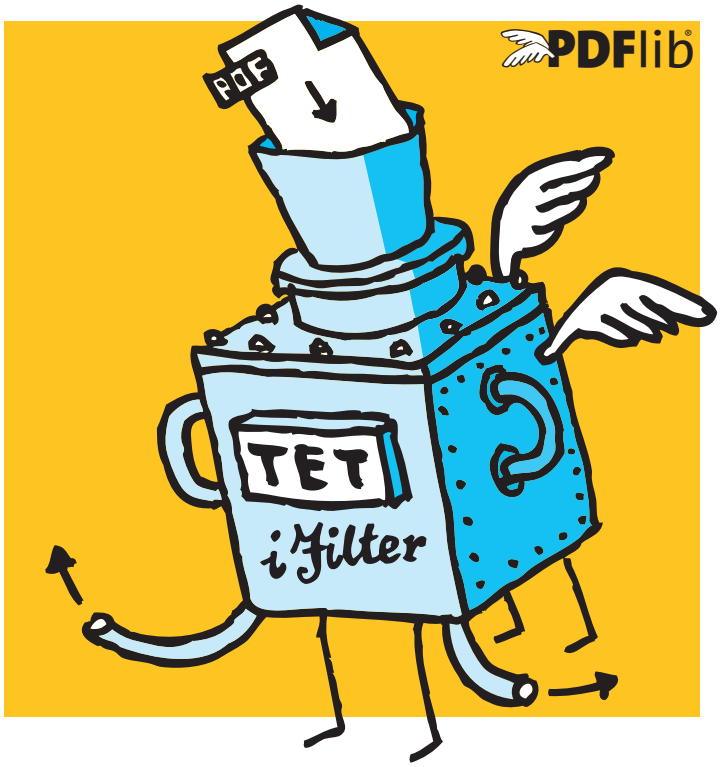


データシート

PDFlib TET PDF IFilter 5

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します。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、特許取得済の PDFlib Text and Image Extraction Toolkit（TET）を基盤としています。TET は、PDF 文書からテキストを確実に抽出する定評ある開発者向け製品です。

TET PDF IFilter は、Microsoft の IFilter インデクシングインタフェースの堅牢な実装です。SharePoint や SQL Server など IFilter インタフェースに対応するあらゆる検索ソフトウェアと連携します。こうした製品では、HTML など特定のファイル形式に対してそれぞれ IFilter という形式独自のフィルタプログラムを使用します。TET PDF IFilter もこのようなプログラムの 1 つであり、PDF 文書を対象としたものです。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ありえます。対話的な検索だけでなく、ユーザーインタフェースなしでプログラム的にクエリを発することもできます。

特許取得済 TET テクノロジーを基盤に

TET PDF IFilter の基盤である 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 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。TET は Adobe Acrobat 用無償プラグインという形でも利用可能です。このプラグインを使って、TET の優れたテキスト・画像抽出を対話的にテスト・評価していただけます。

比類なきさまざまな特長

TET PDF IFilter の特長は以下の通りです：

- 欧米テキスト、日本語・中国語・韓国語（CJK）テキスト、アラビア語・ヘブライ語など右から左へ記述する言語に対応
- 保護されたドキュメントをインデックスし、Acrobat で開けない PDF からもテキストを抽出
- Unicode のフォルディング・分解・正規化に対応
- 実装：スレッドセーフ、高速、堅牢、32 ／ 64 ビット版
- 自動用字系・言語検知で検索精度向上

企業向け PDF 検索

TET PDF IFilter は完全スレッドセーフなネイティブ 32 ／ 64 ビット版として利用可能です。TET PDF IFilter を以下の製品と組み合わせて企業向け PDF 検索ソリューションを実現可能です：

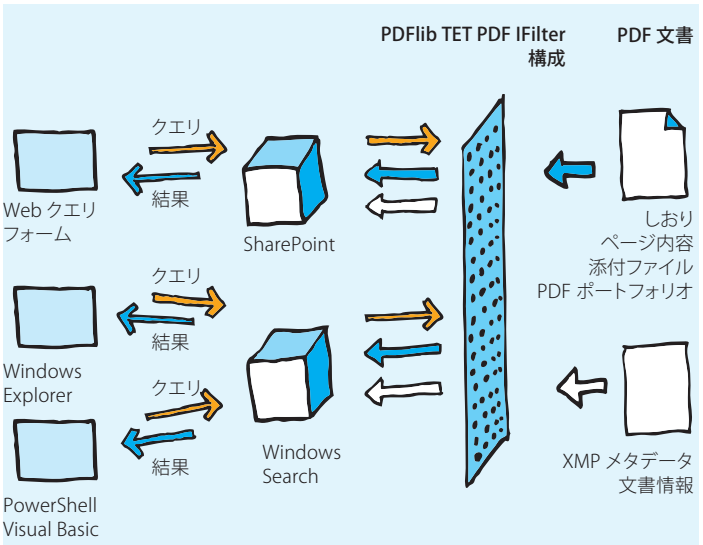
- Microsoft SharePoint Server 2013 およびそれ以前のバージョン
- Microsoft Search Server
- Microsoft SQL Server
- Microsoft Exchange Server
- Microsoft Site Server

TET PDF IFilter はこの他全ての IFilter インタフェース対応 Microsoft・サードパーティ製品とともに利用可能です。

デスクトップ PDF 検索

TET PDF IFilter を利用して、例えば Windows に内蔵の Windows Search をともにデスクトップ PDF 検索を実装することもできます。

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すので、気軽にテスト・評価していただけます。



機能詳細

受け付ける PDF 入力

TET PDF IFilter は全ての重要な種類の PDF 入力に対応しています：

- ISO 32000-1 と -2 を含め、Acrobat DC まで全 PDF バージョン
- 文書を開くのにパスワードを必要としない保護された PDF
- 破損 PDF 文書は修復されます

Unicode 後処理

TET PDF IFilter は様々な Unicode 後処理工程をサポートしており、これを活用して抽出テキストを向上させることができます：

- フォルディング：キャラクタを保持・削除・置換します。例えば句読点や、無関係な用字系のキャラクタを削除します。
- 分解：1 個のキャラクタを、等価な 1 個ないし複数のキャラクタの列へ置き換えます。例えば、和文の半角・全角・縦書きキャラクタや英字の上付き形（ª など）をそれぞれ標準形へ置き換えます。
- テキストを 4 種の Unicode 正規形全てへ変換可能です。例えば、Web テキストやデータベースの要件を満たすよう NFC 形式で出力できます。

国際化

TET PDF IFilter は、欧米テキストだけでなく日本語・中国語・韓国語（CJK）テキストに完全対応しています。あらゆる CJK エンコーディングを認識でき、横書き・縦書きにも対応しています。テキストのロケール ID（言語・リージョン識別子）の自動検知により、Microsoft の単語分割・語幹処理アルゴリズムの結果に改善を加えます。これは特に東アジアテキストで重要です。

ヘブライ語・アラビア語など右から左へ記述する言語にも対応しています。位置依存キャラクタ字形が正規化され、テキストが論理的順序で出力されます。

PDF は単なるページの束にあらず

TET PDF IFilter は PDF 文書を、単なるページ群の他にも多様な情報を含む入れ物として扱います。TET PDF IFilter は PDF 文書内のすべての重要な項目をインデックスします：

- ページ内容
- しおりのテキスト
- メタデータ（下記）
- 埋め込まれた PDF 文書と PDF パッケージ／ポートフォリオを再帰的に処理して、添付の中も検索可能に

XMP 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、Windows のメタデータ用プロパティシステムに対応しています。XMP メタデータと標準またはカスタム文書情報項目群をインデックスします。メタデータインデクシングはいくつかのレベルに設定可能です：

- 文書情報項目群・ダブリンコアフィールド群など共通 XMP プロパティ群を、等価な Windows プロパティへマップします。例：タイトル・テーマ・作者
- TET PDF IFilter は、有用な PDF 独自の仮想プロパティ群を追加します。例：ページ寸法・PDF/A 準拠レベル・フォント名
- 全ての重要な定義済み XMP プロパティを検索できます。
- ユーザー定義 XMP プロパティ群を検索できます。例：会社独自の分類プロパティ、PDF/A 拡張スキーマ

TET PDF IFilter は、メタデータを全文テキストインデックスに内蔵させることもできます。それにより、メタデータ対応のない全文テキスト検索エンジン（SQL Server など）でもメタデータを検索可能になります。

PDFlib ソフトウェア利用の利点

磐石の製品群

世界中の数万人のプログラマーが当社ソフトウェアを使用しています。PDFlib はサーバ運用のためのあらゆる品質・パフォーマンス要求を満たします。PDFlib 製品はすべて、堅牢な 24×7 サーバ運用と無人バッチ処理に適しています。

速度とシンプル性

PDFlib 製品群は驚異的に高速です――秒速数千ページを実現します。そのプログラミングインタフェースは簡明で習得が容易です。

世界中に PDFlib 製品群

当社製品群は世界のあらゆる言語と Unicode に対応しています。世界じゅうのお客様にご利用いただいています。

プロフェッショナルサポート

問題があれば、当社は支援に努めます。ビジネスクリティカルなさまざまな用途の要求を満たす商用サポートを提供しています。サポートを追加すると、最新バージョンへのアクセスと、問題発生時の回答時間保証をご利用いただけます。

ライセンシング

サーバライセンス・組み込み／サイトライセンス・ソースコードライセンスのためのさまざまなライセンシングオプションを提供しています。短時間回答と無償アップデートを伴う拡大技術サポートのためのサポート契約もご利用いただけます。

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 関連技術の最新動向にも迅速に対応しています。

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インフォテックまで。PDFlib 製品のダウンロードや技術情報の入手もインフォテック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。

infoTek
Information Technology for People

インフォテック株式会社 PDFlib セールス担当
〒183-0055 東京都府中市府中町 2-1-7
電話 : 042-358-5777 FAX: 042-358-5801
電子メール : pdflib_sales@infotek.co.jp
製品情報 : http://www.infotek.co.jp