

PRESS RELEASE

New versions of PDFlib TET (Text Extraction Toolkit), PDFlib TET PDF IFilter, and TET Plugin

PDFlib TET 4 Product Family available

The new version of the established TET PDF content extraction engine improves page content analysis, supports right-to-left languages like Arabic and Hebrew, and offers advanced Unicode postprocessing controls.

Munich, July 30, 2010. PDFlib GmbH announces the availability of the new product versions PDFlib TET 4, PDFlib TET PDF IFilter 4, and TET Plugin 4. The new versions offer faster, more efficient and more reliable content extraction from PDF documents.

TET 4.0. PDFlib TET (Text Extraction Toolkit) reliably extracts text, images and metadata from PDF documents. TET makes available the text contents of a PDF as Unicode strings, plus detailed glyph and font information as well as the position on the page. Images are extracted in common raster formats. TET optionally converts PDF documents to an XML-based format called TETML which contains text, image, and metadata as well as resource information. TET supports Chinese, Japanese, and Korean (CJK) text as well as right-to-left languages such as Hebrew and Arabic.

TET contains advanced content analysis algorithms for determining word boundaries, grouping text into columns and removing redundant text. Using the integrated pCOS interface you can retrieve arbitrary objects from the PDF, such as metadata, interactive elements, etc. TET is suitable for server use (thread-safe and robust, no memory leaks, clean exception handling).

New Features in TET 4.0. TET 4.0 offers considerable performance enhancements and is faster for many classes of documents. Especially very large documents up to hundreds of thousands of pages benefit from higher speed and smaller memory consumption.

The results of PDF text extraction are enhanced with improved shadow removal, word boundary detection, dehyphenation, and super- and subscript detection. More workarounds for non-conforming PDF documents improve the robustness of text extraction; the enhanced repair mode can successfully extract text from damaged PDFs.

TET 4 rearranges bidirectional text in Arabic or Hebrew documents to the proper logical order. Unicode postprocessing controls offer folding, decomposition and normalization according to the Unicode standard which is useful to adjust the extracted text according to the requirements of the application.

TET PDF IFilter 4.0. Based on patented TET technology TET PDF IFilter is a robust implementation of Microsoft's IFilter indexing interface. It works with all search and retrieval products which support the IFilter interface, e.g. SharePoint and SQL Server. The new language detection feature automatically assigns the proper natural language to the text, which is important for proper word stemming and therefore improves the search experience.

TET Plugin 4.0. TET is also available as a free plugin for Adobe Acrobat. This plugin allows interactive test and evaluation of TET's superior text extraction. The new TET Plugin supports Unicode syntax for search text and can highlight all search hits on a page.

PDFlib GmbH

Franziska-Bilek-Weg 9

D- 80339 München

+49 • 89 • 452 33 84-0

+49 • 89 • 452 33 84-99

press@pdflib.com

www.pdflib.com

PRESS RELEASE

TET Cookbook. The TET Cookbook is a collection of programming examples which demonstrate the use of TET for various text and image extraction tasks. Several Cookbook samples show how to combine the TET and PDFlib+PDI products in order to enhance PDF documents, e.g. add bookmarks or links based on the text on the page.

Pricing and availability. TET 4 for Windows Server 2003/2008, Linux or Apple Mac OS X Server costs Euro 795 or US-\$ 995. TET for Windows 2000/XP/Vista/7 or Mac OS X desktop costs Euro 295 or US-\$ 375. Additional packages are available for Sun Solaris, IBM AIX and HP-UX as well as for IBM i5/iSeries and zSeries.

TET PDF IFilter 4.0 is freely available for non-commercial use on desktop systems, which provides a convenient basis for test and evaluation. The license fee for Windows Server is Euro 555 or US-\$ 695.

The TET Plugin for Acrobat Professional on Windows and Mac is free for non-commercial use.

About PDFlib GmbH. PDFlib GmbH is completely focused on PDF technology. Customers worldwide use PDFlib products since 1997. The company closely follows development and market trends, such as ISO standards for PDF. PDFlib GmbH products are distributed all over the world with major markets in North America, Europe, and Japan.