

## PRESS RELEASE

*New version of PDFlib TET (Text Extraction Toolkit)*

### PDFlib TET 3 available

*The new version of TET extracts text from PDF documents, retrieves raster image data and tables, and converts PDF documents to XML.*

**Munich, February 10, 2009.** PDFlib GmbH announces the availability of the new product version TET 3. The new version offers a variety of new features, including raster image and table extraction, XML output, a new Cookbook, and a number of TET connectors to interface TET with other software. Various new workarounds make it possible to extract the text from PDF documents where this was previously not possible, e.g. some PDFs generated with the TeX typesetting system.

**Image Extract.** Images on PDF pages can be extracted as TIFF, JPEG, or JPEG 2000 files. Precise geometric information (position, size, and angles) are reported for each image. Fragmented images will be combined to larger images to facilitate repurposing. Image fidelity is guaranteed since no downsampling or color space conversion occurs. This ensures the highest possible image quality.

**Page Layout and Table Detection.** The page content is analyzed to determine text columns. Tables are detected, including cells which span multiple columns. This improves the ordering of the extracted text. Table rows and the contents of each table cell can be identified.

**TETML represents PDF contents in XML.** TET optionally represents the PDF contents in an XML flavor called TETML. It contains a variety of PDF information in a form which can easily be processed with common XML tools. TETML contains the actual text plus optionally font and position information, resource details (fonts, images, colorspace), and metadata. TETML is governed by a corresponding XML schema to make sure that TET always creates consistent and reliable XML output. TETML can be processed with XSLT stylesheets, e.g. to apply certain filters or to convert TETML to other formats. Sample XSLT stylesheets for processing TETML are included in the TET distribution.

**TET Connectors.** TET connectors provide the necessary glue code to interface TET with other software. The following TET connectors make PDF text extraction functionality available for various software environments:

- ▶ The TET Plugin for Adobe Acrobat is a free utility for extracting text and images from PDF. It offers better functionality than Acrobat's built-in tools, and can be used to evaluate TET interactively.
- ▶ TET connector for the Lucene Search Engine
- ▶ TET connector for the Solr Search Server
- ▶ TET connector for Oracle Text

## PRESS RELEASE

- ▶ TET PDF IFilter for Microsoft products is available as a separate product. It extracts text and metadata from PDF documents and makes it available to search and retrieval software on Windows
- ▶ TET connector for MediaWiki

**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.** On Windows Server 2000/2003/2008, Linux or Apple Mac OS X Server TET 3 costs Euro 795 or US-\$ 995. For Windows 2000/XP/Vista or Mac OS X TET costs Euro 159 or US-\$ 199. Additional packages are available for Sun Solaris, IBM AIX und HP-UX.

**About TET.** *The PDFlib Text Extraction Toolkit (TET) is a developer product for reliably extracting text and raster images 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. Raster images are extracted in common formats. TET optionally converts PDF documents to an XML-based format called TETML which contains text and metadata as well as resource information. 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.*

**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.*