

PRESS RELEASE

Neue Version von TET (Text Extraction Toolkit)

PDFlib TET 3 verfügbar

Die neue Version von PDFlib TET extrahiert Text, Rasterbilder und Tabellen und konvertiert PDF-Dokumente in XML.

München, 10. Februar 2009. PDFlib GmbH gibt die Verfügbarkeit von PDFlib TET 3.0 bekannt. Die neue Version bietet zusätzlich zum Extrahieren von Text eine Reihe von Neuerungen, wie etwa die Extraktion von Rasterbildern und Tabellen, sowie einen erweiterten XML-Export. Verschiedene neue Heuristiken ermöglichen es jetzt, auch den Text aus solchen PDF-Dokumenten zu extrahieren, bei denen dies bisher nicht möglich war, etwa manche PDFs, die mit dem Satzsystem TeX erstellt wurden. Mehrere TET-Konnektoren erleichtern die Anbindung an Datenbanken und Suchmaschinen. Das neue TET-Cookbook demonstriert anhand von Programmierbeispielen den Einsatz von TET.

Extrahieren von Rasterbildern. Bilder auf den Seiten einer PDF-Datei lassen sich als TIFF-, JPEG- oder JPEG-2000-Dateien speichern. Zusätzlich werden genaue geometrische Informationen (Position, Größe und Winkel) für jedes Bild zurückgeliefert. Fragmentierte Bilder werden zu einem größeren Bild zusammengesetzt, um die Wiederverwendung zu vereinfachen. Die originalgetreue Darstellung ist dadurch garantiert, dass beim Extrahieren keine Neuberechnung der Auflösung (Downsampling) und keine Farbraumtransformation stattfinden. Damit erhält TET die höchstmögliche Bildqualität.

Seitenlayout und Tabellenerkennung. TET analysiert die Seiteninhalte, um Textspalten zu erkennen. Tabellen werden erkannt – inklusive Tabellenzellen, die sich über mehrere Textspalten erstrecken. Damit lässt sich die Sortierung des extrahierten Textes verbessern. Tabellenzeilen und der Inhalt einzelner Tabellenzellen werden als solche markiert.

TETML stellt PDF-Inhalte als XML dar. Optional kann TET die Inhalte einer PDF-Datei in einer XML-Variante namens TETML darstellen, mit der sich eine Vielzahl der im PDF enthaltenen Informationen in einer Form ausdrücken lässt, die für gebräuchliche XML-Werkzeuge zugänglich ist. TETML enthält den eigentlichen Text, optional ergänzt durch Schriftinformationen und Positionsangaben, Details zu den Ressourcen (Schriften, Bilder, Farbräume) und Metadaten. TETML wird durch ein zugehöriges XML-Schema definiert, so dass TET immer eine konsistente und zuverlässige XML-Ausgabe garantiert. TETML lässt sich mit XSLT-Stylesheets verarbeiten, um Filter anzuwenden oder TETML in andere Formate zu konvertieren. Die TET-Distribution enthält Beispiele für XSLT-Stylesheets zur Verarbeitung von TETML.

TET-Konnektoren. TET-Konnektoren verbinden TET mit anderer Software. Die folgenden TET-Konnektoren erlauben die Textextraktion aus PDF in verschiedenen Softwareumgebungen:

- Das TET Plugin für Adobe Acrobat ist ein kostenloses Tool, mit dem sich Texte und Bilder extrahieren lassen, und das bessere Funktionen als die in Acrobat eingebauten Werkzeuge bietet. Das TET Plugin lässt sich zur interaktiven Evaluierung der TET-Funktionen nutzen.

PRESS RELEASE

- ▶ TET-Konnektor für die Suchmaschine Lucene
- ▶ TET-Konnektor für den Solr Search Server
- ▶ TET-Konnektor für Oracle Text
- ▶ TET PDF IFilter für Microsoft-Umgebungen ist als separates Produkt erhältlich. Der IFilter extrahiert Text und Metadaten aus PDF-Dokumenten und macht sie Such- und Retrieval-Software unter Windows zugänglich.
- ▶ TET-Konnektor für MediaWiki

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 3 für die Plattformen Windows Server 2000/2003/2008, Apple Mac OS X Server oder Linux ist für 795 Euro zu haben. Für Windows 2000/XP/Vista oder Mac OS liegt der Preis für TET bei 159 Euro. Auch für Sun Solaris, IBM AIX und HP-UX stehen Pakete zur Verfügung.

Über TET. PDFlib TET (Text Extraction Toolkit) ist ein Entwicklungswerkzeug, mit dem sich Text und Rasterbilder zuverlässig aus PDF-Dokumenten extrahieren lassen. TET stellt den Text eines PDF-Dokuments als Unicode-Strings zur Verfügung und liefert detaillierte Informationen über Schriften und Zeichen sowie die Position auf der Seite. Rasterbilder werden in gebräuchliche Bilddatenformate extrahiert. Optional kann TET die PDF-Dokumente in ein XML-basierendes Format namens TETML konvertieren, das Text und Metadaten sowie Ressource-Informationen enthält. TET verfügt über einen ausgefeilten Algorithmus 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.

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.