PDFlib TET 3
Text Extraction Toolkit

What is PDFlib 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 raster 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.

With PDFlib TET you can:

► Implement the PDF indexer for a search engine
► Repurpose the text and images in PDFs
► Convert the contents of PDFs to other formats
► Process PDFs based on their contents, e.g. splitting based on headings (requires PDFlib+PDI in addition to TET)

PDFlib TET Features

PDF Input

PDFlib TET supports all PDF versions up to Acrobat 9 (including RC4 and AES encryption). TET can extract Chinese, Japanese, and Korean text. All CJK encodings are recognized; horizontal and vertical writing modes are supported.

Protected documents can be indexed while at the same time respecting access permissions and permission controls.

Unicode

Since text in PDF is usually not encoded in Unicode, PDFlib TET normalizes the text in a PDF document to Unicode:

► TET converts all text contents to Unicode. In C and other non-Unicode aware languages the text is returned in the UTF-8 or UTF-16 formats, and as native strings in Unicode-capable programming languages.
► Ligatures and other multi-character glyphs are decomposed into a sequence of the corresponding Unicode characters.
► Vendor-specific Unicode assignments (PUA characters) are identified, and mapped to characters in the common Unicode area if possible.

► Glyphs without appropriate Unicode mappings are identified as such, and are mapped to a configurable replacement character in order to avoid misinterpretation.
► TET implements various workarounds for problems with specific document creation packages, such as InDesign and TeX documents or PDFs generated on mainframe systems.

Content Analysis and Word Detection

TET includes advanced content analysis algorithms:

► Patented algorithm for determining word boundaries which is required to retrieve proper words
► Recombine the parts of hyphenated words
► Remove duplicate instances of text, e.g. shadow and artificially bolded text
► Recombine paragraphs in reading order
► Reorder text which is scattered over the page

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.

Text Geometry

TET provides precise metrics for the text, such as the position on the page, glyph widths, and text direction. Specific areas on the page can be excluded or included in the text extraction, e.g. to ignore headers and footers or margins.

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.

PDF Analysis

The TET library includes the pCOS interface for querying details about a PDF document, such as document info and XMP metadata, font lists, page size, and many more (see separate datasheet for the pCOS product).

Repair Mode

Various kinds of damaged PDF documents are detected and automatically repaired if possible.

Configuration Options for problematic PDF

TET contains special handling and workarounds for various kinds of PDF where the text cannot be extracted correctly with other products. In addition, it includes various configuration features to improve processing of problem documents:

- ► Unicode mapping can be customized via user-supplied tables for mapping character codes or glyph names to Unicode.
- ► PDFlib FontReporter is an auxiliary tool for analyzing fonts, encodings, and glyphs in PDF. It works as a plugin for Adobe Acrobat. This plugin is freely available for Mac and Windows.
- ► Embedded fonts are analyzed to find additional hints which are useful for Unicode mapping. External font files or system fonts are used to improve text extraction results if a font is not embedded.

Document Domains

PDF documents may contain text in other places than the page contents. While most applications will deal with the page contents only, in many situations other document domains may be relevant as well. TET extracts the text from all of the following document domains:

- ► page contents
- ► predefined and custom document info entries
- ► XMP metadata on document and image level
- ► bookmarks
- ► file attachments and PDF collections/portfolios can be processed recursively
- ► form fields
- ► comments (annotations)
- ► general PDF properties can be queried, such as page count, conformance to standards like PDF/A or PDF/X, etc.

XMP Metadata

TET supports XMP metadata in several ways:

- ► Using the integrated pCOS interface, XMP metadata for the document, individual pages, images, or other parts of the document can be extracted programmatically.
- ► TETML output contains XMP document and image metadata if present in the PDF.
- ► Images extracted in the TIFF or JPEG formats contain image metadata if present in the PDF.

TETML represents PDF Contents as XML

TET optionally represents the PDF contents in an XML flavor called TETML which 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, colorspaces), 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
- ► 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 (see separate datasheet for details).
- ► 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.

A detailed look at TET Features

Unique TET Advantages
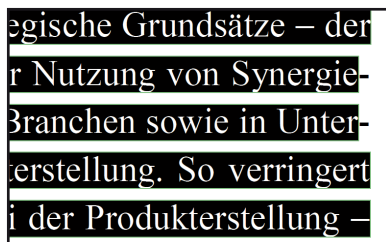
Dehyphenation

TET detects hyphenated words which span multiple lines, removes the hyphen, and combines the individual parts to form a complete word. This is important to make sure that searches for the full word will be successful although only hyphenated parts are present in the document. Dashes (different from hyphens) will be treated separately since they must not be removed.


Other products extract »Inttrroduccttiion«
TET extracts »Introduction«

Shadow and artifical bold Text Detection

Digital documents often contain shadowed text where the shadow effect is achieved by placing the text multiply on the page, using a small offset between the instances of text. Similarly, bold text is often simulated by overprinting the same text multiply. As a result, the document contains the characters in the shadowed or bold word more than once. TET's patented shadow detection algorithm identifies and removes redundant instances of text to avoid excess text extraction. While other software will extract the shadowed or bold text multiply, TET correctly removes the redundant copies. While extra instances of a word will still result in a search engine hit, no more hits would be found if the text is duplicated character by character as in the example.
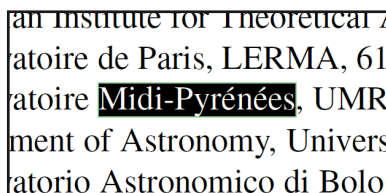

Other products extract »Midi-Pyr´en´ees«
TET extracts »Midi-Pyrénées«

Accented Characters

In many languages accents and other diacritical marks are placed close to other characters to form combined characters. Some typesetting programs, most notably TeX, emit two characters (base character and accent) separately to create a combined character. For example, to create the character $ä$ first the letter a is placed on the page, and then the dieresis character ¨ is placed on top of it. TET detects this situation and recombines both characters to form the appropriate combined character.
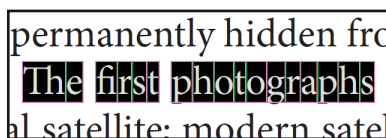
Other products extract » e rst photographs«
TET extracts »The first photographs«

Ligatures

Ligatures combine two or more characters in a single glyph. The most common ligatures are in use for the combinations *fi*, *fl*, and *ffi*; less common ligatures are used for the combinations *Th*, *sp*, *ct*, *st*, and many others. When extracting text from digital documents, ligatures must be analyzed and separated to the constituent characters to allow proper text processing. TET detects ligatures based on many properties and delivers two or more characters as appropriate.

Other products extract 133 tiny little strips
TET extracts a single large image

Image Merging

The images in many PDF documents are broken into smaller pieces by the software producing the PDF. What appears as a single image on the page may actually consist of hundreds or thousands of small fragments. Among others, Microsoft Office applications and TeX are known to produce such documents. TET detects fragmented images and merges the pieces to form a usable larger image. Only with image merging such images can be repurposed in any way.

Many Ways to use TET

TET is available as a programming library (component) for various development environments, and as a command-line tool for batch operations. Both offer similar features, but are suitable for different deployment tasks. Both the TET library and command-line tool can create TETML, TET's XML-based output format.

The TET Programming Library is used...

...for integration into desktop or server applications. Examples for using the library with all supported language bindings are included in the TET package.

The TET Command-line Tool is suited...

...for batch processing PDF documents. It doesn't require any programming, but offers command-line options which can be used to integrate it into complex workflows.

TETML Output is suited...

...for XML-based workflows and developers who are familiar with the wide range of XML processing tools and languages, e.g. XSLT.

TET Connectors are suited...

...for integrating TET in various common software packages, e.g. databases and search engines.

Supported Development Environments

PDFlib TET is everywhere – it runs on practically all computing platforms. We offer 32-bit and 64-bit packages for all common flavors of Windows, Mac OS, Linux and Unix, as well as for IBM eServer iSeries and zSeries systems.

The TET core is written in highly optimized C code for maximum performance and small overhead. Via a simple API (Application Programming Interface) the TET functionality is accessible from a variety of development environments:

► COM for use with VB, ASP, Borland Delphi, etc.

► C and C++

► Java, including servlets and Java Application Server

► .NET for use with C#, VB.NET, ASP.NET, etc.

► Perl

► PHP

► Python

► RPG (IBM eServer iSeries)

Benefits of using PDFlib Software

Rock-solid Products

Tens of thousands of programmers worldwide are working with our software. PDFlib meets all quality and performance requirements for server deployment. All PDFlib products are suitable for robust 24x7 server deployment and unattended batch processing.

Speed and Simplicity

PDFlib products are incredibly fast – up to thousands of pages per second. The programming interface is straightforward and easy to learn.

PDFlib Products all over the World

Our products support all international languages as well as Unicode. They are used by customers in all parts of the world.

Professional Support

If there's a problem, we will try to help. We offer commercial support to meet the requirements of your business-critical applications. By adding support you will have access to the latest versions, and have guaranteed response times should any problems arise.

Licensing

We offer various licensing programs for server licenses, integration and site licenses, and source code licenses. Support contracts for extended technical support with short response times and free updates are also available.

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.

PDFlib®

Contact

Fully functional evaluation versions including documentation and samples are available on our Web site. For more information please contact:

PDFlib GmbH

Franziska-Bilek-Weg 9, 80339 München, Germany

phone +49 • 89 • 452 33 84-0

fax +49 • 89 • 452 33 84-99

sales@pdflib.com

www.pdflib.com

February 2009