

# AHPDFXML 変換ライブラリ V2.0 の紹介

PDFデータをXML形式に変換して再利用

アンテナハウス株式会社

2017/7/7



# AHPDFXML 変換ライブラリ V2.0 の紹介

- \* AHPDFXML変換ライブラリの特徴
- \* AHPDFXML変換ライブラリの用途
- \* 出力フォルダとファイル構成
- \* AHPDFXML形式
- \* AHPDFXML 変換ライブラリの構成
- \* 販売価格

# PDFデータの再利用

PDFデータからコンテンツを取得したい

- \* PDFデータから...

- すべての文字を取得したい  
全文検索や部分検索で利用
- 決められた領域の中の文字を取得したい  
定型フォーマットのPDFから  
領域の中の文字だけ収集
- 添付画像を取得したい

- \* このような要求に答えるには？

# AHPDFXML 変換ライブラリ

PDFデータを解析してXML形式に変換するプログラム

- \* PDFデータを解析して文書構造を生成XML化
  - このXMLがAHPDFXML形式
- \* XML化によりデータの利用が安易
  - XSLTスタイルシートを定義してfoやhtmlへ変換
- \* アンテナハウス独自プログラム
  - AcrobatやMicrosoftOfficeが不要

# PDFデータの独自解析エンジン

セクション、行、表など、意味のある要素にまとめXMLへ出力

## \* PDFデータでは...

- 文字データは見た目順に並んでいない
- 行や段組みと言う概念がない
- 表と言う概念がない

## \* 独自解析エンジンの役割

- 文字は紙上の位置に応じて  
セクションや行としてXMLに出力
- 線分で囲まれた領域は  
表(行、列、セル)としてXMLに出力

# AHPDFXML 変換ライブラリの用途

- \* PDFから文書構造を生成
  - 要素を利用することにより、DocBookなど文書構造を記述するデータに加工
  - XSLTスタイルシートを定義すれば、用途に応じた加工に柔軟に対応
- \* PDFの位置情報を利用することで、任意の範囲に含まれるテキストを抽出
- \* 表のデータを業務用のデータベースに取り込む
  - CSV形式でデスクトップ処理も可能

# AHPDFXML 変換ライブラリ

## 出力フォルダとファイル構成(例)

- \* sample.xml

- 本文情報

- \* sample\_style.xml

- 文字のフォント情報ID ahp:s-id=“s3”、セル、表、行の情報の参照元

- \* sample\_catalog.xml

- 画像ファイルのID ahp:file-id=“f1”の参照元

- \* 各種画像ファイル

# AHPDFXML形式

PDFデータをXMLで表現したもの

## \* XMLで表現する要素

- ページ(サイズ、マージン)
- セクション(用紙上の位置、マージン、スタイル)
- パラグラフ(用紙上の位置、スタイル)
- 行(用紙上の位置)
- 文字(用紙上の位置、文字情報、フォント情報)
- 表(用紙上の位置、行数、列数、セル情報)
- 画像(用紙上の位置、外部ファイルとして出力)



# AHPDFXML 出力例(ページ)

## 【 ahp:page 】

```
<ahp:page ahp:width="595.200012"  
ahp:height="841.919983" ahp:margin-l="45.120003"  
ahp:margin-r="51.600037" ahp:margin-  
t="94.000000" ahp:margin-b="52.799988"
```

ページの幅 ahp:width, 高さ ahp:height

マージン ahp:margin-l,r,t,b

# AHPDFXML 出力例(セクション)

## 【 ahp:section 】

```
<ahp:section ahp:l="72.000000" ahp:r="523.200012"  
ahp:t="0.100000" ahp:b="747.919983" ahp:margin-  
l="50.000000" ahp:margin-r="50.000000" ahp:margin-  
t="50.000000" ahp:margin-b="50.000000" ahp:writing-  
mode="horizontal" ahp:section-column-count="1"
```

用紙上のセクションの座標 ahp:l,r,t,b

セクションのマージン ahp:margin-l,r,t,b

文字の流れ ahp:writing-mode

段数 ahp:section-column-count

# AHPDFXML 出力例(パラグラフ)

【 ahp:p 】

```
<ahp:p ahp:l="81.599998" ahp:r="335.958252"  
ahp:t="353.280060" ahp:b="371.040070"
```

用紙上のパラグラフの座標 ahp:l,r,t,b  
パラグラフの子要素として行を出力

# AHPDFXML 出力例(行)

【 ahp:line 】

```
<ahp:line ahp:l="81.599998"  
ahp:r="335.958252" ahp:t="353.280060"  
ahp:b="371.040070"
```

用紙上の行の座標 ahp:l,r,t,b  
行の子要素として文字を出力

# AHPDFXML 出力例(文字)

## 【 ahp:run 】

```
<ahp:run ahp:l="305.579987" ahp:r="341.698792"  
ahp:t="165.419128" ahp:b="183.419128"  
ahp:page-no="1" ahp:ope-no="27"  
ahp:s-id="s3" ahp:z-order="1">材料</ahp:run>
```

用紙上の文字列の座標 ahp:l,r,t,b

フォント情報はスタイルXMLのスタイルID ahp:s-id="s3"  
を参照。文字列は“材料”

# AHPDFXML 出力例(表)

## 【 ahp:table 】

```
<ahp:table ahp:l="307.799988" ahp:r="543.599976"  
ahp:t="555.560059" ahp:b="593.400024" ahp:table-  
column-count="3" ahp:table-row-count="1"
```

用紙上の表の座標 ahp:l,r,t,b

表の列数 ahp:table-column-count

表の行数 ahp:table-row-count

表の子要素として表の行を出力

# AHPDFXML 出力例(表の行と列)

## 【 ahp:row 】

```
<ahp:row ahp:l="307.799988" ahp:r="543.599976"  
ahp:t="555.560059" ahp:b="593.400024">
```

```
<ahp:cell ahp:l="307.799988" ahp:r="329.121979"  
ahp:t="555.560059" ahp:b="593.400024" ahp:s-id="s37" />
```

```
<ahp:cell ahp:l="329.121979" ahp:r="389.279999"  
ahp:t="555.560059" ahp:b="593.400024" ahp:s-id="s38"
```

用紙上の表の行の座標 ahp:row の ahp:l,r,t,b  
表の行上の列(セル)の座標 ahp:cell の ahp:l,r,t,b  
列(セル)の子要素として文字を出力

# AHPDFXML 出力例(画像)

## 【 ahp:frame 】

```
<ahp:frame ahp:l="90.300003" ahp:r="290.479553"  
ahp:t="139.619690" ahp:b="289.619690"  
ahp:frame-type="image" ahp:z-order="0"  
ahp:page-no="1" ahp:ope-no="4" ahp:file-id="f1"
```

用紙上の画像の座標 ahp:l,r,t,b

画像ファイルは、カタログXMLのファイルID ahp:file-id="f1"を参照



# AHPDFXML 出力例(フォント情報)

## 【 ahp:font 】

```
<ahp:font ahp:name="MS-PMincho"  
ahp:encoding="Identity-H" ahp:size="10.500000"  
ahp:pitch="variable">  
<ahp:color ahp:rgb="#000000"/></ahp:font>
```

フォントファミリー ahp:font ahp:name

エンコーディング ahp:encoding

サイズ ahp:size

ピッチ ahp:pitch

フォントの色 ahp:color の ahp:rgb

# AHPDFXML 変換ライブラリの構成

- \* AHPDFXML仕様書：AHPDFXMLの仕様について説明したHTML形式のドキュメント。
- \* コマンドラインインターフェース 仕様書
- \* C++インターフェース仕様書：本ライブラリをアプリケーションから呼び出すためのインターフェースについて説明
- \* 実行形式：ライブラリ本体のバイナリモジュール一式

# AHPDFXML 変換ライブラリのデモ

XSLTスタイルシートを定義してfoやhtmlなどへ変換

## \* html へ変換

- Sample.pdf を AHPDFXML形式へ変換
- AHPDFXML形式をXSLTを使いhtmlへ変換
- 変換した html をブラウザで表示

# AHPDFXML 変換ライブラリ

## \* 販売価格

- サーバライセンス: ¥600,000 (税別)
- スタンドアロンライセンス: ¥120,000 (税別)
- 開発ライセンス: ¥400,000 (税別)

## \* 製品Webサイト

- 製品トップページ  
<http://www.antenna.co.jp/pdfxml/index.html>
- 評価版のお申込み  
<http://www.antenna.co.jp/pdfxml/eval.html>