datasheet

PDFlib pCOS 3

PDF Information Retrieval Tool

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