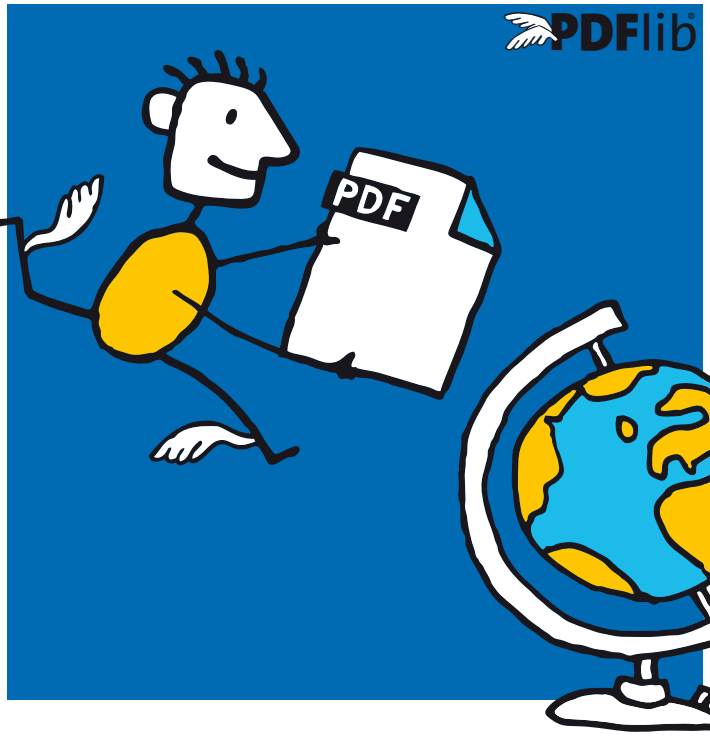


datasheet

PDFlib Products in the Real World

www.pdflib.com

What is PDFlib?

PDFlib is the leading developer toolbox for generating and manipulating files in the Portable Document Format (PDF).

PDFlib's main targets are dynamic PDF creation on a Web server or any other server system, and to implement »Save as PDF« in existing applications. You can use PDFlib to dynamically create PDF documents from database contents, similar to dynamic Web pages.

PDFlib has proven itself in a wide range of other use cases as well.

Application programmers need only decent graphics or print output experience to be able to use PDFlib quickly. Since PDFlib frees you from the technicalities of the PDF file format, you can focus on acquiring the data and arranging text, graphics, and images on the page.

The PDFlib Product Family

The PDFlib product family comprises the following products:

PDFlib offers all functions required to generate PDF documents with text, graphics, images, and interactive elements such as annotations or bookmarks.

PDFlib+PDI includes all PDFlib functions plus the PDF Import Library (PDI). With PDI you can open existing PDF documents and incorporate some pages into the PDFlib output.

PDFlib Personalization Server (PPS) includes PDFlib+PDI plus additional functions for variable data processing using PDFlib Blocks. PPS makes applications independent of any layout changes.

»Save as PDF« for Applications

I work with a software development company and want to implement a »Save as PDF« feature in our applications.

PDFlib easily integrates into all kinds of applications to enable reliable and high-quality PDF output. Many well-known developers of graphics programs, geographical information systems (GIS), prepress and DTP applications and from many other domains rely on PDFlib to add PDF generation features to their products.

Invoices for an Online Shop

How can I create PDF invoices dynamically in my online shop?

Dynamic invoice generation is one of the most popular PDFlib scenarios. The generated PDF invoices can be viewed in the Web browser, made available for separate download, or e-mailed to the user.
Use PDFlib to place transaction data (customer details, item list, prices, etc.) on a PDF page. Add images, such as a company logo, in a variety of image formats. Use PDFlib+PDI to incorporate existing PDF material, for example company stationery as background.

Mail Merge

How can I merge personal data into an existing PDF document to create mass mailings?

PDFlib+PDI imports one or more pages of an existing PDF and adds individual text and images to create individual letters. The programmer adds code for retrieving text or graphics from a text file or database. A single large PDF containing all letters can be produced for printing, or many personalized small PDFs for e-mailing to the recipients.
If you need more flexibility because slightly different mailings must be produced or changes in the page design occur frequently, you can use the PDFlib Personalization Server (PPS). This facilitates both the designer's and the programmer's job when it comes to variable data processing.

Invoices and Reports from Office Applications

I'm unsatisfied with the look of invoices and reports created by our office applications. How can I create nice PDF documents?

PDFlib can be attached to common office applications. You can add PDF capability to MS Office and other applications with the popular Visual Basic scripting language. Use PDFlib to create invoices from a database in order to print or e-mail them to customers. Use PDFlib+PDI to incorporate PDF company stationery. Make PDF processing even more efficient by deploying the PDFlib Personalization Server (PPS).

Commercial Printing

Can I use PDFlib to prepare prepress data for commercial printing?

Customers use PDFlib to build systems for creating, assembling, or personalizing PDF documents for commercial printing. In many cases these production systems are accessible via a Web browser. The PDFlib product family supports a variety of features for the graphic arts industry, including color management with ICC profiles, CMYK color, spot colors with built-in PANTONE® and HKS® tables, and PDF/X-3, PDF/X-4 or PDF/X-5 conforming output.

Mass Generation of Phone Bills

I am responsible for creating the monthly phone bills at a major telecommunications provider. We plan to migrate from paper-based bills to online PDFs and distribute them via e-mail or Web.

PDFlib has a proven track record in mission-critical environments. Even with several millions of bills in each run you won't experience performance or reliability problems. PDFlib works on any kind of server, including midrange and mainframe systems.

Spice up existing PDFs

Can I add stamps and page numbers to existing PDF pages?

First, you import the pages from your PDF with PDFlib+PDI. Then you add a transparent stamp, running page numbers, barcodes, company logo, or any other content. You can even add interactive elements including links, form fields, bookmarks, etc. With these features you can approach PDF problems with a postprocessing solution.

Other PDFlib GmbH Products

PDFlib TET. Extract text and images from any PDF and normalize it to Unicode. TET includes patented content analysis algorithms for identifying word boundaries or dehyphenating text, and much more.

PDFlib TET PDF IFilter. Extracts the text and metadata of PDF documents and makes it available to search and retrieval software on Windows.

PDFlib PLOP. Linearize, optimize, and protect PDF documents, or add XMP metadata.

PDFlib PLOP DS. Apply digital signatures to PDF documents.

PDFlib pCOS. Query any kind of information from PDF.

Supported Development Environments

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

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

- COM for use with VB, ASP, Borland Delphi, etc.
- C and C++
- Cobol (IBM eServer zSeries)
- Java, including servlets and Java Application Server
- .NET for use with C#, VB.NET, ASP.NET, etc.
- PHP hypertext processor
- Perl
- Python
- REALbasic
- RPG (IBM eServer iSeries)
- Ruby
- Tcl

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

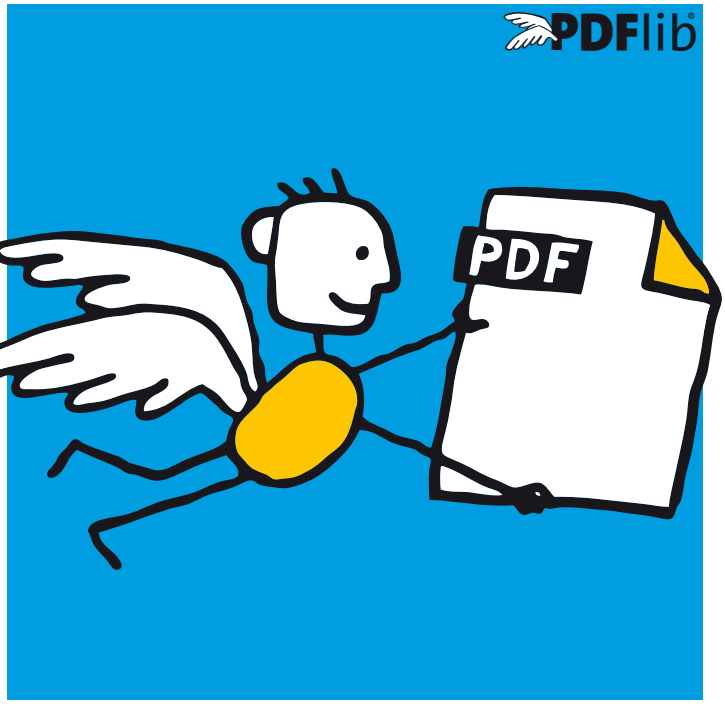
PDFlib GmbH is completely focused on PDF technology. Customers worldwide use PDFlib products since 1997. The company closely follows development and market trends, such as ISO standards for PDF. PDFlib GmbH products are distributed all over the world with major markets in North America, Europe, and Japan.

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

www.pdflib.com



datasheet

PDFlib, PDFlib+PDI, PDFlib Personalization Server (PPS) 8

www.pdflib.com

What is PDFlib?

PDFlib is the leading developer toolbox for generating and manipulating files in the Portable Document Format (PDF). PDFlib's main targets are dynamic PDF creation on a Web server or any other server system, and to implement »Save as PDF« in existing applications. You can use PDFlib to dynamically create PDF documents from database contents, similar to dynamic Web pages. PDFlib has proven itself in a wide range of other use cases as well. Application programmers need only decent graphics or print output experience to be able to use PDFlib quickly. Since PDFlib frees you from the technicalities of the PDF file format, you can focus on acquiring the data and arranging text, graphics, and images on the page.

The PDFlib product family is available in three different flavors: PDFlib, PDFlib+PDI (PDF Import), and the PDFlib Personalization Server (PPS).

PDFlib

PDFlib offers all functions required to generate PDF documents with text, graphics, images, and interactive elements such as annotations or bookmarks. Use PDFlib for the following tasks:

- Add »Save as PDF« capability to your application
- Create PDF documents on a Web server in real time
- Create database reports in PDF
- Create PDF/X-1/3/4/5 documents for commercial printing
- Convert TIFF, JPEG, or other image formats to PDF
- Create PDF/A for archiving

PDFlib+PDI (PDF Import)

PDFlib+PDI includes all PDFlib functions, plus the PDF Import Library (PDI). With PDI you can open existing PDF documents and incorporate some pages into the PDFlib output. Use PDFlib+PDI for all PDFlib tasks plus the following:

- Impose multiple PDF pages on a single sheet for printing
- Add text, such as headers, footers, stamps, or page numbers to existing PDF pages
- Place images, e.g. company logo, on existing pages
- Add barcodes to existing PDF pages
- Assemble existing PDF pages
- Add content to PDF/X or PDF/A documents

PDFlib Personalization Server (PPS)

The PDFlib Personalization Server (PPS) includes PDFlib+PDI plus additional functions for variable data processing using PDFlib Blocks. PPS makes applications independent from layout changes. The designer creates the page layout and converts it to PDF. She takes into account areas as placeholders for variable text and images. In Acrobat she drags a rectangular Block for each area using the PDFlib Block Plugin. Each Block contains a variety of Block properties, such as font size, color, image scaling. The PDFlib Block Plugin offers a Preview feature which shows the results of filling Blocks according to their properties.

The programmer writes code to fill PDFlib Blocks with text, images, or PDF pages. He doesn't need to know the formatting or position of a Block. Use PPS for all PDFlib+PDI tasks plus the following:

- Customize direct mailings with text and images
- Fill templates for transactional and statement processing
- Personalize promotional material with address data
- Generate individual parts catalogs from a database
- Produce customized documentation for multiple similar products

What's new in PDFlib 8?

New PDF Features for Acrobat 9

PDFlib supports various PDF features according to Acrobat 9 (technically: PDF 1.7 Adobe extension level 3):

- External graphical content (Reference XObjects)
- Layer variants (also called layer configurations)
- PDF Portfolios
- Georeferenced PDF
- AES-256 encryption and Unicode passwords
- PDFlib+PDI and PPS can import and process Acrobat 9 data.

Font Handling and Text Output

Quite a number of new typographical features can be found in PDFlib 8:

- Complex script shaping and bidirectional formatting for Arabic, Thai, Hindi, and many other writing systems
- Fallback fonts
- OpenType layout features, e.g. ligatures and swash characters
- Retain fonts across documents
- SING fonts for CJK Gaiji characters
- Redesigned font engine
- Wrap text around image clipping paths
- Text on a path

PDFlib Block Plugin and the PDFlib Personalization Server

The PDFlib Block Plugin is used to prepare PDF documents for Block filling (personalization) with the PDFlib Personalization Server (PPS). New features:

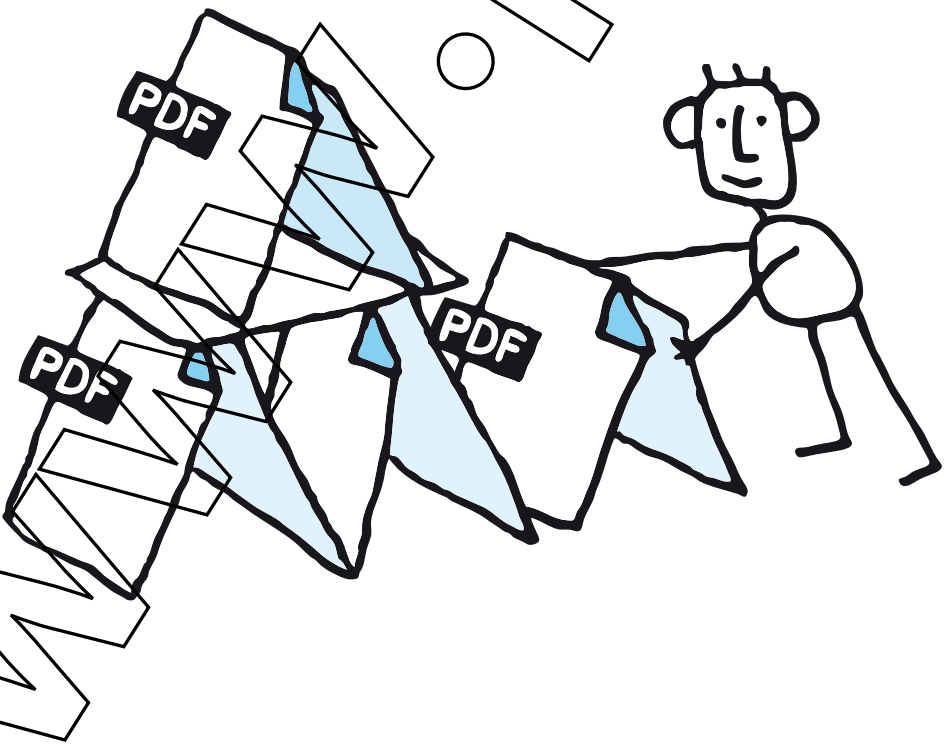
- Preview PPS Block processing in Acrobat
- Redesigned user interface
- Snap-to-grid for quickly layout out Blocks in a raster
- Additional Block properties, e.g. for transparency
- Clone PDF/A or PDF/X status of the Block container
- Leverage PDFlib 8 features with Blocks

Other important features

There are a number of other important new features, details can be found in the product documentation:

- Reusable path objects
- PDF/X-4 and PDF/X-5
- Alpha channel in TIFF and PNG images
- JBIG2-compressed images
- Compressed object streams and cross-reference streams
- Built-in PANTONE® Goe™ color libraries

PDFlib 8 also introduces a variety of improvements in existing functions.



www.pdflib.com

Common Features in PDFlib, PDFlib+PDI, and the PDFlib Personalization Server

PDF output	Generate PDF documents on disk file or directly in memory (for Web servers)
	High-volume output and arbitrary PDF file size (even beyond 10 GB)
	Suspend/resume and insert page features to create pages out of order
PDF flavors	PDF 1.3 – PDF 1.7ext3* (Acrobat 4–9) including ISO 32000-1 (=PDF 1.7)
	Linearized (web-optimized) PDF for byteserving over the Web
	Tagged PDF for accessibility and reflow
	Marked Content for adding application-specific data or alternate text without Tagging*
ISO standards	ISO 15930: PDF/X for the graphic arts industry*
	ISO 19005: PDF/A for archiving
	ISO 32000: standardized version of PDF 1.7*
Graphics	Common vector graphics primitives: lines, curves, arcs, ellipses*, rectangles, etc.
	Smooth shadings (color blends), pattern fills and strokes
	Transparency (opacity) and blend modes
	External graphical content (Reference XObjects) for variable data printing*
	Reusable path objects and clipping paths imported from images*
Layers	Optional page content which can selectively be displayed
	Annotations and form fields can be placed on layers
	Layers can be locked, automatically activated depending on zoom factor, etc.
	Layer variants* (production-safe groups of layers) for PDF/X-4 and PDF/X-5
Fonts	TrueType (TTF and TTC) and PostScript Type 1 fonts (PFB and PFA, plus LWFN on the Mac)
	OpenType fonts with PostScript or TrueType outlines (TTF, OTF)
	Support for dozens of OpenType layout features for Western and CJK text output, e.g. ligatures, small caps, old-style numerals, swash characters, simplified/traditional forms, vertical alternates*
	Directly use fonts which are installed on the Windows or Mac system (»host fonts«)
	Font embedding for all font types; subsetting for TrueType, OpenType, and Type 3 fonts
	User-defined (Type 3) fonts for bitmap fonts or custom logos
	EUDC and SING* fonts (glyphlets) for CJK Gaiji characters
	Fallback fonts (pull missing glyphs from an auxiliary font)*
	Retain fonts across documents to increase performance*
Text output	Text output in different fonts; underlined, overlined, and strikeout text
	Glyphs in a font can be addressed by numerical value, Unicode value, or glyph name*
	Kerning for improved character spacing
	Artificial bold, italic, and shadow* text
	Create text on a path*
	Proportional widths for standard CJK fonts*
	Configurable replacement of missing glyphs
Internationalization	Unicode strings for page content, interactive elements, and file names*; UTF-8, UTF-16, and UTF-32 formats
	Support for a variety of 8-bit and legacy multi-byte CJK encodings (e.g. Shift-JIS; Big5)
	Fetch code pages from the system (Windows, IBM eServer iSeries and zSeries)
	Standard and custom CJK fonts and CMaps for Chinese, Japanese, and Korean text
	Vertical writing mode for Chinese, Japanese, and Korean text
	Character shaping for complex scripts, e.g. Arabic, Thai, Devanagari*
	Bidirectional text formatting for right-to-left scripts, e.g. Arabic and Hebrew*
	Embed Unicode information in PDF for proper text extraction in Acrobat

www.pdflib.com

Images	Embed BMP, GIF, PNG, TIFF, JBIG2*, JPEG, JPEG 2000*, and CCITT raster images
	Automatic detection of image file formats
	Query image information (pixel size, resolution, ICC profile, clipping path, etc.)*
	Interpret clipping paths in TIFF and JPEG images
	Interpret alpha channel (transparency) in TIFF and PNG images*
	Image masks (transparent images with a color applied), colorize images with a spot color
Color	Grayscale, RGB (numerical, hexadecimal strings, HTML color names), CMYK, CIE Lab color
	Integrated support for PANTONE® colors (incl. PANTONE® Goe™)* and HKS® colors
	User-defined spot colors
Color management	ICC-based color with ICC profiles; support for ICC 4 profiles*
	Rendering intent for text, graphics, and raster images
	Default gray, RGB, and CMYK color spaces to remap device-dependent colors
	ICC profiles as output intent for PDF/A and PDF/X
Archiving	PDF/A-1a and PDF/A-1b (ISO 19005-1)
	XMP extension schemas for PDF/A-1
Graphic arts	PDF/X-1a, PDF/X-3, PDF/X-4*, PDF/X-4p*, PDF/X-5p*, PDF/X-5pg* (ISO 15930)
	Embedded or externally referenced* output intent ICC profile
	External graphical content (referenced pages) for PDF/X-5p and PDF/X-5pg*
	Create OPI 1.3 and OPI 2.0 information for imported images
	Separation information (PlateColor)
	Settings for text knockout, overprinting etc.
Textflow Formatting	Format text into one or more rectangular or arbitrarily shaped areas with hyphenation (user-supplied hyphenation points required), font and color changes, justification methods, tabs, leaders, control commands; wrap text around images
	Advanced line-breaking with language-specific processing
	Flexible image placement and formatting
	Wrap text around images or image clipping paths*
Table formatting	Table formatter places rows and columns, and automatically calculates their sizes according to a variety of user preferences. Tables can be split across multiple pages.
	Table cells can hold single- or multi-line text, images, PDF pages, path objects, annotations, and form fields
	Table cells can be formatted with ruling and shading options
	Flexible stamping function
	Matchbox concept for referencing the coordinates of placed images or other objects
Security	Encrypt PDF output with RC4 (40/128 bit) or AES encryption algorithms (128/256* bit)
	Unicode passwords*
	Specify permission settings (e.g. printing or copying not allowed)
Interactive elements	Create form fields with all field options and JavaScript
	Create actions for bookmarks, annotations, page open/close and other events
	Create bookmarks with a variety of options and controls
	Page transition effects, such as shades and mosaic
	Create all PDF annotation types, such as PDF links, launch links (other document types), Web links
	Named destinations for links, bookmarks, and document open action
	Create page labels (symbolic names for pages)
Multimedia	Embed 3D animations in PDF
GeoPDF	Create PDF with geospatial reference information*
Tagged PDF	Create Tagged PDF and structure information for accessibility, page reflow, and improved content repurposing; links and other annotations can be integrated in the document structure

www.pdflib.com

Metadata	Document information: common fields (Title, Subject, Author, Keywords) and user-defined fields
	Create XMP metadata from document info fields or from client-supplied XMP streams
	Process XMP image metadata in TIFF, JPEG, and JPEG 2000 images*
Programming	Language bindings for Cobol, COM, C, C++*, Java, .NET, Perl, PHP, Python, REALbasic, RPG, Ruby, Tcl
	Virtual file system for supplying data in memory, e.g., images from a database

** New or considerably improved in PDFlib/PDFlib+PDI/PPS 8*

Additional Features in PDFlib+PDI and the PDFlib Personalization Server

PDF input (PDI)	Import pages from existing PDF documents
	Import all PDF versions up to PDF 1.7 extension level 3 (Acrobat 9)
	Import documents which are encrypted with any of PDF's standard encryption algorithms (master password required)*
	Query information about imported pages*
	Clone page geometry of imported pages (e.g. BleedBox, TrimBox, CropBox)*
	Delete redundant objects (e.g. identical fonts) across multiple imported PDF documents
	Repair malformed input PDF documents*
	Copy PDF/A or PDF/X output intent from imported PDF documents
pCOS interface	pCOS interface for querying details about imported PDF documents*

**New or considerably improved in PDFlib+PDI and PPS 8*

Additional Features in the PDFlib Personalization Server

Variable Data Processing (PPS)	PDF personalization with PDFlib Blocks for text, image, and PDF data
PDFlib Block Plugin	PDFlib Block plugin for creating PDFlib Blocks interactively in Acrobat on Windows and Mac
	Redesigned user interface*
	Preview PPS Block filling in Acrobat*
	Snap-to-grid for interactively creating or editing Blocks in Acrobat*
	Clone PDF/X or PDF/A properties of the Block container*
	Convert PDF form fields to PDFlib Blocks for automated filling
	Textflow Blocks can be linked so that one Block holds the overflow text of a previous Block
	List of PANTONE® and HKS® spot color names integrated in the Block plugin*

**New or considerably improved in PPS 8*

www.pdflib.com

Supported Development Environments

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

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

- COM for use with VB, ASP, Borland Delphi, etc.
- C and C++
- Cobol (IBM eServer zSeries)
- Java, including servlets and Java Application Server
- .NET for use with C#, VB.NET, ASP.NET, etc.
- PHP
- Perl
- Python
- REALbasic
- RPG (IBM eServer iSeries)
- Ruby
- Tcl

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.

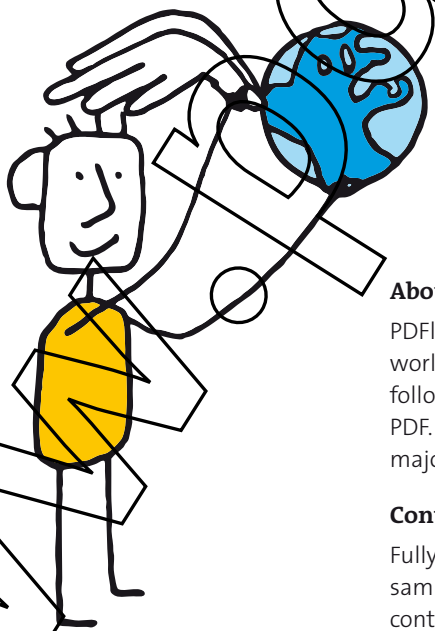
Professional Support

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.

www.pdflib.com



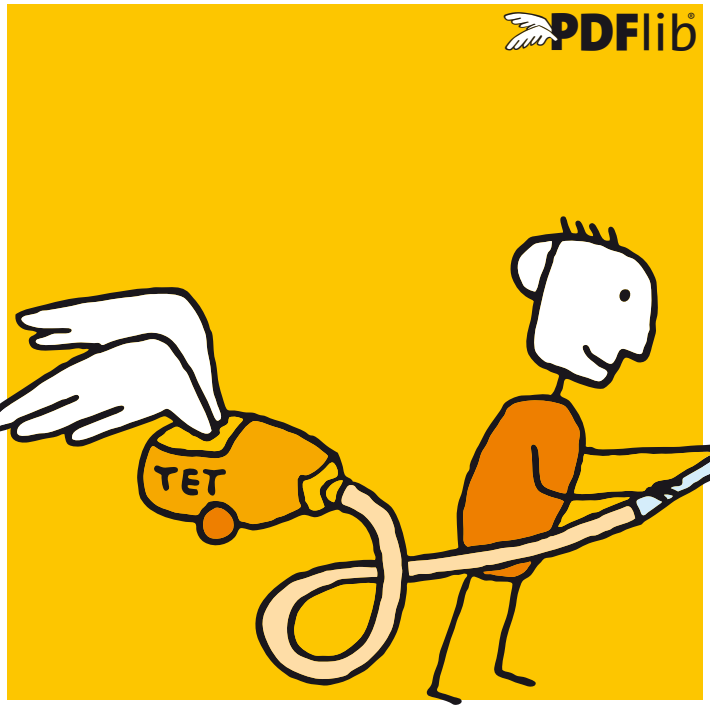
About PDFlib GmbH

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



datasheet

PDFlib TET 4

Text Extraction Toolkit

www.pdflib.com

What is PDFlib TET?

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. 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

Accepted PDF Input

TET supports all relevant flavors of PDF input:

- All PDF versions up to Acrobat 9, including ISO 32000-1
- Protected PDFs which do not require a password for opening the document
- Damaged PDF documents will be repaired

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.
- 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 (dehyphenation)
- Remove duplicate instances of text, e.g. shadow and artificially bolded text
- Recombine paragraphs in reading order
- Correctly order 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.

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).

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.

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.

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 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. 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, 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.

The following fragment shows TETML output with glyph details:

```
<Word>
 <Text>PDFlib</Text>
 <Box llx="111.48" lly="636.33" urx="161.14" ury="654.33">
 <Glyph font="F1" size="18" x="111.48" y="636.33" width="9.65">P</Glyph>
 <Glyph font="F1" size="18" x="121.12" y="636.33" width="11.88">D</Glyph>
 <Glyph font="F1" size="18" x="133.00" y="636.33" width="8.33">F</Glyph>
 <Glyph font="F1" size="18" x="141.33" y="636.33" width="4.88">l</Glyph>
 <Glyph font="F1" size="18" x="146.21" y="636.33" width="4.88">i</Glyph>
 <Glyph font="F1" size="18" x="151.08" y="636.33" width="10.06">b</Glyph>
 </Box>
 </Word>
```

TET Connectors

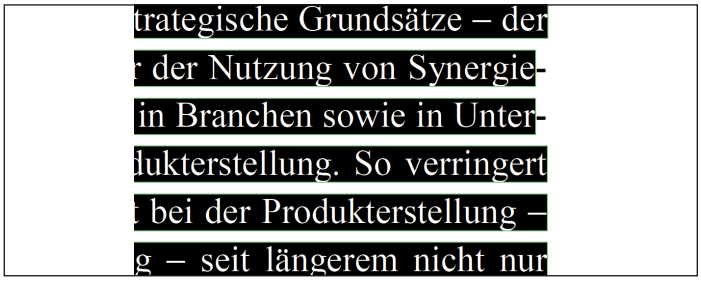
TET connectors provide the necessary glue code to interface TET with other software. The following TET connectors make PDF text extraction functionality available for various software environments:

- TET connector for the Lucene Search Engine
- TET connector for the Solr Search Server
- TET connector for Oracle Text
- TET connector for MediaWiki
- 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 Cookbook

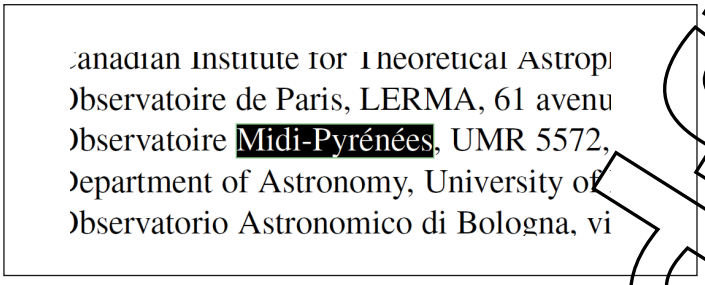
The TET Cookbook is a collection of programming examples which demonstrate the use of TET for various text and image extraction tasks. Several Cookbook samples show how to combine the TET and PDFlib+PDI products in order to process and enhance PDF documents, e.g. add bookmarks or links based on the text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

www.pdflib.com

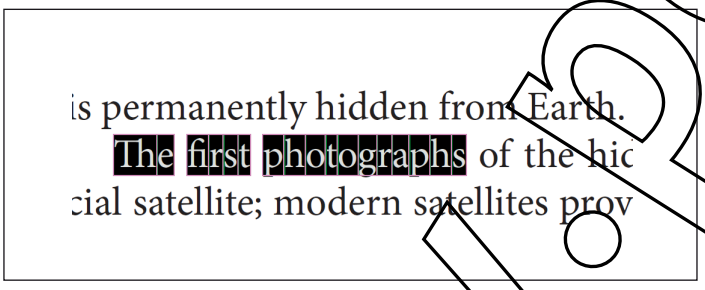


TET correctly removes the hyphen, but keeps the dash.

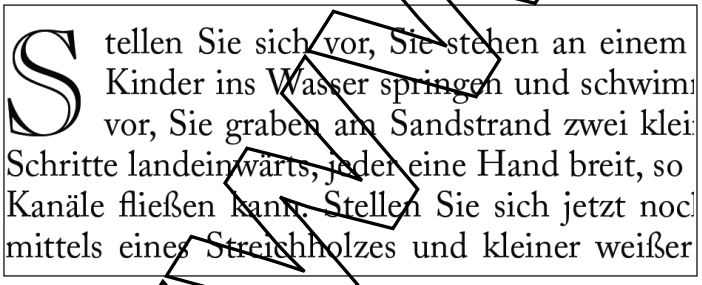
Other products extract »Inttrroduccttiion«.
TET correctly extracts »Introduction«.



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



Other products extract » e rst photographs«.
TET correctly extracts »The first photographs«.



Other products extract two words: the drop cap »S« and »tellen«.
TET correctly extracts the single word »Stellen«.

Challenges with PDF Text Extraction

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.

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 extracts 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.

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.

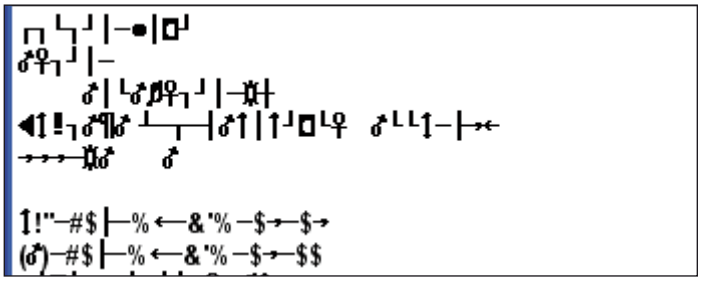
Ligatures

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 and delivers two or more characters as appropriate.

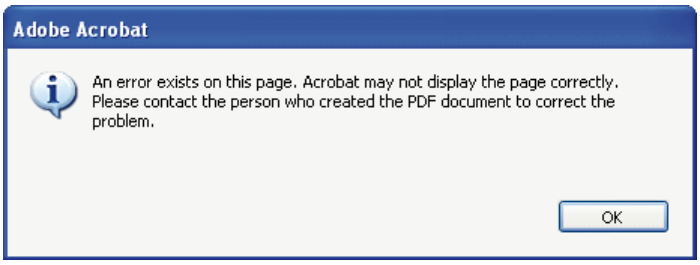
Drop Caps

Drop caps are large initial characters at the beginning of a paragraph where the top of the initial aligns with the top of the line, and the remainder of the character drops down several lines. Drop caps are used to emphasize the start of a paragraph. If they are not treated properly the initial word is extracted in two parts: the single initial character and the remainder of the word.

www.pdflib.com

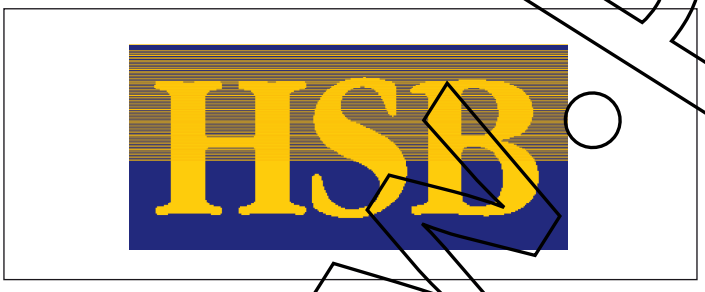


Other products extract unusable garbage, while TET delivers text.



The page contents are not even displayed in Acrobat, but TET still correctly extracts the text.

TET reorders the visual mixture of right-to-left and left-to-right text to create proper logical text output.



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

Challenges with PDF Text Extraction

Unicode Mapping

Unicode mapping forms the foundation of PDF text extraction: every glyph on the page must be assigned the corresponding Unicode value. PDF complicates this tasks by supporting a variety of font and encoding variants which may or may not provide the information required to assign proper Unicode values. In the worst case the document does not provide enough information with the result that no usable text can be extracted from the document.

TET's patented Unicode mapping algorithm implements a cascaded algorithm which takes all available pieces of information in order to determine Unicode values. For many problematic documents TET extracts proper Unicode text where other products deliver only unusable garbage.

Damaged PDF Documents

PDF documents may get damaged because of transmission errors or other problems. TET's repair mode recovers many kinds of damaged PDFs. Sometimes PDF documents are damaged so heavily that the pages cannot even be displayed in Acrobat. Even in such extreme cases TET often delivers the page contents of the document.

Bidirectional Text with Arabic and Hebrew

PDF does not encode logical text, but is simply a container for glyphs on the page. Text in the Arabic and Hebrew script runs from right to left. Since it often contains left-to-right inserts such as numbers or names in Western languages, text must be interpreted in both directions – hence the term »bidirectional«. Arabic poses additional challenges since the characters can be used in up to four different contextual forms. These shaped forms of characters must be normalized to the corresponding standard (isolated) form.

Challenges with PDF Image Extraction

Color Spaces and Compression

Raster image data in PDF may be encoded in any combination of eleven color spaces and nine compression filters, but common image file formats such as JPEG and TIFF support only a subset of those. TET's image extractor carefully balances the characteristics of the PDF image with the capabilities of the image output format. Regardless of the internal structure of the PDF image, the pixel image will be extracted in one of the common image file formats.

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.

www.pdflib.com

Many Ways to use TET

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

TET offers the following deployment options:

- The TET programming library (component) is used for integration into desktop or server applications. Examples for using the library 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.

The TET Family of Products

The TET family comprises the following products:

- The TET core product as described in this datasheet.
- TET PDF IFilter is available as a separate product. It is suitable for use with Microsoft search products, e.g. Windows Search, SharePoint and SQL Server (see separate datasheet for details).
- The TET Plugin for Adobe Acrobat is a free utility for extracting text and images from PDF. It can be used to evaluate TET interactively.

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 i5/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
- REALbasic
- RPG (IBM i5/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.

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

www.pdflib.com

datasheet

PDFlib PLOP 4.1

PDF Linearization, Optimization, Protection

www.pdflib.com

What is PDFlib PLOP?

PDFlib PLOP is a versatile tool for linearizing, optimizing, repairing, analyzing, encrypting and decrypting PDF documents. PLOP linearization and optimization features create efficient and small PDF documents for fast Web delivery. PLOP protection features encrypt or decrypt PDF documents and apply or remove permission settings, such as »printing not allowed« or »content extraction not allowed«. PLOP's repair mode automatically detects damaged PDF documents and fixes the problems if possible. PLOP analysis features can be used to query arbitrary properties of a PDF document. Document info entries and XMP metadata can be retrieved and set in a PDF/A and PDF/X conforming manner.

PDFlib PLOP Features

Linearization

With PDFlib PLOP you can linearize a PDF document for fast delivery over the Web (byteserving). Byteserving increases the perceived download speed since the first page is already visible while the remainder of the document is downloaded in the background.

Optimization

PLOP can significantly reduce the file size of a PDF document without affecting quality. It achieves this by removing unnecessary or redundant identical objects, such as repeatedly embedded fonts, images, identical ICC color profiles, etc.

Protection

PLOP can apply user and master passwords, and set access permissions to prevent the document from being printed with Acrobat, disallow text extraction or modification, etc.

PLOP supports all relevant PDF encryption methods including strong AES-256 encryption and Unicode passwords for Acrobat X. With PLOP you can:

- encrypt a PDF document with user or master password, or both;
- remove PDF encryption (if you know the master password);
- add or remove permission settings, e.g. »text extraction not allowed« (if you know the master password);
- query information about the security status, encryption scheme, permission settings, and document info fields

Repair Mode

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

PDF Analysis

The PLOP 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).

XMP Metadata

Metadata (»data about data«) is an important topic in many areas of application software. XMP (Extensible Metadata Platform) is an XML-based framework with many predefined metadata properties. XMP is integrated in Acrobat/PDF, and much more powerful than simple document info entries. XMP is required for PDF/A and other ISO standards. Many industry groups have published XMP-based recommendations for vertical applications.

With PLOP you can insert XMP metadata in PDF documents and extract XMP from PDF. Inserted XMP will be validated to make sure that valid output can be created. If the input document conforms to the PDF/A-1 standard, the user-supplied XMP must conform to the XMP rules set forth in PDF/A.

XMP insertion with PLOP can be used in the following and other situations (sample XMP is contained in the PLOP distribution):

- Add XMP metadata to PDF/A-1 documents, including support for XMP extension schemas as defined in the PDF/A-1 standard.
- Add XMP metadata describing the scanning process for digitized legacy documents, also according to PDF/A-1.
- Add XMP metadata according to the Ghent Workgroup (GWG) Ad Ticket scheme.
- Add company-specific XMP metadata.

Document Info Entries

With PLOP you can add new document information entries or replace the values of existing info entries. Both predefined and custom entries can be set. If the input document contains XMP document metadata, all predefined info entries will automatically be synchronized to the XMP metadata in order to keep the metadata consistent (this is a requirement of PDF/A-1).

PDF Versions and Standards

PLOP supports all PDF versions up to Acrobat X, including PDF 1.7 (ISO 32000). PLOP is PDF/A-aware: if the input document conforms to the PDF/A standard (ISO 19005-1), the output document is guaranteed to still comply with PDF/A. PLOP fully supports XMP extension schemas as required by PDF/A-1. Similarly, PLOP is aware of PDF/X-1a/3/4/5 (ISO 15930).
The ability to insert PDF/A-conforming XMP metadata in PDF documents is an important advantage of PLOP.

PLOP Library or Command-Line Tool?

PLOP is available as a programming library (component) for various development environments, and as a command-line tool for batch operations. The library and the command-line tool offer similar features, but are suitable for different deployment tasks.

The PLOP programming library is used...

...for integration into your desktop or server application. Examples for using the library with all supported language bindings are included in the PLOP package. Since the PLOP library accepts PDF input documents from a disk file or directly in memory, it can easily be combined with other products.

The PLOP command-line tool is suited...

...for batch processing PDF documents. It doesn't require any programming, but offers powerful command-line options which can be used to integrate it into complex workflows. The PLOP command-line tool can also be called from environments which do not support the use of the PLOP library.

Supported Development Environments

PDFlib PLOP 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 i5/iSeries and zSeries systems. The PLOP core is written in highly optimized C code for maximum performance and small overhead. Via a simple API (Application Programming Interface) the PLOP 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 on i5/iSeries

PLOP DS for digitally signing PDF

The extended version PLOP DS supports all features of PLOP, plus the ability to apply digital signatures to PDF documents. Please see the separate PLOP DS datasheet for more information.

Benefits of using PDFlib Software

Rock-solid Products

Tens of thousands of programmers worldwide successfully use 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 options 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

www.pdflib.com

datasheet

PDFlib pCOS 3

PDF Information Retrieval Tool

www.pdflib.com

What is PDFlib pCOS?

PDFlib pCOS provides a simple and elegant facility for retrieving any information from a PDF document which is not part of the page contents. For example, PDF metadata, interactive elements (links, form fields, etc.), or page dimensions can easily be queried with pCOS.

With pCOS you can extract a variety of interesting items and create output for different purposes. By processing multiple PDF documents with a single call you can easily create summaries of document info entries, page formats, fonts, or any other property. Combined with tabular output this provides a powerful PDF administration tool.

There are many application scenarios for the PDF Information Retrieval Tool PDFlib pCOS within PDF workflows, but you can also use PDFlib pCOS as a tool for learning or debugging PDF. Here are some typical situations:

- Check incoming documents for predefined criteria
- Identify problem files in a large collection
- Create metadata summaries for document management
- quality assurance before publishing documents
- document retrieval and repository workflows
- summarize the bookmarks
- extract components of PDF documents, e.g. ICC profiles
- Check PDFs for security problems (JavaScript etc.)

The pCOS retrieval interface is included in other PDFlib GmbH products: if you use PDFlib+PDI, PDFlib Personalization Server, TET or PLOP you also have access to the pCOS interface. If you need access to text or images on the page use our product PDFlib TET for PDF content extraction.

pCOS Cookbook

The pCOS Cookbook is a collection of programming examples which demonstrate the use of pCOS for various PDF retrieval tasks. The Cookbook is available on the PDFlib Web site and includes sample code, input documents and sample output.

PDFlib pCOS Features

Supported Input

PDFlib pCOS supports all flavors of PDF input:

- All PDF versions up to Acrobat X, including ISO 32000
- Encrypted documents (password may be required)
- Sophisticated security model: even if you don't know the password, you can query certain pieces of information as long as this doesn't violate the document author's intentions
- Damaged PDF input documents will be repaired if possible

Information Retrieval

PDFlib pCOS offers a simple query interface. With PDFlib pCOS you can extract a variety of interesting items, such as:

- Document info entries and XMP metadata
- General information: linearization and tagged PDF status, encryption details and permission settings, number of pages and fonts
- Fonts with name, embedding status, etc.
- Image data, such as bit depth, color space, compression
- Color space details
- Target URLs and coordinates of Web links
- Bookmarks and the corresponding page numbers, e.g. to create a table of contents
- Form field data: full field names, contents, position, etc.
- Page size, CropBox, page rotation
- Status of ISO standards: PDF/X, PDF/A, PDF/UA, PDF/E, and PDF/VT
- Geospatial reference information
- List or extract file attachments
- Layer names, page labels, article threads
- Annotation details
- List all comments along with the reviewer's name
- Digital signature details: name of signature field(s), signed/unsigned, name of signer, date and reason of signature
- Extract ICC output intent profiles from PDF/X or PDF/A documents
- Block properties for PDFlib Personalization Server
- JavaScript on document, page, annotation, or field level

Output Formats

PDFlib pCOS can create output for different purposes:

- Plain text output
- Unicode text output in UTF-8 or UTF-16 formats
- Tabular output for processing with a spreadsheet/database
- Binary data, e.g. ICC profiles or file attachments
- User-defined output formats for custom post-processing

pCOS Paths: Simple Syntax for PDF Objects

Instead of getting bogged down by complex tree structures, e.g. for bookmarks or form fields, you can easily access PDF objects by using the simple pCOS path syntax. It offers convenient shortcuts for accessing commonly used PDF objects, such as pages, fonts, bookmarks, form fields etc.

pCOS Library or Command-Line Tool?

pCOS is available as a programming library (component) for many development environments, and as a command-line tool for batch operations. Both offer similar features, but are suitable for different deployment tasks.

The pCOS 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 pCOS package.

The pCOS command-line tool is suited...

...for batch processing PDF documents. It doesn't require any programming, but offers powerful command-line options which can be used to integrate it into complex workflows. The pCOS command-line tool extends the features of the library:

- Simple retrieval of common PDF elements, such as bookmarks, annotations, metadata, form fields, etc.
- Extended mode for querying more complex objects and customizing the output format
- Extract data items such as file attachments, ICC profiles, etc.
- Emit information as comma-separated values or a user-defined format for import into a spreadsheet or database
- Recursion feature for dumping composite PDF objects, such as dictionaries and arrays

Supported Development Environments

PDFlib pCOS 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 X, Linux and Unix.

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

- COM for use with VB, ASP and many other languages
- C and C++
- Java, including servlets and Java Application Server
- .NET for use with C#, VB.NET, ASP.NET, etc.
- Perl
- PHP
- Python

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. 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

www.pdflib.com