PDFlib TET 2
Text Extraction Toolkit

PDFlib® data sheet

What is PDFlib TET?

The PDFlib Text Extraction Toolkit (TET) is a developer product for reliably extracting text 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.

In addition, TET contains advanced content analysis algorithms for determining word boundaries, grouping text into columns and removing redundant text, such as shadows or artificially bolded text. Using the auxiliary pCOS interface you can retrieve arbitrary objects from the PDF, such as metadata, interactive elements, etc. With PDFlib TET you can:

- Implement a search engine for processing PDF
- Extract text from PDFs, e.g. to store it in a database
- Convert text contents of PDFs to other formats, such as XML
- Process PDFs based on their contents

PDFlib TET Features

Supported PDF Input

PDFlib TET supports all relevant flavors of PDF input:

- All PDF versions up to PDF 1.7 (Acrobat 8)
- All font and encoding types: base 14 fonts, TrueType, PostScript, OpenType, CID (composite) fonts
- RC4 and AES encryption
- Damaged PDF input documents will be repaired if possible

Unicode

Since text in PDF is usually not encoded in Unicode, PDFlib TET normalizes the text from 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 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.

Full Support for Chinese, Japanese, and Korean

TET includes full support for extracting Chinese, Japanese, and Korean text. All predefined CJK CMaps (encodings) are recognized; horizontal and vertical writing modes are supported.

Content Analysis and Word Detection

TET can be used to retrieve low-level glyph information, but also includes advanced content analysis algorithms:

- Detect word boundaries to retrieve proper words
- Recombine the parts of hyphenated words
- Remove duplicate instances of text, e.g. shadow and artificially bolded text
- Recombine paragraphs into reading order
- Reorder text which is scattered over the page
- Reconstruct lines of 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.

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 runs as a plugin for Adobe Acrobat 5–8. This plugin is freely available for Mac and Windows.
- Embedded fonts are parsed to find additional hints which are useful for Unicode mapping. External font files or system fonts can be used to improve text extraction results if a font is not embedded.

pCOS Interface for simple Access to PDF Objects

TET includes the auxiliary pCOS programming interface. PDFlib pCOS is a PDF information retrieval tool which provides a simple and elegant facility for retrieving any information from a PDF document which is not part of the page contents, such as PDF metadata, interactive elements (e.g. links), page dimensions, and many more.

TET Library or Command-Line Tool?

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.

TET is also available as a free plugin for Adobe Acrobat: the TET Plugin can be used to interactively take advantage of the benefits of TET, and conveniently explore TET options.

The TET programming library is used...

...for integration into desktop or server application. 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. The TET command-line tool extends the features of the library:

- In addition to creating plain text it can convert PDF documents to XML.
- The TET command-line tool can also be called from environments which do not support the use of the TET library.

PDFlib TET PDF IFilter

Note that TET is also available as an IFilter for use with Windows search and retrieval products. Please ask for the separate product »PDFlib TET PDF IFilter«.

Supported Development Environments

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

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 hypertext processor
- 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 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 for software developers. 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 2008