How PDF Works

Gary Staas
ByteSizeBooks.com
www.pdfdream.com

Outline

How PDF represents a document

File structure—4 parts

Document structure—elements that build a document

Painting pages—portray any document

Dictionaries—flexible data structure used throughout PDF

Acrobat Forms—built from dictionaries and annotations

PDF Specification

PDF 1.4

- *PDF Reference, third edition*
  Adobe Portable Document Format
  Version 1.4
- http://partners.adobe.com/asn/developer/acrosdk/docs/filefmtspecs/PDFReference.pdf

Structure levels

File structure

- Four parts, one contains most data

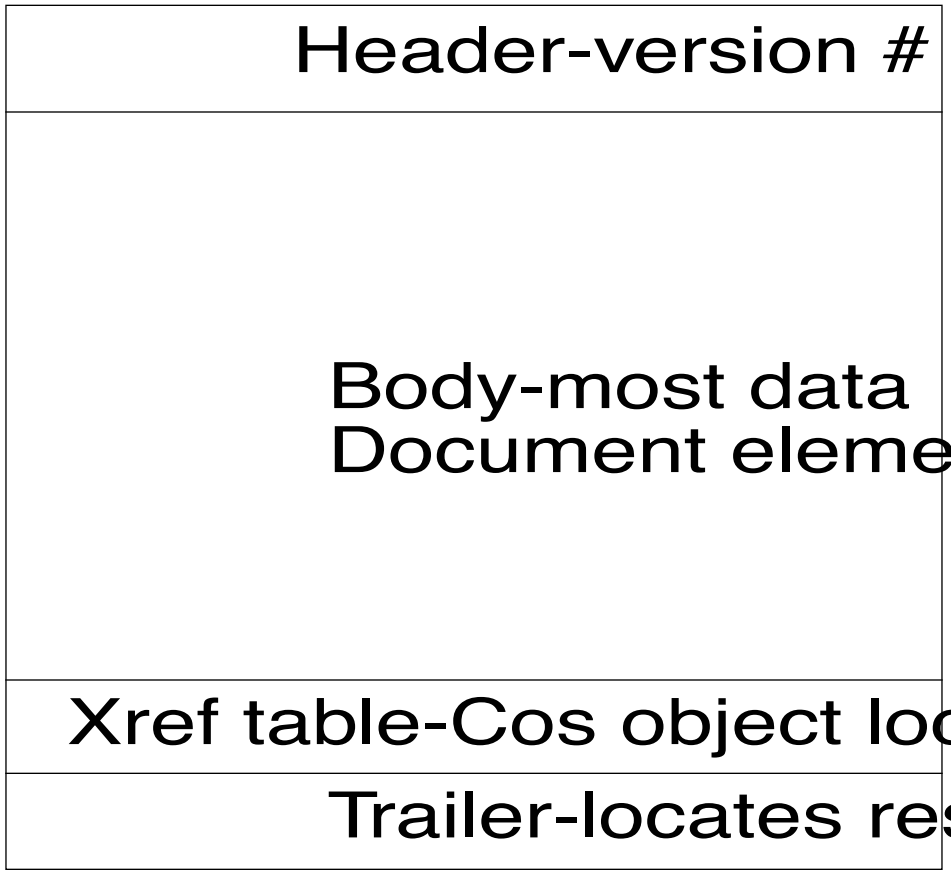
Document structure

- How document elements represented
  - Pages, annotations, bookmarks…

Cos objects

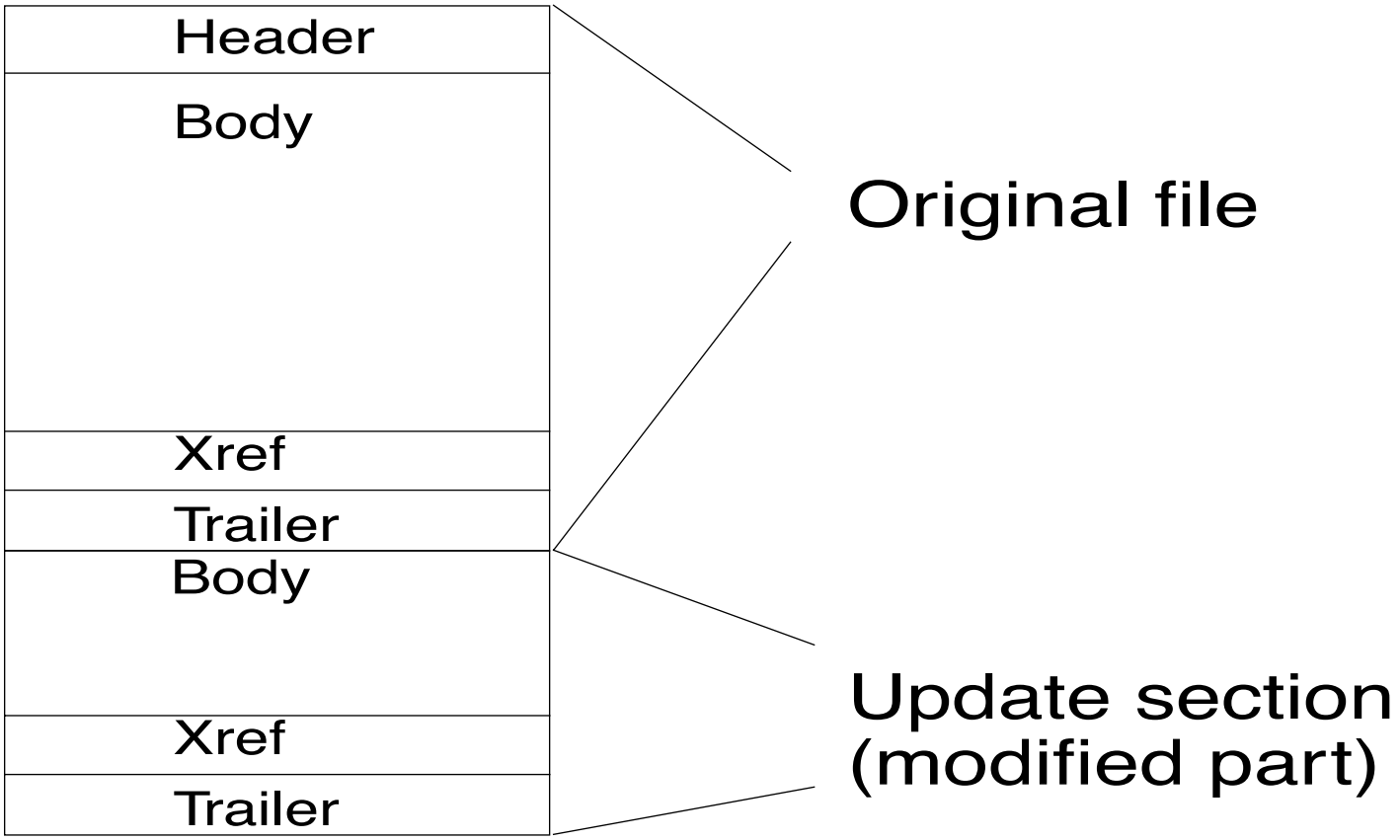
- Building blocks of document elements
- Same as PostScript objects
- Objects may be used repeatedly
- Dictionary—important object type

File structure



May have 1000 bytes before Header

Incremental update



Incremental update

One update section added for *each* save

File gets *bigger* after each save

Can revert to previous versions by chopping off update sections

- Digital signatures uses this feature

“Save As...” compacts file

- Combines update sections and original file into one Body, Xref, and Trailer

Observing File structure

Use text editor

- Wordpad, MS Word, BBEdit…
- Don’t change line endings!

Modify file

- Put data before header

Document structure

Document comprised of various elements

- Catalog/root
- Pages
- Annotations
- Bookmarks

Each document element has its format defined in *PDF Reference*

Example: Info dictionary

```
<<
/CreationDate (D:20010329220824Z)
/ModDate (D:20011006152637-07'00')
/Producer (Acrobat Distiller 5.0)
/Title (Acrobat SDK Release Notes)
/Creator (FrameMaker 5.5.6p145)
/Author (Adobe Developer Support)
>>
```

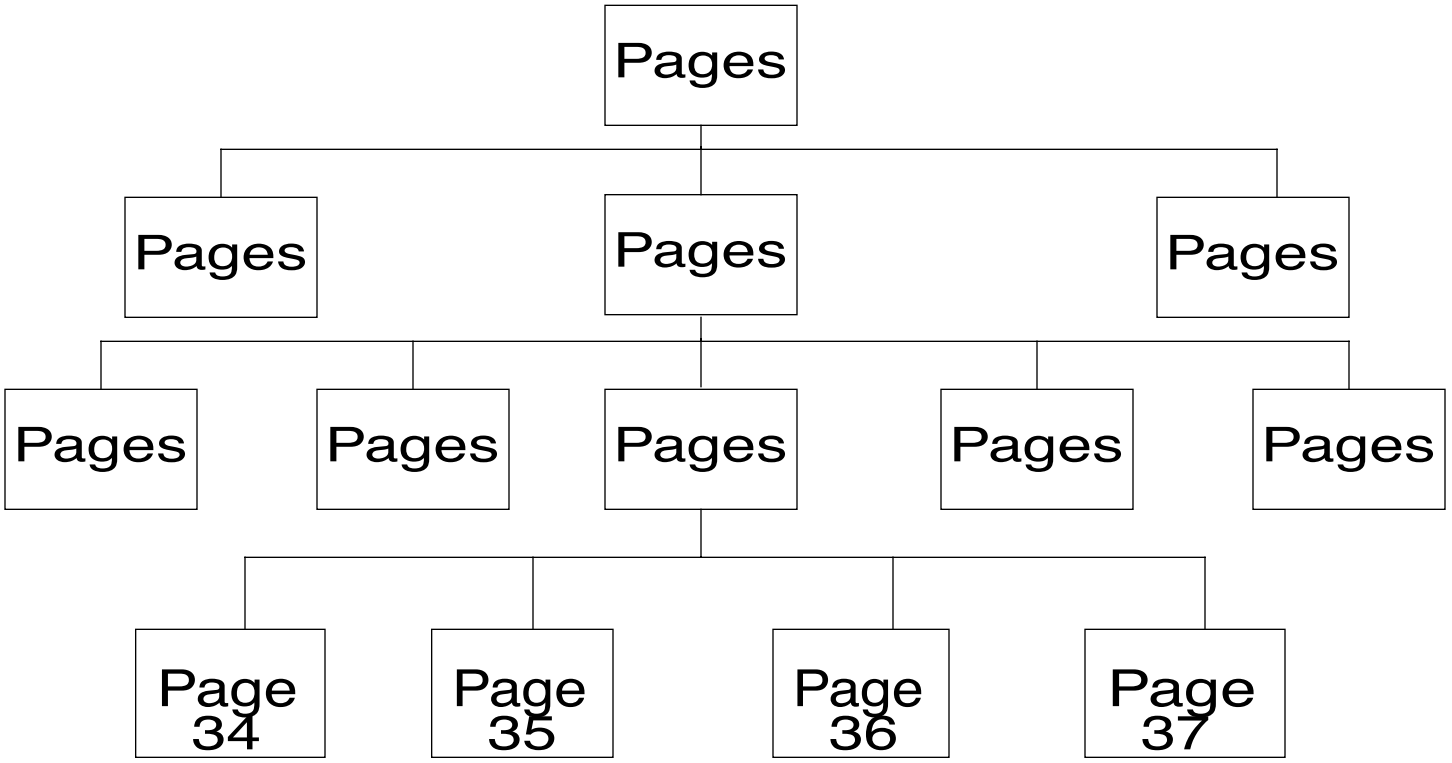
Document structure

Trailer points to 2 things—

- Info dictionary
- Catalog (root) contains document objects
  - Page tree—all doc pages—typically most of data
  - Viewer preferences—show/hide toolbar...
  - Page labels
  - Form information
  - Bookmark (outline) tree
  - Everything else!

Page tree

Allows random page access



Page

Each page *fully self contained* for *page independence*

Refer to everything needed for page in *one* place

- Know where to find page data
  - Cos objects—byte displacements in file from cross reference table

Page components

Contents—page description—*visible* part

Resources

- Info needed to render page, e.g., fonts

Thumbnail

- Bitmap of page

Crop box

Annotations

- Additional data on page
- Standard-built-in annots
- Custom annots, e.g., Acrobat Form fields

Cos objects

Document elements are built from Cos objects

Types

- Number—integer and real
- String—text
- Array—list
- Dictionary—key-value pairs
- Stream—contains data stream

Cos objects—Dictionary

Unordered set of *key-value pairs*

Database

Inherently extensible

Can represent many data structures

Many things in PDF file are dictionaries

- Info dictionary
- Pages
- Annotations

Much of *PDF Ref* is dictionary definitions

Undefined entries ignored by Acrobat

- Can easily add custom data

Cos objects—Dictionary

Example: Text annotation

```
<<
/Type /Annot
/Subtype /Text
/Rect [266 116 430 204]
/Contents (Data of text annot)
>>
```

Cos objects—Stream

Two parts:

- dictionary describing stream, e.g., length
- data, which is usually compressed

Streams in PDF file

- Page contents
- Images

Examining Document structure

Annotation dictionary specification
PDF Ref, Section 8.4 *Annotations*

- Provides names of keys

Modify annotations

- Flags with annotation attributes
  - invisible - 32
- Add custom data

Acrobat can repair file—redundancy

Acrobat can't repair everything!

Can't do this with Touchup tool

Page contents — Visible part of page

Represents *any* document's appearance

Appearance created with set of operators that make marks on page

Descended from PostScript

Types of marks on pages

- Text
- Paths
- Images

Contents is ordered list of drawing operations

Page drawn in order of list

Paint characteristics

Pages marked with “paint”

Objects can hide objects below them

- PDF 1.4 *transparency*—combine layers

Text is font based—vector graphics

- Text and lines are *not* bitmaps/images
- Resolution independent

Images—bitmaps

Page marking operators like PostScript’s and Illustrator’s

- Illustrator 9.0+ native format is PDF 1.4

Operator format

Prefix notation

```
<operands> operator
```

operator typically 1 or 2 letters

operands typically numbers, strings

Types of operators

- Set painting color
- Draw paths—lines and curves
- Draw text
- Draw images

73 drawing operators

Color

Specify color space and coordinates

Separate colors for path *stroke* and *fill*

Basic color spaces

- Gray, RGB, CMYK
- Can be Device or Calibrated

Other colorspaces

- ICC Based
- Pattern

Paths

Draw lines and curves
Can stroke and/or fill path

Example: Draw and fill rectangle

```
1 0 0 RG                 Red stroke in RGB
0 1 0 rg                 Green fill in RGB
200 100 50 60 re         Draw rectangle
b                        Stroke and fill
```

Text

Text is characters in a font

Text outlines can be stroked/filled

Example: Text line

Text “PDF sample” in 12 point Times

```
12 /F1 Tf (PDF sample) Tj
```

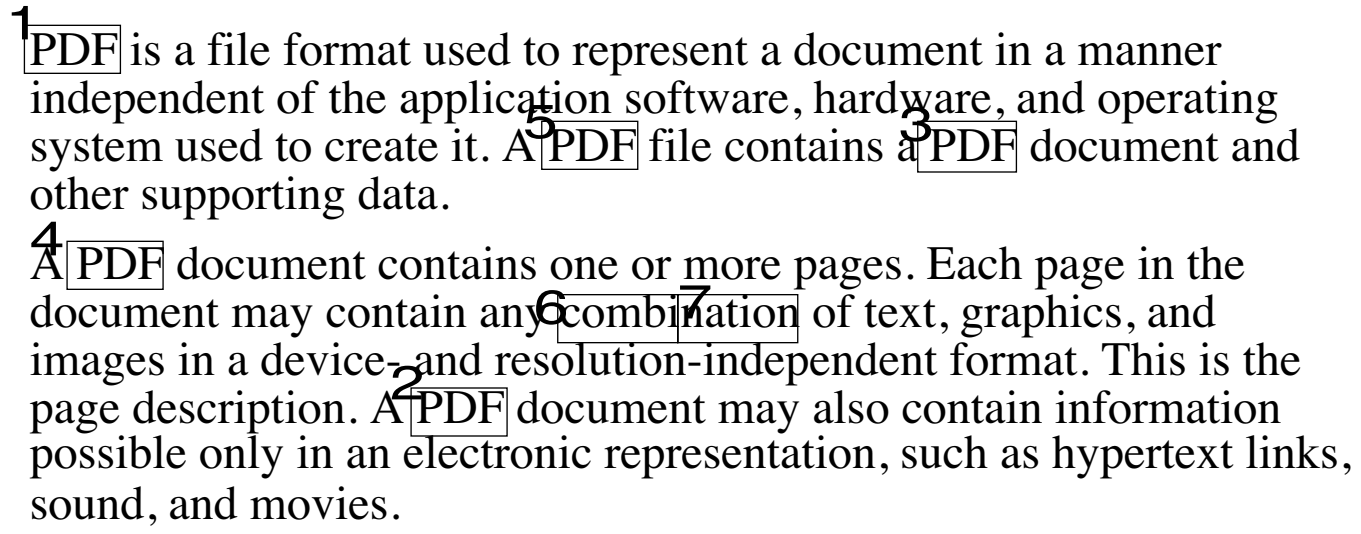
/F1 is name of font resource specifying font attributes, such as Times-Roman

Variety of text operators

- Text positioning
- Character and word spacing
- T- operators from Illustrator

Text ordering

Text doesn't need to be in *any* order



- Page contents could draw all the "PDF" words first
- Words can be broken up
- Page could be represented in *many* ways
- Tagged PDF enforces reading order—reflow

Resources

Additional information needed to draw page

Referenced by page contents

Font

Color space

Images

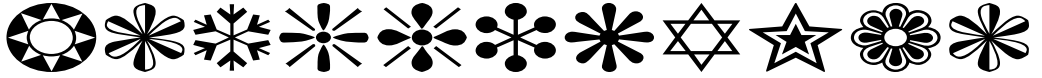
Font

Font is *glyph* description plus *encoding*

Glyphs

- Actual character or ligature shapes

Parisian abcdefghijklmnopqrstuvwxyz

ZapfDingbats 

Symbol αβχδεφγηιφκλμνοπθ

Woodtype 

- Font contains drawing operators for glyph description in defined format

Encoding

- One glyph for each character code
- Most fonts map byte code to glyph
- Many encodings close to ASCII

Data compression and encryption

Compressed data—human unreadable

- Visible data compressed
- Infrastructure not compressed
- Dictionaries *not* compressed
- Decoded without password

Encrypted data—human unreadable

- Hide contents
- Only data streams and strings encrypted
- Need password to decode

Infrastructure *not* all encrypted

- Dictionary keys *not* encrypted—data may not be either

Acrobat Interactive Forms

Catalog object has AcroForm dictionary

- Fields list of root fields
- NeedAppearances create appearances for fields without one
- CO calculation order of fields

Field dictionary

Each form field is a dictionary

Field can have Kids

Field identified by a name

- Fully qualified name from ancestors

  `applicant.address.city`

Field is a *Widget* annotation if it has appearance

Field dictionary attributes

FT Field type
- Button, Text, Choice, Signature

T partial field name

V Value
- variable format, depending on field type

DV Default value

Ff Form flags
- Read-only, Required, No-export
- Other flags, depending on field type

Kids Children fields

AP Appearances—drawn with same operators that draw page appearance

Easily examining PDF internals

Enfocus Browser plug-in

- Shows Cos object structure
- Mac and Windows

Look at file

PDF Reference Highlights

Chapter 3 *Syntax*

- 3.2 *Objects*—Cos objects
- 3.4 *File Structure*

Chapter 4 *Graphics*

- 4.1 *Graphics Objects—drawing overview*
- 4.4 *Path Construction and Painting*
- 4.8 *Images*

Chapter 5 *Text*

- 5.1 *Organization and Use of Fonts*
- 5.3 *Text Objects—how text is drawn*
- 5.5 *Simple Fonts—basic font structure*

Chapter 8 *Interactive Features*

- 8.4 *Annotations—dictionary and types*
- 8.6 *Interactive Forms—field meanings*

Summary

Three structure levels
- File, Document, Cos object

Pages have contents and resources referenced in one place

Marking operators draw page

Dictionaries are everywhere

Can examine and alter PDF file without Acrobat

Acrobat form fields are dictionaries

PDF Ref tells you dictionary keys and structure of document elements

How PDF Works

gstaas@pdfdream.com
www.pdfdream.com