

datasheet

PDFlib

TET PDF IFilter 4

企業向けPDF検索

Windows版

PDFlib TET PDF IFilter とは

TET PDF IFilter は、PDF 文書からテキストとメタデータを抽出し、Windows 検索ソフトウェアで利用できるようにする製品です。これにより、PDF 文章をローカルのデスクトップ上や企業のサーバ上、Web 上から検索できるようになります。TET PDF IFilter は、特許を取得した PDFlib Text Extraction Toolkit (TET) をベースにしています。TET は、PDF 文書からテキストを確実に抽出するための確立された開発者向け製品です。

TET PDF IFilter は、Microsoft 社の IFilter インデクシングインタフェースを堅牢に実装しています。SharePoint や SQL Server 等、IFilter インタフェースに対応するすべての検索ソフトウェアと連携します。こうした製品では、HTML 等の特定のファイル形式に対して、それぞれ IFilter と呼ばれる専用のフィルタプログラムを使用します。TET PDF IFilter もフィルタプログラムの一つで、PDF 文章を対象としています。文書を検索するためのユーザーインタフェースとしては、Windows Explorer、Web やデータベースフロントエンド、クエリスクリプト、カスタムアプリケーションが考えられます。対話的な検索だけでなく、ユーザーインタフェースのないプログラムからでも利用することができます。

TET テクノロジーの利用

TET PDF IFilter のベースである PDFlib TET は、2002 年に初めてリリースされて以来、サーバとデスクトップ環境で世界中のお客様に利用されています。TET には、ページ内容やメタデータをテキストとして取得するだけでなく、XML 形式で提供することもできます。また、TET は Adobe Acrobat 用の無償プラグインという形でも利用することができます。このプラグインを使えば、TET の優れたテキスト抽出と画像抽出を、対話的にテスト、評価することができます。

TET PDF IFilter の特長

TET PDF IFilter の特長は以下の通りです。

- ▶ 欧米テキスト、中国語、日本語、韓国語 (CJK) テキスト、右から左に記述するアラビア語、ヘブライ語などのテキストをサポート
- ▶ 保護されたドキュメントのインデックス化や Acrobat では開けない PDF からの抽出
- ▶ ユニコードのホールディング、デコンポジション、ノーマライゼーションをサポート
- ▶ スレッドセーフ、高速、堅牢で、32 ビット版、64 ビット版をご用意
- ▶ 検索精度向上のために言語や文字体系を自動検知

企業向け PDF 検索

TET PDF IFilter は 32 ビット版でも 64 ビット版でも、完全なスレッドセーフとして利用することができます。TET PDF IFilter を以下の製品と組み合わせて利用することで、企業向け PDF 検索ソリューションを実現することができます。

- ▶ Microsoft SharePoint Server
- ▶ Microsoft Search Server
- ▶ Microsoft SQL Server
- ▶ Microsoft Exchange Server
- ▶ Microsoft Site Server

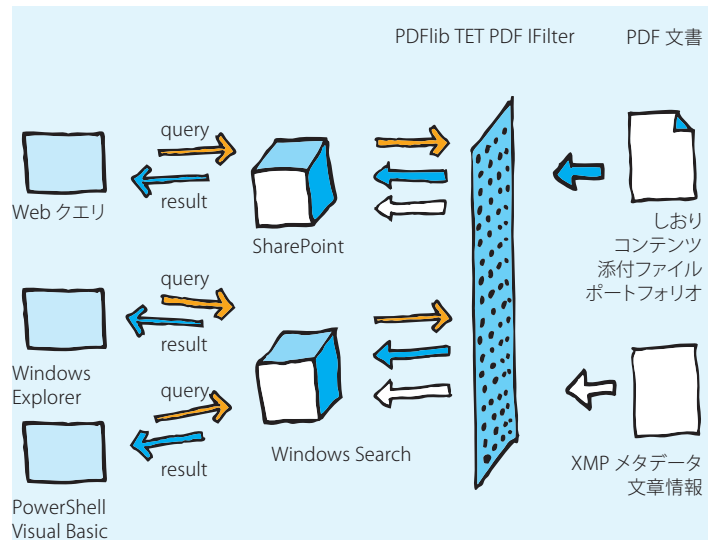
上記製品の他にも、IFilter インタフェースをサポートするすべての Microsoft 社製品、サードパーティ製品で利用できます。

デスクトップ PDF 検索

TET PDF IFilter は以下の製品と組み合わせることで、デスクトップ向け PDF 検索として利用することもできます。

- ▶ Windows Vista/7 に内蔵されている (もしくは Windows XP 向けの無償アドオンの) Windows Search
- ▶ Windows Indexing Service

TET PDF IFilter は、デスクトップ OS での非商用利用であれば無償で利用することができますので、十分にテスト、評価することができます。



機能詳細

対応する PDF

TET PDF IFilter は、あらゆる種類の PDF の入力に対応しています。

- ▶ ISO 32000-1 を含む、Acrobat 9 までのすべての PDF バージョン
- ▶ 表示用パスワードを必要としない暗号化された PDF
- ▶ 破損した PDF 文書も修復

Unicode への後処理

TET PDF IFilter は Unicode への様々な後処理をサポートし、より良い検索結果が得られるようにします。

- ▶ ホールディングは文字の保持や削除、置換を行います。例えば、検索と無関係な句読点や不要な文字を削除します。
- ▶ デコンポジションは文字を一字ないし複数の等価な文字に置き換えます。例えば、中国語の文字を標準的で等価な Unicode 文字に置き換えます。
- ▶ ノーマライゼーションはテキストをすべて 4 バイトの Unicode に変換します。例えば、データベースの要件に合うように NFC 形式で出力します。

国際化

TET PDF IFilter は、欧米テキストに加え中国語、日本語、韓国語 (CJK) テキストをフルサポートします。あらゆる CJK エンコーディングを認識でき、横書き、縦書きの出力モードにも対応しています。テキストの言語、リージョン識別子といったロケール ID の自動検知により、Microsoft 社の単語分割や語幹処理の精度が向上します。これは、東アジア諸語のテキストでは特に重要なことです。

ヘブライ語やアラビア語等の右から左に記述する言語もサポートしています。文脈上に沿って文字をノーマライゼーションし、テキストを論理的な順序に配置します。

PDF が持つコンテンツ以外の有用な情報

TET PDF IFilter は、PDF 文書をページ内容以上の情報を持つものとして扱います。TET PDF IFilter は、PDF 文書内の以下の項目にインデックスを付けます。

- ▶ ページ内容
- ▶ テキストしおり
- ▶ メタデータ (以下に記載)
- ▶ 添付 PDF 文書内のテキストも検索できるように、添付された PDF 文書や PDF パッケージノポートフォリオを再帰的に処理

XMP 文書メタデータと文書情報

TET PDF IFilter の高度なメタデータ実装は、メタデータ向けの Windows プロパティシステムをサポートしており、標準の、またはカスタムの文書情報項目だけでなく、XMP メタデータ (Adobe 社のリッチな XML ベースのメタデータの記述言語) にもインデックスを付けることができます。メタデータのインデクシングは、以下のレベルに設定することができます。

- ▶ タイトル、テーマ、作者などを表す文書情報項目やダブリンコアフィールド、その他共通の XMP プロパティを、等価な Windows プロパティに割り当てる
- ▶ ページサイズ、PDF/A 準拠レベル、フォント名等の有用な情報を、PDF 固有の仮想プロパティとして追加する
- ▶ すべての定義済み XMP プロパティを検索可能にする
- ▶ 社内独自の分類プロパティ、PDF/A 拡張スキーマ等の、ユーザー定義の XMP プロパティを検索可能にする

TET PDF IFilter はオプションとして、メタデータをフルテキストインデックスに統合することもできます。それにより、SQL Server のようなメタデータをサポートしていないフルテキスト検索エンジンでも、メタデータを検索できるようになります。

PDFlib の特長

世界的な導入実績と信頼性

世界 100 カ国以上で 20,000 ライセンスを超える導入実績がある PDF 文書処理ライブラリの定番ソフトウェアです。

使いやすい API を提供

PDF の詳細を意識することなく、製品ファミリーに共通する使いやすいインターフェースや操作性で PDF 文書の生成や処理を行うことができます。

事前に評価、開発が可能

PDFlib はダウンロードして無償で評価することができます。評価版は一部の制限を除いて製品の全機能を使用でき、納得いくまで評価した後で購入することができます。

効率的で安定した動作

PDFlib は、コンパクトなコードとして設計、開発されており、資源消費やオーバーヘッドが少なく高速かつ安定的に動作します。またスレッドセーフな設計のためマルチスレッド環境でも安心して利用することができます。

リーズナブルな価格体系

クライアント数に依存せず管理の容易なシンプルかつリーズナブルな価格のライセンス体系をご提供しています。

総合的な PDF 文書処理機能を実現

PDFlib、PLOG、TET、pCOS の併用により総合的な PDF 文書処理を実現できます。

安心のサポート

テクスタイル、2003 年から PDFlib 社との直接契約による正規リセラーとして PDFlib の販売を行っています。同社との強力なチャンネルを活用した迅速で正確な製品サポートをご提供しています。



PDFlib GmbH について

PDFlib の開発元である PDFlib GmbH は PDF テクノロジーにフォーカスしたドイツのソフトウェア会社です。1997 年に PDFlib を発表して以来、同製品ファミリーの充実を図り、PDF 関連技術の最新同行に迅速に対応してきています。

購入及びお問い合わせ

日本での PDFlib のご購入及びお問い合わせはテクスタイルまで。評価版のダウンロードや PDFlib 技術情報の入手もテクスタイルのウェブサイトで行えます。お見積りやその他ご質問については下記までお問い合わせください。

TechStyle

株式会社テクスタイル

104-0042 東京都中央区入船

3-7-1 星和京橋ビル別館

電話 : +81-3-6222-0063, FAX: +81-3-6222-3372

電子メール : pdflib@techstyle.jp

製品情報 : <http://pdflib.techstyle.jp>