Action Plan Background: PDF 1.7

Author: Carol C.H. Chou, FCLA
Release Date: 1/26/2009

Preface

PDF is a general document representation language which provides platform independent and device independent document representation. It uses the imaging model of the once popular PostScript page description language as the underlying model.

1 General Description

1.1 Format Name: PDF (Adobe Portable Document Format)

1.2 Version: 1.7

1.3 MIME media type name: application

1.4 MIME subtype: pdf

1.5 Short Description: a page description language; the native file format of Adobe Acrobat 8.0.

1.6 Common Extensions: .pdf

1.7 Color depth: 1, 2, 4, 8, or 16 bit (produces a maximum of 2, 4, 16, 256 and 65536 colors respectively per color component).

1.8 Color Space: Eleven color spaces including:

- four device-independent CIE-based: CalGray, CalRGB, Lab, and ICCBased.
- three device-dependent color spaces: DeviceRGB, DeviceCMYK, and DeviceGray
- four special color spaces: Pattern, Indexed, Separation, and DeviceN

1.9 Compression: PDF supports various compression 'filters' including:

- JPEG and JPEG 2000 compression for color and grayscale images

- CCITT (Group3 or Group 4), JBIG2 and run-length compression for monochrome images
- LZW (Lempel-Ziv-Welch) and Flate compression for text, graphics and images

1.10 Progressive Display: can stream over a network if the PDF is linearized

1.11 Animation: no (uses other applications to play movies)

1.12 Magic number(s): %PDF-1.7

1.13 Latin-text Encoding: Can be Unicode, PDFDocEncoding (see 4.4.1 of [3]), MacRomanEncoding, MacExpertEncoding, WinAnsiEncoding, StandardEncoding

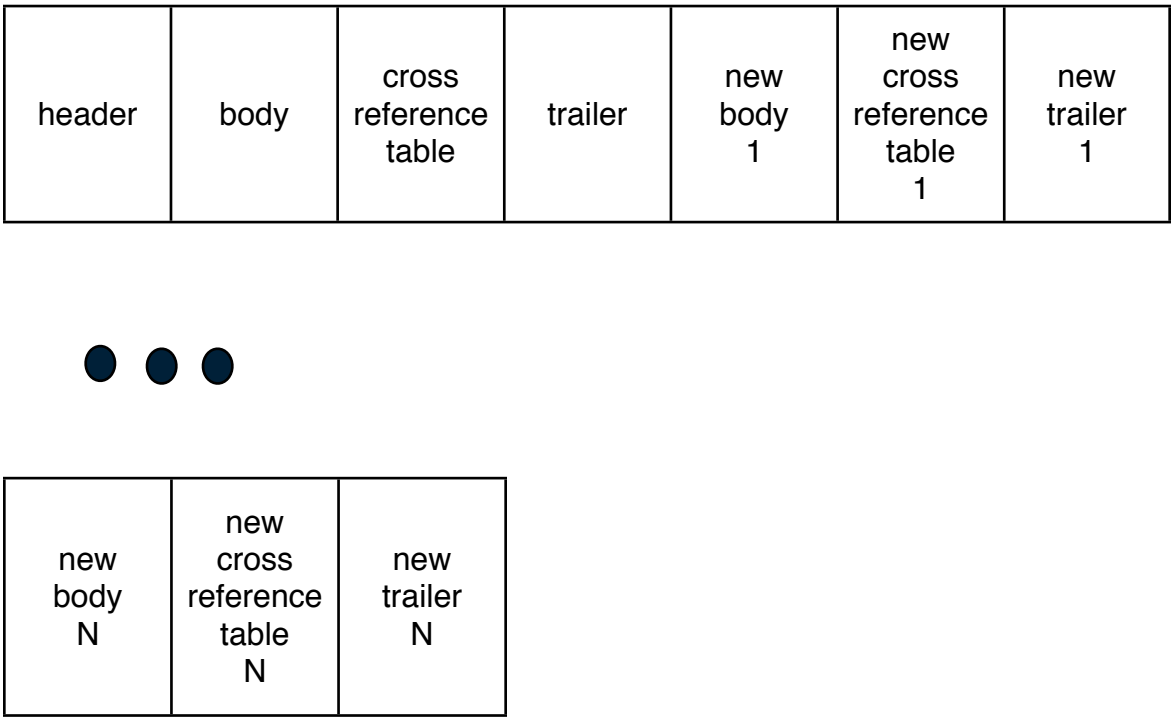
1.14 Specification Requirements

A PDF file is a collection of numbered objects that can refer to each other by their numbers [King 2004]. A PDF file is initially comprised of the following four elements:

- header: a one-line header identifying the version of the PDF specification the document conforms (the magic number). The header must appear in the first 1024 bytes of the PDF file.
- body: containing those objects that constitute the PDF document. There are five basic object types: number (integer and fractional), boolean, string, name and null, as well as three compound object types: arrays, dictionaries and streams.
- cross-reference table: providing random access to numbered objects within a PDF file such that the entire file need not be read in order to access a particular object. The cross-reference table specifies the exact byte offset within the file for each numbered object. Therefore, after the cross-reference table has been read, any object can be read directly from physical device without reading other information.
- trailer: specifying the location of the cross-reference table and certain special objects within the PDF file body. This allows applications to quickly find the cross-reference table. Hence, PDF applications should read a PDF file from the end. The trailer section ends with the end of file marker "%%EOF" which should appear in the last 1024 bytes of the file.

The initial structure could be subsequently modified by appending additional elements to the end of the file. For each subsequent update, any new or changed objects are added in the new body section with a new cross-reference table and a new trailer appended to the end of the file. This

incremental update feature makes it possible to update the document without rewriting the entire file.



In addition to the structural requirements, a PDF file is required to include the character metrics for all fonts used in the document. This allows for font substitution that maintains the correct character spacing.

2 Contents and Features

2.1 Essential and Distinguishing Characteristics

PDF is a highly structured page description language based on PostScript, as well as a binary file format (PDF 1.1 and later). A PDF document is a hierarchy of 'objects' - text, pages, forms, images, sounds, movies, annotation, scripts, and higher-level application data in a platform-independent, device independent and resolution-independent file. The content and layout of the document is specified as part of the file, but not the logical/semantic structure.

PDF 1.7 adds several enhancements on 3D graphic model which was introduced in PDF 1.6.

2.2 Internal Technical Metadata

Technical metadata element (G = general file metadata, F = formatspecific metadata)	***Obligation (R = required by spec., S= Information given by spec., O = Optional but described in spec., X = described by publication external to spec.)***
specification version [G]	R
linearization [F]	O
encryption (is this document encrypted and/ or password protected?) [F]	O
use thumbnail image [F]	O
contain image [F]	O
use embedded font [F]	O
End of File Marker [G]	R
cross reference table [F]	R
catalog dictionary of the document [F]	R
number of pages [F]	O
page layout [F]	O
page mode [F]	O
outline[F]	O
document open actions [F]	O
tagged PDF [F]	O
natural language [F]	O
XML metadata [F]	O
document information dictionary [F]	O
producer [F]	O
creation date [F]	O
modification date [F]	O
compression filters used in each page [F]	O
number of images in each page [F]	O

annotations associated with the page [F]	O
trigger events (actions to be performed when the page is opened or closed) [F]	O

****TODO: add extension metadata**

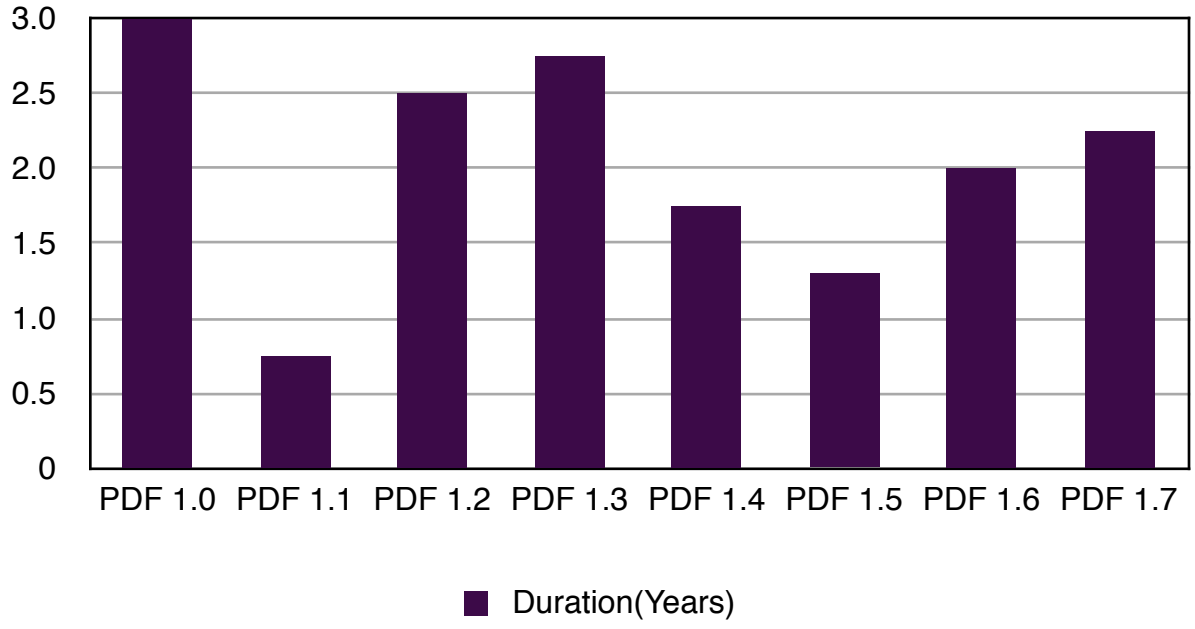
3 Usefulness

3.1 Version Duration: 5 months and continuing

3.2 History of Prior Versions Duration:

- - 1993: PDF 1.0
- March 1996: PDF 1.1, it is the native format of Adobe Acrobat 2.0. It added the support for Encryption, Articles, Transitions, Calibrated Color, improved Hyperlinks and Actions.
- November 1996: PDF 1.2, the native format for Acrobat 3.0. It added the supports for Prepress features, External Streams, Flate and LZW compression, Forms, Asian Fonts, more Annotations and linearization.
- March1999: PDF 1.3, PDF 1.3 added support for PostScript 3 imaging model, ICC color, LogicalStructure, JavaScript, more Annotations, Digital Signatures and embedded file streams.
- December 2001: PDF 1.4, PDF 1.4 added the support for Transparency, 128-bit encryption, XML metadata, Tagged PDF, embedded file streams, links to external PDFs and enhancements on form features.
- August 2003: PDF 1.5, the native format for Acrobat 6.0. PDF 1.5 added the support for JPEG 2000 compression and Object stream compression, 16 bit image support, crypt filter (encryption on the stream level), ability to selectively view or hide content in a PDF file (Option Content), slide show presentation, and enhancements on digital signature, interactive forms and multimedia playback and embedding.
- November 2004: PDF 1.6, the native format for Acrobat 7.0. PDF 1.6 added the ability to embed 3D artwork in PDF files (for AUTOCAD and VISIO software), support for AES encryption algorithms, measurement property, user property, watermark annotation for graphic to be printed on a fixed size and position on a page, embedding open-type font and enhancements on markup annotations, digital signature and embedded file attachment.
- November 2006: PDF 1.7, the native file format for Acrobat 8.0. PDF 1.7 includes many enhancements on 3D artwork, additional markup annotations to support additional interactive features, additional tags for tagged PDF, and enhancements on digital signatures.
- 2008: PDF 1.7, Adobe Extension Level 3, the native file format for Acrobat 9.0 and Adobe LiveCycle ES 8.2.

Figure 1: Duration in Years of PDF Versions



3.3 Expected Newer Versions:
Historically, the release of the new version of PDF specification has coincided with the new release of Adobe Acrobat.

In July 2008, PDF 1.7 becomes an ISO standard (ISO 320000-1) [4]. ISO 320000-1 is equivalent to PDF 1.7 specification. According to Adobe, future versions of PDF specification will be published by ISO. Hence, Adobe does not produce PDF 1.8 specification. New PDF features for Acrobat 9.0 are submitted as extensions to the ISO 320000. PDFs containing new extended features still carry PDF version number 1.7, with extension level indicating which extension the PDF producer followed. There are currently three extension levels defined in ISO 320000. Extension Level 3 describes the extensions used for Adobe 9.0 and LiveCycle ES 8.2[1].

3.4 Existence of Publicly Available Complete Specifications:
Adobe publishes PDF specifications on its website. Older PDF specifications such as 1.2 and earlier are not available on its website. Some PDF specifications (version 1.3, 1.4 and 1.6) are

[1] http://www.adobe.com/devnet/pdf/pdf_reference.html

also available in print. While each new specification attempts to incorporate information about older PDF versions, the older specifications provide additional useful information. Some of these older specifications can be located elsewhere on the Internet.

ISO 320000 can be purchased directly from ISO. Per agreement with ISO, Adobe also offer PDF 1.7 specification on Adobe website. The PDF 1.7 specification is technically identical to the ISO 320000 specification but it is not an official ISO document.

All the information needed to completely understand the PDF format is not completely contained in the PDF specifications. There are numerous supplementary specifications that are relevant, many of which are available on Adobe's website. These include but are not limited to the specifications for digital signatures, encryption algorithm, JavaScript objects, fonts, compression algorithms, character encoding, color management and the PostScript language.

3.5 Specifications-controlling Body: Adobe Systems Incorporated(1.1-1.6); ISO (1.7 and later)

3.6 Related Legal Issues:
Adobe owns the copyright to the PDF format and the data structures and operators contained in the PDF format. Adobe gives anyone copyright permission to create software that reads or writes PDF files in a way compliant with the PDF specification.

Adobe owns many patents on PDF. To encourage third party implementation on PDF, Adobe grants royalty-free, non-exclusive use of patent license to PDF implementations which are compliant with the specification .

Some of the encryption methods, digital signatures, and compression algorithms that are supported by PDF are copyrighted. Software that decrypt or encrypt any of these algorithms may be subject to other technology licensing.

3.7 Application and Platform Support:
PDF 1.7 is the native format of Adobe Acrobat 8 and Adobe family products. In addition, there are many third party PDF readers that can read and create PDF 1.7. A search on planet PDF uncovers 867 applications for viewing, creating and manipulating PDFs. Many of these applications are open source software. In addition, OSX has built-in PDF printing capabilities. Many popular applications such as OpenOffice, Microsoft Office, Google Doc, etc also support PDF.

Adobe provides free PDF readers for various desktop and mobile operating systems including Windows (9X/XP/2K/Vista), Macintosh (Mac OS 10.2-10.5), Linux, IBM AIX, Solaris (9/10), HP-UX, OS/2 Warp, Palm OS, Pocket PC, Symbian OS, etc. Adobe does not always provide new PDF readers on every supported platform. Acrobat Readers support for operation systems other than Windows and Macintosh has been lagging. As an example, Adobe skipped the Linux support for Acrobat 6 and then added the Linux support for Acrobat 7.0 and 8.0, after it saw a wider adoption on Linux system [Millard]. Acrobat Readers support for HP-UX, AIX, Solaris is still only available with version 7.0 [2].

When the PDF major number is incremented (ex: PDF 1.7 is superseded by PDF 2.0), existing PDF readers will be unlikely to read the newer files. When the minor number changes (ex: PDF 1.6 is superseded by PDF 1.7), a PDF viewer designed to read PDF 1.6s should still be able to read PDF 1.7s. PDF readers are supposed to ignore codes that they don't understand. Essentially this prevents older PDF readers from becoming obsolete as new PDF versions emerge.

3.8 Limitations:

The only limitation of a PDF file, due to its design, is that file size is limited to 10^10 bytes (approx. 10GB) because 10 digits are allocated to byte offsets. There are other limitations posed by PDF readers, such as the number of dictionary entries or length of a name object. See Table C. 1 [7] for details.

3.9 Perceived Popularity:

PDFs have become the de facto standard for electronic exchange of documents, especially those containing text and images. The Adobe Acrobat Reader is freely downloadable from the WWW. It also works as a web browser plug-in. To date, over 700 million copies of Adobe Acrobat Reader have been distributed from Adobe [1]. PDF has become so ubiquitous that the majority of US government agencies have standardized on the PDF format [2]. FDA (Food and Drug Administration) makes PDF the required format for submitting drug approval and U.S. Federal Court requests electronic case filing to be in PDF format [6]. Internal Revenue Service also uses PDF to distribute tax forms.

3.10 Market Competitions

Since 2005, Microsoft has developed its own fixed document format "XML Paper Specification (XPS)". XPS is essentially a ZIP-archive containing XMLs describing the document layouts, textual content, 2-D graphics, fonts, etc. Just like PDF, XPS preserves the appearance of the

---

[2] "PDF Reference and Adobe Extensions to the PDF Specification", http://www.adobe.com/products/reader/systemreqs/#90mac

documents regardless of the device rendering the documents. Unlike PDF which is based on the legacy Postscript format, XPS is XML-based. Since XPS serves as a final document product and offers many functionalities that PDF offers, this puts XPS in direct competition with PDF.

Microsoft has since built XPS into its own Windows Vista operation system and Internet Explorer 7. In July 2007, Microsoft submitted XPS specification to ECMA International to be published as ECMA standard. XPS is currently under reviewed by ECMA technical committee (TC45) with another draft coming up in January 2009 [9]. It is likely XPS may follow the trail of OOXML to eventually become an ISO standard.

Whether XPS will replace PDF as a document printing format will need to be closely monitored.

4 Related Formats

4.1 Specification Variations:

PDF-Archive or PDF/A (Portable Document Format / Archive)
PDF/A was initiated by the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES), the Association for Information and Image Management, International (AIIM International), and the administrative office of U.S. Courts. The purpose is to develop an international standard that defines the use of PDF for archiving and preserving documents. The first part of PDF/A, PDF/A-1 (ISO 19005-1), is based on PDF 1.4 specification. PDF/A-1 is essentially a subset of PDF 1.4, leaving out things deemed to hinder document preservation such as linearization, encryption, device dependent color space, scripting, pointers to external content and embedded multimedia. In December 2005, ISO approved and published the PDF/A-1 (ISO 19005-1) standard.

A new part to the PDF/A, PDF/A-2 (ISO 19005-1, Part-2), is currently under development by the ISO Technical Committee, TC 171. PDF/A-2 will address those features added since PDF 1.4 including those features in PDF 1.5, PDF 1.6 and PDF 1.7 [5].

PDF/X family (PDF/X-1, PDF/X-1a. PDF/X-2, PDF/X-3, PDF/X-4)
These are specifications based on PDF tailored for the print publishing industry. They are subsets of PDF specifications. PDF/X family formats eliminate those elements that make it hard to prepare documents for printing. The PDF/X family formats has been approved as ISO standard (ISO 15930).

PDF/E (PDF/E-1)
PDF engineering (PDF/E) working group was formed by AIIM (international authority on enterprise content management) and NPES (the association for supplier in printing, publishing and converting technology) to promote standard for the use of PDF format in engineering workflows. In 2007, PDF/E was published as an ISO standard (ISO 24517-1).

FDF (Forms Data Format)
FDFs is the file format used for interactive form data. It is a simplified version of PDFs and is described in PDF 1.2 and later specifications. They are used for exporting or importing form data, especially across a network. FDFs are based on PDFs - they have the same syntax, object types, and a similar file structure. However, it does not require to contain the cross-reference table and cannot be incrementally updated. It also has simpler document structure. They have the extension '.fdf'.

PJTF (Portable Job Ticket Format)
PJTF is an Adobe specification based on PDFs, that is used to specify the instructions and the location of the contents needed to execute a print job. They may be embedded in PDF files, or they may be in stand-alone files. If it is a standalone file, it will have the header %JTF-1.x where 1.x is replaced with a specific version number (1.2 if it conforms with the PDF 1.2 specification, for example).

XFDF (XML Forms Data Format)
XFDF is the XML version of Forms Data Format (FDF). It is a format used for representing form data and annotations in PDF documents. XFDF conforms to the XML standard and can be transformed to other formats by XML tools that support XLST. XML validation tools can be used to ensure that a XFDF file conforms to the XFDF Schema [8].

5 Summary and Conclusions

The PDF format is very popular, in part because PDF readers are so accessible. PDF readers are free of charge, the readers also integrate well with popular web browsers and they are available for most operating systems. PDFs tend to have smaller file sizes than images of the same content, and they can be linearized for network streaming. They have an enormous amount of supported features, including embedded multimedia, scripts, CAD drawing and many security features. While the increased feature set has sustained PDF's dominant role in document presentation, it inherently presents some challenges in digital archiving. Some features such as references to external resources, JavaScript, alternative content, etc are not suitable for long-term preservation. Hence, Florida Digital Archive implements a normalization strategy to hedge our bets on ensuring the long-term accessibility of PDFs in our archive.

PDF/A is perceived to be an ideal normalization format for PDF 1.X because documents comforming to PDF/A are considered to be archive friendly and do not require massive storage. Unfortunately, PDF to PDF/A-1 conversion tools are still in their nascent stages and are not yet ready for automation. Alternatively, Florida Digital Archive normalizes every PDF into TIFFs where each TIFF represents a page in the PDF. Though the created TIFFs are not stored due to current storage constraints, the PDF to TIFFs creation process can be invoked later on if any PDF in the archive becomes inaccesible. In the mean time, Florida Digital Archive will periodically monitor and assess PDF to PDF/A-1 conversion tools for possible future adoption for PDF to PDF/A normalization. In addition, Florida Digital Archive will also monitor the development of PDF/A-2 standardization for possible normalization on PDFs versioned 1.5 and after.

The increasingly popular XML poses as a possible successor format to PDF. Many commercial and open source office applications have adopted XML as the basis of their office formats. For example, OpenOffice uses XML as the foundation of its OpenDocument file format. It consists of a zip file containing XML files and stand-alone images. Microsoft also develops its own document format based on XML specification, Office Open XML, and has published its Office Open Document formats. Both OpenDocument and Office Open Document format specification have also published by ISO as ISO standards. XML has become one of the most popular formats used by many office software. XML-based document formats may be potential succesors for PDF when reader support for PDF becomes less prominent in the future.

8 References

[1] Adobe Systems Incorporated, “Adobe Company fact sheet”, 2008, (http://www.adobe.com/aboutadobe/pressroom/pdfs/fastfacts.pdf)

[2] Adobe System Incorporated, “Government Customer Stories”, http://www.adobe.com/government/customers.html

[3] Adobe Systems Incorporated, “Portable Job Ticket Format”, Technical Note #5620, Version 1.1, Adobe Developer Support, April 2, 1999.

[4] Adobe Systems Incorporated, “PDF Reference and Adobe Extensions to the PDF Specification”, http://www.adobe.com/devnet/pdf/pdf_reference.html

[5] PDF Tools AG, “PDF/A - The Basics”, January 2007, www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf

[6] Case Management/Electronic Case Files, http://www.uscourts.gov/cmecf/cmecf_about.html.

[7] Adobe Systems Incorporated, “PDF Reference: Adobe Portable Document Format Version 1.6”, November, 2004.

[8] Adobe Systems Incorporated, “XML Form Data Format Specification version 2.0”, July, 2003.

[9] TC46 - XML Paper Specification(XPS), http://www.ecma-international.org/memento/TC46-M.htm