A Short Introduction to PDF

Peter Fischer, ZITI, Uni Heidelberg

pdf vs. Postscript

- pdf describes *graphics, text & document structure*
- It uses *vector graphics*, very similar to postscript
- *Some* differences to postscript:
  - pdf is *NO* programming language
    - – Less powerful
    - \+ But easier to interpret, embed,…
  - pdf has some new graphic features (like transparency)
  - pdf documents are *structured* (with random access to parts)
    - \+ Faster page rendering
    - – Hard coded addresses in file are difficult to calculate by hand
  - Pages sizes can be defined
  - Navigation between pages is possible
  - A directory (with tumbnails) can be set up
  - ...
  - Changes can be stored, files can be encrypted, 3D extensions,...

Why Use & Know About PDF ?

- It is widely used
- Can be directly integrated into pdflatex

- May want to change file content directly
- May need to ‘repair’ broken files
- May want to create simple, compact files directly

- Problem:
  - Calculating the indices in the X-ref table by hand is difficult

- Help:
  - The program ‘pdf toolkit’ *pdftk* contains commands to re-build the X-ref table:
    `pdftk broken.pdf output fixed.pdf`
  - Can also be used to rotate, split or merge,… files
  - See www.pdflabs.com/tools/pdftk-the-pdf-toolkit

PostScript or PDF?

- Pros Postscript
  - Full programming language
  - Simple file structure

- Pros PDF
  - Better / more viewers
  - Better integration with other tools (more ‚modern‘)
  - More graphics features (transparency, ..)
  - Navigation

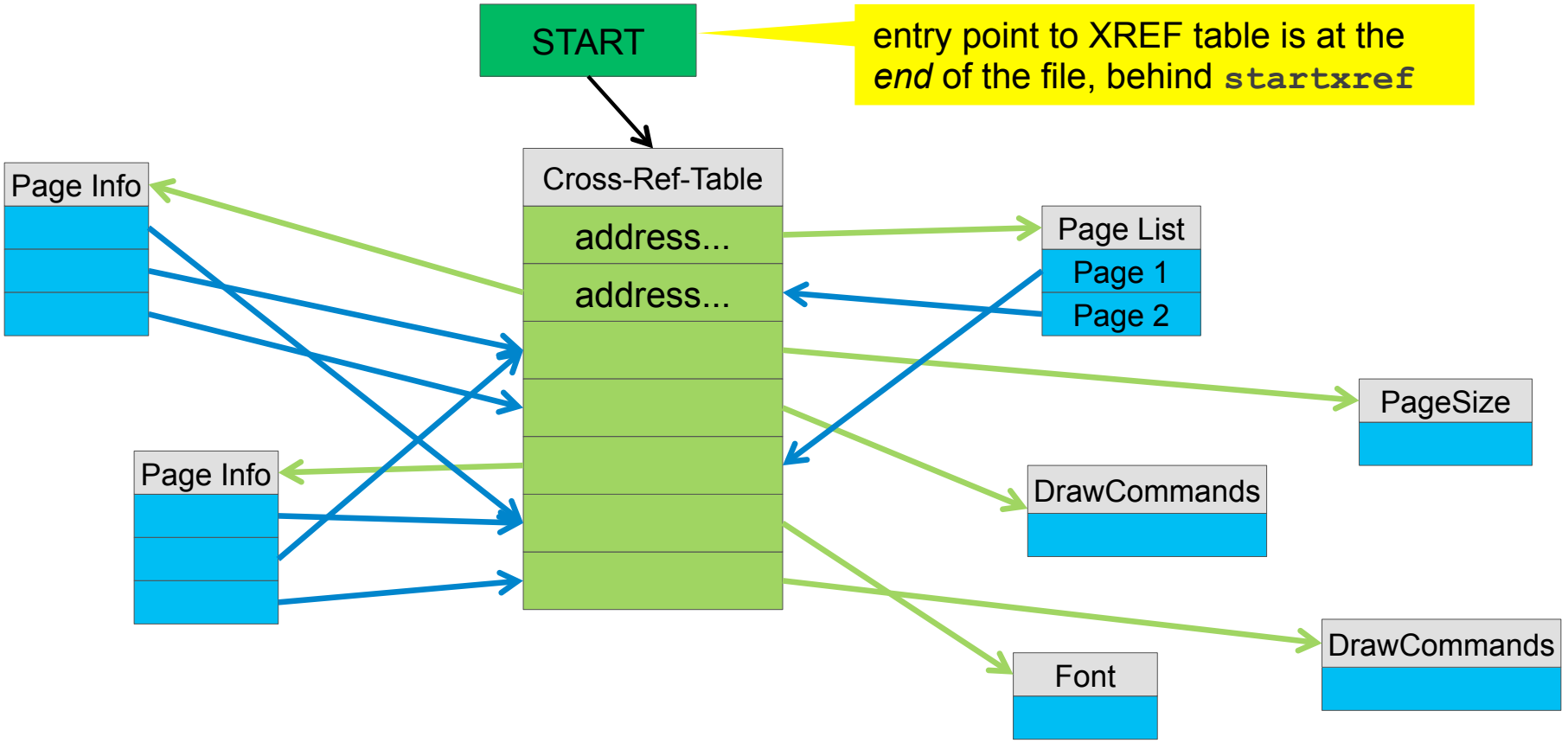
Where to get Documentation?

- @ Adobe:
  http://www.adobe.com/devnet/pdf.html

- Main Document:
  - ‘Adobe PDF Reference’ 31 MB, 1300 pages (!)

Conceptual File Structure

- Most objects are accessed via a *Cross-Reference Table*
  - It contains the *absolute position* of objects in the file
- The document structure is defined by linking objects



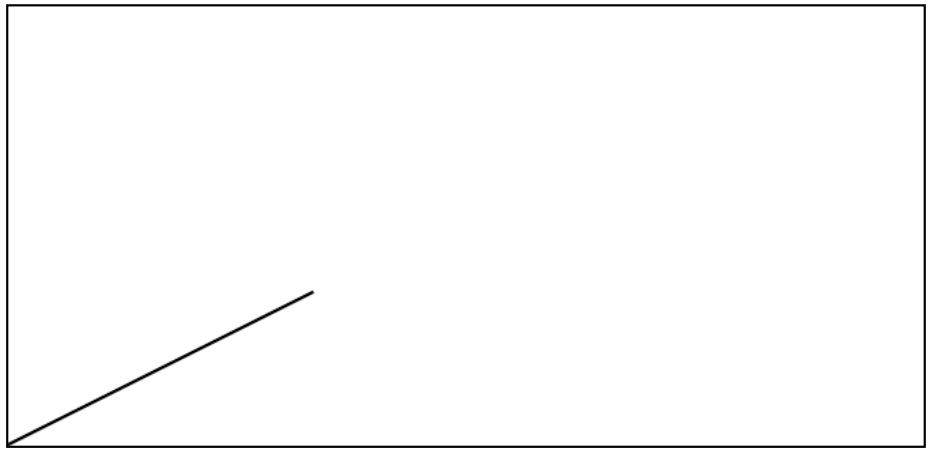
A ‘real’ simple pdf file (Sample_Minimal_Line.pdf)

```
%PDF-1.1
1 0 obj << /Type /Catalog    /Pages 2 0 R                                  >> endobj
2 0 obj << /Type /Pages      /Count 1          /Kids [3 0 R]               >> endobj
3 0 obj << /Type /Page       /Parent 2 0 R     /MediaBox [0 0 300 144]
           /Contents 4 0 R   /Resources << >>                              >> endobj
4 0 obj << /Length 16 >> stream 0 0 m 100 50 l S endstream                    endobj

xref 0 5
0000000000 65535 f
0000000009 00000 n
0000000063 00000 n
0000000128 00000 n
0000000242 00000 n

trailer <<
  /Size 5
  /Root 1 0 R
>>
startxref 311
%%EOF
```

Drawing commands are here:

```
0 0 move
100 50 line
stroke
```

Output (without the frame):

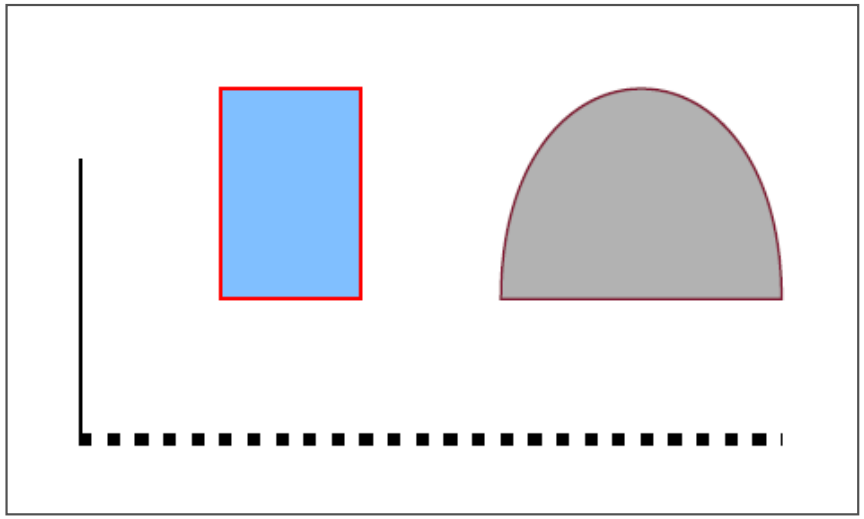


Calculating the Addresses (same file as before)

```
Offset(d)  00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15
00000000   25 50 44 46 2D 31 2E 31 0A 31 20 30 20 6F 62 6A  %PDF-1.1■1 0 obj
00000016   20 3C 3C 0A 20 20 2F 54 79 70 65 20 2F 43 61 74   <<■  /Type /Cat
00000032   61 6C 6F 67 0A 20 20 2F 50 61 67 65 73 20 32 20  alog■  /Pages 2
00000048   30 20 52 0A 3E 3E 20 65 6E 64 6F 62 6A 20 0A 32  0 R■>> endobj ■2
00000064   20 30 20 6F 62 6A 20 20 3C 3C 0A 20 20 2F 54 79   0 obj  <<■  /Ty
00000080   70 65 20 2F 50 61 67 65 73 0A 20 20 2F 43 6F 75  pe /Pages■  /Cou
00000096   6E 74 20 31 0A 20 20 2F 4B 69 64 73 20 5B 33 20  nt 1■  /Kids [3
00000112   30 20 52 5D 0A 3E 3E 20 65 6E 64 6F 62 6A 20 0A  0 R]■>> endobj ■
00000128   33 20 30 20 6F 62 6A 20 3C 3C 0A 20 20 2F 54 79  3 0 obj <<■  /Ty
00000144   70 65 20 2F 50 61 67 65 0A 20 20 2F 50 61 72 65  pe /Page■  /Pare
00000160   6E 74 20 32 20 30 20 52 0A 20 20 2F 4D 65 64 69  nt 2 0 R■  /Medi
00000176   61 42 6F 78 20 5B 30 20 30 20 33 30 30 20 31 34  aBox [0 0 300 14
00000192   34 5D 0A 20 20 2F 52 65 73 6F 75 72 63 65 73 20  4]■  /Resources
00000208   3C 3C 20 3E 3E 0A 20 20 2F 43 6F 6E 74 65 6E 74  << >>■  /Content
00000224   73 20 34 20 30 20 52 0A 3E 3E 20 65 6E 64 6F 62  s 4 0 R■>> endob
00000240   6A 0A 34 20 30 20 6F 62 6A 0A 20 20 3C 3C 20 2F  j■4 0 obj■  << /
00000256   4C 65 6E 67 74 68 20 31 36 20 3E 3E 0A 73 74 72  Length 16 >>■str
00000272   65 61 6D 0A 30 20 30 20 6D 20 31 30 30 20 35 30  eam■0 0 m 100 50
00000288   20 6C 20 53 0A 65 6E 64 73 74 72 65 61 6D 0A 65   l S■endstream■e
00000304   6E 64 6F 62 6A 0A 0A 78 72 65 66 20 30 20 35 0A  ndobj■■xref 0 5■
00000320   30 30 30 30 30 30 30 30 30 30 20 36 35 35 33 35  0000000000 65535
00000336   20 66 20 0A 30 30 30 30 30 30 30 30 30 30 39 20   f ■0000000009
00000352   30 30 30 30 20 6E 20 0A 30 30 30 30 30 30 30 30  0000 n ■00000000
00000368   36 33 20 30 30 30 30 30 20 6E 20 0A 30 30 30 30  63 00000 n ■0000
00000384   30 30 30 31 32 38 20 30 30 30 30 30 20 6E 20 0A  000128 00000 n ■
00000400   30 30 30 30 30 30 30 32 34 32 20 30 30 30 30 30  0000000242 00000
00000416   20 6E 20 0A 0A 74 72 61 69 6C 65 72 20 3C 3C 0A   n ■■trailer <<■
00000432   20 20 2F 53 69 7A 65 20 35 0A 20 20 2F 52 6F 6F    /Size 5■  /Roo
00000448   74 20 31 20 30 20 52 0A 3E 3E 0A 73 74 61 72 74  t 1 0 R■>>■start
00000464   78 72 65 66 20 33 31 31 0A 25 25 45 4F 46 0A     xref 311■%%EOF■
```

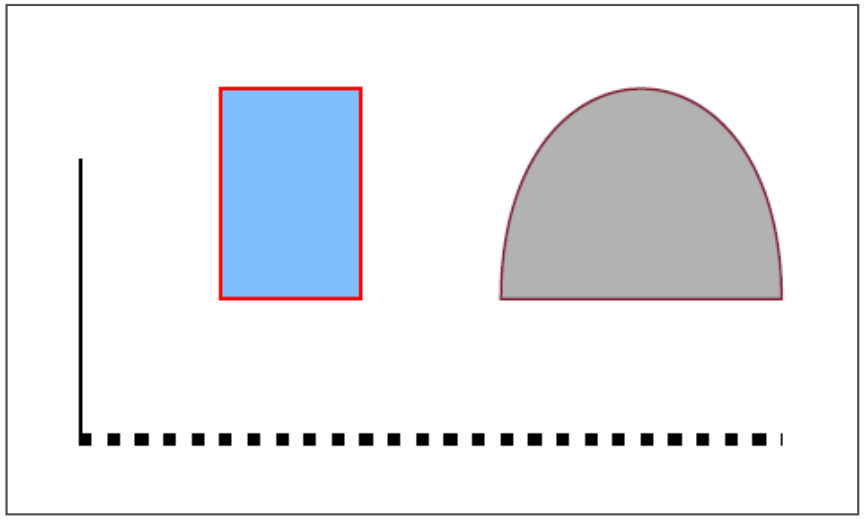
Another pdf file (Sample_Simple_Graphics.pdf)

```
%PDF-1. 4
1 0 obj << /Type /Catalog /Outlines 2 0 R /Pages 3 0 R >> endobj
2 0 obj << /Type /Outlines /Count 0 >> endobj
3 0 obj << /Type /Pages /Kids [ 4 0 R ] /Count 1 >> endobj
4 0 obj << /Type /Page /Parent 3 0 R /MediaBox [ 0 0 612 792 ] /Contents 5 0 R /
  Resources << /ProcSet 6 0 R >> >> endobj
5 0 obj << /Length 883 >> stream
150 250 m 150 350 l S                          % Draw a black line segment
4 w [ 4 6 ] 0 d 150 250 m 400 250 l S          % Draw a thicker, dashed line
[ ] 0 d 1 w 1.0 0.0 0.0 RG 0.5 0.75 1.0 rg     % Set colors
200 300 50 75 re B                             % Draw a rectangle
0.5 0.1 0.2 RG 0.7 g                           % set some more colors
300 300 m 300 400 400 400 400 300 c b          % Draw a curve
endstream endobj
6 0 obj [ /PDF ] endobj

xref 0 7
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000300 00000 n
0000001532 00000 n

trailer << /Size 7 /Root 1 0 R >>
startxref 1556
%%EOF
```

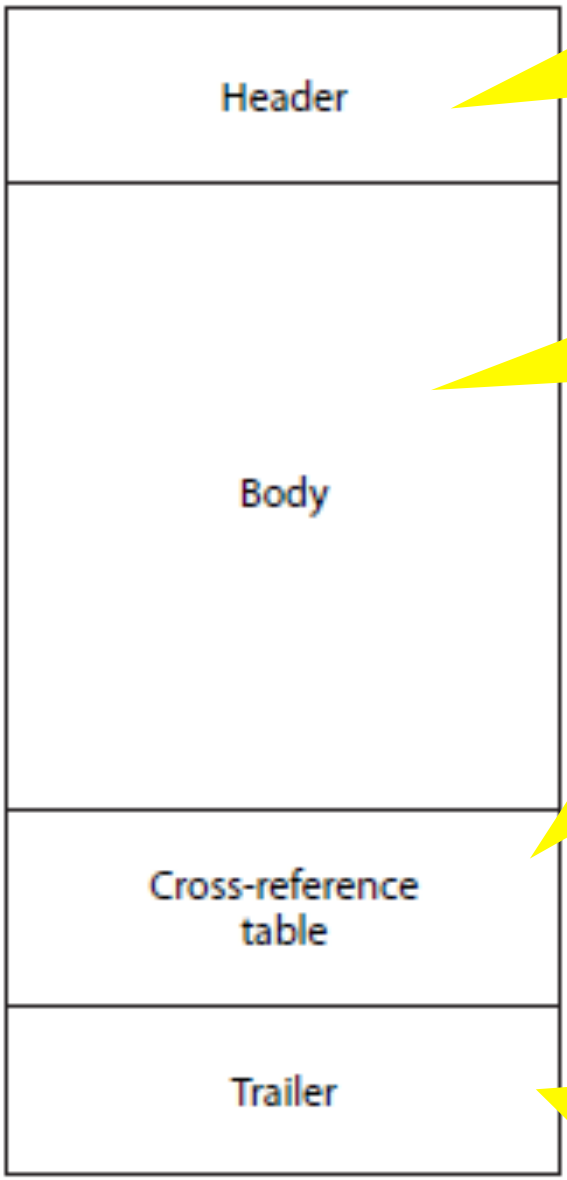


Linking of objects:

```
%PDF-1. 4
1 0 obj << /Type /Catalog /Outlines 2 0 R /Pages 3 0 R >> endobj
2 0 obj << /Type /Outlines /Count 0 >> endobj
3 0 obj << /Type /Pages /Kids [ 4 0 R ] /Count 1 >> endobj
4 0 obj << /Type /Page /Parent 3 0 R /MediaBox [ 0 0 612 792 ] /Contents 5 0 R /
  Resources << /ProcSet 6 0 R >> >> endobj
5 0 obj << /Length 883 >> stream
150 250 m 150 350 l S                          % Draw a black line segment
4 w [ 4 6 ] 0 d 150 250 m 400 250 l S          % Draw a thicker, dashed line
[ ] 0 d 1 w 1.0 0.0 0.0 RG 0.5 0.75 1.0 rg     % Set width, reset dash, set colors
200 300 50 75 re B                             % Draw a rectangle (stroke & fill)
0.6 0.1 0.2 RG 0.7 g                           % set some more colors
300 300 m 300 400 400 400 400 300 c b          % Draw a curve
endstream endobj
6 0 obj [ /PDF ] endobj

xref 0 7
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000300 00000 n
0000001532 00000 n

trailer << /Size 7 /Root 1 0 R >>
startxref 1556
%%EOF
```



Document Structure



- Version
- Binary header
  (so that external tools recognize file as binary)

```
%PDF-1.4
%????
```

- Indirect (‘linked’) objects
  see later

Blanks **required** here...

- keyword xref
- ID of first entry
- number of entries
- for each entry:
  - address version free/used

```
xref 0 5
0000000000 65535 f
0000000009 00000 n
0000000063 00000 n
0000000128 00000 n
0000000242 00000 n
```

- Size of X-ref table (entries)
- ID of ‘start’ object
- Other information...
- Location of X-ref table
- End of File marker

```
Trailer <<
  /Size 7
  /Root 1 0 R >>
startxref 1556
%%EOF
```

Object Types

Type	Comment	Marker	Example
Boolean		-	true, false
Number	Integer or float Format 1.3e+3 is NOT supported	-	13 1.45
String	‚Literal string' : characters ‚Hexadecimal string': hex values	(…) <…>	(test) <3EFF00>
Name	Unique (string) Identifiers	/…	/test /Catalog
Array	Can contain mixed object types	[…]	[0 1 /test false (test) (false)]
Dictionary	- Contains Key/Value pairs - Keys must be Names - Values can be anything, including nested dictionaries…	<<…>>	<< /Type /Example /Data [1 0 0] /ShowFlag false >>
Stream	- Byte data - Can have arbitrary length - MUST be indirect object (see later) - Comes with a Dictionary - MUST start new line after 'stream'		<< /Length value /… other values >> stream … endstream
Null Object	Equivalent to ‚no entry'	-	null

Labelling Objects – Indirect Objects

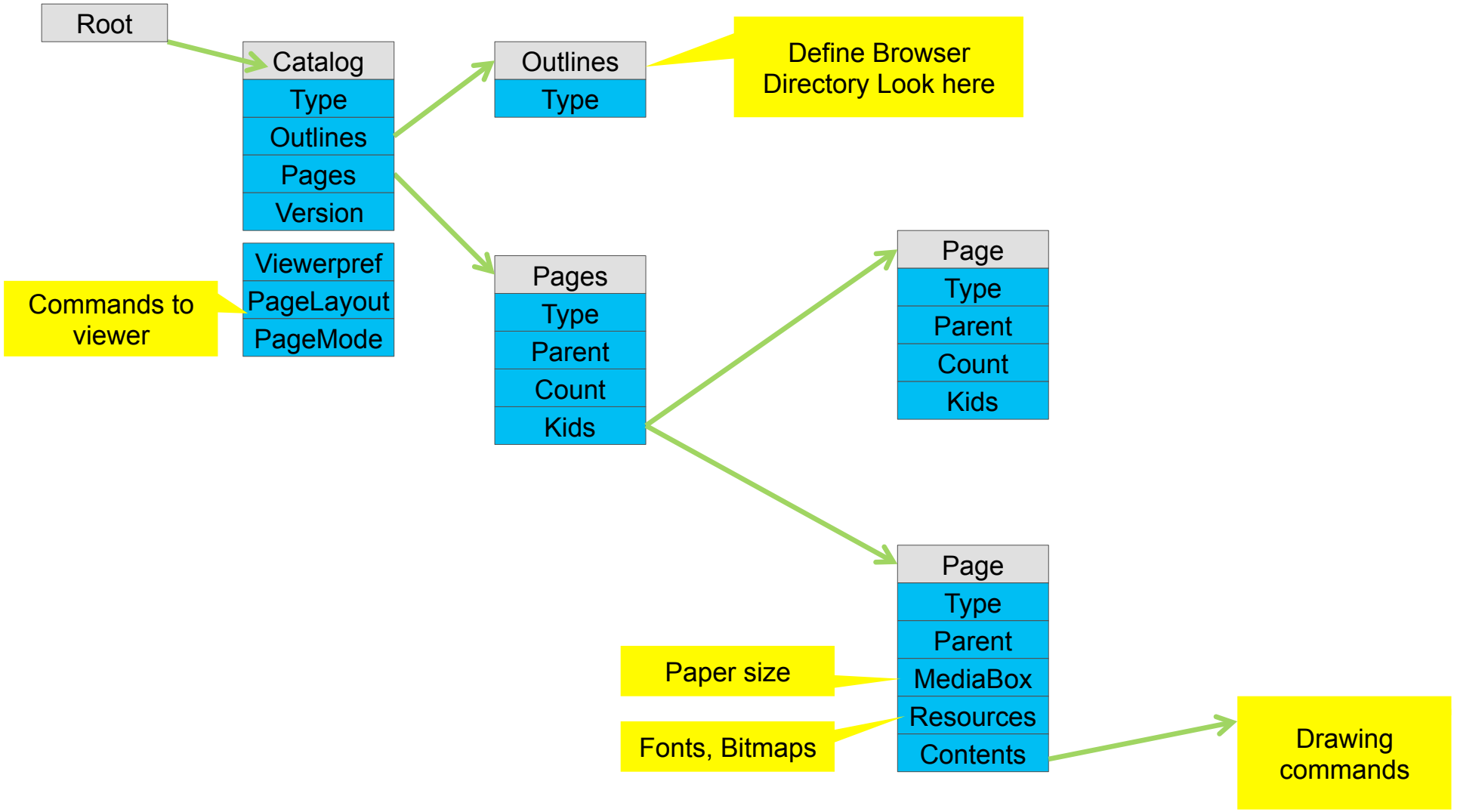
- Objects can be labelled.
  They are then called ‚*indirect objects*‘
- Definition:
  - **ID GEN obj** …data… **endobj**
  - Identifier **ID** is integer > 0. It must be unique.
    Every **ID** must be listed in Crossref-Table
  - Integer generation marker **GEN** is ≥ 0
    It *can* number versions.
  - **obj**, **endobj** are keywords

```
1 0 obj
...
endobj

2 0 obj
...
endobj
```

- To refer (‘point’) to an (indirect) object, use the indirect reference:
  - **ID GEN R**

```
...
/Pages 1 0 R
...
```

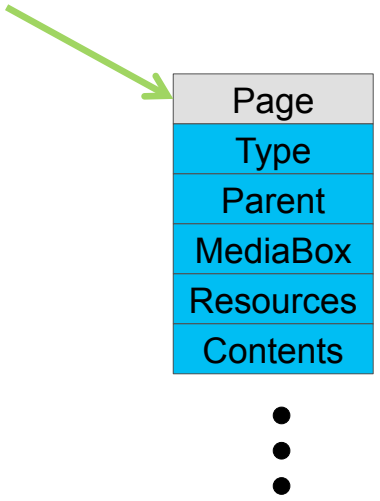
Objects for a Typical Page



Examples

- Page Node

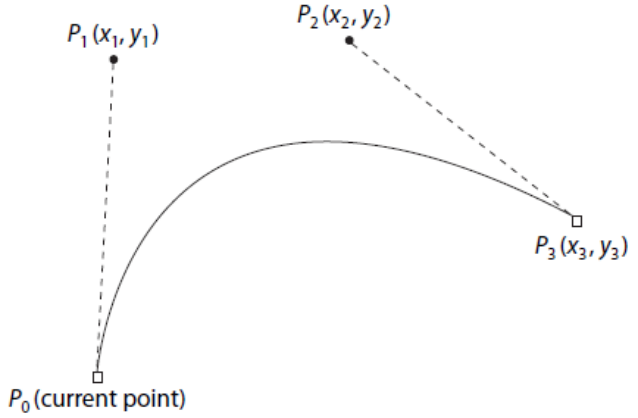
```
3 0 obj
   << /Type /Page
      /Parent 4 0 R
      /MediaBox [0 0 612 792]
      /Resources << /Font << /F3 7 0 R
                             /F5 9 0 R
                             /F7 11 0 R
                          >>
                    /ProcSet [/PDF]
                 >>
      /Contents 12 0 R
      /Thumb 14 0 R
      /Annots [ 23 0 R
                24 0 R
              ]
   >>
endobj
```



Drawing Commands

- Very similar to Postscript, but short hand notations
- RPN

`value` **`w`**	set line width
`x y` **`m`**	moveto
`x y` **`l`**	lineto
`h`	closepath
`x y x y x y` **`c`**	cubic bezier curve
`x y w h` **`re`**	rectangle
`S`	stroke
`s`	close & stroke
`f`	fill
`B`	stroke & fill
`...`	...



Sample File

- See file *Sample_2Pages.pdf*

- 12 indirect Objects
  - 1: Catalog
  - 2: Info
  - 3: ViewerPreferences (tell browser to show doc. Title)
  - 4: Pages (2 Kids)
  - 5: Outlines (top)
  - 6: Large Page
  - 7: Small Page
  - 8: Resources (both pages)
  - 9: Content page 1
  - 10: Content Page 2
  - 11: First Outline entry
  - 12: Last Outline entry

pdftk

- pdftk is a software tool to do simple, usefull manipulations
- Uncompress a file:
  - pdftk inputfile.pdf output outputfile.pdf uncompress
- …