehmensweite Suche in PDF-Dokumenten

TET PDF IFilter ist in thread-sicheren nativen 32- und 64-Bit-Versionen verfügbar. Unternehmensweite Lösungen zur Textsuche lassen sich in Kombination mit folgenden Produkten implementieren:

- ▶ Microsoft Office SharePoint Server
- ▶ Microsoft Search Server
- ▶ Microsoft SQL Server
- ▶ Microsoft Exchange Server
- ▶ Microsoft Site Server

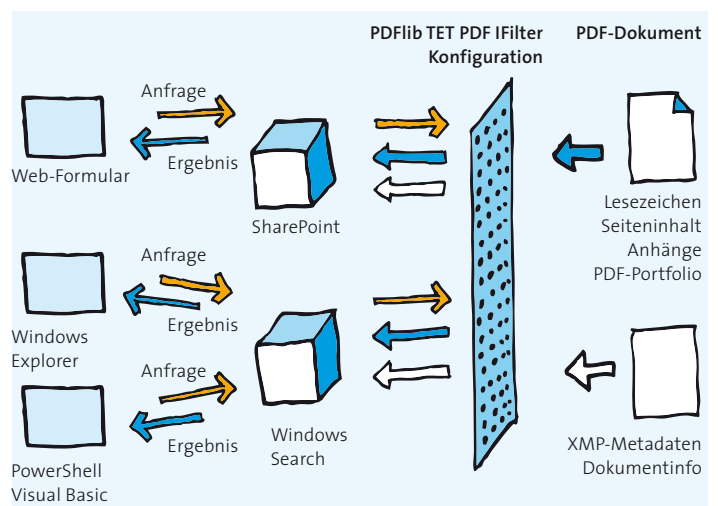
TET PDF IFilter ist mit allen Produkten von Microsoft und anderen Herstellern einsetzbar, die die IFilter-Schnittstelle unterstützen.

### Desktop-Suche in PDF-Dokumenten

TET PDF IFilter unterstützt auch die Suche nach PDF-Daten auf dem Desktop-Computer, zum Beispiel mit folgenden Produkten:

- ▶ Windows Search ist in Windows Vista/7 integriert, aber auch als kostenloser Zusatz für Windows XP verfügbar
- ▶ Windows Indexing Service

TET PDF IFilter ist für den nicht-kommerziellen Einsatz auf Desktop-Systemen kostenlos verfügbar und bietet damit eine bequeme Basis zum Testen und Evaluieren.



## Funktionalität im Detail

### Akzeptierte PDF-Eingabe

TET PDF IFilter verarbeitet alle gängigen Varianten von PDF:

- ▶ Alle PDF-Versionen bis Acrobat 9, inklusive ISO 32000-1
- ▶ Geschützte PDFs, die zum Öffnen kein Kennwort erfordern
- ▶ Beschädigte PDF-Eingabedokumente werden repariert

### Unicode-Nachbearbeitung

TET PDF IFilter unterstützt Nachbearbeitung zur Verbesserung der Suchergebnisse:

- ▶ Foldings erhalten, entfernen oder ersetzen Zeichen, um z.B. Interpunktionszeichen oder Zeichen aus einem irrelevanten Schriftsystem herauszufiltern.
- ▶ Decompositions ersetzen ein Zeichen mit einer äquivalenten Folge von einem oder mehreren anderen Zeichen, z.B. beim Ersetzen eines chinesischen Zeichens durch das kanonisch äquivalente Unicode-Zeichen.
- ▶ Text kann in alle vier Unicode-Normalformen konvertiert werden, z.B. um Texte in NFC-Form auszugeben, damit sie den Anforderungen einer Datenbank entsprechen.

### Internationalisierung

Neben westlichem Text unterstützt TET PDF IFilter chinesischen, japanischen und koreanischen (CJK) Text. Alle CJK-Kodierungen werden erkannt; horizontale und vertikale Schreibrichtung werden korrekt behandelt. Die automatische Erkennung von Sprache bzw. Schriftsystem des extrahierten Textes (Locale ID) verbessert die Ergebnisse von Microsofts Algorithmen zur Bestimmung von Wortgrenzen und Wortstämmen, was insbesondere bei ostasiatischem Text wichtig ist.

Von rechts nach links laufende Schriften wie Hebräisch und Arabisch werden auch unterstützt. Dabei normalisiert TET PDF IFilter kontextabhängige Zeichenformen und sortiert den Text in logische Reihenfolge um.

### PDF enthält mehr als nur Seiten

TET PDF IFilter behandelt PDF-Dokumente als Container für weit mehr als nur die Seiteninhalte und indiziert alle relevanten Elemente eines PDF-Dokuments:

- ▶ Seiteninhalte
- ▶ Text in Lesezeichen
- ▶ Metadaten (siehe unten)
- ▶ Eingebettete PDFs und PDF-Pakete werden rekursiv verarbeitet, so dass sich auch Text in PDF-Dateianhängen durchsuchen lässt.

### XMP-Metadaten und Dokumentinfelder

Die leistungsfähige Metadaten-Implementierung von TET PDF IFilter unterstützt das Property-System von Windows für Metadaten. TET PDF IFilter indiziert XMP-Metadaten sowie Standard- und benutzerdefinierte Dokumentinfelder. Die Indizierung der Metadaten lässt sich auf verschiedenen Ebenen konfigurieren:

- ▶ Dokumentinfelder, Dublin-Core-Felder und andere gängige XMP-Properties werden auf entsprechende Windows-Properties wie *Title*, *Subject* oder *Author* abgebildet.
- ▶ TET PDF IFilter ergänzt nützliche PDF-spezifische Pseudo-Properties wie Seitengröße, PDF/A-Konformitätslevel oder Fontnamen.
- ▶ Nach allen relevanten vordefinierten XMP-Properties kann gesucht werden.
- ▶ Die Suche umfasst auch benutzerdefinierte XMP-Properties wie firmenspezifische Klassifizierungen oder PDF/A-Extension-Schemas.

TET PDF IFilter bietet optional die Möglichkeit, Metadaten in den indizierten Rohtext zu integrieren. Damit können auch Volltextsuchmaschinen ohne Metadaten-Unterstützung (z.B. SQL Server) nach Metadaten suchen.

## Vorteile von PDFlib-Software

### Zuverlässig

Weltweit arbeiten viele Tausend Programmierer erfolgreich mit unserer Software. PDFlib-Produkte erfüllen alle Qualitäts- und Geschwindigkeitskriterien für den Einsatz auf großen Servern. Alle Produkte sind für den zuverlässigen, unbeaufsichtigten 24-Stunden-Betrieb ausgelegt.

### Schnell und einfach

PDFlib-Produkte sind unglaublich schnell – bis zu Tausenden von Seiten pro Sekunde. Die Programmierschnittstelle ist übersichtlich und einfach zu erlernen.

### PDFlib-Produkte sind überall

Unsere Produkte unterstützen alle internationalen Sprachen sowie Unicode. Sie werden von Kunden in der ganzen Welt eingesetzt.

### Professioneller Support

Bei Problemen bietet Ihnen unser Support-Team professionelle Unterstützung. Um den reibungslosen Ablauf unternehmenskritischer Anwendungen zu gewährleisten, können Sie Ihre Software-Lizenz durch einen Supportvertrag ergänzen. Ein Supportvertrag garantiert Ihnen kurze Antwortzeiten und Zugang zu den jeweils neuesten Versionen.

### Lizenzierung

Bei der Lizenzierung können Sie zwischen verschiedenen Modellen für Server-, Integrations-, Firmen- sowie Quellcodelizenzen wählen. Ergänzend bieten wir Supportverträge für umfangreichen technischen Support mit kurzen Reaktionszeiten und kostenlosen Software-Aktualisierungen an.

### Über PDFlib GmbH

PDFlib GmbH ist auf die Entwicklung von PDF-Technologie spezialisiert. PDFlib-Produkte sind seit 1997 weltweit im Einsatz. Das Unternehmen berücksichtigt wichtige technologische Trends, etwa ISO-Standards für PDF. PDFlib GmbH vertreibt alle Produkte weltweit, wobei Nordamerika, Europa und Japan die wichtigsten Märkte darstellen.

### Kontakt

Evaluierungsversionen mit vollem Funktionsumfang und Dokumentation sowie Beispielen sind auf unserer Webseite verfügbar. Weitere Informationen erhalten Sie unter:



#### PDFlib GmbH

Franziska-Bilek-Weg 9, D-80339 München  
Tel. +49 • 89 • 452 33 84-0, Fax +49 • 89 • 452 33 84-99  
sales@pdflib.com  
www.pdflib.com