Inside PDF

Lecture @21C3

The Portable Document Format

A Short Introduction

Maik Musall <maik@musall.de>
CCC Erlangen

Overview

- History of PDF and it's relation to PostScript
- Licenses and legal issues
- File format syntax and semantics
- Display model
- Images and vector graphics
- Text and Font management
- Encryption and compression
- Overview of Tools and libraries

History of PDF versions

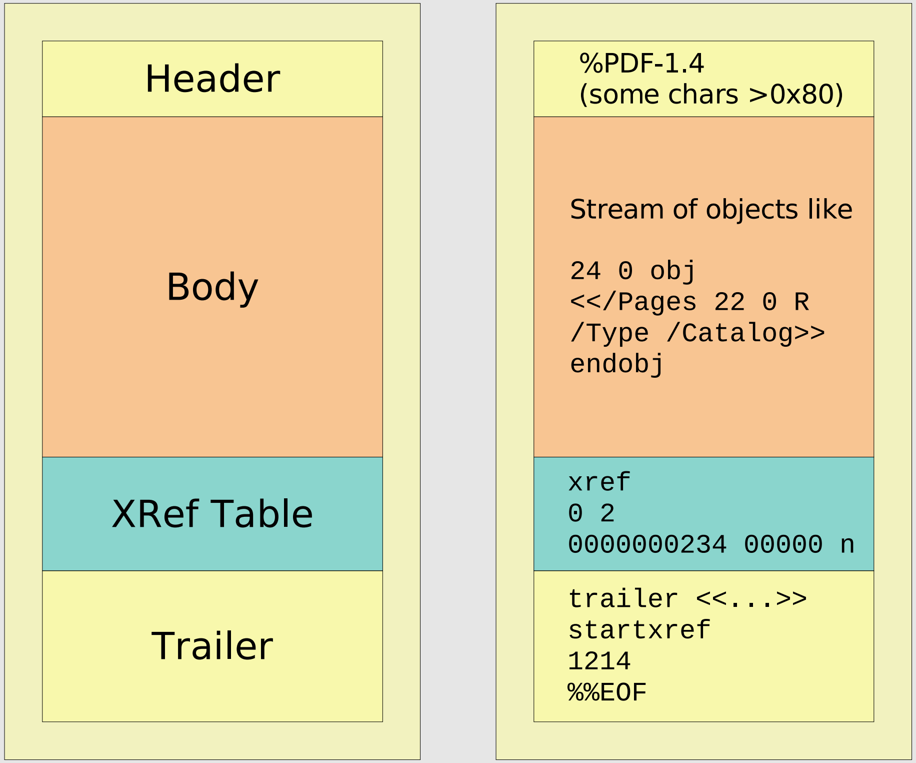
• The past:	PDF 1.0 (1993) to 1.4
• The present:	PDF 1.5, a contribution to storage and bandwidth
• The near future:	PDF 1.6, the 3D bloat
• The prepress world:	PDF/X (ISO standards)
• The archiver's vision:	PDF/A (upcoming ISO)

PDF and PostScript

- PS is a programming language (special domain, but turing-complete). PDF is not.
- PDF is just a data structure and provides random access to all contained objects.
- PDF supports interactive features (forms, annotations, JavaScript, open actions etc.)
- PDF shares the PS imaging model.
- Both will produce the same output when printed.
- Both have similar licenses that include permission for free use, but prohibit cloning the format.
- Lots of PDF features cannot be represented in PS.

PDF Syntax: General File Structure

- A file is read starting at the end.
- Incremental Updates may be appended at the end, leading to several body, xref and trailer sections.
- A PDF can be written all ASCII, if needed.
- Single-Pass File Generation is possible.



```
%PDF-1.4
(some chars >0x80)
```

Stream of objects like

```
24 0 obj
<</Pages 22 0 R
/Type /Catalog>>
endobj
```

```
xref
0 2
0000000234 00000 n
```

```
trailer <<...>>
startxref
1214
%%EOF
```

PDF Syntax: Object types

• Bool	`true     false`
• Numbers	`0  1  5.4  -.002`
• Strings	`(Hello World)   <4D617465>`
• Names	`/Type  /Pages`
• Arrays	`[ obj obj obj ]`
• Dictionaries	`<</Key1 val1 /Key2 val2>>`
• Streams	`<<...>> stream...endstream`
• The null Object	`null`
• Indirect Objects	`665 0 R`

- EOL is flexible (CR „Mac“, LF „Unix“, CRLF „DOS“)
- Filters may be used to encode streams.
- PDF 1.5 introduces object streams.

PDF Encryption

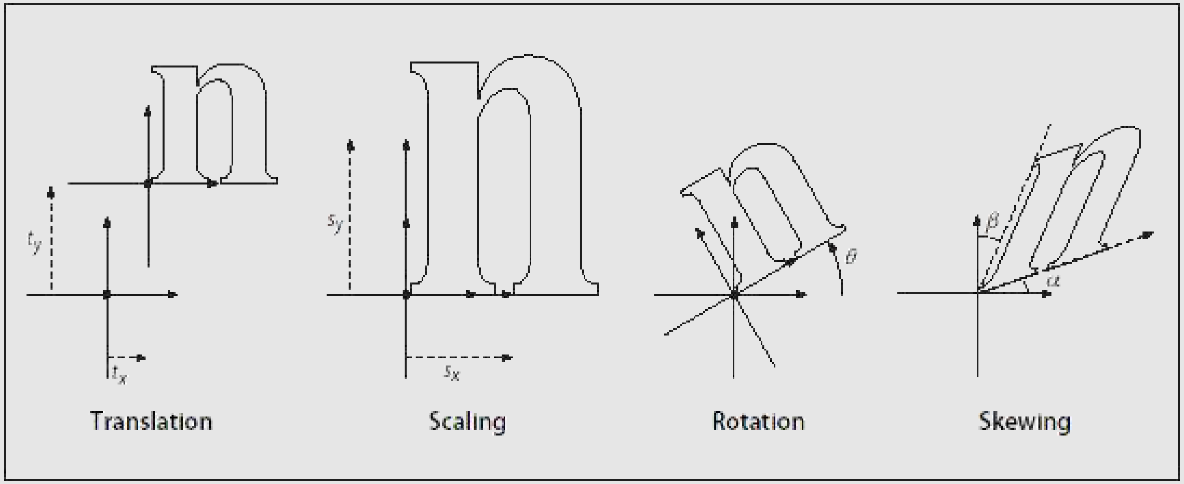
- All strings and streams go through the cipher (more selectively since PDF 1.5)
- up to PDF 1.3: RC4, 40 Bit
- since PDF 1.4: RC4, up to 128 Bit
- since PDF 1.4: unpublished algo (U.S. export law, no longer in use)
- since PDF 1.6: AES
- PDF 1.3 (spec 1.5): PKCS#7 (RFC 2315)
- PDF spec requires implementators to honor document access restriction settings.

PDF Document Structure

- Everything starts at a root object (`/Type /Catalog`)
- Pages are organized in a tree of objects
- Trees are also used for Names, Outlines (a.k.a. bookmarks), Logical structure, ...
- Tree nodes can contain data that is inherited to their child nodes (e.g. physical page dimensions).
- A most basic PDF document will contain:
  * Header, XRef table, trailer
  * `/Catalog`, `/Pages`, one `/Page`, page content stream
- Content streams use a language resembling PostScript

PDF Display model (1)

- Model is identical to that of PostScript, existing implementations can be reused (so does ghostscript)
- Three types of content are common: Text, Bitmap images, and vector graphics
- A set of coordinate systems is used to transform between user space and devices with different resolutions.
- Relations of coordinate systems are described using transformation matrices



PDF Display model (2)

- A Graphics State stack machine is used to manage changes in CTM, color, overprint, clipping, line patterns, transparency (PDF 1.4) etc.
- Color spaces can be RGB, CMYK, Gray, ICC, Indexed, and a few others, grouped in Device-, CIE- and special color space groups
- Each content object can be reused several times within the document.

PDF Text

- Text state knows: char spacing, word spacing, horizontal scaling, leading, font name, font size, rendering mode, rise and knockout.
- Font types are: Type0 (composite), Type1 (PS font program), Type3 (arbitrary graphics operators), TrueType, CIDFonts.
- Choice of embedding levels: name only, glyphs, complete font program
- The 14 PS standard fonts (Helvetica, Courier, Times in different styles, Symbol, ZapfDingbats) are considered built-in and required by every PDF processor to provide on it's own.
- Each text object can have a different encoding.

PDF Bitmap Images

- Bitmap Images are stored in stream objects
- Each one can have it's own resolution, dimension, depth, color space, compression.
- Depth: 1, 2, 4, 8 or (PDF 1.5) 16 bits per component
- All filters can be applied as for every stream: ASCIIHex, ASCII85, LZW, Flate, RLE, CCITT, JBIG2 (PDF 1.4), DCT (Jpeg), JPX (Jpeg2000, PDF 1.5), Crypt (PDF 1.5).
- An image may be present in several representations, e.g. a low-resolution image for fast screen viewing and a very-high-resolution image for printing.

PDF Vector Graphics

- Arbitrary „Paths“ can be painted using Bézier curves.
- Paths can overlap, using transparency features (since PDF 1.4).
- One Path can function as a clip/crop mask for another one.
- Paths can create fill patterns.
- Paths can even define Type3 font glyphs.

- Other object (annotation) types include Sounds, Movies, and 3D objects (PDF 1.6).

PDF Metadata

- Since PDF 1.4, a document may include metadata in XML format
- The XML semantics use the XMP (Extensible Metadata Platform) technology.
- XMP is a RDF application
- „XMP is an important piece that brings the Semantic Web closer to realization.“ *(Eric Miller, W3C Semantic Web Activity Lead)*

Tools and Libraries

- **Adobe PDF Library** (datalogics.com): can do everything, but quite expensive
- C/C++: **pdflib** (pdflib.com), free and commercial variants available. Good for creation, processing limited to copying whole pages.
- Java: **iText** (lowagie.com): very promising, still some flaws with PDF 1.5/1.4 hybrid updates a few months ago, but quickly developing, gcj compatible
- **Apache FOP** (xml.apache.org/fop), an XSL-FO implementation transforming XML to PDF
- Nothing fits all purposes, most tools have a special domain (creation, conversion, split/concat etc.)

PDF Information Resources

- Adobe Specification and Resources:
  http://partners.adobe.com/asn/techresources.jsp
- PDF/X: http://www.pdfx.info
- Forums: http://www.planetpdf.com
- Tools: http://www.pdf-tools.com
- Portal: http://www.pdfzone.com
- Usenet: news://comp.text.pdf

Thanks for listening!

- Contact:
  <maik@musall.de>
  Congress DECT: M-A-I-K (6-2-4-5)

THAT CONCLUDES MY TWO-HOUR PRESENTATION. ANY QUESTIONS?
scottadams@aol.com
www.dilbert.com
DID YOU INTEND THE PRESENTATION TO BE INCOMPREHENSIBLE, OR DO YOU HAVE SOME SORT OF RARE "POWER-POINT" DISABILITY?
© 2003 United Feature Syndicate, Inc.
8/9/03
ARE THERE ANY QUESTIONS ABOUT THE CONTENT?
THERE WAS CONTENT?

© 2003 United Feature Syndicate, Inc.