PDF Overview

Everyting you wanted to know about PDF, but were afraid to ask!

Leonard Rosenthol
Senior Software Engineer
Digital Applications, Inc.

PDF - Portable Document Format

- File format designed specifically for electronic distribution of “final form documents”
- Created by Adobe in 1992-1993, as part of their Acrobat product.
- PDF is an open public format with specs available from Adobe at <http://partners.adobe.com/asn/developer/acrosdk/DOCS/pdfspec.pdf>.

PDF - What's in there?

- PostScript/Adobe imaging model
  - Text & graphics in a device & resolution independent manner
- Bitmap Images
- Annotations
  - Text notes, Hyperlinks, "MarkUp", Movies, Sounds & Widgets
- Forms

PDF - What's NOT in there?

- PDF is NOT Postscript!
  - Non-printable elements (hyperlinks, etc.)
  - No programming language constructs
  - Strict file structure allowing for random access
  - Presence of font metrics for viewing fidelity
- A PDF file can not be directly interpreted by a PS interpreter, though conversion of PDF page descriptions to PS is simplified.

A bit of history

- PDF 1.0 - Acrobat 1.0
  - Postscript imaging model
- PDF 1.1 - Acrobat 2.0
  - Security, annotations, binary files
- PDF 1.2 - Acrobat 3.0
  - Interaction, movies/sounds, forms, CJK, web
- PDF 1.3 - Acrobat 4.0
  - Structure, Digital Sigs, embedding, JavaScript, RTL, color seps, PS3

Peeling the layers ofPDF

- PDF file
  - physical container in a file system containing the PDF document and other data
- PDF document (aka page description)
  - Contains one or more pages, where each page consists of text, graphics and/or images as well as hyperlinks, sounds, etc.
- “other data”
  - PDF version, object catalog, etc.

Properties of PDF

- Adobe Imaging Model
- Portability
- Compression/Encryption
- Font Independence
- Random Access
- Incremental Update
- Extensibility

Adobe Imaging Model

- Same model as Postscript, where a page is drawn by “placing paint” on a selected area
  - “figures” can be letter shapes, regions defined by lines and curves or sampled images
  - Paint can be any color (specified in variable color spaces)
  - Figures can be clipped to any other figure/shape
  - Figures are “overlayed” on each other, in the order they exist in the page description.

Portability

- PDF files are binary, all 8 bits can be used - though support for 7 bit files exists
  - ASCII-85 when needing to encode to 7 bits
- Single document format regardless of platform
- Non-Roman language support via standard encodings (SJIS, KCS7, etc.) as well as Unicode

Compression/Encryption

- Support for a number of industry standard algorithms
  - JPEG (for color & grayscale images)
  - CCITT Group 3 & 4, LZW, RLE for monochrome images
  - LZW & Flate (ZIP) for text, graphics, etc.
  - RC4 (at 40 bits, moving to 56)

Font Independence

- Use of Font Descriptors
  - Name, character metrics, styles
- If a font exists, it is used. Otherwise a multiple master font is used using the font descriptor as the basis.
- Symbol fonts must be included, since they can’t be simulated with MM’s.

Random Access

- Cross reference table maintains lists of pages, objects on a page, etc.
- Xref is stored at the end of the document, allowing for single pass creation and ease of location
  - Except in the case of linearized documents designed for byte-serving (ie. dynamic serving via the web)

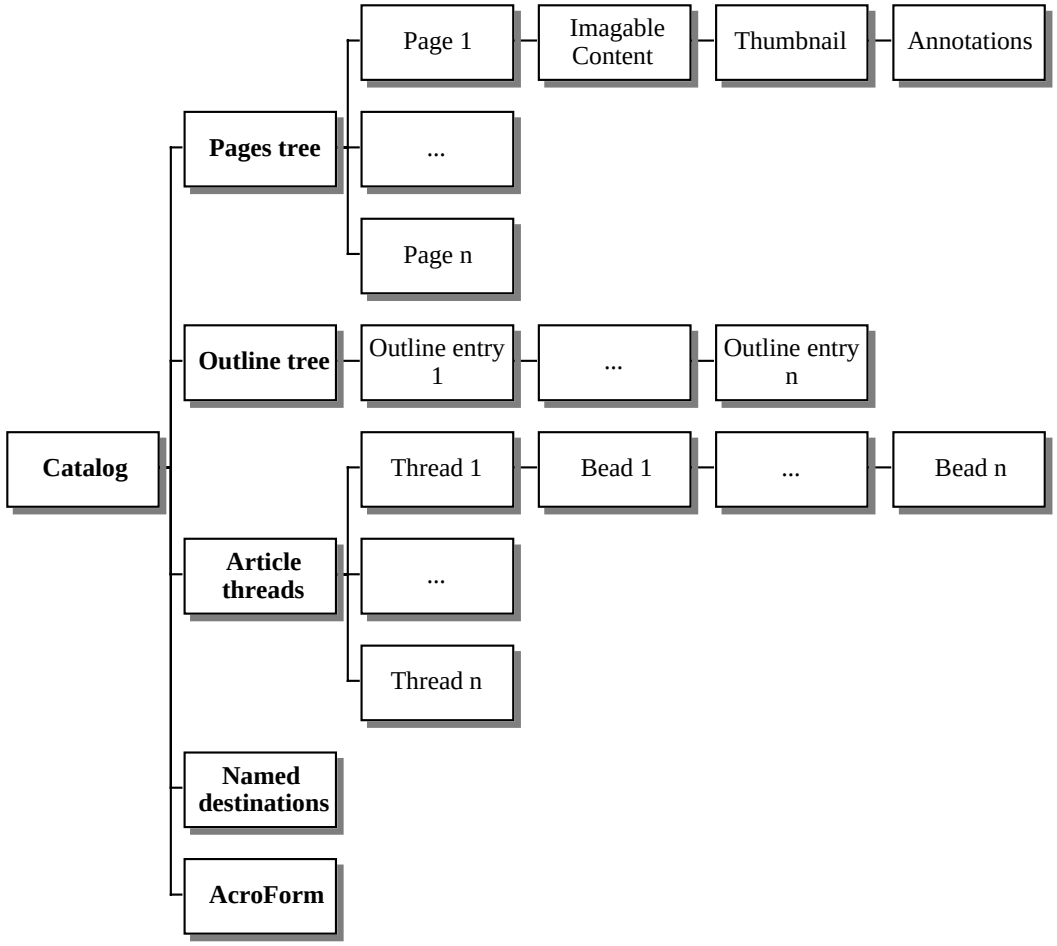
Incremental Update

- Modifications are written to the end of the file, leaving the original data intact
- A new xref table is written containing the new/modified data, and a link back to the old xref.
- Since original data is still present, support fo r multiple undos across save boundaries can be supported.

Extensibility

- As seen by the features added to PDF since 1.0, you can see that new features can easily be added to PDF w/o breaking backwards compatibility. A viewer will simply ignore an object that it doesn't understand.

Structure of a PDF document



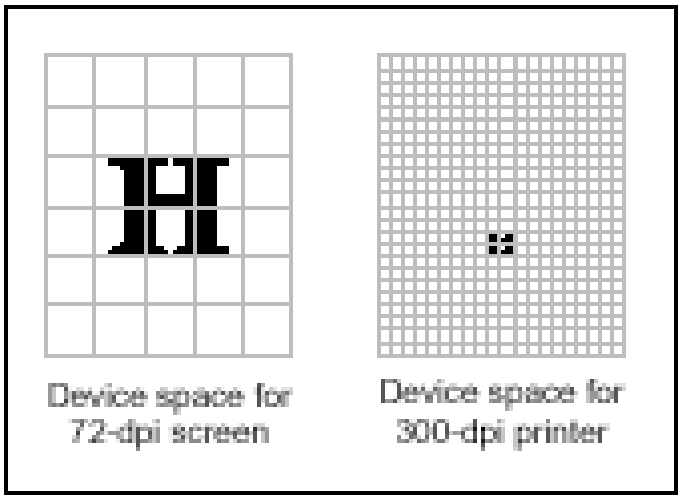
Adobe Imaging Model in Depth

Coordinate Systems

- A coordinate system defines the canvas on which all drawing takes place
  - Position
  - Orientation
  - Size
- Multiple coord systems are used by PDF, and they all interrelate in the final rendering of the page

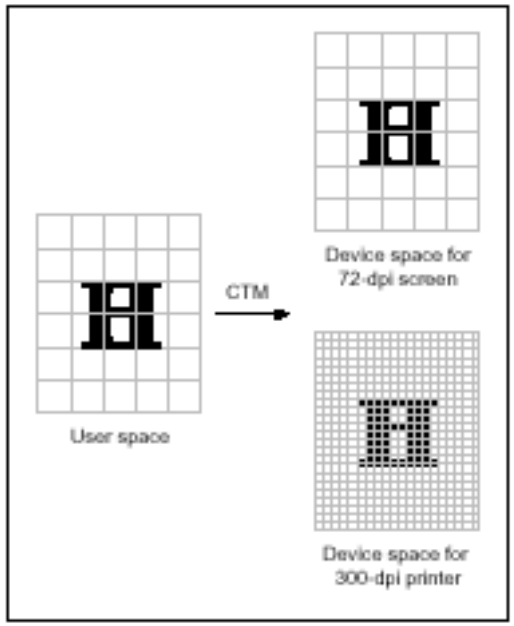
Coordinate Systems (cont.)

- Device Space
  - Output device (screen, printer, etc.)
    - Differences in resolution (72 dpi vs. 600 dpi)
    - Differences in origin (top left, bottom left)



Coordinate Systems (cont.)

- User Space
  - A coordinate system that stays the same regardless of the output device
  - The Current Transformation Matrix (CTM) specifies the transformation from user space to device space
  - Default user space is 72 units per inch (aka “a point”) with the origin at bottom left



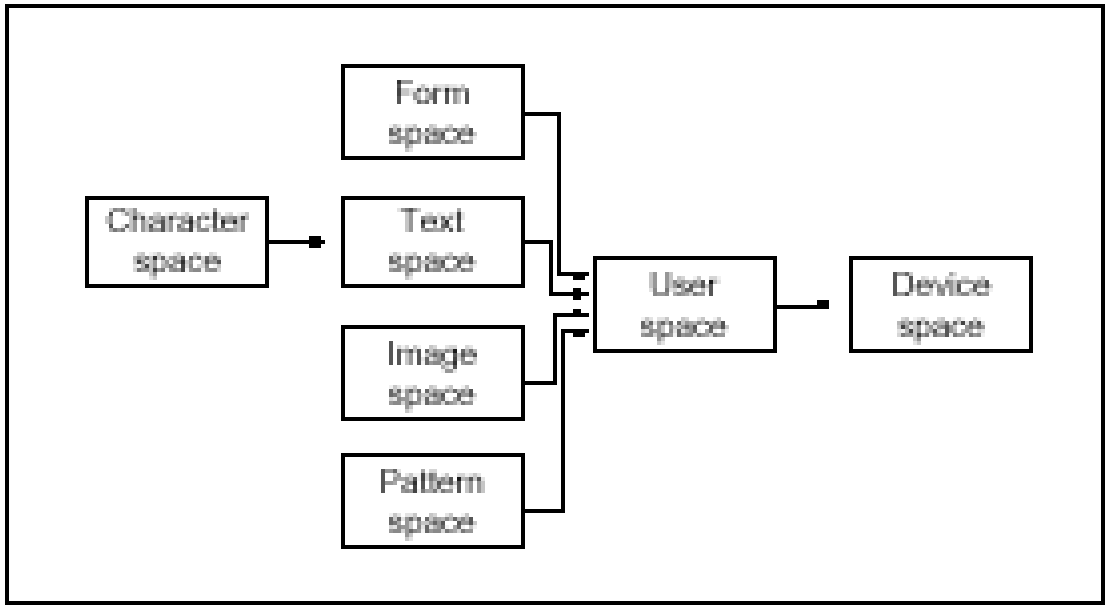
Coordinate Systems (cont.)

- Text Space
  - Coordinates for text are in text space, the “text matrix” defines the transform to user space
- Character Space
  - Each glyph in a font is in character space.
  - Predefined based on the type of font
- Image Space
  - All images are here, and the transform to user space is unchangable! (images are 1 x 1)

Coordinate Systems (cont.)

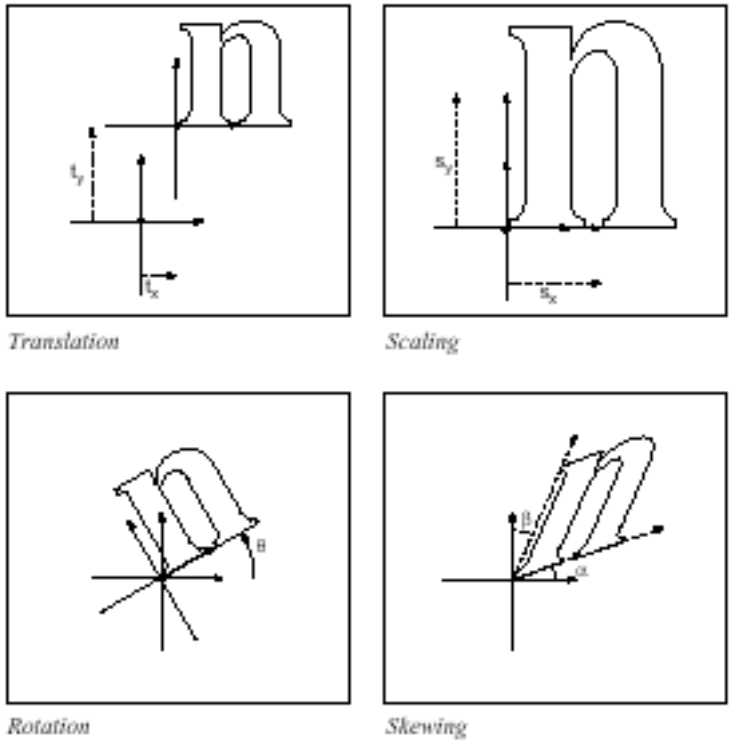
- Form Space
  - Form XObjects (NOT buttons, fields, etc.) are in form space. Each Form XObject defines it's own transform to user space
- Pattern Space
  - Each pattern defines it's own transform to user space

Coordinate Systems Relationships

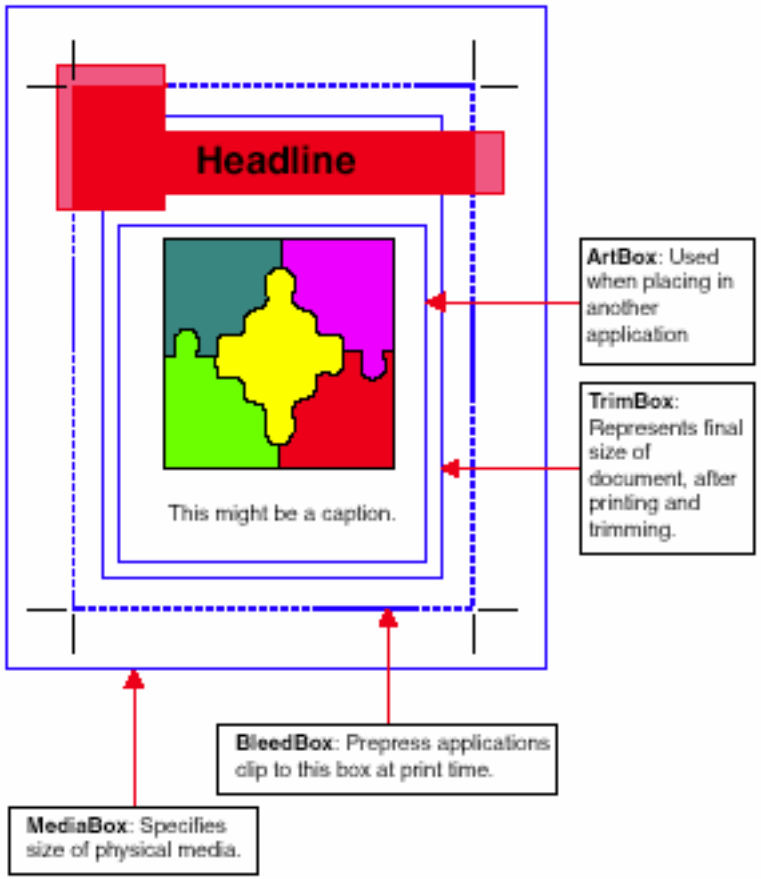


Transformations

- Translation
- Scaling
- Rotation
- Skew



Page “Layout”



Fonts

- The following types of fonts are supported
  - Type 1 (either full, multimaster or subset)
  - Type 3
  - TrueType (full or subset)
  - Type 0
  - CIDFont Type 0 & 2
- There are 14 fonts that are “built-in”, including the Courier, Helvetica and Times families, Symbol & ITC Zapf Dingbats.

Color Spaces

- PDF Supports 11 color spaces
  - 3 device dependant
    - DeviceGray, DeviceRGB & DeviceCMYK
  - 4 device independent
    - CalGray, CalRGB, Lab & ICCBased
  - 4 special
    - Indexed, Pattern, Separation & DeviceN

Patterns

- Two types
  - Tiling
    - A sequence of explicit marking instructions which are repeated once per tile
  - Smooth shading
    - Description of the desired effect in terms of transitions between colors across a certain area
- All patterns are treated as colors

Images

- PDF support a number of different types
  - Image masks
  - Grayscale images (1, 2, 4, & 8 bit)
  - Color images (1, 2, 4, & 8 bits per component)
    - Number of components is determined by color space
- Images are usually either in “PDF bits”, JPEG or TIFF formats.

Graphic Objects

- Paths
- Text
- Image
- External

Graphic Objects (cont.)

- Paths
  - Arbitrary shape made of straight lines, rectangles, and cubic curves.
  - May intersect itself and may have disconnected sections and holes.
  - Filled, stroked, and/or a clipping path.
- Text
  - One or more character strings
  - Filled, stroked, and/or a clipping path

Graphic Objects (cont.)

- Image
  - "bitmap" using a specified color model
- External
  - Images
  - Forms
  - PS language fragments

Graphics State

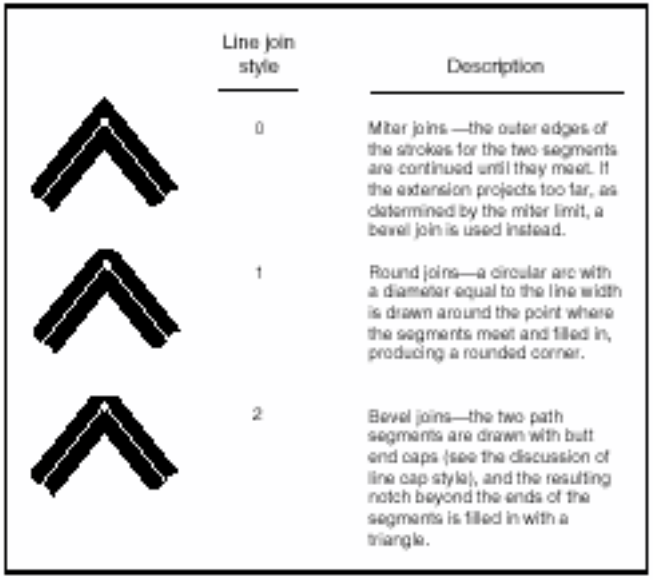
- Kind of like a GrafPort in that it maintains the current “state” of the graphics world.
- States are “grouped” by kind
  - Text, Color, General and Special
- Allows for user extensions to preserve “app-specific” state
  - Eg. Adobe Illustrator saves things like layer & extended text info

Current Point

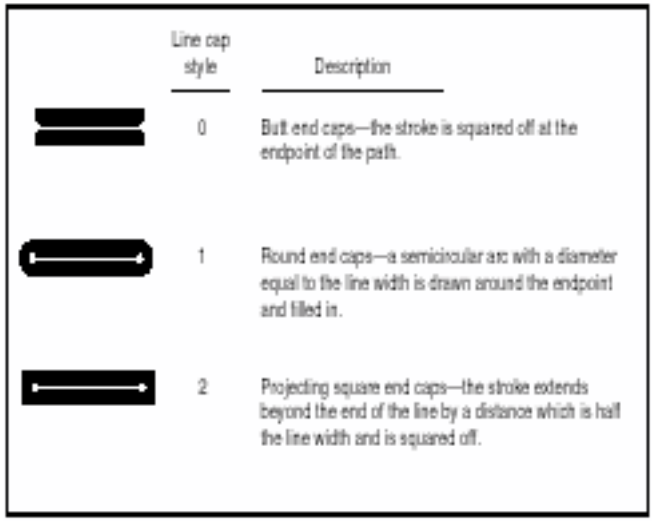
- This is the current pen location used for drawing and text operations. Works just like the current pen location in QuickDraw - EXCEPT that after a paint operation, the current point becomes undefined.

Lines

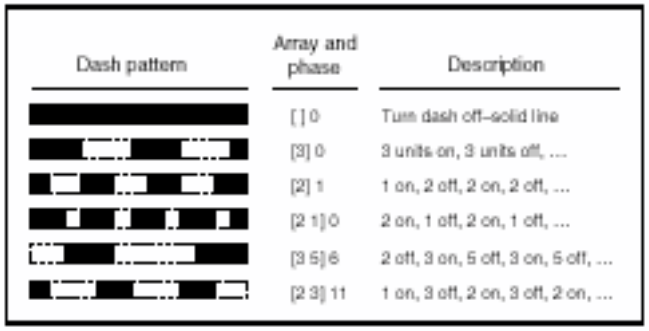
- Cap styles
- Dash patterns
- Join styles
- Width



	Line join style	Description
	0	Miter joins—the outer edges of the strokes for the two segments are continued until they meet. If the extension projects too far, as determined by the miter limit, a bevel join is used instead.
	1	Round joins—a circular arc with a diameter equal to the line width is drawn around the point where the segments meet and filled in, producing a rounded corner.
	2	Bevel joins—the two path segments are drawn with butt end caps (see the discussion of line cap style), and the resulting notch beyond the ends of the segments is filled in with a triangle.



	Line cap style	Description
	0	Butt end caps—the stroke is squared off at the endpoint of the path.
	1	Round end caps—a semicircular arc with a diameter equal to the line width is drawn around the endpoint and filled in.
	2	Projecting square end caps—the stroke extends beyond the end of the line by a distance which is half the line width and is squared off.



Dash pattern	Array and phase	Description
	[ ] 0	Turn dash off–solid line
	[3] 0	3 units on, 3 units off, ...
	[2] 1	1 on, 2 off, 2 on, 2 off, ...
	[2 1] 0	2 on, 1 off, 2 on, 1 off, ...
	[3 5] 6	2 off, 3 on, 5 off, 3 on, 5 off, ...
	[2 3] 11	1 on, 3 off, 2 on, 3 off, 2 on, ...

Colors

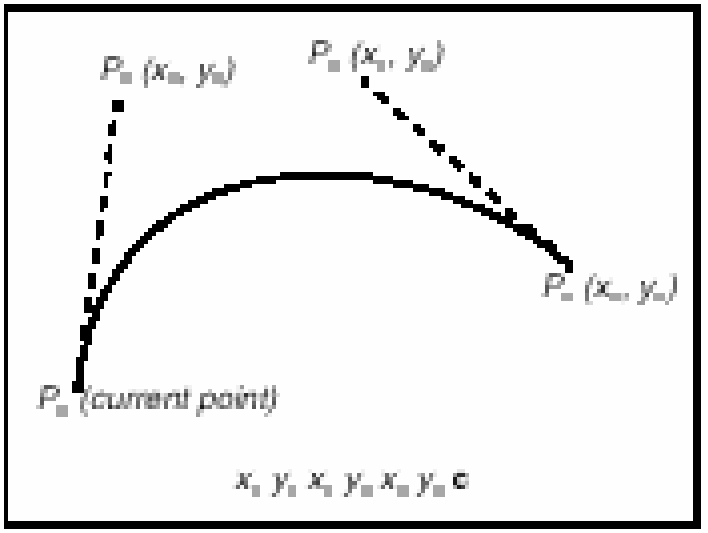
- Fill
- Stroke
- Color Space

Paths

- A path consists of a series of path segments describing where marks are to appear on the page.
- A path may be composed of one or more disconnected sections (subpaths)
  - Eg. Parallel line segments
- Segments may be straight or curved

Curves

- Curves are cubic Bézier curves represented by two end points and two control points



Painting a path

- Paths are stroked, filled or both.
- Painting completely obscures anything already drawn in the same location
- Stroking draws a line along the path given the current graphics state.
- Filling paints the enclosing path.

Figure 8.10 *Non-zero winding number rule*

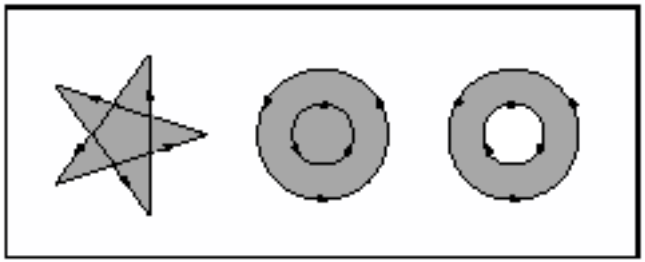
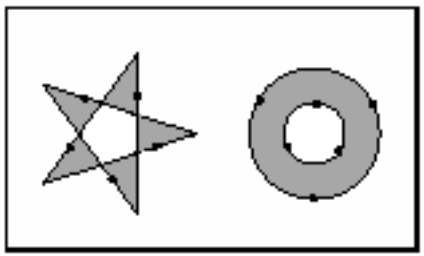


Figure 8.11 *Even–odd rule*



Text

- Text state is based on nine parameters
  - Character spacing
  - Word spacing
  - Horizontal spacing
  - Leading
  - Font
  - Size
  - Transform matrix
  - Rendering mode
  - “Text rise”

Text Rendering Modes

	Rendering mode	Description
	0	Fill text
	1	Stroke text
	2	Fill then stroke text
	3	Text with no fill and no stroke (invisible)
	4	Fill text and add it to the clipping path
	5	Stroke text and add it to the clipping path
	6	Fill then stroke text and add it to the clipping path
	7	Add text to the clipping path

Text Rise

- Specifies the amount to move the baseline up or down from the default location.
- Subscripts, Superscripts or just general adjustment

Form XObject

- A form xobject is kind of like a “PICT”. It’s a named self-contained description of text, graphics or sample images that can be drawn more than once on a single page, or on multiple pages.

What else you got?

Annotations

- Text Notes
- Hypertext Links
  - Inter-document, intra-document, URI's
- Movies
- Sounds
- Widgets
- Trap Networks
  - Used by the PJTF (Portable Job Ticket Format)

Actions

- These are things that can be attached to certain objects/events in a document
  - Open doc, close doc, view page, mouse enter/leave, annotations, form elements, etc.
- They come in a few types:
  - GoTo, GoToR(emote), URI
  - Launch
  - Sound, Movie
  - JavaScript

Embedded Files

- You can embed an entire file (or any type!) inside a PDF document
- Useful for having a single file that contains everything related. A Launch command, for example, might reference the embedded doc rather than an external reference.

AcroForms

- A PDF file may contain AT MOST ONE AcroForm, though that form may contain any number of fields located on any page.
- You can also dynamically import./export sets of fields from the file
- There a number of predefined field types
  - Button (checkbox/radio/push)
  - Text
  - Choice (popup, combo or list)

AcroForms (cont.)

- Fields can be typed (integer, string, boolean) and marked read-only.
- Fields can be “calculated”, such that a JavaScript will be autoexecuted when a “related” field is modified.
- Form data can be “submitted” via either FDF (Forms Data Format) documents or via an HTTP get or post.

Structured PDF

- This is a feature added as part of the AcroSpider/Web Capture project now available in Acrobat 4.05

- Allows for storing information about the logical structure (such as HTML tag levels) of a document along side the layout.
- Useful for improved searching & indexing type operations

Online Resources

- Adobe Developer Association
  - <http://partners.adobe.com/asn/developer/sdks.html>
- PDFZone
  - <http://www.pdfzone.com/>
- AcroBuddies
  - <http://www.acrobuddies.com/>