White Paper

PDF Primer

From the Understanding PDF White Papers

PDF Tools AG

- **What is PDF and what is it good for?**
- **How does PDF manage content?**
- **How is a PDF file structured?**
- **What are its capabilities?**
- **What are its limitations?**

pdf-tools.com

Premium PDF Technology

Version: 1.0

Date: October 6, 2005

Contents

Overview

Introduction

PDF files are prevalent in virtually all market segments worldwide. Most people know and use the term "PDF" but have misconceptions what it can and can't do. The purpose of this White Paper is to explain what PDF actually is and what it's strengths and limitations are.

What does PDF stand for?

PDF is an abbreviation for "Portable Document Format". PDF is a proprietary format of Adobe Systems Inc.

"Portable Document Format" from Adobe Systems Inc.

Why was PDF designed?

PDF was created in the early 1990's as a new platform independent file format with the following goals:

- Exchange and view electronic documents
- Represent text and graphics in a resolution independent manner
- Optimize documents for (web) viewing
- Enhance with interactive features

Why is PDF attractive?

PDF is attractive as an electronic document format for a variety of reasons:

- **Portability** - PDF is platform independent, e.g. a PDF file created in a Windows application can be subsequently processed on a UNIX server and then viewed on a Macintosh computer.
- **Electronic Document with added Features** - PDF builds on the very successful PostScript page description language by adding many features such as random access, compression, encryption and interactive navigation features to PostScript's underlying imaging model.
- **Industry Standard** – PDF has become the de-facto standard for the electronic exchange of documents. In addition, PDF is now the industry standard for the representation of printed material in electronic prepress systems. Private corporations, government agencies, and educational institutions are redesigning their business processes by replacing paper-based workflows with an electronic exchange of information.
- **Free Viewer -** One main reason why PDF managed to expand so quickly in the market is because Adobe's PDF reader has been available at no cost, virtually since PDF format was introduced. Only their PDF creation and manipulation applications must be purchased.

PDF is portable over different platforms.

It is enhanced with many features and has become a de-facto standard.

Features

PDF Features

PDF offers several features that make it so versatile and in some cases unique:

- **Graphics separated from rendering** - PDF separates graphics (shapes and colors) from rendering (raster output device). The appearance of pages is specified in the PDF file in a device-independent way. Rendering the pages (e.g. for viewing or printing) can be optimized based on the output devices' specific characteristics.
- **Compression** - PDF objects, especially images, can be highly compressed with different compression algorithms without a visible loss of quality. A PDF file can be a fraction the size of the original file.
- **Font Management** - All fonts used in a PDF file can be embedded in the file, guaranteeing that the text will look exactly the same when the file is reproduced. To save space fonts can be subsetted, i.e. the fonts only contain those parts that are really needed.
- **Single Pass File Generation** - When creating a PDF file, the physical order of the objects in the file is irrelevant, so that the objects do not need to be first organized in a preliminary processing step. This makes it possible to generate a PDF file in one single processing action.
- **Random Access** - PDF files can be randomly accessed. For example, if you want to view page 733 in a 800-page document, PDF can identify and load the objects needed to display page 733 first. You do not have wait for the entire file to load before the page can be viewed.
- **Security** (Encryption, Digital Signatures) - PDF supports different levels of encryption, access control, and digital signatures. This makes it attractive for processing sensitive documents that are sent over the internet or used in web browser applications. PDF documents can be encrypted such that their contents cannot be reconstructed without knowing the password.
- **Incremental Update** - If you append a PDF document, the changed objects are simply added to the end of the PDF file. The entire PDF file is not regenerated from scratch. Small amendments (e.g. adding a watermark or a text correction) can be easily made with a very short processing time. Larger changes can however lead to larger file sizes.
- **Extensibility (Document Interchange)** - PDF files contain numerous features that do not affect the final appearance of a document, but are useful for the interchange of documents among application. The inclusion of metadata in the file (e.g. title, author, creation date, modification date etc.) and file identifiers (for reliable reference from one PDF file to another) are two such examples.

Rendering a PDF file can be optimized based on the output devices' specific characteristics.

PDF can be highly compressed and offers a wide spectrum of security features.

PDF information can be randomly accessed.

Updates are incremental - the PDF file does not have to be regenerated from scratch.

PDF Page Content

Page Description Language

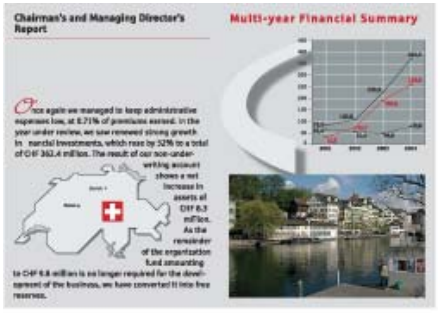
PDF is a page description language, i.e. it describes how a page looks so that it can be reproduced for viewing and printing. The language resembles Postscript, but is much simpler to allow for more efficient processing. For example, it does not contain control structures like loops and "if" statements.

PDF Page Content Elements

Basically PDF recognizes three types of page content elements:

- Text (fonts programs)
- Graphic paths (lines and curves)
- Images (raster samples)

The picture to the right shows examples of the three PDF content types.



Content Objects

PDF uses objects and object types to describe the content. Every string of text and all graphics and images are defined by one or several objects, created from one or more object types.

- Text Objects. Text objects are defined by a number of attributes including font and font size, a string of characters, and a position on a page. PDF does not recognize nor store objects for line breaks, headers, paragraphs, indentation etc. (i.e. paragraph formatting operators used in word processing applications like Microsoft Word). Text is broken down into fragments as small as single characters but not more than one line. The fragments can be randomly stored and are like pieces of a puzzle that all have to be placed in their correct location on the page to complete its appearance.
- Graphic Path Objects. A graphic path object is an arbitrary shape made up of straight lines, rectangles, and cubic Bézier curves. A graphic path object ends with one or more painting operators that specify whether the path is stroked, filled, used as a clipping boundary or some combination of these operations.
- PDF Image Objects. A PDF-specific image format is used for embedding images in a PDF file. This format is independent of the input image format. For example, scanned pages in TIFF format or GIF images that are converted to PDF are newly packaged into PDF image format. Once an image has been converted to PDF image format, it is usually not possible to determine what the original image format was. It is however possible to export PDF images into raster image formats, provided the raster image format supports all features of the image (e.g. transparency).

Unlike word processors, text is not continuous in a paragraph.

Text is defined line- or character-wise by attributes, e.g. font and font size, a string of characters and a location on a page.

Images that are imported into a PDF file are converted to PDF image format. They are not stored as TIFF, GIF etc. images.

PDF File Structure

Logical vs. Physical File Structure

So what does a PDF file look like? To answer this question, we have to differentiate between the logical document structure and the physical structure of a PDF file.

Logical Document Structure

The logical structure of a PDF document refers to how the document is reproduced for viewing or printing:

Page 1

object 1

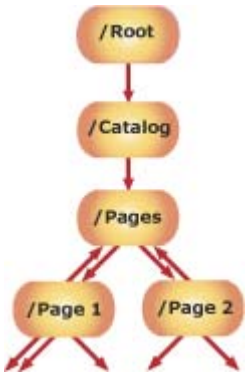
object 2

Page 2

object 3

object 4

Page 3 etc..



Logical PDF Document Structure

The logical document structure differs greatly from the physical PDF file structure.

Physical File Structure

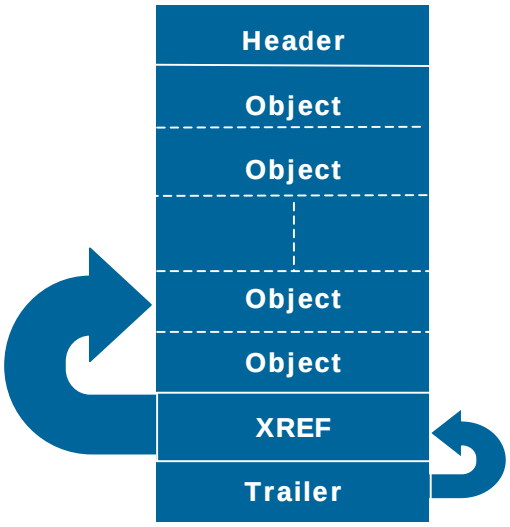
The physical structure of a PDF file is quite different from it's logical document structure. A PDF file consists of a header, objects, a cross-reference table, and a trailer.

The objects on a page are not stored in any particular order (e.g. top to bottom). Their location in the physical file is usually completely random.

Accessing a PDF File

When accessing a PDF file (e.g. to render for viewing or printing), the header is read first. This identifies the file as PDF. Next the trailer is read. The trailer points to the cross-reference table, which then points to the objects containing pages, fonts, text, images etc.

Pages are rendered by randomly retrieving all of the objects required on that page and displaying them according to their x and y coordinates.



Physical PDF File Structure

Objects are stored in random order.

The trailer and the cross reference table are needed before a PDF page can be accessed.

PDF Capabilities

Strengths

There are a number of key strengths to PDF that have helped it become so popular in such a short time, and to remain popular despite the speed at which the IT world is evolving.

- PDF is a de-facto standard. It is used throughout the world thanks to its portability, comparatively small file size, and the availability of free viewers.
- PDF was designed in part for the internet and, with the explosive expansion of the internet during the past 10 years, PDF has established itself as the format of choice for documents available on-line.
- Multi-page PDF files can be optimized for fast web viewing. A PDF file can start by loading a specific page, for example page 6537 in a 10'000 page document. Page 6537 will then be displayed first, before the preceding 6536 pages are loaded by the viewer.
- PDF files can be compressed to extremely small sizes. Images in particular can be greatly optimized.
- PDF readers can still be obtained for free, unless you want to integrate a PDF viewer into an application.
- PDF supports multi-page documents in all paper sizes.
- Adobe has created an excellent PDF Specification. It goes into great depth and contains explicit details on all aspects of PDF. This makes it possible for software development companies to create their own PDF programming tools.
- There are a large number of application opportunities with PDF. Tools are available for creating and manipulating PDF files in numerous manners: on-the-fly document generation, merge & split, stamp, extract content, encrypt, convert, view, print, form-filling ...
- And finally, PDF has evolved into more than just a document format. Advanced features and modern technologies like multimedia, JavaScript, XML, forms processing, compression, custom encryption etc. can be used with or embedded into a PDF file, making PDF a powerful, interactive and intelligent file format.

- De-facto standard
- Designed for Web
- Free Reader
- Multi-page documents
- Numerous developer tools
- Newest technology

PDF Limitations

Weaknesses

PDF also has its weaknesses and limitations.

- PDF is still the intellectual property of Adobe. Adobe writes the specification and decides when new releases and new functions are going to be added.
- New versions are coming out more rapidly. Each new version brings not only new features but unfortunately possible incompatibilities with older versions. The reduced time between version releases is beginning to compound this problem. For example, if you are still using Adobe® Acrobat 4, you won't be able to open a lot of PDF documents that are being optimized with the current Acrobat version.
- PDF contains some design problems, most notably with mixing syntax and semantics. The design errors are a potential problem for the future as new features and technologies are integrated into PDF.
- PDF is beginning to offer too many 'foreign' formats/technologies for embedded objects (PostScript, Fonts, XML, etc.). Each of these formats can cause corruption and undesirable / unexpected effects.
- PDF is not necessarily WYSIWYG (what you see is what you get). This is particularly true in the areas of colors and fonts. PDF files may look a lot different than the presentation in their original (non-PDF) document format, although there are subsets of the PDF Specification, such as PDF/X, which claim to guarantee an exact replica.
- PDF is not easy to process due to certain design issues, and includes a huge number of technologies which have to be mastered. A deep understanding of PDF technology is required for developing quality solutions.
- PDF doesn't recognize paragraphs, formatting, headers, footers, indentations, broken words (line-breaks) etc. This makes it difficult to convert a PDF file back into a formatted Microsoft Word file, for example. Comparing PDF files is also especially challenging due to text being stored in fragments on a page, and not sequentially or as part of a sentence or paragraph.

- Adobe proprietary
- New versions mean new incompatibilities
- Inherent design issues
- Complicated technologies
- Not WYSIWYG
- Paragraph formatting not stored

Summary

Where to go from here?

The goal of this White Paper was to present an initial introduction into the world of PDF.

Hopefully you now have a better understanding of what PDF is, how it is structured, and what some of the main strengths and weaknesses are.

Further White Papers

If you would like to read more about specific PDF technologies, there are numerous web portals and white papers that could help you out. PDF Tools AG (http://www.pdf-tools.com) is publishing a complete series of White Papers dealing with a variety of PDF technologies. The topics being covered include:

- PDF Color and Color Spaces
- PDF Text and Fonts
- PDF Encryption and Digital Signatures
- PDF Compression and Optimization
- PDF Validation, Repair and Recovery

White papers can help you further your knowledge on PDF.

Numerous companies offer courses and in-house training sessions.

These White Papers are or will be published on our website as well as on a number of other PDF and white paper portals.

Training Sessions

You also might be interested in attending a course or workshop, or having someone visit your business to conduct in-house training. Several companies and organisations offer courses and training sessions on PDF technology. They can be easily found by searching the web or by visiting a PDF-dedicated portal.

PDF Conferences

There are also PDF-specific and document management conferences with special PDF tracks aimed at a variety of user groups, e.g. end-users, programmers, consultants, integrators. These conferences offer an opportunity to gain valuable information about PDF in a short time, and are often coupled together with PDF workshops and tutorials.

PDF conferences offer an opportunity to gain valuable information about PDF and often include workshops or tutorials.