

PDFlib TET 4

テキスト抽出ツールキット



PDFlib TET とは

PDFlib Text Extraction Tool Kit (TET) 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 文字列として取得することができるほか、グリフやフォントに関する詳しい情報や、ページ上の位置を知ることができます。ラスタ画像は、広く用いられている画像形式で抽出されます。TET は、TETML という XML ベースの書式を実装しており、必要に応じ PDF 文書を TETML 形式に変換することができます。TETML はテキストやメタデータに加えリソース情報を保持することができます。

TET は、高度な内容分析アルゴリズムを実装し、単語境界の検出、テキストの段組認識、冗長テキストの除去などの処理を効率的に実現します。また pCOS インタフェースにより、PDF からメタデータやインタラクティブ要素等任意のオブジェクトを抽出することができます。

PDFlib TET の利用例：

- ▶ 検索エンジンの PDF 文書のインデクシング
- ▶ 既存 PDF 文書内のテキストや画像の再利用
- ▶ PDF 文書の内容の異なる形式への変換
- ▶ PDF 文書を解析し、その内容に応じて異なる処理を実施。
例えば、見出しによる文書の分割 (TET の他、PDFlib+PDI が必要となります)

PDFlib TET の機能

対応する PDF

PDFlib TET は様々な 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

PDF 内のテキストは通常、Unicode でエンコードされていないので、PDFlib TET は PDF 文書内のテキストを、次のように Unicode により正規化します。

- ▶ TET は、すべてのテキストコンテンツを Unicode へ変換します。C などの 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。
- ▶ 合字などの複数文字グリフは、対応する Unicode 文字列に分解します。

- ▶ 適切な Unicode マッピングのないグリフを認識した場合、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します。
- ▶ InDesign や Tex の文書或いはメインフレームシステム上で生成された PDF など特定の文書作成環境に起因する問題に対して TET ではさまざまな回避策を実装しています。

内容分析と単語の検出*

TET は、次のような高度な分析アルゴリズムを有しています。

- ▶ 適切な単語抽出に必須の単語境界決定アルゴリズム (特許技術)
- ▶ ハイフネーションされた単語の各部分の再結合 (デハイフネーション)
- ▶ 影付きや太字化等のテキストの重複インスタンスの除去
- ▶ 段落の読み順による再結合
- ▶ ページ上に分散したテキストを正しい順序に並べ替え

ページレイアウトと表組みの検出*

ページ内容を分析し、段組を割り出します。複数列をまたぐセルも含め表組みを検出します。本処理により抽出テキストの順序の決定が容易になり、表の行や各表のセルの内容を特定することができます。

幾何情報

TET は、ページ上の位置、グリフの幅、テキストの向きなど、テキストの正確な幾何情報を提供します。ページの特定の領域を指定してテキスト抽出の対象から除外したり、逆に指定部分のみからテキスト抽出することができます。たとえばヘッダー、フッターや余白を除外することができます。

画像抽出

PDF ページ上の画像を、TIFF、JPEG または JPEG 2000 ファイルとして抽出できます。各画像について、正確な幾何情報 (位置、寸法及び角度) を取得できます。分割されている画像を大きな画像に結合して再利用することができます。ダウンサンプリングや色空間の変換は行われないので、画像の忠実度が保証され、最高の画像品質が保証されます。

PDF の分析

TET ライブラリには pCOS インタフェースが含まれ、PDF 文書の文書情報、XMP メタデータ、フォントリストやページ寸法などさまざまな詳細情報を取得できます。(pCOS 製品については pCOS のデータシートを参照してください)

* 本機能は、主に欧文処理を対象としており、日本語処理については制限があります。

問題を含む PDF に対するオプション設定

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 に対して、特殊な処理や回避策を実現します。さらに、問題文書の処理を改善するためのさまざまな設定機能を備えています。

- ▶ 文字コードまたはグリフ名を Unicode へマッピングするテーブルをユーザーが設定することによって、Unicode マッピングをカスタマイズすることができます。
- ▶ PDFlib FontReporter は、PDF 内のフォント、エンコーディング及びグリフを分析する補助ツールで、Adobe Acrobat のプラグインとして動作します。このプラグインには Mac 版と Windows 版があり、無料で使用できます。
- ▶ Unicode マッピングに有効な情報を得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またはシステムフォントを用い、テキスト抽出結果を改善します。

Unicode への後処理

TET 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。

- ▶ フォ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。
- ▶ 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、半角カタカナや機種依存文字（「㊦」など）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。

文書の領域

PDF 文書では、ページコンテンツ以外の場所にもテキストがあります。多くのアプリケーションはページコンテンツしか扱いませんが、文書のその他の領域が必要な場面も多くあります。TET は、以下の文書領域全てからテキストを抽出することができます。

- ▶ ページコンテンツ
- ▶ 定義済み及びカスタム文書情報項目
- ▶ 文書と画像レベルの XMP メタデータ
- ▶ しおり
- ▶ ファイル添付と PDF ポートフォリオの再帰的処理
- ▶ フォームフィールド
- ▶ コメント（注釈）
- ▶ ページ数や PDF/A・PDF/X 等標準への準拠状態など一般的 PDF プロパティ

XMP メタデータ

TET は、以下のような形式で XMP メタデータをサポートしています。

- ▶ 内蔵 pCOS インタフェースを用い、文書、各ページ、画像または文書の他の部分の XMP メタデータをプログラムにより抽出する。
- ▶ XMP 文書や画像メタデータが PDF 文書内に存在する場合には TETML 出力にこれを含める。
- ▶ 画像メタデータが PDF 文書内に存在する場合には TIFF または JPEG 形式で抽出された画像にこれを含む。

TETML：PDF 内容を XML で表現

TET では、PDF コンテンツを TETML という一種の XML で表現することができます。TETML で表現されたさまざまな PDF 情報は広く用いられている XML ツールで容易に処理することができます。TETML にはテキスト本体のほか、フォント、位置情報、画像・カラー

スペースなどリソースの詳細及びメタデータを含めることができます。

TETML は、対応する XML スキーマに規定されており、TET はつねに、一貫性と信頼性を具えた XML 出力を生成します。フィルタリングや書式の変換などのために XSLT スタイルシートで TETML を処理することも可能です。TET には、TETML を処理するサンプル XSLT スタイルシートが添付されています。

以下に示すのはグリフの詳細の一部を TETML で表したものです。

```
<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">I</Glyph>
  <Glyph font="F1" size="18" x="146.21" y="636.33" width="4.88">K</Glyph>
  <Glyph font="F1" size="18" x="151.08" y="636.33" width="10.06">B</Glyph>
</Box>
</Word>
```

TET コネクタ

TET コネクタは、TET を他のソフトウェアと連携するのに必要な接続プログラムです。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
 - ▶ Oracle Text 用 TET コネクタ
 - ▶ MediaWiki 用 TET コネクタ
 - ▶ Microsoft 製品用 TET PDF IFilter
- 別製品として提供されます。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利用可能にします（詳しくは製品説明文書をご覧ください。）

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成例集です。ページ上のテキストに応じてしおりやリンクを追加するなど、TET と PDFlib+PDI を組み合わせて PDF 文書を改良する方法を示したサンプルもあります。

strategische Grundsätze – der
 : der Nutzung von Synergie-
 in Branchen sowie in Unter-
 lukterstellung. So verringert
 bei der Produkterstellung –
 g – seit längerem nicht nur

ハイフンは除去されますが、ダッシュは温存されます

Introduction

他製品による抽出結果：「Inttrroduccttiion」

TETによる抽出結果：「Introduction」

Canadian Institute for Theoretical Astrop
 Observatoire de Paris, LERMA, 61 avenu
 Observatoire **Midi-Pyrénées**, UMR 5572,
 Department of Astronomy, University of
 Observatorio Astronomico di Bologna, vi

他製品による抽出結果：「MIDI-PYR'EN'EES」

TETによる抽出結果：「Midi-Pyrénées」

is permanently hidden from Earth.
The first photographs of the hic
 cial satellite; modern satellites prov

他製品による抽出結果：「e rst photographs」

TETによる抽出結果：「The first photographs」

Stellen Sie sich vor, Sie stehen an einem
 Kinder ins Wasser springen und schwim
 vor, Sie graben am Sandstrand zwei klei
 Schritte landeinwärts, jeder eine Hand breit, so
 Kanäle fließen kann. Stellen Sie sich jetzt noc
 mittels eines Streichholzes und kleiner weißer

他製品による抽出結果：「S」と「tellen」の2単語

TETによる抽出結果：「Stellen」の1単語

PDF からテキストの抽出に挑戦*

ハイフンの除去

TETは、複数行にわたるハイフネーションされた単語を検出してハイフンを除去し、部分どうしを結合して単語を復元します。これは、文書内で単語がハイフンで分割されていても単語が正しく検索にかかるようにするための重要な処理です。ハイフンと異なるダッシュは除去しないよう区別して扱われます。

影付き・太字テキストの検出

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えています。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、影付きや太字の箇所のキャラクタは、文書内に複数個含まれています。TETの影付き検出アルゴリズム（特許取得済）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しますが、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、例のように文字毎に重複しているケースでは検索結果に含まれないことになります。

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、合成キャラクタを作るものがあります。たとえばキャラクタäを作るには、まず文字aをページ上に配置し、その頭に分音記号¨を配置します。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。

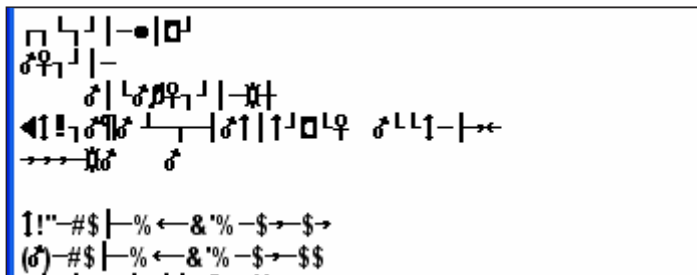
合字

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、合字を分析してキャラクタ列に分解することで、適切な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。

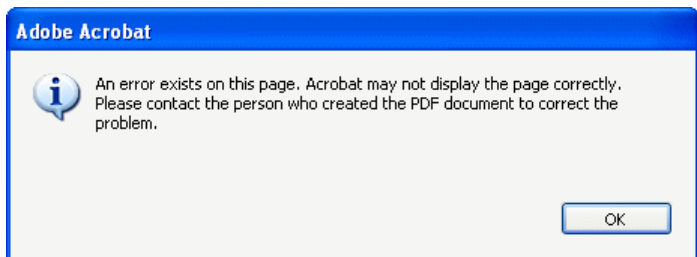
ドロップキャップ

ドロップキャップは、一番初めの段落の一字目を大きな文字で表現したものです。ドロップキャップスは段落の開始を強調したいときによく使います。ドロップキャップスを適切に扱わないと、単語の一字目とそれ以降の文字を別々の単語として抽出してしまうでしょう。

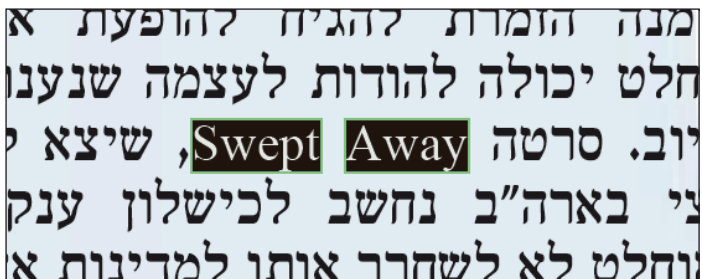
* 本機能は、主に欧文処理を対象としており、日本語処理については制限があります。



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



Acrobat でページ内容を表示できない場合でも、TET なら正確に文字を抽出します



TET は左向きのテキストと右向きのテキストが混在している場合でも、正しい順序に並び替えます



他製品による抽出結果：細切れの画像 133 個
TET による抽出結果：大きな画像 1 つ

テキストの抽出に挑戦*

Unicode マッピング

Unicode マッピングは PDF からテキストを抽出するための基礎であり、グリフごとに対応する Unicode 値を割り当てなければなりません。PDF は様々なフォントやエンコーディングをサポートしており、中には適切な Unicode を割り当てるための情報を持たないものもあるため、Unicode マッピングは複雑な作業です。最悪の場合、PDF 文書から使い物にならないテキストしか抽出できず、十分な情報が得られない場合もあります。

TET の特許技術である Unicode マッピングアルゴリズムは Unicode 値を決定するための情報すべてを使うような数珠つなぎの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抽出できない場合でも、TET は適切な Unicode テキストを抽出します。

破損した PDF

変換エラーなどによって、PDF が破損する場合があります。TET の修復モードは PDF の破損の多くを復元します。PDF の破損が激しく、Acrobat で表示できないような極端なケースにおいても、TET はページ内容を抽出します。

アラビア語やヘブライ語による双方向テキスト

PDF は論理的なテキストをエンコードせずに、単純にグリフとして内包しています。アラビア語やヘブライ語で書かれたテキストは右から左に進みます。また、数字や欧米言語の名前といった右から左に進みます。そのため文字が挿入される「bidirectional」と呼ばれる状態になることもあり、両方向に解釈しなければなりません。また他の課題として、アラビア語の文字は文脈によって最大 4 つの形をとることがあげられます。このような文字に対しては、対応する標準形（独立形）に正規化する必要があります。

画像抽出に挑戦

色空間と圧縮

PDF 内のラスタ画像データは、11 の色空間と 9 の圧縮フィルタの任意の組み合わせでエンコードされている可能性があります。JPEG や TIFF など一般的な画像ファイル形式は、それらのサブセットしかサポートしていません。TET の画像抽出は PDF 画像の特性と出力形式の機能のバランスを慎重にとります。PDF 画像の内部構造に関係なく、ピクセル画像は、共通の画像ファイル形式で抽出されます。

画像の結合

多くの PDF 文書では、それを作ったソフトウェアによって、中の画像が細かく分解されています。ページ上で 1 つの画像として見えていても、実は数百・数千の断片の寄せ集めとすることがあります。とくに、Microsoft Office アプリケーションや TeX がこうした文書を作ることで知られています。TET は、断片化した画像を検出して結合し、大きく利用可能な画像として復元します。このような画像を利用するためには結合は必須の処理といえます。

* 本機能は、主に欧文処理を対象としており、日本語処理については制限があります。

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（コンポーネント）としても、バッチ処理向けに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。TET ライブラリ及びコマンドラインツール共に、TETML（XML ベースの出力形式）を生成することができます。TET は以下のように使い分けられます。

- ▶ TET プログラミングライブラリ（コンポーネント）は、デスクトップ上やサーバー上のアプリケーションに組み込んで利用します。ライブラリの使用例は製品に添付されています。さらに多様な作成例がテックスタイルでは Web サイトの TET クックブックに掲載されています。
- ▶ TET コマンドラインツールは PDF 文書のバッチ処理に適しています。プログラミングを全く必要とせず、コマンドラインオプションによる処理を複雑なワークフローに組み込むことができます。
- ▶ XSLT などさまざまな XML 処理ツールや言語に通じた開発者であれば、TETML による出力を用いて XML ベースのワークフローを実現することができます。
- ▶ TET コネクタは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 を容易に統合できます。

TET 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
当データシートで述べてきた製品です。
- ▶ TET PDF IFilter
Windows Search、SharePoint、SQL Server などの Microsoft 社製検索製品に適した製品です。詳しくは TET PDF IFilter のデータシートをご覧ください。
- ▶ TET Plugin
Adobe Acrobat 用のプラグインで、PDF からテキストや画像の抽出を無償行えるユーティリティです。TET の性能を検証していただくことができます。

対応開発環境

PDFlib TET is everywhere. TET は事実上、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。32 ビット /64 ビットバージョンの Windows をはじめ Mac OS、Linux、Unix、さらに IBM i5/iSeries・zSeries メインフレーム版も提供しています。

パフォーマンスの最大化を図りオーバーヘッドを小さくするため TET の中核部分は高度に最適化された C コードで書かれています。平易な 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、PDFlib の機能は、次のようなさまざまな開発環境から利用することができます。

- ▶ COM（VB・ASP・Borland Delphi 等での使用）
- ▶ C・C++
- ▶ Java（サーブレット・Java Application Server を含む）
- ▶ .NET（C#・VB.NET・ASP.NET 等での使用）
- ▶ Perl
- ▶ PHP
- ▶ Python
- ▶ REALbasic
- ▶ RPG（IBM i5/iSeries）

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でご提供しています。

総合的な PDF 文書処理機能を実現

PDFlib、PLOP、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

テックスタイルでは、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ックスタイルまで。評価版のダウンロードや PDFlib 技術情報の入手もテッ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

株式会社テックスタイル

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>