Action Plan Background: PDF 1.3

Author: Andrea Goethals, FCLA
Last Revision Date: 3/10/03

1 General Description

1.1 Format Name: PDF (Adobe Portable Document Format)

1.2 Version: 1.3

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 4.0

1.6 Common Extensions: .pdf

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

1.8 Color Space: 11 color spaces: Device-independent CIE-based (CalGray, CalRGB, Lab, and ICCBased), device-dependent color spaces (DeviceRGB, DeviceCMYK, DeviceGray), special color spaces (Pattern, Indexed, Separation, and DeviceN)

1.9 Compression: Can use different compression 'filters' on different parts of the file: JPEG (for color and grayscale images), CCITT Group3 or Group 4 (for monochrome images), LZW (for text, graphics and images), Flate (for text, graphics and images), Run length encoding (for monochrome 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) or equivalent: %PDF-

1.13 Encoding: Can be Unicode, PDFDocEncoding (see 4.4.1 of [Adobe, 1999]), MacRomanEncoding, MacExpertEncoding, WinAnsiEncoding, StandardEncoding

2 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 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 1.3 supports optional advanced features such as embedding audio and video, 'linearizing' for network streaming (PDF 1.2 and later), additional colorspaces, RC4 encryption for strings and streams (proprietary algorithm) controlling access permissions, digital signatures (using any signature handler),forms, embedded file streams, embedded fonts, embedded XML metadata streams, and support for embedded JavaScript.

3 Usefulness

3.1 Version Duration: 2.75 years

3.2 History of Prior Versions Duration:
- 1993: PDF 1.0
- March 1996: PDF 1.1 (native format of Adobe Acrobat 2.0)(Added Encryption, Articles, Transitions, Calibrated Color, improved Hyperlinks, Actions)
- November 1996: PDF 1.2 (native format for Acrobat 3.0) (Added Prepress features, External Streams, Flate and LZW compression, Forms, Asian Fonts, more Annotations, linearization)
- March1999: PDF 1.3 (Added PostScript 3 imaging model, ICC color, Logical Structure, JavaScript, more Annotations, Digital Signatures, embedded file streams)

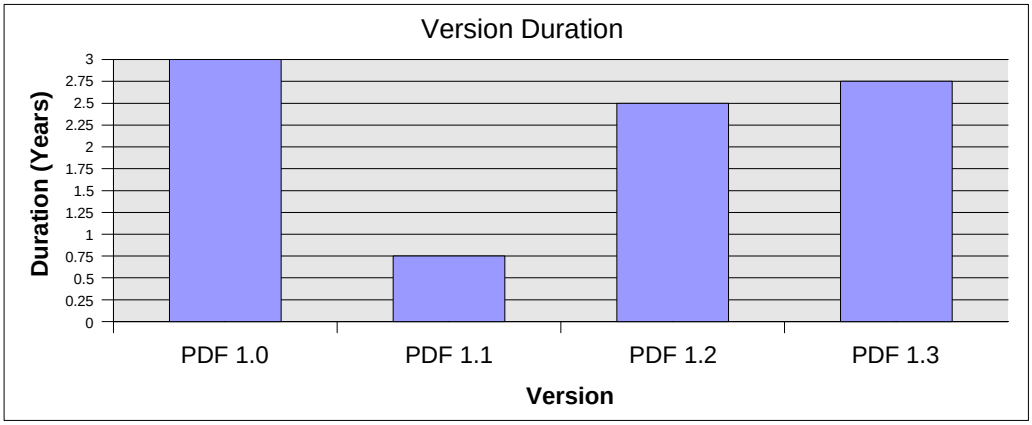

Figure 1: Duration in Years of PDF Versions

3.3 Expected Newer Versions: A newer version (PDF 1.4) became available in 2001.

3.4 Existence of Publicly Available Complete Specifications:

Adobe provides freely-downloadable PDF specifications on its website for the most recent PDF specifications. However, Adobe does not keep older PDF specifications (such as the PDF 1.3 specification) on its website. While each new specification attempts to incorporate information about older PDF versions, the older specifications provide additional and useful information. Some of these older specification versions can be located elsewhere on the Internet.

The information needed to completely understand PDFs are not self-contained in the PDF specification. There are numerous supplementary specifications that are relevant, many of which are available on Adobe's website. These include but are not limited to specifications for digital

signatures, JavaScript objects, fonts, compression algorithms, character encoding, color management and the PostScript language.

3.5 Specifications-controlling Body: Adobe Systems Incorporated

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.

[Adobe, 2001b] explicitly states that a requirement of the copyright permission stated above is that software that "accepts input in the form of the Portable Document Format must respect the access permissions specified in that document." These access permissions are maintained through encryption. The user has the ability to set the following access permissions: opening the file, printing, changing the document, selecting text and graphics, and adding/changing annotations and forms fields.

Some of the encryption methods, digital signatures, and compression algorithms that are supported by PDFs are copyrighted. Software that is created to decrypt or encrypt any of these algorithms may be subject to license these technologies.

3.7 Application and Platform Support:

PDF 1.3 is the native format of Adobe Acrobat 4. It can also be read by Adobe Acrobat 5, as well as many third party PDF readers. PDFzone.com lists 62 applications for viewing, creating and manipulating PDFs. Some of these applications are open source software.

When the PDF major number is incremented (ex: PDF 1.5 is superseded by PDF 2.0), existing PDF readers will be unlikely to read the newer files. When the minor number changes (ex: PDF 1.3 is superseded by PDF 1.4), a PDF viewer designed to read PDF 1.3s should still be able to read PDF 1.4s. 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 [Adobe, 2001b] 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, and works well as a web browser plug-in. There are free PDF readers for every popular and not-so-popular operating system (Win9x/NT/XP/2K, Macintosh, MacOSX, Linux, AIX, Solaris, SGI IRIX, HP-UX, OS/2, etc).

4 Related Formats

4.1 Specification Variations:

FDF (Forms Data Format)
FDFs are an Adobe specification 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. They have the extension '.fdf'.

PJTF (Portable Job Ticket Format)
An Adobe specification based on PDFs, and 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 1.2 (if it conforms with the PDF 1.2 specification, for example).

PDF-Archive or PDF/A (Portable Document Format / Archive)
A proposed format based on PDF 1.2 by the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES), and the Association for Information and Image Management, International (AIIM International). The intention is to develop an international standard that defines the use of PDF for archiving and preserving documents. It is a subset of the specification, leaving out things deemed to hinder document preservation (linearization, encryption, embedded multimedia, etc.). See http://www.aiim.org/standards.asp?ID=25013 for more details.

PDF/X family (PDF/X-1, PDF/X-1a. PDF/X-2, PDF/X-3)
These are specifications based on PDF tailored for the print publishing industry. They are subsets of PDF specifications 1.2, 1.3 and 1.4 that eliminate elements that make it hard to prepare documents for printing. Some of these formats have been ratified by ANSI and/or ISO.

5 Summary and Conclusions

The PDF format is very popular, in part because PDF readers are so accessible. PDF readers are free, the readers integrate well with popular web browsers and they are available for most operating systems. PDFs are also popular because of their special features. 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 and scripts, and security features. In short, no other *single* media format can support all the features that a PDF can. There are so many existing documents in PDF format, that it is likely that PDF will be supported for a long time.

Despite its popularity, it cannot be ignored that PDF is a proprietary format. The duration of each PDF version does not fit a pattern of increasing format stability, like exists for the TIFF format. There are no indications that PDF's features are becoming 'fixed' anytime soon. Instead, we are seeing increasing complexity with each new PDF version. The increasing size of the specification with each new PDF version is a good indication of this (see Figure 2).

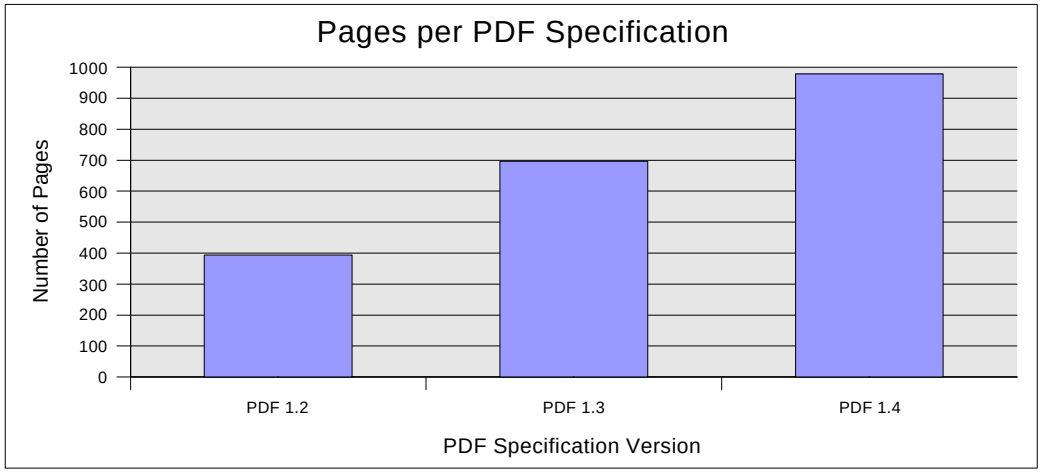

Figure 2: Increasing Size of PDF Specifications. Note that this does not include any of the supplementary specifications for PostScript, etc.

The complexity of the PDF specification is evident by comparing the size of the PDF 1.4 specification in comparison to other formats' specifications (see Figure 3). Because of the complexity of the PDF specification, third party software tends not to implement as many of the PDF features as Adobe's PDF software. Essentially this leads to Adobe dominating the PDF software market, tightly coupling the PDF format with the stability of Adobe as a company. As long as Adobe can stay a profitable company, Adobe will support the PDF 1.x format because so many of their products are based on it.

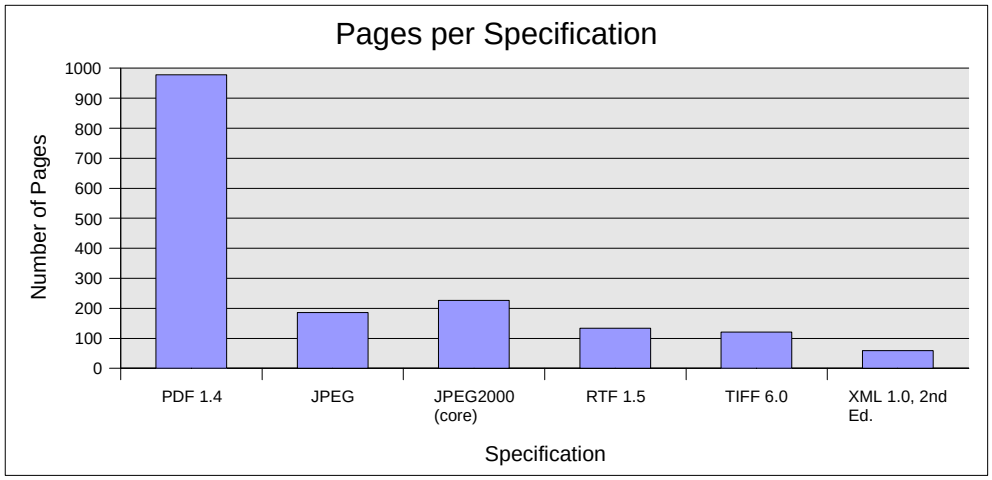

Figure 3: Size of PDF 1.4 specification as compared to other specifications.

6 Discussion

It is safe to assume that every format will become obsolete at some point. What will the successor format to PDF look like? It is possible that it could be XML-based. There is a developing trend in office software to use XML-based formats. The format used by StarOffice, and now OpenOffice, is XML-based. It consists of a zip file containing XML files and stand-alone images. Microsoft's Office 11, due out later this year, will use XML as its native format. It is not clear yet how it will handle binary data, like images. PDFs are more complex than word

processing formats, but they have similarities. They are both container formats that can contain text, images and presentation/styles information.

Current XML technologies by themselves can not replace the functionality of the PDF format. XML is not designed to hold binary data, like PDFs are (PDF 1.1 and later). There is no standard XML 'packaging' technology that can keep together the text, formatting information, images, etc. In summary, XML does not currently work well for electronic exchange of diverse multimedia content - PDF does. However, if the intellectual content contained within a PDF were extracted, XML could be used, in combination with other formats, to store the intellectual content as a normalized version for long-term archival purposes. The benefit to storing the contents this way is that the separate files could be recomposed into future formats fairly easily, especially if PDF's successor format is XML-based. Additionally, the image/audio/video media is separated from the text to allow separate migration schedules.

The largest barrier to implementing the XML-based normalized version is in extracting the intellectual content (text, images, etc.) from PDFs. Adobe has a PDF library that could be used to extract this content, however it is very expensive[1] [Gonzalez, 2003] and has licensing issues which would prohibit the sharing of archive software based on the library. There are no known existing non-proprietary libraries for extracting text and other media from PDFs that work well. There has been progress made in this area however, see [IDRSolutions, 2003] and [Litchfield, 2003], that could be built upon.

6 References

[Adobe, 1999a] Adobe Systems Incorporated, “Portable Document Format Reference Manual”, Version 1.3, March 11, 1999.

[Adobe, 1999b] Adobe Systems Incorporated, “Portable Document Format Reference Manual”, Version 1.3, Second Edition, Addison-Wesley, July 03, 2000.

[Adobe, 2001a] Adobe Systems Incorporated, “Portable Document Format: Changes from Version 1.3 to1.4”, Technical Note #5409, Preliminary, Adobe Systems Incorporated: San Jose, CA; June, 2001.

[Adobe, 2001b] Adobe Systems Incorporated, “PDF Reference: Adobe Portable Document Format Version 1.4”, Third Edition, Addison-Wesley: Boston, December, 2001.

[Gonzalez, 2003] Gonzalez, Cynthia. Email correspondence, Account Manager, Adobe PDF Library; February 11, 2003.

[IDRSolutions, 2003] IDRSolutions. JPedal website, 2003.
http://www.jpedal.org/

[Litchfield, 2003] Litchfield, Ben. PDFBox website, 2003.
http://www.pdfbox.org/

1A one-time fee of $25,000 USD plus annual support and maintenance charges of $5,000 USD per platform

[Ockerbloom, 2001] Ockerbloom, John Mark. “Archiving and Preserving PDF Files”, RLG DigiNews: Volume 5, Number 1, February 15, 2001. http://www.rlg.org/preserv/diginews/diginews5-1.html