PDFの基本技術知識

2008年6月
アンテナハウス株式会社
http://www.antenna.co.jp
小林　徳滋
koba@antenna.co.jp

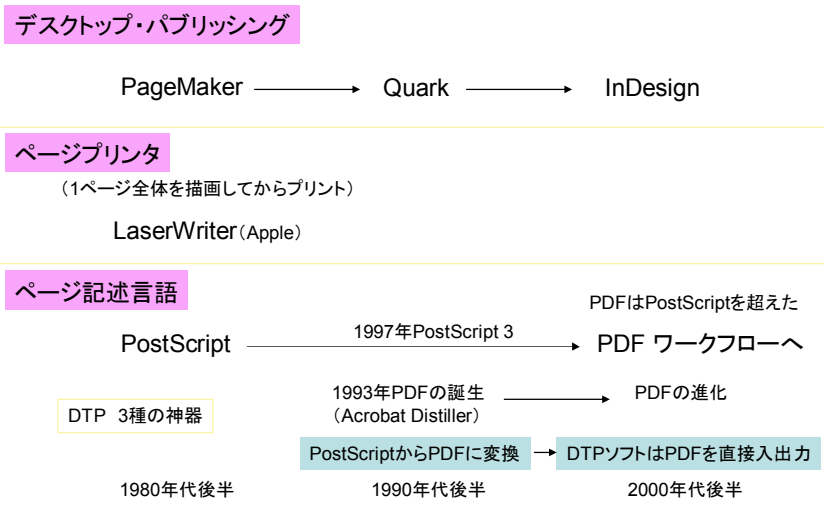
PDFの仕様

- PDF Reference (アドビが発行)最新は1.7
- ISO 32000-1
  - PDF Reference 1.7ベース、ISO標準化承認済
  - 現在、印刷のための準備中
- ISO PDF/A：PDF長期保存するための仕様
  - PDF/Aに準拠したPDFファイル生成は難しい
- ISO PDF/X：PDFを印刷用途のために情報交換するための仕様

PDFの作り方

- PDFの作り方には2通りある
  - デジタル生まれのPDF（こちらが起源）
    - アプリケーション・ソフトからPDFを出力
  - 紙から作成するPDF（PDFの本質ではない）
    - 紙をスキャナで読み、PDFに変換
- 同じPDFでも特性が根本的に異なる
  - デジタル生まれのPDFはベクトル・データ（中心）
  - 紙から生まれたPDFはイメージ

起源はデジタル印刷技術



デジタル生まれのPDF

原理：紙に出力するのと完全に同じ内容をPDFに出力して受け渡す

プリンタがコンピュータに繋がっているとき

印刷会社やサービス会社などの印刷機を使うとき

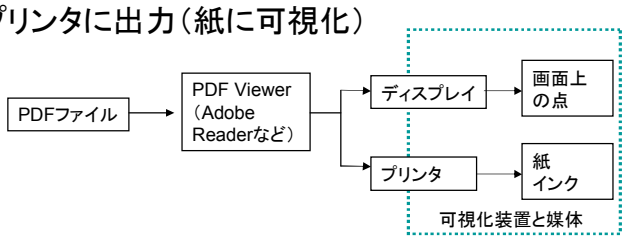
印刷会社・印刷装置

紙から作ったPDF

- 紙のPDF化は印刷起源のPDFとは別のもの
  - 拡大・縮小に弱い
- 紙をスキャナーで読み取り画像化
  ⇒その画像をPDFに埋め込んだもの
- PDFにすることで、複数ページの管理、メタデータをつけての管理が簡単になる
- 透明テキスト付きPDF
  - OCRで認識したテキストを画像の上に配置し、テキスト検索可能にしたもの

PDFの可視化

- PDFは画像（イメージ）の一種ではありません。
  - オブジェクト辞書とデータを記述したファイル
- PDF ReaderでPDFファイルを読み
  - オブジェクト辞書とデータを解読して可視化する
    - 画面
    - プリンタに出力（紙に可視化）



PDFの可視化結果保証

- アプリケーションからPDF出力して作成し、そのPDFを画面に表示・印刷したもの（A)と元のアプリケーションで画面に表示・印刷したもの（B)が、同等であるかどうか？
- AとBの同等性を常に保証できるか？
- 現実には、トラブルが起きることがある。
- この問題が起きない対策のために、PDF/A、PDF/Xが規定されていると言っても良い。

PDFファイル構造

作成直後のPDF

ヘッダ　例）*%PDF-1.5*
本体
相互参照表
トレイラ

PDFであることを識別するための情報

PDFの本体情報

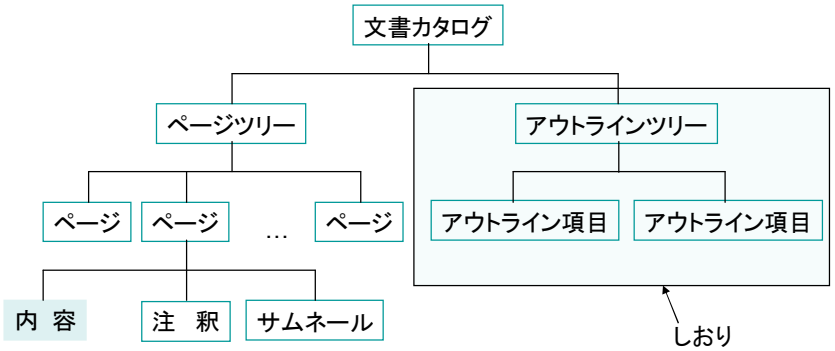
PDFの本体にランダムアクセスするための情報

PDFファイルは最後にファイルサイズ、カタログ情報、暗号辞書などが登録されている

標準ではトレイラが最後にあるため、Adobe Readerなどの利用アプリケーションは通常、PDFファイルの一番後ろから読まねばならない。

このため、ファイルの容量が大きい（ページ数の多い）PDFをWeb経由で表示しようとすると、全部ダウンロードするまで、画面には、内容がまったく表示できません。
【オプション】Web表示用に最適化（リニアライズ）で、ランダムアクセス用の情報を先頭に複製

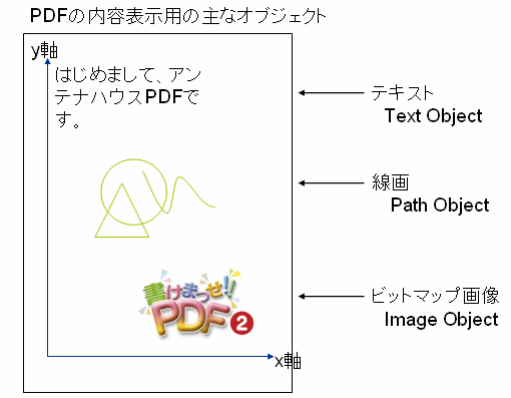
PDF本体の構造（抜粋）



①PDFの内容は頁単位になっている（ワープロ文書（例：Word）などとは異質）
②しおりの情報は、アウトラインツリーとして別管理
③注釈は1頁毎に管理されていて、かつ、頁の内容とは別管理

PDFのページ内容

- PDFには1頁毎にページの内容を描画するための情報が保存されています。

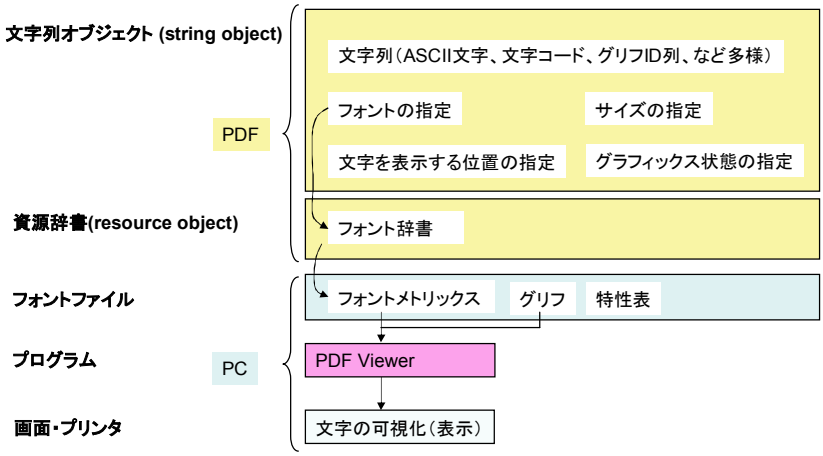


線画オブジェクトは2次元座標系の上に数学的な直線・曲線（パス）として表現されることができます。そうしたパスに線幅指定、色指定したり、パスで囲む領域を塗り潰したりすることで、図形が表現されます。

文字はビットマップとしてドットの塗り潰しで表すか、あるいは、文字の輪郭（アウトライン）を曲線で表して囲まれた部分を塗り潰すアウトラインフォントの方式で表現します。

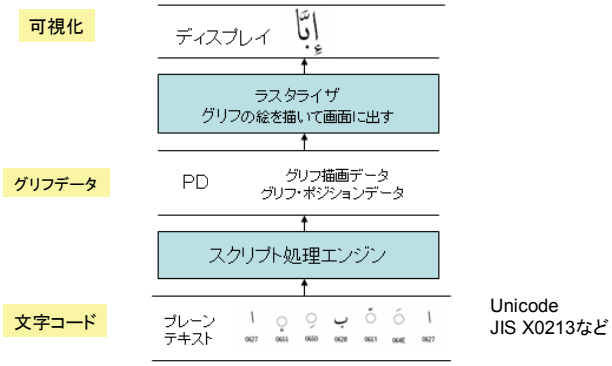
PDFファイルは、オブジェクトを規定するデジタルデータの塊

PDFでの文字表示



文字コードの可視化

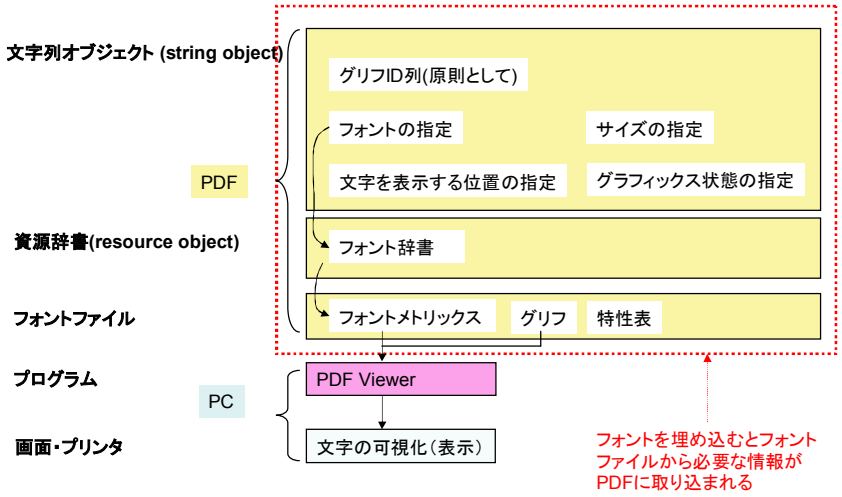
- 文字コードの並びは、フォントのグリフデータを用いて可視化される。



フォント埋め込み

- フォントを埋め込まないPDFの表示
  - 受信相手のシステム上のフォントを使う。
  - Adobe Readerは必ずしもそうなっていない。
  - 特に海外へ送るときは注意が必要。
    - 例）日本語の文字が少しでも入っていると英語のAdobe Readerではまったく表示できない。Windows\Fontsにフォントがあってもだめ。
- フォントを埋め込んだPDF
  - PDFにフォントのサブセットが添付される。全文字ではなく使っている文字だけが原則。

フォントを埋め込んだPDF



イメージの扱い

- PDF Ref 1.7 Section 4.8 Images で定義
  - image Xobject（4.8.4）と inline-image (4.8.6)がある
- 4.8.4 Image Dictionaries
  - イメージの基本的属性を辞書で定義
    - 縦横のピクセル数
    - ピクセル単位のビット数
    - カラースペース
    - 圧縮方法(フィルタ)
    - 透過色

PDF Reference 1.7 pp343-344より引用

例:
幅256×高さ256、8ビットのイメージ
カラー空間:DeviceRGB
ページの左下(45,240)の位置に配置され
ユーザ空間の単位132の幅と高さに
スケーリングされて配置される

```
Example 4.28
20 0 obj % Page object
<< /Type /Page
/Parent 1 0 R
/Resources 21 0 R
/MediaBox [ 0 0 612 792 ]
/Contents 23 0 R
>>
endobj
21 0 obj% Resource dictionary for page
<< /ProcSet [ /PDF /ImageB ]
/XObject << /Im1 22 0 R >>
>>
endobj
```

右上へ続く

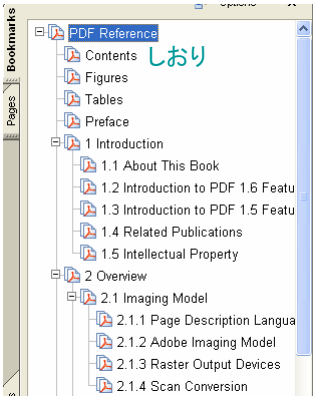
```
22 0 obj% Image XObject
<< /Type /XObject
/Subtype /Image
/Width 256
/Height 256
/ColorSpace /DeviceGray
/BitsPerComponent 8
/Length 83183
/Filter /ASCII85Decode
>>
stream
9LhZI9h\GY9i+bb;,p:e;G9SP92/)X9MJ>^:f14d;,U(X8P
;cO;G9e];c$=k9Mn\]
… Image data representing 65,536 samples …
8P;cO;G9e];c$=k9Mn\]~>
endstream
Endobj

23 0 obj% Contents of page
<< /Length 56 >>
stream
Q % Save graphics state
132 0 0 132 45 140 cm % Translate to (45,140)
and scale by 132
/Im1 Do % Paint image
Q % Restore graphics state
endstream
endobj
```

イメージの実体データ

PDFのしおり

- しおり(Bookmark、アウトライン項目)
- しおりを階層化したツリーがアウトライン
  - 文書構造を表示する目次になる
- ドキュメント・カタログにて本体のページとは別に管理される



アウトライン

注釈（コメント）など

- 注釈（コメント注釈など）は付加情報で本文と別に管理されています。
- 注釈の種類
  - Text annotation
  - Link annotation
  - Free text annotation
  - Line annotation
  - Widget annotation （フィールドの概観）
  - . . .

対話フォーム

- ユーザ対話データを保管
  - 電子署名は対話フォーム・データの一種
- 対話フォーム辞書
  - フィールド辞書を規定
  - フィールド辞書のタイプは次の4つ
    - ボタン・フィールド
    - テキスト・フィールド
    - 選択肢フィールド
    - 署名フィールド ⟶ 電子署名のデータを保管

PDFのセキュリティ

- 方式
  - パスワード・セキュリティ
    - 閲覧制限パスワード（ユーザパスワード）
    - 編集制限パスワード（オーナーパスワード）
  - 公開鍵暗号方式セキュリティ
    - 公開鍵を使って暗号化。秘密鍵保有者のみ閲覧可。
- アルゴリズム
  - RC4：40ビット、128ビット
  - AES暗号
- セキュリティハンドラに実装する

PDFと電子署名

- 署名フィールドを使う。
  - 未署名の署名フィールド
  - 署名済みの署名フィールド
- 署名済みPDFの署名データ（ハッシュ値）は署名フィールドの中の署名辞書に保管される。
- 署名の概観
  - 署名には、Widget注釈機能を使って、外観を添えることができる。概観のない署名（不可視署名）も可能。
- 署名を検証した結果も、外観に表示できる。
- 署名後のPDFは増分更新する。
  - 署名を次々に追加していくことも可能。