

PRESS RELEASE

Neue Versionen von PDFlib TET, PDFlib TET PDF IFilter und TET Plugin

PDFlib TET 4 Produktfamilie verfügbar

Die neue Version der bewährten TET-Engine zur Extraktion von PDF-Inhalten bietet verbesserte Analyse von Seiteninhalten, unterstützt von rechts nach links laufende Schriftsysteme wie Arabisch und Hebräisch sowie ausgefeilte Unicode-Nachbearbeitungsfunktionen.

München, 30. Juli 2010. PDFlib GmbH gibt die Verfügbarkeit der neuen Produktversionen PDFlib TET 4.0, PDFlib TET PDF IFilter 4.0 und TET Plugin 4.0 bekannt. Mit den neuen Versionen lassen sich Inhalte noch schneller, effizienter und zuverlässiger aus PDF-Dokumenten extrahieren.

TET 4. PDFlib TET (Text Extraction Toolkit) extrahiert zuverlässig Text, Bilder und Metadaten aus PDF-Dokumenten. TET stellt den Text eines PDF-Dokuments als Unicode-Strings zur Verfügung und liefert detaillierte Informationen über Fonts und Zeichen sowie die Position auf der Seite. Rasterbilder werden in gebräuchliche Bilddatenformate extrahiert. Optional kann TET die PDF-Dokumente in ein XML-basiertes Format namens TETML konvertieren, das Text und Metadaten sowie Ressource-Informationen enthält. Neben westlichem Text unterstützt TET japanischen, chinesischen und koreanischen (CJK) Text sowie von rechts nach links laufende Schriften wie Hebräisch und Arabisch.

TET verfügt über ausgefeilte Algorithmen zur Inhaltsanalyse und kann damit Wortgrenzen erkennen, Text zu Spalten zusammenfassen oder redundanten Text entfernen, zum Beispiel Schatteneffekte oder künstliche Fettschrift. Mit der pCOS-Schnittstelle können Sie zudem beliebige Objekte aus einem PDF-Dokument abfragen, zum Beispiel Metadaten oder interaktive Elemente. TET ist für den Serverbetrieb bestimmt (thread-sicher, robust, keine memory-leaks und sauberes Exception-Handling).

Neue Funktionen in TET 4.0. Die neue Version bietet erheblich bessere Performance und arbeitet für viele Dokumententypen erheblich schneller. Insbesondere sehr große Dokumente mit bis zu Hunderttausenden von Seiten profitieren von höherer Geschwindigkeit und geringerem Speicherplatzbedarf. Den Extraktionsergebnissen kommt zugute, dass TET 4 noch besser Schatten, Wortgrenzen, Trennungen und hoch oder tief gestellten Text erkennt. Zahlreiche neue Workarounds für nicht standardkonforme PDF-Dokumente machen die Textextraktion noch robuster. Ein erweiterter Reparaturmodus erlaubt, auch aus beschädigten PDF-Daten Inhalte zu extrahieren. TET 4 bringt bidirektionalen Text in arabischen oder hebräischen Dokumenten in die richtige logische Reihenfolge. Unicode-Nachbearbeitungsfunktionen wie Folding, Decomposition und Normalisierung sind nützlich, um die Textextraktion an die Erfordernisse anderer Programme anzupassen.

TET PDF IFilter 4.0. Basierend auf der patentierten Technologie von TET bietet TET PDF IFilter eine stabile Implementierung der Microsoft IFilter-Schnittstelle zur Volltextindizierung. Die Software arbeitet mit allen Suchprodukten zur Textabfrage zusammen, die die IFilter-Schnittstelle unterstützen, z.B. SharePoint oder SQL Server. Die neue Spracherkennung in TET PDF IFilter 4.0 ordnet dem Text automatisch die richtige natürliche Sprache zu, was für das Word Stemming (Reduzierung eines Suchbegriffs auf den Wortstamm) notwendig ist. Word Stemming macht die Suche deutlich komfortabler.

PRESS RELEASE

TET Plugin 4.0. TET ist auch als kostenloses Plugin für Adobe Acrobat verfügbar; damit können Anwender die hervorragende Text- und Bildextraktion von TET interaktiv testen und evaluieren. Das neue Plugin unterstützt Unicode-Syntax für Suchabfragen und kann alle Treffer auf einer Seite gleichzeitig hervorheben.

TET Cookbook. Das TET Cookbook ist eine Sammlung von Programmierbeispielen, die den Einsatz von TET bei verschiedensten Aufgabenstellungen der Text- und Bildextraktion demonstrieren. Zahlreiche Cookbook-Beispiele zeigen auch, wie sich TET und PDFlib+PDI kombinieren lassen, um PDF-Dokumente anzureichern, etwa durch Lesezeichen oder Links, die auf Basis des Textinhalts erzeugt werden.

Preise und Verfügbarkeit. PDFlib TET 4 für Windows Server 2003/2008, Apple Mac OS X Server oder Linux ist für 795 Euro zu haben. Für Windows 2000/XP/Vista/7 oder Mac OS Desktop liegt der Preis für TET bei 295 Euro. Auch für Sun Solaris, IBM AIX und HP-UX sowie IBM i5/iSeries und zSeries stehen Pakete zur Verfügung.

TET PDF IFilter 4.0 ist für den nicht-kommerziellen Einsatz auf Desktop-Systemen kostenlos verfügbar und bietet damit eine bequeme Basis zum Testen und Evaluieren. Die Lizenzgebühr für TET PDF IFilter auf Windows Server liegt bei 555 Euro.

Das TET Plugin für Acrobat Professional auf Windows und Mac ist für nicht-kommerzielle Nutzung kostenlos verfügbar.

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.