PDF and Accessibility

Mark Gavin

Appligent, Inc.

January 11, 2005

Agenda

1. What is PDF?
    a. What is it not?
    b. What are its Limitations?
2. Basic Drawing in PDF.
3. PDF Reference

What is PDF?

What is it not?

What are its Limitations?

What is PDF?

"THE ADOBE PORTABLE DOCUMENT FORMAT (PDF) is a file format for representing documents in a manner independent of the application software, hardware, and operating system used to create them and of the output device on which they are to be displayed or printed."

PDF Reference 1.6 Chapter 2

What is Portable Document Format

1. A binary file format based on the Postscript language developed by Adobe Systems.
2. Designed for device independent and resolution independent viewing of page based documents.
3. Free format database file which can contain almost any type of additional data.
4. Electronic Paper.

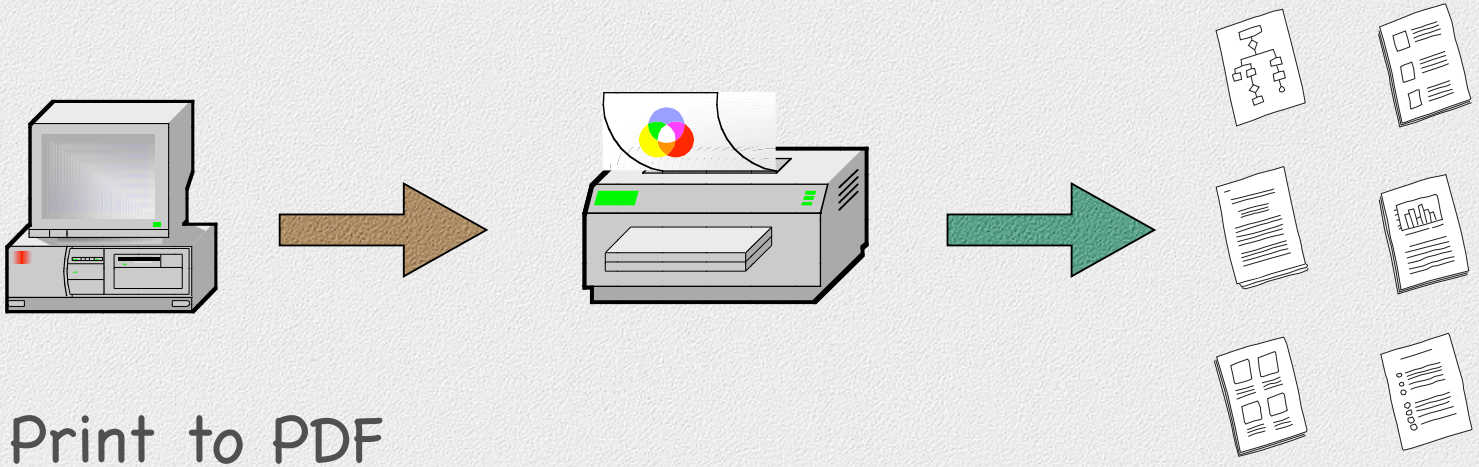
PDF is Not Acrobat

1. Acrobat is a commercial application, from Adobe, which can View and Annotate PDF file.
2. There are other commercial and open source PDF viewers available.
    a. XPDF
    b. Mac OS X
    c. Global Graphics PDF Editor

What PDF is not

1. A PDF file is not an ASCII text file.
2. PDF is not a stream like file format which flows in the natural read order of the document.
    a. The content of page 2 does not necessarily follow the content of page 1.
    b. The content of the second paragraph of a page may not follow the content of the first paragraph.

Creating a PDF File



1. Print to PDF
    a. Postscript to PDF
    b. Printer Driver to PDF
2. Postscript to PDF with PDFMarks
3. Direct Application Support for Saving to PDF.

Limitation of Print

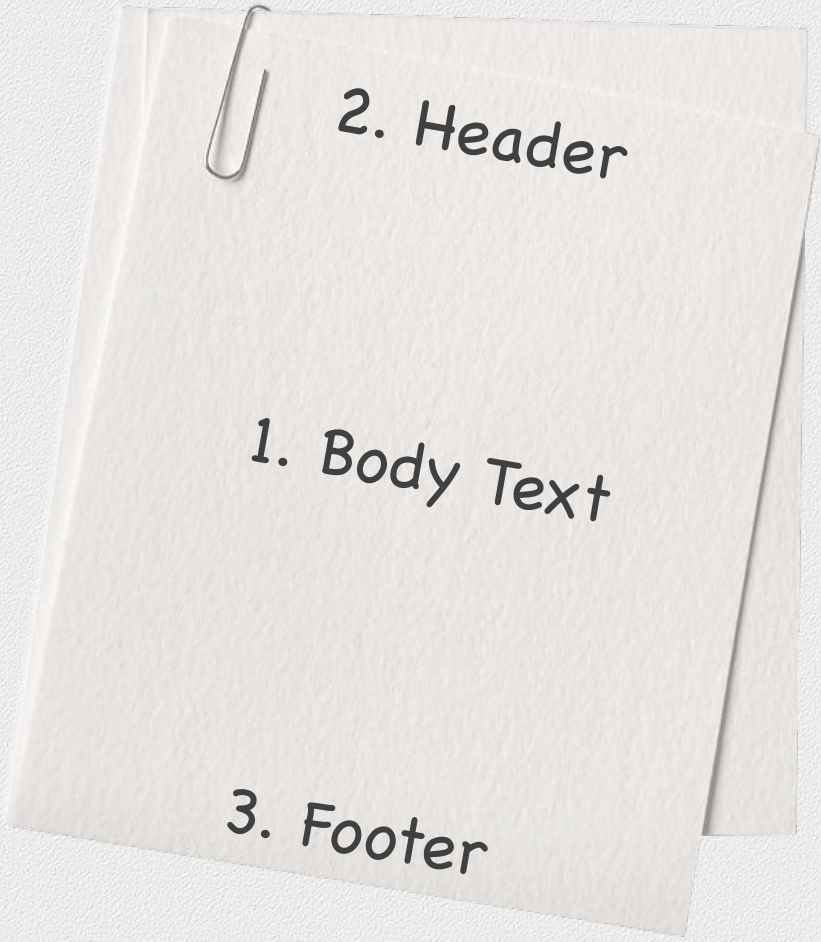
Information is, almost always, lost printing to PDF.

Why? Information from the source application is being changed into a graphic.

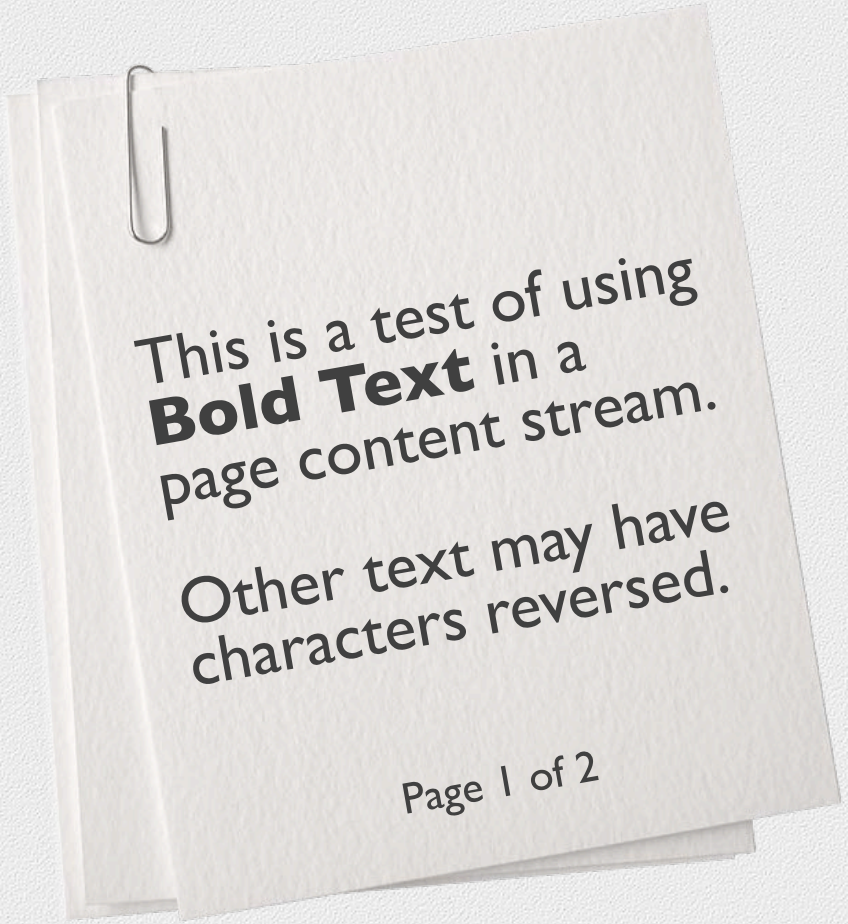
PDF - Electronic Paper

Information is not always placed in the PDF document in a natural read order.

For example; the Body text can be drawn to a page first, followed by the Header then the Footer.



Electronic Paper



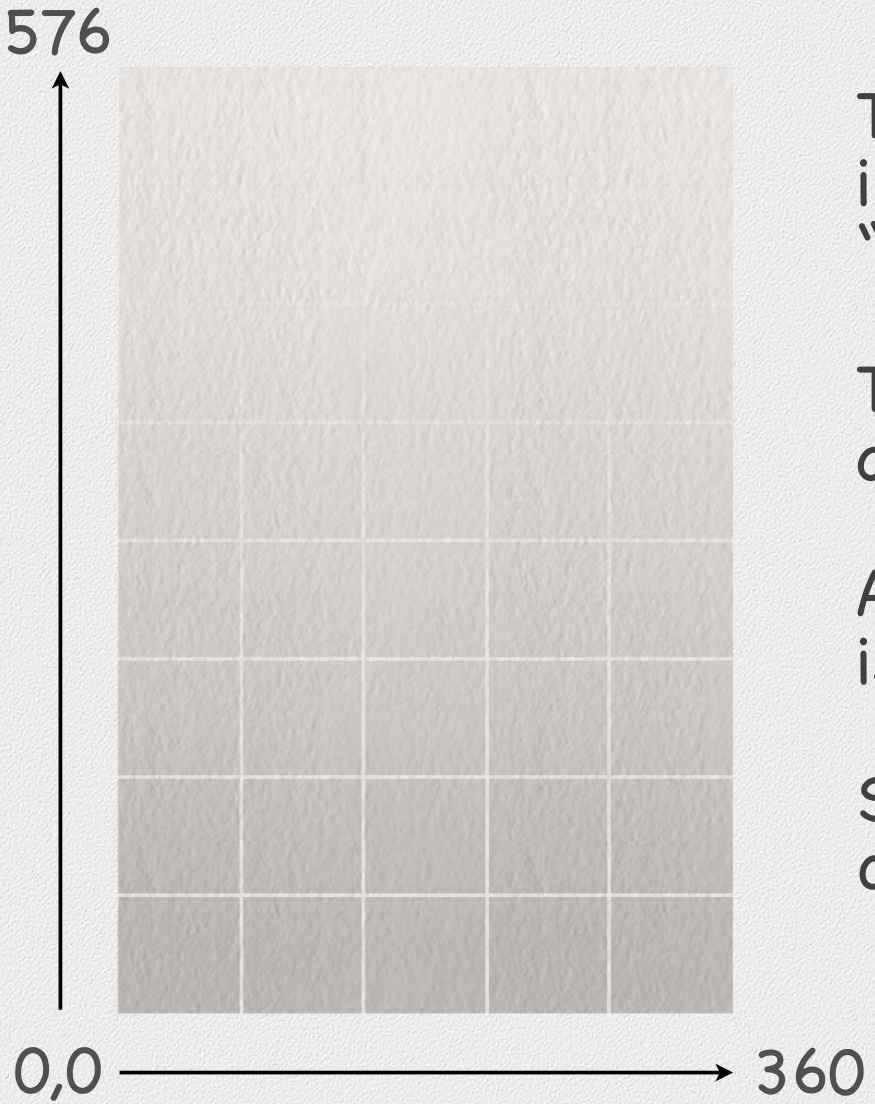
To make the issue even more complex; body text is not always placed in the order one would expect.

Page 1 of 2 ←
This is a test of using in a page content stream.
Other text may have characters desrever.
Bold Text

Basic Drawing in PDF

How to Draw using PDF

The Page



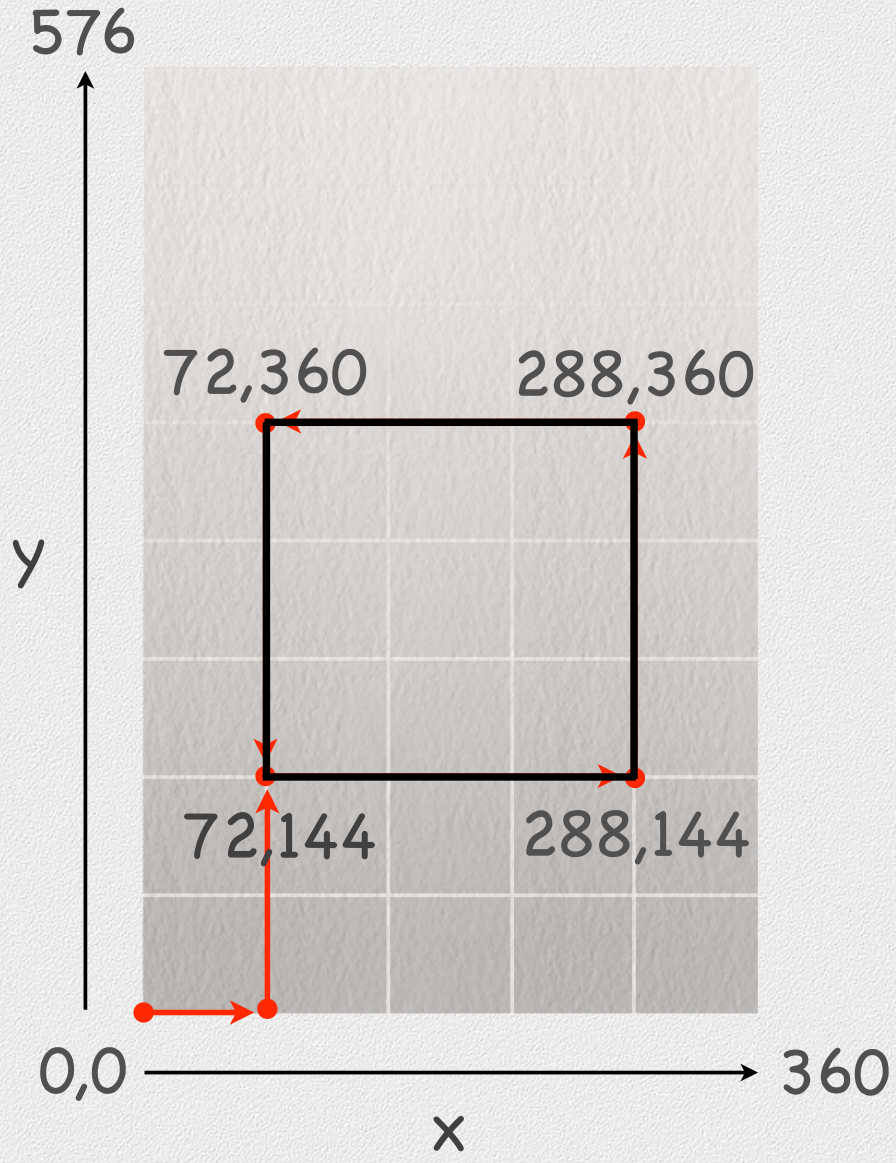
The Measure of distance in PDF is the printers "point".

There are 72 points in one inch.

An 8 1/2 by 11 inch page is 612 by 792 points.

Standard Cartesian coordinate system.

Basic Drawing



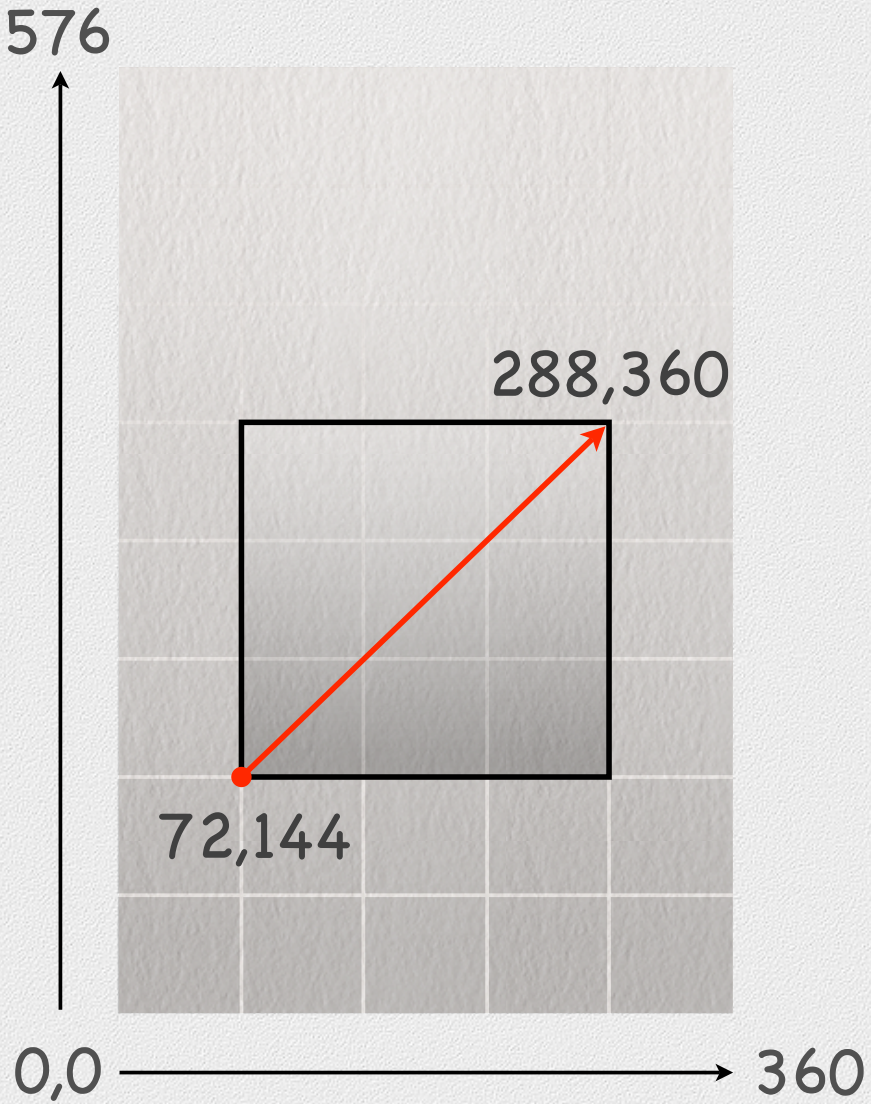
Drawing starts from 0,0

Move to 72x and 144y

Draw Line 288,144
Draw Line 288,360
Draw Line 72,360
Draw Line 72,144

Finally Stroke the path.

Basic Drawing 2



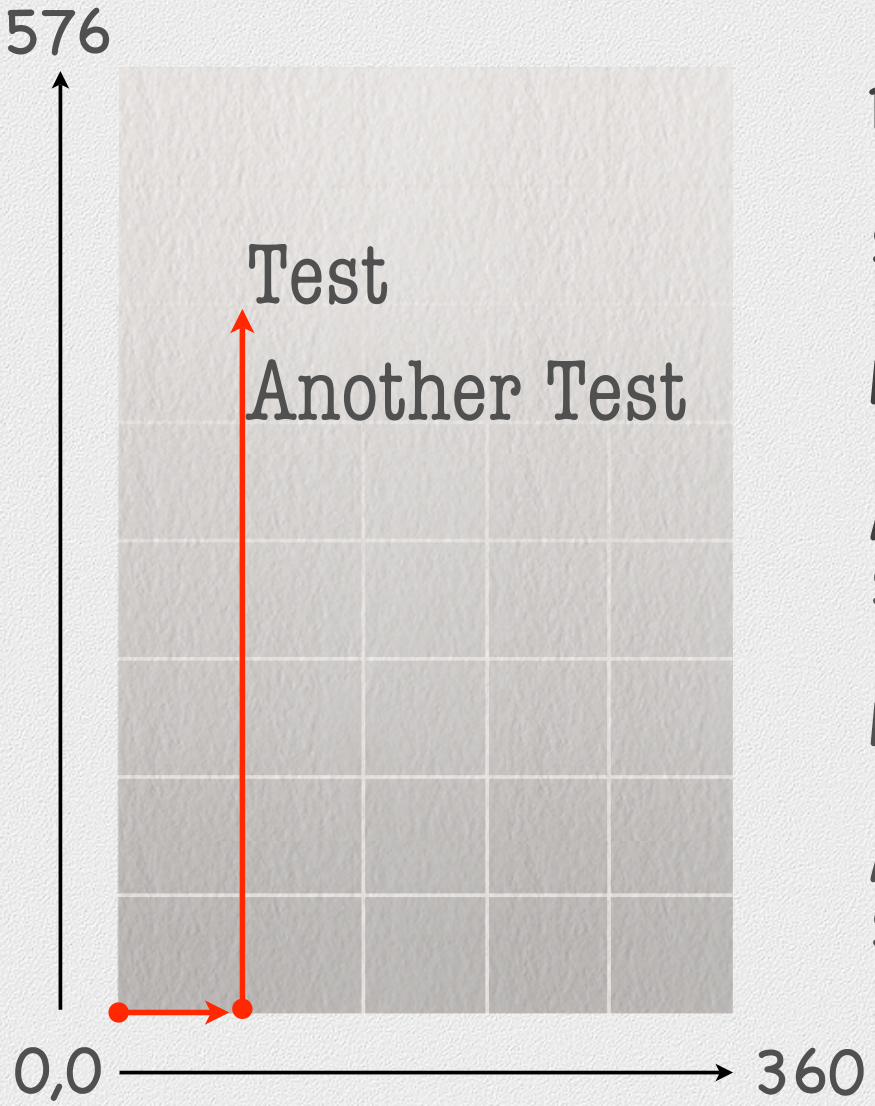
Drawing starts from 0,0

Define the Top, Left and Bottom, Right corners of a rectangle

Finally Stroke the path.

Two different ways to represent exactly the same visual information on the page.

Drawing Text



Drawing starts from 0,0

Set the Font

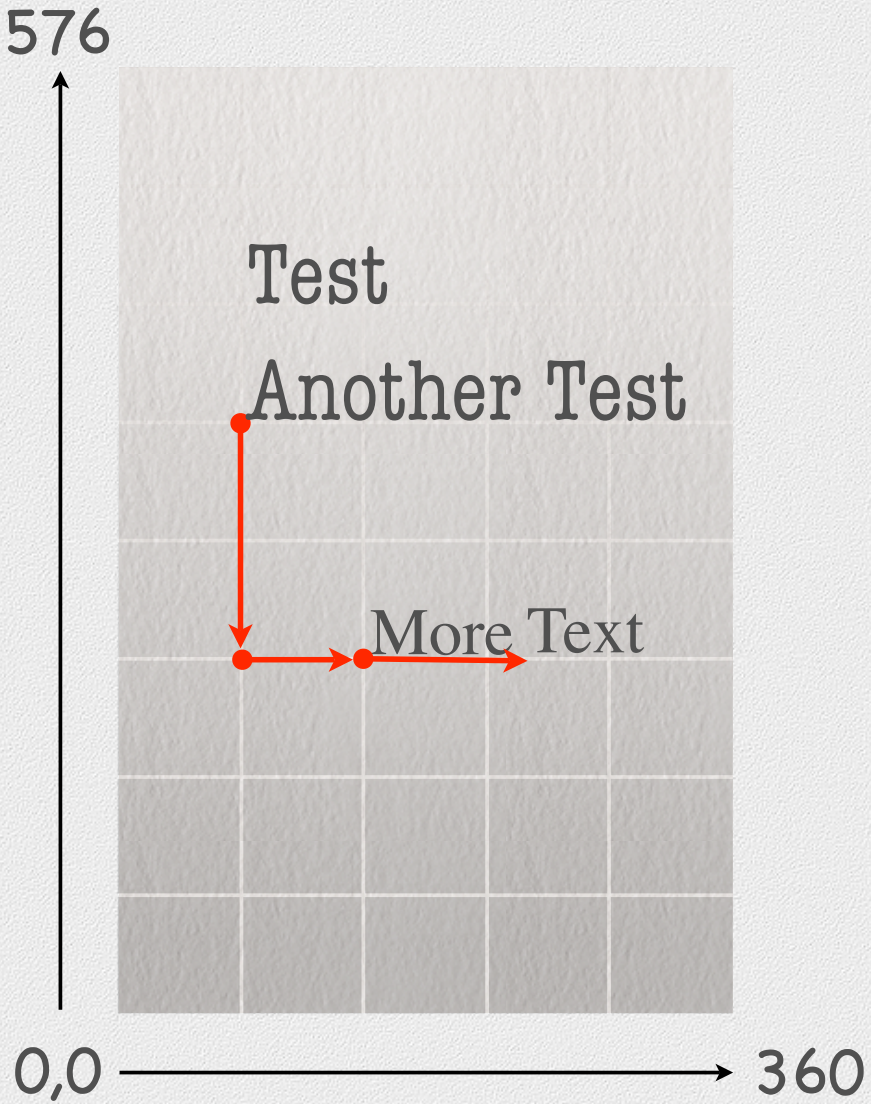
Move to 72x and 432y

Add Text (Test)
Show Text

Move to next line

Add Text (Another Test)
Show Text

Drawing More Text



Drawing starts from the beginning of the last line of text

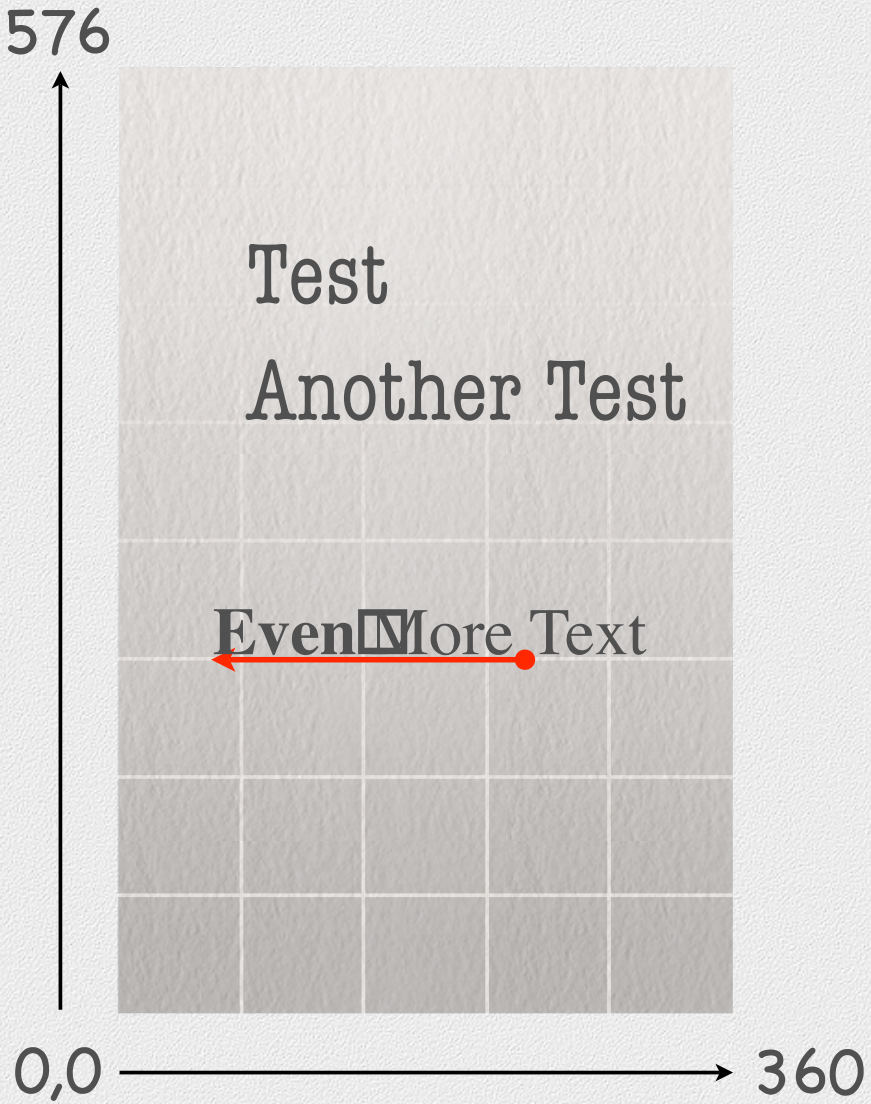
Move +x and -y

Set the Font & Size

Add Text (More)
Show Text

Move +x
Add Text (Text)
Show Text

Even More Text



Drawing starts from the beginning of the last line of text

Move -x

Set the Font & Size

Add Text (Even)
Show Text

Content stream
MoreTextEven

Not Runs of Text

1. Information is spread all about in a PDF file. There is no way to easily determine the text reading order.
2. PDF has no concept of a carriage return or line break.
3. Words may or may not be separated by space characters.
4. Individual characters can simply be thought of as small pictures placed on a drawing page.
5. Characters do not have to be encoded using any known character encoding.

PDF Reference

The Document

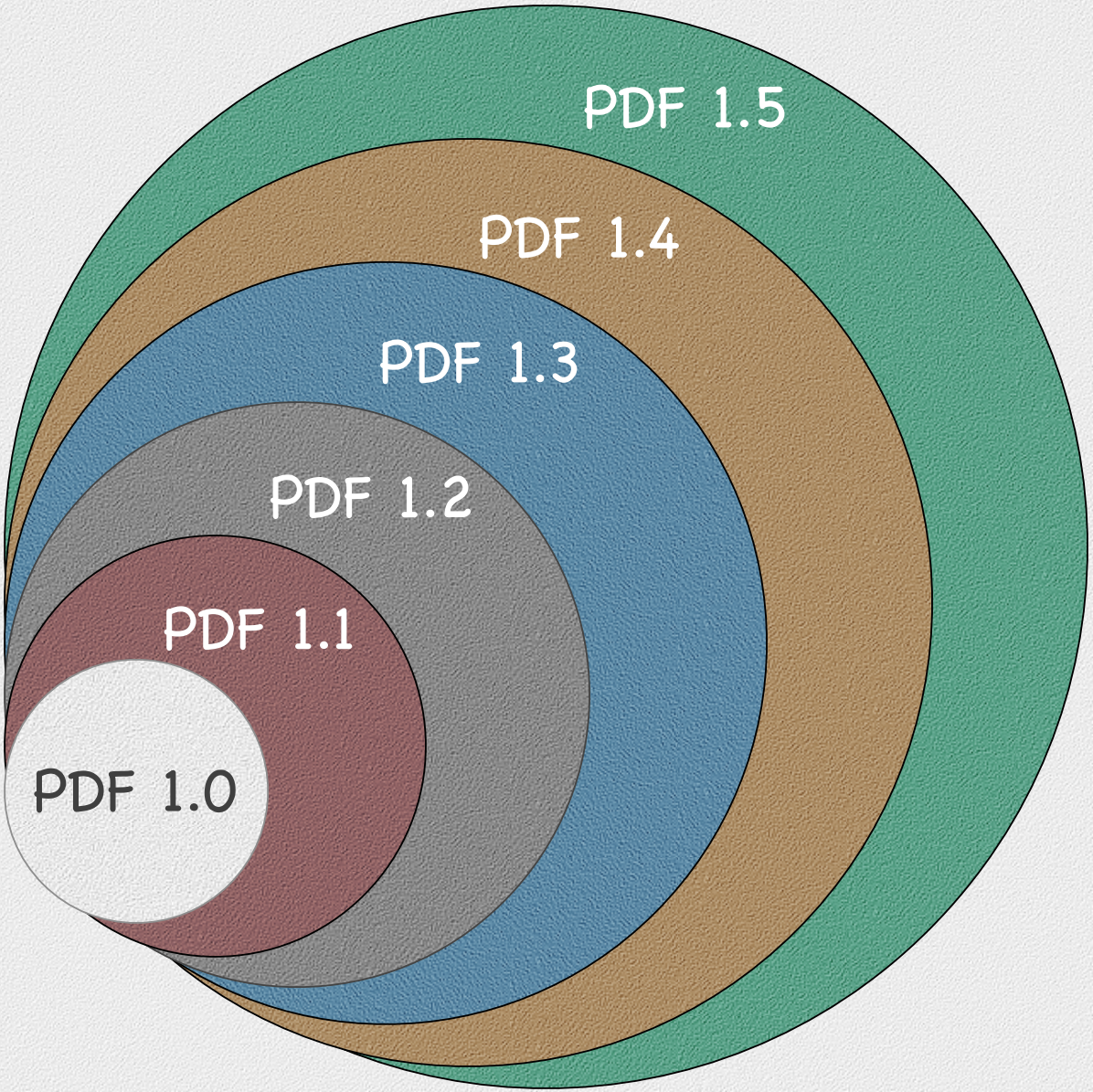
&

What it Contains

PDF Reference

Version	Year	Pages	Printed
PDF 1.0	6/1993	214	Yes
PDF 1.1	4/1995	50	No
PDF 1.2	11/1996	394	No
PDF 1.3	7/2000	679	Yes
PDF 1.4	12/2001	945	Yes
PDF 1.5	≈8/2003	1149	No
PDF 1.6	1/2005	1213	?

PDF Versions



Contents

Syntax	Everything about the syntax of PDF at the object, file, and document level.
Graphics	Graphics Operators used in PDF content streams to describe the appearance of pages.
Text	Representing characters with glyphs from fonts.
Rendering	Describes the facilities for controlling how shapes and colors are rendered on the output device.
Transparency	Extends the Adobe imaging model to include the notion of transparency. (PDF 1.4)
Interactive Features	Allow a user to interactwith a document on the screen, using the mouse and keyboard.
Multimedia	Support for embedding and playing multimedia content.
Document Interchange	Higher-level information that is useful for the interchange of documents among applications.

Document Interchange

Metadata Streams	PDF 1.4
Marked Content	PDF 1.2
Logical Structure	PDF 1.3
Tagged PDF	PDF 1.4
Accessibility Support	PDF 1.4

Accessibility support has been added over the years and is very dependent on Marked Content, Logical Structure and Tagged PDF.

Metadata and Marked Content

1. Metadata
    a. Document Information Dictionary
    b. Metadata Streams - flexible XML metadata container
2. Marked Content - tag a piece of content
    a. Property Lists - private application data
    b. Marked Content and Clipping - some path and text objects are only used to clip other objects.

Logical Structure

1. Structure Hierarchy
2. Structure Types
3. Structure Content
    a. Marked-Content Sequences as Content Items
    b. PDF Objects as Content Items
    c. Finding Structure Elements from Content Items
4. Structure Attributes

Logical Structure

1. Organize Document Contents
    a. Chapters and Sections
    b. Figures and Tables
    c. Footnotes and Bibliography
    d. Table of Contents, Index and Colophon
2. Stored Separate from the Visible Content
    a. Logical Structure Independent of the order and location of graphics content on the page

Tagged PDF

1. Tagged PDF and Page Content
    a. Page Content Artifacts
    b. Page Content Order
    c. Extraction of Character Properties
    d. Identifying Word Breaks
2. Basic Layout Model
3. Standard Structure Types
4. Standard Structure Attributes

Tagged PDF

1. Text Extraction
2. Reflow of Text and Graphics
3. Searching, Indexing and Spell-Checking
4. Conversion to other file formats
5. Accessibility

Accessibility Support

1. Natural Language Specification
    a. Language Identifiers
    b. Language Specification Hierarchy
    c. Multi-Language Text Arrays
2. Alternate Descriptions
3. Replacement Text
4. Expansion of Abbreviations and Acronyms

Accessibility Support

1. Specify the Language used in the document.
2. Text Descriptions for Images.
3. Replacement text for Ligatures and Illuminated characters.
4. Expand Abbreviations and Acronyms.
5. Enable Proper Vocalization by Screen Readers or Text-to-Speech Engines.

Accessibility Support

"The core of this support lies in the ability to determine the logical order of content in a PDF document, independently of the content's appearance or layout, through logical structure and Tagged PDF..."

PDF Reference 1.6 Section 10.8

Build Better PDF Solutions