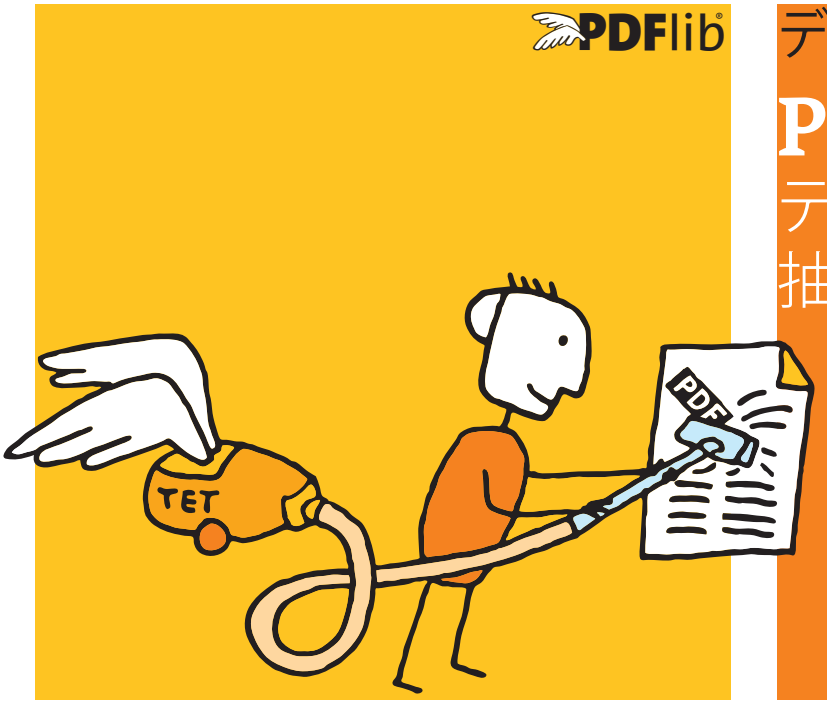


データシート

PDFlib TET 5

テキスト・画像
抽出ツールキット

PDFlib TET とは

PDFlib TET（テキスト・画像抽出ツールキット）は、PDF 文書からテキスト・画像・メタデータを確実に抽出します。TET を利用すると、PDF のテキスト内容を Unicode 文字列として取得できるだけでなく、詳細な色・グリフ・フォント情報やページ上の位置を知ることもできます。ラスタ画像を、広く用いられているさまざまな画像形式で抽出できます。TET を用いて、PDF を、テキストとメタデータ全てリソース情報を内容とする TETML という XML ベースの形式へ変換することも可能です。

TET は、単語区切りを検出し、テキストの段組を認識し、冗長なテキストを除去するための高度な内容解析アルゴリズムを実装しています。内蔵の pCOS インタフェースを用いて、PDF から、メタデータやインタラクティブ要素など任意のオブジェクトを取得できます。

PDFlib TET でできること：

- 検索エンジン用の PDF インデクサを実装
- PDF 内のテキストと画像を再利用
- PDF の内容をさまざまな他形式へ変換
- PDF をその内容に応じて処理。例：見出しで分割（TET に加えて PDFlib+PDI が必要です）
- ページ上の特定の位置が空かどうかをチェック。例：バーコードやスタンプを配置するため

PDFlib TET のさまざまな機能

対応する PDF 入力

TET は全ての重要な種類の PDF の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

PDF 内のテキストは通常、Unicode で符号化されていませんので、PDFlib TET は PDF 文書内のテキストを Unicode へ正規化します：

- TET は全テキスト内容を Unicode へ 変換します。C をはじめとする Unicode 非対応言語ではテキストは UTF-8 または UTF-16 形式で返され、Unicode 対応のプログラミング言語ではネイティブ文字列として返されます。
- 合字をはじめとする複数キャラクタグリフは、照応する Unicode キャラクタ列へ分解されます。
- 適切な Unicode マッピングのないグリフは認識され、誤解釈防止のため、設定可能な置き換えキャラクタへマップされます。
- TET は、個別の文書作成環境、例えば InDesign や TeX の文書や、メインフレームシステムで生成された PDF で起こる問題に対して、さまざまな回避策を実装しています。

内容解析と単語検出

TET は、特許になっている内容解析アルゴリズムを複数有しています：

- 単語を正しく抽出するために単語区切りを割り出し
- ハイフンで分割された単語を再結合（デハイフネーション）
- 影や太字化などが施されたテキストの重複インスタンスを除去
- 段落群を読み順に再結合
- ページ上に分散したテキストを正しく並べ替え

ページレイアウト・表組み検出

ページ内容を解析し、段組を割り出します。表組みを検出し、複数列にわたるセルも認識します。これにより、抽出テキストの並べ替えが向上します。表行と、各表セルの内容を識別できます。

幾何情報

TET はテキストの正確な幾何情報を提供します。例えばページ上の位置、グリフ幅、テキストの向きなどです。ページの特定領域をテキスト抽出から除外したり含めたりすることも可能です。これにより、例えばヘッダ・フッタや余白を無視できます。

テキスト色

TET は、PDF ページ記述内の色情報を解析して各グリフの正確な色情報を返します。これを活用して、例えば見出しをはじめとする強調テキストを識別できます。

画像抽出

PDF ページ上の画像を TIFF・JBIG2・JPEG・JPEG 2000 ファイルとして抽出できます。各画像の正確な幾何情報（位置・寸法・角度）が取得されます。分割されている画像は大きな画像として結合されますので再利用も容易です。ダウンサンプリングも色変換も一切行われませんので画像の忠実性が保証されます。これにより、求める最高の画像品質が必ず実現されます。

PDF 解析

TET ライブラリは、PDF 文書に関するさまざまな詳細をクエリするための pCOS インタフェースを内蔵しています。文書情報・XMP メタデータ・フォントリスト・ページ寸法をはじめとする多様な情報を取得できます（pCOS 製品のデータシートを別途ご参照下さい）。

問題を含む PDF に対するさまざまな設定オプション

TET は、他の製品では正しくテキストを抽出できないようなさまざまな種類の PDF に対する特別な処理や回避策を有しています。さらに、問題のある文書の処理を向上させる多様な設定機能を備えています：

- 文字コードかグリフ名を Unicode へマップするテーブルをユーザーが与えることによって Unicode マッピングをカスタマイズできます。
- PDFlib FontReporter：PDF 内のフォント・エンコーディング・グリフを解析する補助ツールです。Adobe Acrobat のプラグインとして動作します。このプラグインは無償で、OS X 版と Windows 版があります。
- Unicode マッピングのためのさらなるヒントを見つけるために埋め込みフォントを解析します。フォントが埋め込まれていないときは、外部フォントファイルかシステムフォントを用いてテキスト抽出結果を改善します。

Unicode 後処理

TET 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。

さまざまな文書領域

PDF 文書は、ページ内容以外にもいろいろな所にテキストを持てます。ページ内容のみを扱う事例が多いですが、他の文書領域も重要な場合は多々あります。TET は、以下の文書領域全てからテキストを抽出します：

- ページ内容
- 定義済み・カスタム文書情報項目
- 文書レベル・画像レベルの XMP メタデータ
- しおり
- ファイル添付・PDF ポートフォリオを再帰的に処理可能
- フォームフィールド
- コメント（注釈）
- さまざまな一般 PDF プロパティをクエリ可能。例：ページ数、PDF/A・PDF/X などの標準準拠

XMP メタデータ

TET は複数の方式で XMP メタデータに対応しています：

- 内蔵の pCOS インタフェースを用いて、文書・個別ページ・画像、または文書のその他の部分に対する XMP メタデータをプログラム的に抽出可能。
- XMP 文書・画像メタデータが PDF 内にあれば TETML 出力の内容に含まれる。
- TIFF または JPEG 形式で抽出された画像には XMP 画像メタデータが、PDF 内にあったなら含まれる。

TETML：PDF 内容を XML で表現

TET は PDF 内容を、TETML という一種の XML で表現することもできます。TETML で表現された多様な PDF 情報は、広く用いられている各種 XML ツールで容易に処理できます。TETML はテキスト本体を内容とするだけでなく、フォント・位置情報・リソース情報（フォント・画像・色空間）・メタデータを含むこともできます。

TETML は、フォームフィールド・注釈・しおりといったインタラクティブ要素も含んでいます。TETML を用いて、JavaScript や色空間の内容、ICC プロファイルや出力インテントを解析することも可能です。

TETML は、照応する XML スキーマで規定されていますので、TET はつねに、一貫性と信頼性をそなえた XML 出力を生成します。TETML を XSLT スタイルシートで処理することも可能です。それによって例えば、何らかのフィルタを適用したり、TETML を他の形式へ変換したりすることができます。TETML を処理するためのサンプル XSLT スタイルシートが TET ディストリビューションに含まれています。

以下に、グリフの内容を持った TETML 出力の一部を示します：

```
<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>
```

さまざまな TET コネクタ

TET コネクタは、TET を他のソフトウェアとインタフェースさせるために必要な接続コードを提供します。以下の TET コネクタにより、PDF テキスト抽出機能が各種ソフトウェア環境で利用可能になります：

- Lucene 検索エンジン用 TET コネクタ
- Solr 検索サーバ用 TET コネクタ
- TIKA ツールキット用 TET コネクタ
- Oracle Text 用 TET コネクタ
- MediaWiki 用 TET コネクタ
- 各種 Microsoft 製品用には、別製品である PDFlib TET PDF IFilter をお使い下さい。PDF 文書からテキストとメタデータを抽出し、Windows 上の検索・抽出ソフトウェアでの利用を可能にします（詳しくは別途データシートをご覧下さい）。

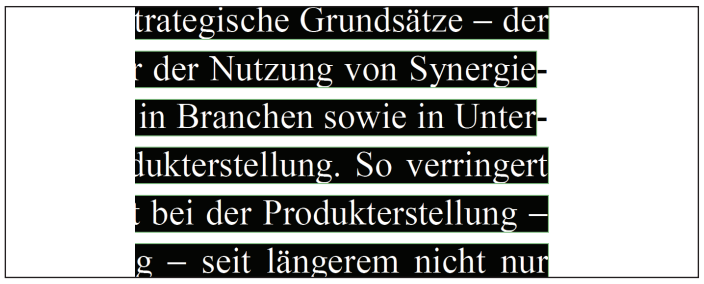
TET クックブック

TET クックブックは、さまざまなテキスト・画像抽出タスクにおける TET の使用法を示したプログラミング作例集です。いくつかのクックブックサンプルは、TET と PDFlib+PDI を組み合わせて PDF 文書を改良する方法を示しています。これにより例えば、ページ上のテキストに基づいてしおりやリンクを追加することが可能です。

PDF テキスト抽出のさまざまな課題

ハイフン除去

TET は、単語がハイフンで区切られて複数行にわたっているのを検出し、そのハイフンを除去して各部を結合して元の単語へ戻します。これは、文書の中にハイフンで区切られた単語しかなくてもこの単語全体を必ず正しく検索されるようにするために重要な処理です。ダーシ（ハイフンとは異なる）については、除去してはいけませんので、別に処理されます。



TET はハイフンを正しく除去し、ダーシを温存します。

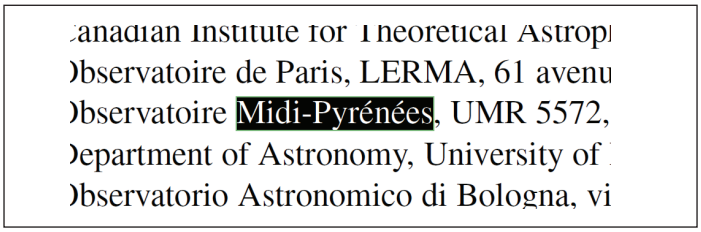
影付き・太字テキスト検出

電子文書では影付きテキストがよく使われますが、これは、同じテキストを少しずらして複数回ページ上に配置することで影付き効果を得ています。同様に太字テキストもたいていは、同じテキストを複数個重ねることで太字に見せかけています。その結果、影付きや太字の単語のキャラクタは、文書内に複数回含まれています。TET の特許取得済の影付き検出アルゴリズムは、重複したテキストを特定して除去することで、余分なテキスト抽出を防止します。他のソフトウェアでは影付きや太字は重複して抽出されてしまいますが、TET では重複が正しく除去されます。単語全体が重複しているなら検索エンジンでヒットしますが、例のようにキャラクタ毎に重複してるケースでは検索結果に含まれないことになります。

他の製品では「Inttrroduccttiion」と抽出されます。
TET は正しく「Introduction」と抽出します。

アクセント付きキャラクタ

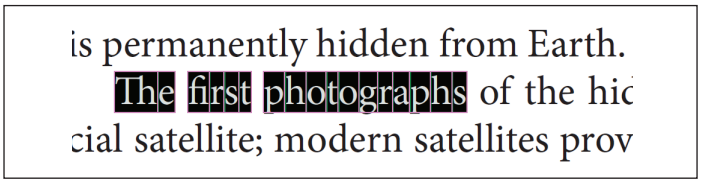
多くの言語では、アクセント等の発音区別記号を他キャラクタのそばに配して合成キャラクタを形成します。TeX に代表される特定の組版ソフトウェアでは、ベースキャラクタとアクセントの 2 つのキャラクタを別々に出力して合成キャラクタを形成します。例えばキャラクタ ä を作るのに、まず文字 a がページ上に配置され、その頭に分音記号 ¨ が配置されます。TET はこうした状況を検出し、2 つのキャラクタを再合成して適切な合成キャラクタを復元します。



他の製品では「Midi-Pyr´en´ees」と抽出されます。
TET は正しく「Midi-Pyrénées」と抽出します。

合字

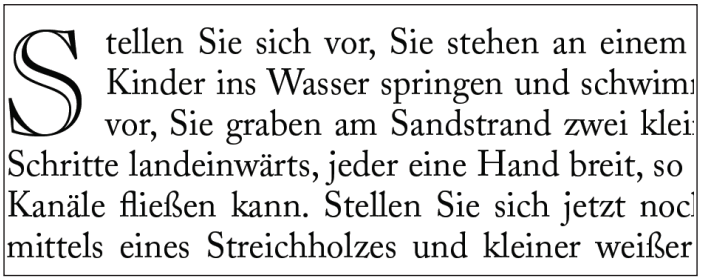
合字は、複数のキャラクタを 1 つのグリフに合体したものです。よく見られる合字は *fi*・*fl*・*ffi* ですが、ほかにも *Th*・*sp*・*ct*・*st* 等あまり目にしない合字が数多くあります。電子文書からテキストを抽出する際には、合字を解析してキャラクタ列に分解することで、正しいテキスト処理を可能にする必要があります。TET は合字を検出し、適切な複数キャラクタとして出力します。



他の製品では「 e rst photographs」と抽出されます。
TET は正しく「The first photographs」と抽出します。

ドロップキャップ

ドロップキャップは、先頭段落の 1 字目を大きな文字で表現したものです。この 1 字目の上端はその行の上端に揃い、この字の残りは複数行にわたってぶら下がります。ドロップキャップは段落の開始を強調するために用いられます。ドロップキャップを正しく扱わないと、単語の 1 字目とその後の文字列を別々に 2 つの単語として抽出してしまうでしょう。



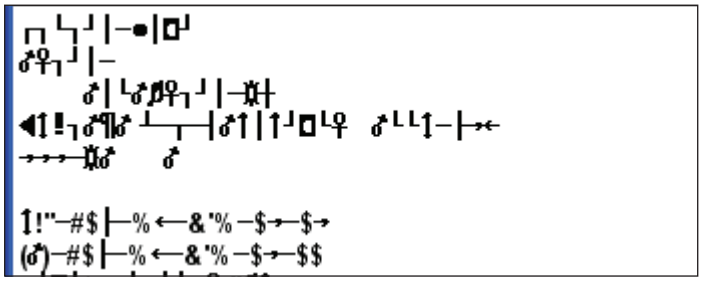
他の製品ではドロップキャップ「S」と「tellen」の 2 つの単語として抽出されます。
TET は正しく 1 単語「Stellen」として抽出します。

PDF テキスト抽出のさまざまな課題

Unicode マッピング

Unicode マッピングは PDF テキスト抽出の基礎を成す部分です：ページ上のすべてのグリフに、照応する Unicode 値を割り当てる必要があります。PDF は多様な種類のフォントやエンコーディングに対応しており、それらの中には正しい Unicode 値を割り当てるために必要な情報を提供するものもあればしないものもあるので、Unicode マッピングは複雑なタスクとなります。最悪の場合、文書が十分な情報を提供しなかったために、文書から使い物になるテキストを一切抽出できない結果に終わります。

TET の特許取得済の Unicode マッピングアルゴリズムは、得られる情報を全て活用して Unicode 値を決定するカスケード型アルゴリズムを実装しています。問題を抱える多くの文書に対して、他の製品が使い物にならないゴミしか出力しない場合でも、TET は正しい Unicode テキストを抽出します。



他製品は使い物にならないゴミを抽出しますが、TET はテキストを抽出します。

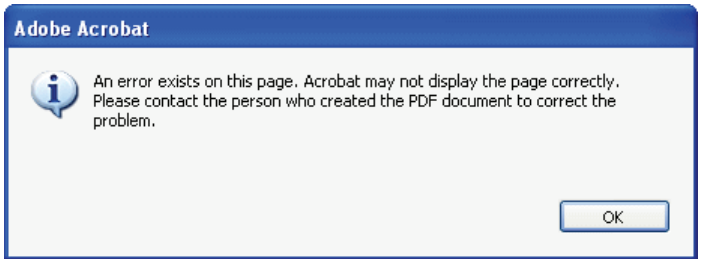
アラビア文字・ヘブライ文字の双方向テキスト

PDF は、論理的なテキストを符号化したものではなく、単にページ上のグリフ群の入れ物です。アラビア文字・ヘブライ文字のテキストは右から左へ進行します。その中にしばしば数や欧米言語での名前などが左から右へ挿入されていますので、テキストを双方向に解釈する必要があります。アラビア文字にはさらに、キャラクタが文脈によって 4 種類もの位置依存形をとりうるという課題もあります。これらの表示形を、照応する標準形（独立形）へ正規化する必要があります。

TET は、右から左へのテキストと左から右へのテキストが視覚的に混在しているのを並べ替えて、正しい論理テキスト出力を生成します。

破損 PDF 文書

PDF 文書は伝送誤りなどによって破損する場合があります。TET の修復モードはさまざまな種類の破損 PDF を復元します。ときには PDF 文書の破損が著しく Acrobat でページ表示もできない場合もあります。このような極端な場合でも TET はしばしばその文書のページ内容を出力します。



ページ内容を Acrobat で表示もできないのに、TET はそのテキストを正確に抽出します。

画像抽出のさまざまな課題

色空間と圧縮

PDF 内のラスタ画像データの符号化には、11 種の色空間と 9 種の圧縮フィルタの組み合わせがありえますが、JPEG や TIFF など一般的な画像ファイル形式がこれらの部分集合のみに対応しています。TET の画像エンジンは、PDF 画像の諸特性と出力形式の諸機能とのバランスをとります。PDF 画像の内部構造にかかわらず、ピクセル画像はいずれかの一般的な画像ファイル形式で抽出されます。

スポットカラー

PDF 内の画像は、CMYK プロセスカラーだけでなくカスタムスポットカラーも用いている場合があります。技術的にはこれらの色空間は Separation（シングルチャンネル）と DeviceN（マルチプルチャンネル）として知られています。

TET は TIFF 出力を、追加のスポットカラーチャンネル群とともに生成します。これは、高い色忠実性を必要とする、色変換を一切受け入れられない用途のために意図されているものです。DeviceN カラーによる画像が一般的な CMYK プロセスカラー群の部分集合しか含んでいない場合には（シアンとマゼンタのみなど）、プレーンな CMYK 出力を生成できるよう、欠けているプロセスチャンネルが追加されます。

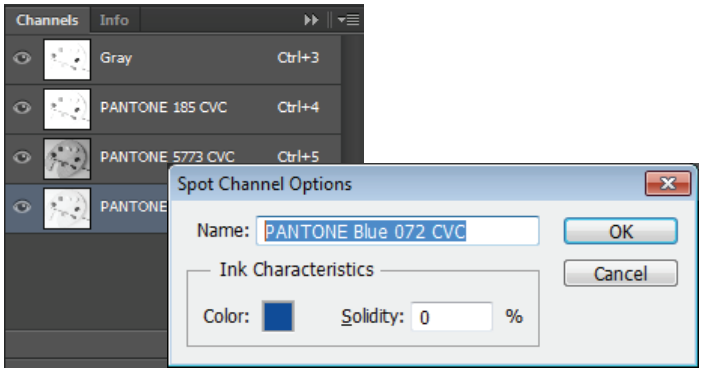
ただし、スポットカラーチャンネル群を取り扱えず、プレーンな TIFF 出力に限定されている用途もあります。この場合、処理を支援するために、ただ 1 つのスポットカラーチャンネルをグレースケール TIFF として出力するよう TET に命じることも可能です。

断片化画像の結合

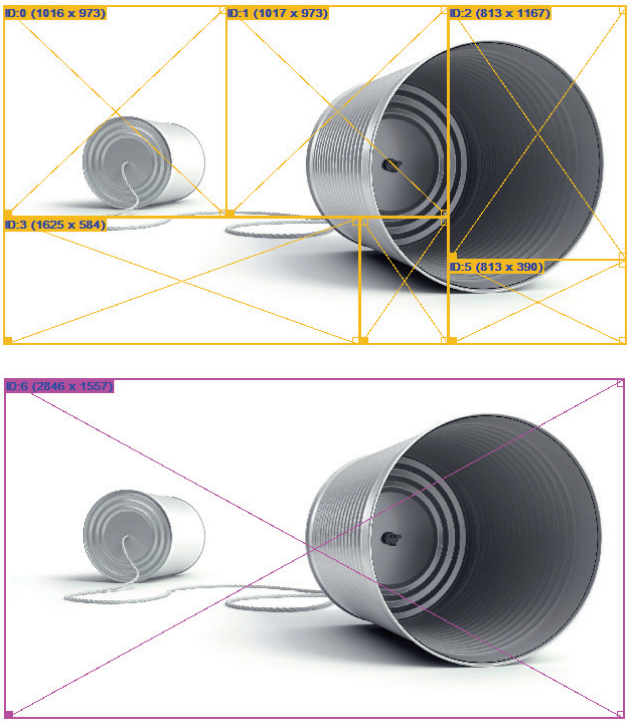
多くの PDF 文書において画像は、その PDF を生成したソフトウェアによって細かく断片化されています。ページ上で 1 つの画像として見えていても、実は多数の断片の寄せ集めということがあります。例えば Microsoft Office アプリケーション群や TeX はしばしば、数百から数千の細かな断片でできた著しく断片化した画像を生成します。Adobe InDesign はしばしば画像を、さまざまな大きさの断片へ分割します。TET は断片化画像を検出し、その断片群を結合して、利用可能な大きな画像を形成します。画像結合をして初めて、断片化画像を有効に再利用することが可能になります。

デバイス独立	CIEベース	特色
DeviceGray	CalGray	Indexed
DeviceRGB	CalRGB	Pattern
DeviceCMYK	Lab	Separation
	ICCBased	DeviceN

TET は、PDF 内で使われうるあらゆる色空間を処理します。



Photoshop が、抽出された TIFF 画像のスポットカラーチャンネルをチャンネルウィンドウに表示（左）。いずれかのアイコンをダブルクリックするとその代替色が表示されます（下）。



画像が小さな部分へ分割されていますが（上）、TET はこれをただ 1 つの、再利用可能な画像として抽出します（下）。

TET の多様な利用形態

TET は、各種開発環境用のプログラミングライブラリとしても、バッチ処理用のコマンドラインツールとしても利用できます。両者は同等の機能を提供しますが、実装目的に応じて使い分けられます。この TET ライブラリと TET コマンドラインツールはいずれも、TET の XML ベースの出力形式である TETML を生成できます。

TET は以下の実装オプションを提供します：

- TET プログラミングライブラリ（コンポーネント）：デスクトップまたはサーバアプリケーションに組み込まれて使用されます。このライブラリの使用例は TET パッケージに含まれています。
- TET コマンドラインツール：PDF 文書のバッチ処理に適しています。プログラミングを一切必要とせずに複雑なワークフローに組み込むために利用できるコマンドラインオプション群を提供します。
- TETML 出力：XML ベースのワークフローと、XSLT などさまざまな XML 処理ツールや言語に通じた開発者に適しています。
- TET コネクタ：データベースや検索エンジンなどさまざまな汎用ソフトウェアパッケージに TET を統合するのに適しています。

TET 製品ファミリー

TET ファミリーには以下の製品があります：

- TET コア製品：本データシートで記述している製品です。
- TET PDF IFilter：別製品として利用可能です。Windows Search・SharePoint・SQL Server など Microsoft 検索製品群での使用に適しています（詳しくは別途データシートをご覧ください）。
- Adobe Acrobat 用 TET Plugin：PDF からテキストと画像を抽出するための無償ユーティリティです。これを使って TET を対話的に評価していただけます。

対応開発環境

PDFlib TET はどこにでも—事実上すべてのコンピューティングプラットフォーム上で動作します。Windows・OS X・Linux・Unix の広く使われているすべての種類と IBM i5/iSeries・zSeries メインフレームのための 32 ビット／64 ビットパッケージを提供しています。TET は、iOS・Android を含むモバイルシステムでも利用できます。

TET のコアは、パフォーマンス最大化とオーバーヘッド最小化のために高度に最適化された C・C++ コードで書かれています。シンプルな API（アプリケーションプログラミングインタフェース）を通じて、TET の機能は、多様な開発環境から利用可能です：

- COM（VB・ASP などで使用）
- C・C++
- Java（サーブレット・Java Application Server を含む）
- .NET（C#・VB.NET・ASP.NET などで使用）
- Objective-C（OS X・iOS）
- Perl
- PHP
- Python
- REALbasic/Xojo
- RPG（IBM i5/iSeries）
- Ruby（Ruby on Rails を含む）

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