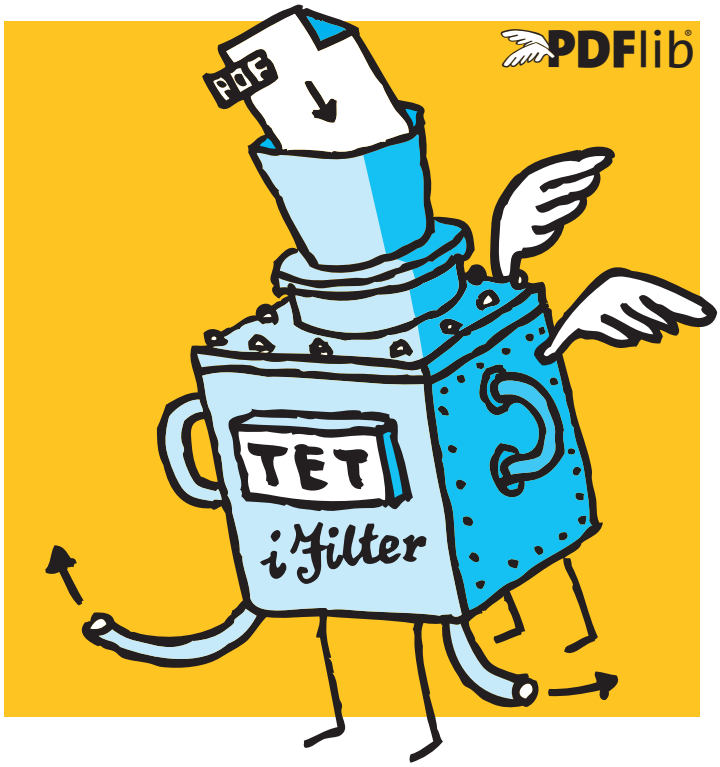


datasheet

PDFlib TET PDF IFilter 4.1

Enterprise PDF Search for Windows

What is PDFlib TET PDF IFilter?

TET PDF IFilter extracts text and metadata from PDF documents and makes it available to search and retrieval software on Windows. This allows PDF documents to be searched on the local desktop, a corporate server, or the Web. TET PDF IFilter is based on the patented PDFlib Text Extraction Toolkit (TET), which is an established developer product for reliably extracting text from PDF documents.

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. Such products use format-specific filter programs – called IFilters – for particular file formats, e.g. HTML. TET PDF IFilter is such a program, aimed at PDF documents. The user interface for searching the documents may be the Windows Explorer, a Web or database frontend, a query script, or a custom application. As an alternative to interactive searches, queries can also be submitted programmatically without any user interface.

Based on patented TET Technology

PDFlib TET, the basis of TET PDF IFilter, was first released in 2002, and has been used by customers worldwide in server and desktop environments. As an alternative to extracting PDF page contents and metadata as raw text, TET can supply the document contents in XML format. TET is also available as a free plugin for Adobe Acrobat; this plugin allows interactive test and evaluation of TET's superior text and image extraction.

Unique Advantages

TET PDF IFilter offers the following advantages:

- Supports Western text, Chinese, Japanese, and Korean (CJK) text and right-to-left languages such as Arabic and Hebrew
- Indexes protected documents and extracts text even from PDFs where Acrobat fails
- Supports Unicode folding, decomposition, and normalization
- Deployment: thread-safe, fast and robust, 32- and 64-bit versions
- Automatic script and language detection for improved search

Enterprise PDF Search

TET PDF IFilter is available in fully thread-safe native 32- and 64-bit versions. You can implement enterprise PDF search solutions with TET PDF IFilter and the following products:

- Microsoft SharePoint Server and FAST Search Server
- Microsoft Search Server
- Microsoft SQL Server
- Microsoft Exchange Server
- Microsoft Site Server

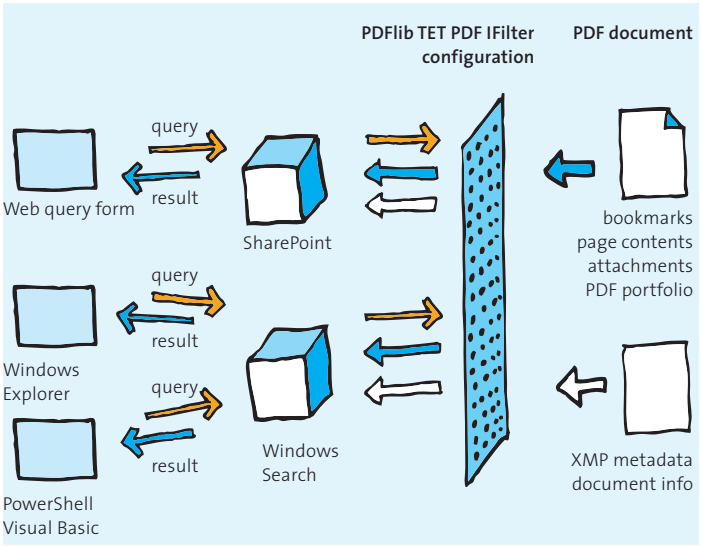
TET PDF IFilter can be used with all other Microsoft and third-party products which support the IFilter interface.

Desktop PDF Search

TET PDF IFilter can also be used to implement desktop PDF search, e.g. with the following products:

- Windows Search is integrated in Windows Vista/7; also available as free add-on for Windows XP
- Windows Indexing Service

TET PDF IFilter is free for non-commercial use on desktop operating systems, which provides a convenient basis for test and evaluation.



Feature Details

Accepted PDF Input

TET PDF IFilter supports all relevant flavors of PDF input:

- All PDF versions up to Acrobat X, including ISO 32000-1
- Protected PDFs which do not require a password for opening the document
- Damaged PDF documents will be repaired

Unicode Postprocessing

TET supports various Unicode postprocessing steps which can be used to improve the extracted text:

- Foldings preserve, remove or replace characters, e.g. remove punctuation or characters from irrelevant scripts.
- Decompositions replace a character with an equivalent sequence of one or more other characters, e.g. replace narrow, wide or vertical Japanese characters or Latin superscript (e.g. ª) variants with their respective standard counterparts.
- Text can be converted to all four Unicode normalization forms, e.g. emit NFC form to meet the requirements for web text or a database.

Internationalization

In addition to Western text TET PDF IFilter fully supports Chinese, Japanese, and Korean (CJK) text. All CJK encodings are recognized; horizontal and vertical writing modes are supported. Automatic detection of the locale ID (language and region identifier) of the text improves the results of Microsoft's word breaking and stemming algorithms, which is especially important for East Asian text.

Right-to-left languages such as Hebrew and Arabic are also supported. Contextual character forms are normalized and the text is delivered in logical order.

PDF is more than just a Bunch of Pages

TET PDF IFilter treats PDF documents as containers which may contain much more information than only plain pages. TET PDF IFilter indexes all relevant items in PDF documents:

- Page contents
- Text in bookmarks
- Metadata (see below)
- Embedded PDFs and PDF packages/portfolios are processed recursively so that attachments can also be searched.

XMP Metadata and Document Info

The advanced metadata implementation in TET PDF IFilter supports the Windows property system for metadata. It indexes XMP metadata as well as standard or custom document info entries. Metadata indexing can be configured on several levels:

- Document info entries, Dublin Core fields and other common XMP properties are mapped to equivalent Windows properties, e.g. *Title*, *Subject*, *Author*.
- TET PDF IFilter adds useful PDF-specific pseudo properties, e.g. page size, PDF/A conformance level, font names.
- All relevant predefined XMP properties can be searched.
- User-defined XMP properties can be searched, e.g. company-specific classification properties, PDF/A extension schemas.

TET PDF IFilter optionally integrates metadata in the full text index. As a result, even full text search engines without metadata support (e.g. SQL Server) can search for metadata.

Benefits of using PDFlib Software

Rock-solid Products

Tens of thousands of programmers worldwide are working with our software. PDFlib products meet all quality and performance requirements for server deployment. All 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.

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