改訂履歴

日付	変更内容
2005 年 9 月 30 日	改行コードの指定・出力に関わる変更をした。 Shift-JIS という誤った記述を Shift_JIS へ修正した。
2005 年 8 月 25 日	改行を出力するコードを指定可能とする。 本文アンダーラインで示す箇所を改訂。

1. プログラムの目的

固定型帳票の PDF ファイルから、ページ内の指定した位置の項目（複数）を取り出して XML ファイルにする。

1.1. 対象 PDF ファイル

下記の条件を満たすこと

① Adobe の PDFReference 記載の仕様に準拠すること。

② PDF データには文字コードが設定されていること。ファントが埋め込まれた PDF で文字コードがないと、Acrobat で表示できても、テキストを抽出できません。

③ PDF にはパスワードが掛かっていないこと。また、印刷不可、テキスト抽出不可の制限がかかっていないこと。

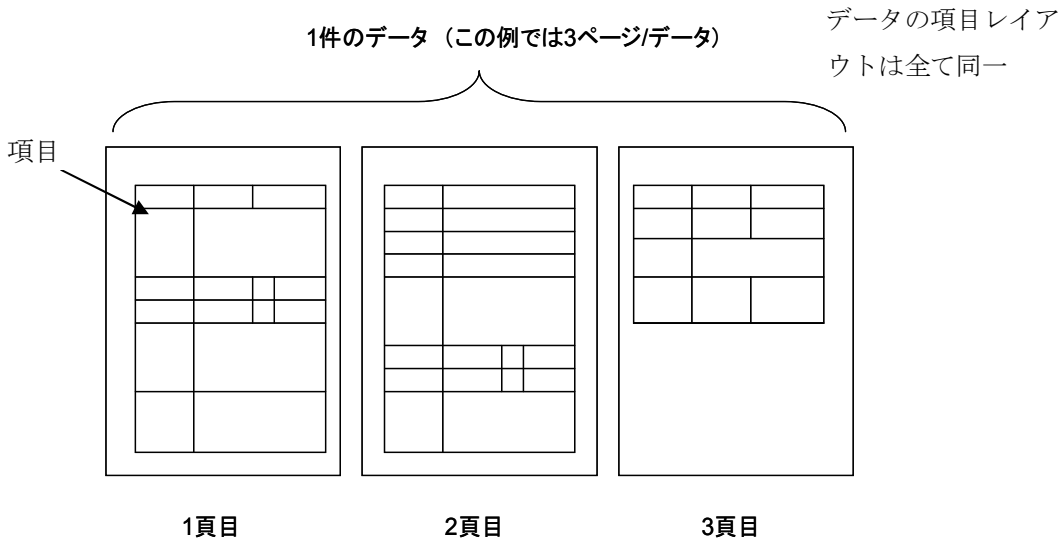
1.2. PDF の帳票レイアウト

次のような条件を満たす固定型帳票の PDF ファイルを対象とする。

① １件のデータは複数ページ（任意ページ数）からなる。

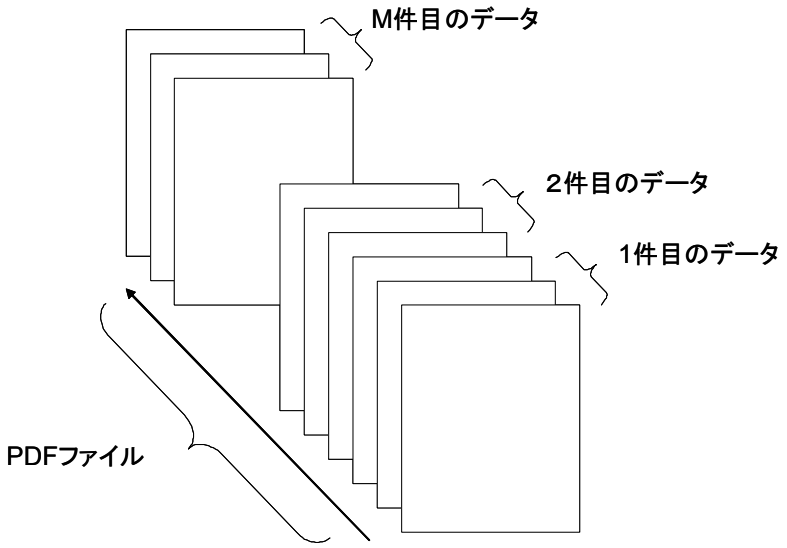
② １件のデータは任意の数の項目からなる。項目の印刷位置は PDF ファイル内のすべてのデータで同じ位置である。すなわち、項目の有効無効、あるいは、項目のテキスト文字数の長さによって項目の印刷位置は変化しない。

③ 但し、これらの条件を満たすかどうかを、入力 PDF の内容をチェックして判定することはできない。これらの条件を満たさない PDF でもテキスト抽出するが結果の妥当性を保証しない。



④ １件の PDF ファイルには任意の数のデータ（仮に M 件とする）が保存されている。ひとつの PDF ファイル内では、すべてのデータのページ数は固定（仮に N とする）である。その PDF フ

ファイルの総ページ数は 1 件のデータあたりのページ数×データ件数＝N×M となる。



1.3. 複数 PDF ファイル一括処理

①　　指定したフォルダの中の複数の PDF ファイルを一括処理することができる。

②　　一括処理する PDF ファイルの中のデータの帳票レイアウトは同一でなければならない。すなわち､1 件のデータあたりのページ数が同じで､各ページ内での項目の印刷位置が同一であること。

③　　但し、入力 PDF の中のデータが前項の条件を満たすかどうかはプログラムでチェックできないので、入力 PDF が不正な場合、出力結果の妥当性は保証されない。

1.4. 抽出対象項目の指定方法

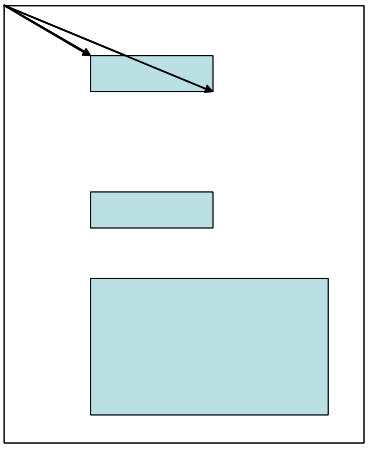
1.4.1. 抽出対象項目の条件

(1)項目の範囲

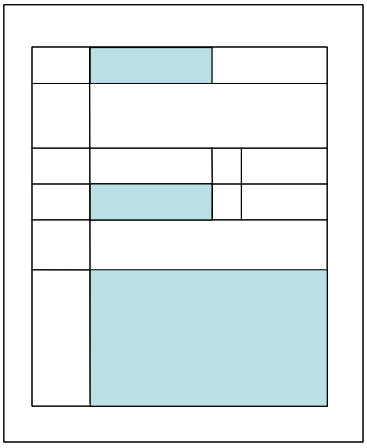
①　　1 件のデータが N ページから構成されるとして、ひとつの項目は 1 ページに入ること(1 項目は複数ページにまたがらない)。

②　　ひとつの項目の印刷範囲は、用紙の左上を原点とし、左上座標と右下座標を指定する矩形の範囲で指定できる。

③　　1 件のデータの項目の中から任意の数の項目を抽出対象として指定することができる｡但し、抽出対象項目は抽出対象 PDF の全データに対して共通に指定する。



テキストを抽出したい項目の印刷範囲(左上、右下)を用紙の左上を原点とする座標値で指定



データの1頁目

(2)抽出対象テキストの条件

① 指定した項目印刷範囲内に含まれるテキスト（文字列）を抽出する。

② 抽出対象テキストは項目印刷範囲に文字が含まれるものとする。

③ 項目内のテキストは項目印刷範囲に左上から横書きで記入されている。

④ ひとつの項目印刷範囲内のテキストは 1 行または複数行から構成される。

⑤ 項目印刷範囲左端から行の先頭の文字までの間の間隔（空白）は無視する。

⑥ 行の最後の文字から印刷範囲右端までの間隔（空白）は無視する。

⑦ 項目印刷範囲の中の 2 行目以降をテキストに出力する際は、先頭文字の前に、出力設定ファイルで指定したコードを使って、改行を示す情報を出力する。

⑧ 前項の改行を示す情報は、その改行を表す文字コードで出力する。

項目印刷範囲の指定が正しく、かつ、PDF データが正しく出力されていれば項目印刷範囲をまたがるテキストは存在しないはずである。しかし、PDF ファイルにテキストをどのように出力するかは、PDF ファイルを作成したアプリケーション次第なので、そのような暗黙の前提はなりたたないかもしれない。例えば、空白を使って位置ぞろえして、次のように出力されている場合もあるだろう。