# Action Plan Background: PDF 1.5

**Author:** Carol Chou, FCLA

**Release Date:** 5/20/2005

## 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 and layers it with the document structure and interactive navigation features.

## 1 General Description

**1.1 Format Name:** PDF (Adobe Portable Document Format)

**1.2 Version:** 1.5

**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 5.0 and Adobe Illustrator 10.0

**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:** 11 color spaces including:

- three device-independent CIE-based: CalGray, CalRGB, Lab, and ICCBased.
- three device-dependent color spaces: DeviceRGB, DeviceCMYK, 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-1.5

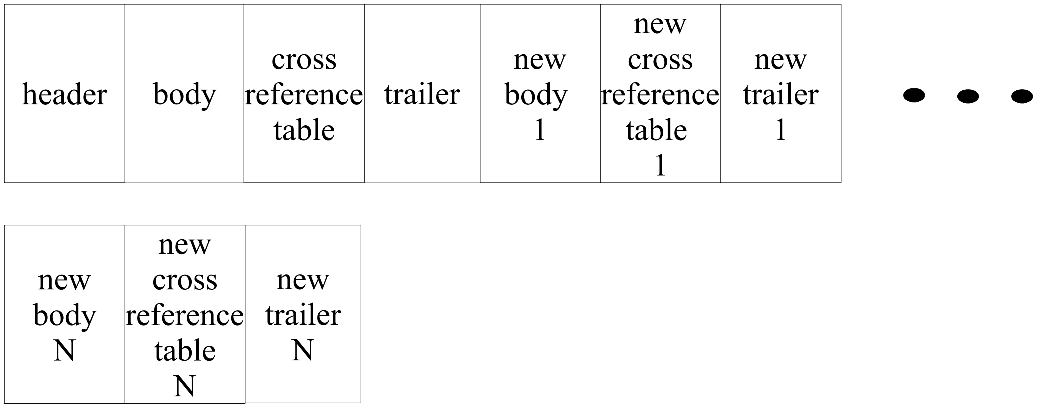
**1.13 Latin-text Encoding:** Can be Unicode, PDFDocEncoding (see 4.4.1 of [Adobe, 1999]), MacRomanEncoding, MacExpertEncoding, WinAnsiEncoding, StandardEncoding

### 1.13 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 1.5 adds the ability to display image using JPEG 2000 compression, to selectively view or hide contents in a PDF file, and adds 16 bits images support. Like earlier versions, PDF1.5 also supports optional advanced features such as embedding audio and video, 'linearizing' for network streaming (PDF 1.2 and later), encryption controlling access permissions (including support for public-key encryption technology), forms, digital signatures, embedded file streams(could be encrypted even if it is in an unencrypted document), embedded fonts, embedded XML meta-data streams, transparency, and trigger events.

### 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 |
| contains image [F] | O |
| use embedded font [F] | O |
| End of File Marker [G] | R |

| ***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.)*** |
|---|---|
| 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 |

## 3 Usefulness

**3.1 Version Duration:** 1 year, 4 months

**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.

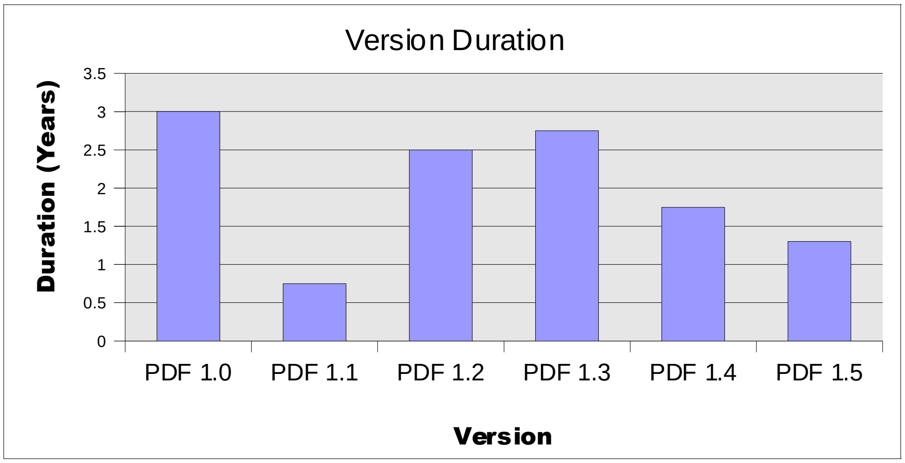


**Figure 1: Duration in Years of PDF Versions**

### 3.3 Expected Newer Versions:

A newer version PDF 1.6 has become available on November 2004.

### 3.4 Existence of Publicly Available Complete Specifications:

Adobe publishes newer versions of PDF specifications (including 1.5) which can be freely downloaded on its website.  Some PDF specifications (version 1.3 and 1.4) are also available in print.  The older PDF specifications such as 1.2 and earlier are not available on its website.  While each new specification attempts to incorporate information about older PDF versions, the older specifications provide additional useful information.  Some of these older specification can be located elsewhere on the Internet.

All the information needed to completely understand PDF format is not completely 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 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

### 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, 2003] 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 make reasonable efforts to ensure the software they create respect the access permissions and permissions controls listed in Table 3.20 of this specification to the extent that they are used in any particular document." These access permissions are maintained through encryption. The user has the ability to set the following access permissions: opening the file, printing, modifying the document, copying and extracting text and graphics, adding/changing annotations and forms fields.

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

### 3.7 Application and Platform Support:

PDF 1.5 is the native format of Adobe Acrobat 6 and Adobe family products. In addition, there are many third party PDF readers that can read and create PDF 1.5. PDFzone.com lists 262 applications for viewing, creating and manipulating PDFs. Many of these applications are open source software.

Adobe provides free PDF readers for various desktop and mobile operating systems including Windows (Win9x/NT/XP/2K), Macintosh (Mac OS 7.X/8.X/9.X/10.X), Linux, IBM AIX, Solaris, HP-UX, OS/2 Warp, Palm OS, Pocket PC, Symbian OS, ..etc with support for many Asian, European and Middle East languages. Adobe does not always provide new PDF readers on every supported platform. In fact, 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 support back on Linux for the latest version of Acrobat (7.0) after it sees a wider adoption on Linux system [Millard]. Acrobat Readers support for HP-UX, AIX, Solaris is still stuck with version 5.0.

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 1.4 is superseded by PDF 1.5), a PDF viewer designed to read PDF 1.4s should still be able to read PDF 1.5s. 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, 2004] 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. To date, over 500 million copies of Adobe Acrobat Reader have been distributed from Adobe [Adobe]. PDF has become so ubiquitous that approximately 2400 government agencies have standardized on the PDF format. FDA (Food and Drug Administration) makes PDF the required format for submitting drug approval and US. Federal Court requests electronic case filing to be in PDF format [CM/ECF]. The IRS also uses PDF to distribute the tax form. Over 1.5 billion PDF forms have been downloaded from IRS web site since 1998.

The largest user of the PDF format is the IRS, who has 350,000 Acrobat seats [Enfocus, 2002].

### 3.10 Market Competition

The latest SEC (Security and Exchange Commission) filing from Adobe [Adobe, 2005] releases the existing and proposed file formats that they perceive to be in direct or indirect competition with PDFs. Office applications or content creation tools that use HTML, Macromedia Flash, Macromedia FlashPaper, Microsoft Word, Microsoft Infopath, TIFFs and/or XML-based formats for electronic document publication indirectly compete with PDFs. Adobe is particularly agitated by Microsoft's increasing development on document collaboration and management. With the upcoming release of Microsoft's next generation operating system, codenamed Longhorn, and the anticipated incorporation of electronic document capabilities into Longhorn, Adobe believes PDF will face increasing competition from Microsoft.

The biggest strength of PDF is its ability to preserve the look of the document across different printers and platforms. Though Microsoft Word is popular on MS Windows systems, it does not always maintain the document appearance across different platforms. Macromedia Flash is one of Adobe's major competitors. With the purchase of Macromedia by Adobe [Adobe, 2005a], it is possible that PDF and Flashpaper would provide better integration with each other in the near future [Fluckinger]. The biggest strength of Flashpaper lies on its faster download speed for web browsing and better multimedia capabilities compared with PDF. Whether the purchase of Flashpaper will bring major changes in PDF is an issue that should be closely watched .

## 4 Related Formats

### 4.1 Specification Variations:

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).

PDF-Archive or PDF/A (Portable Document Format / Archive)
A proposed format based on PDF 1.4 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 PDF 1.4 specification, leaving out things deemed to hinder document preservation such as linearization, encryption, device dependent color space, pointer to external content and embedded multimedia[1].

PDF/A requires all metadata properties pertaining to a PDF file be embedded as XML packages. The format of XML representing the metadata is defined as part of the XMP (Extensible Metadata Platform) framework. PDF/A mandates the use of XMP metadata for basic descriptive and identifying metadata [ISO 2003].

Gammon stated the possibility of PDF/A being approved as ISO standard by the end of 2005 if it goes all according to plan [Gammon 2004] .

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. They eliminate those elements that make it hard to prepare documents for printing. Some of these formats have been ratified by ANSI and/or ISO.

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 [Adobe,2003a].

### 4.2 Other ongoing working group:

PDF/E
PDF engineering (PDF/E) working group has been formed by AIIM (international authority on enterprise content management) and NPES (the association for supplier in printing, publishing and converting technology) to promote a new standard for the use of

[1] See http://www.aiim.org/standards.asp?ID=25013 for more details.

PDF format in engineering workflows. The draft of the new standard is expected to be proposed to ISO with anticipated publication date of mid-2006 [AIIM 2004]. At this moment, there is still no draft specification for the standard.

## 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 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. Although the PDF specification is openly published, it is controlled entirely by Adobe. Over the years, Adobe has expanded PDF features to include support for more application domains such as multimedia, encryption, security, and CAD drawing. This has resulted in increasing complexity of the PDF specification. Though the increased feature set has sustained PDF's dominant role in document presentation, it inherently presents some challenges in digital archiving. In addition, some PDF critics have long argued about PDF's suitability for on-line viewing because of its longer download time compared to other formats such as FlashPaper.

The duration of each PDF version does not fit a pattern of increasing format stability. 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).

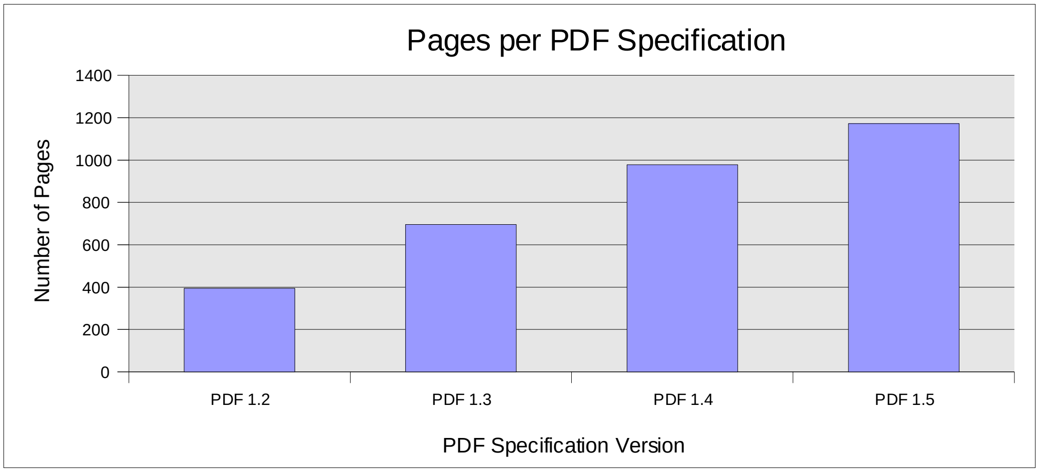


Figure 2: Increasing Size of PDF Specifications. Note that this does not include any of the supplementary specifications for PostScript, etc.

The complexity of the PDF specification is evident by comparing the size of the PDF 1.5 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.

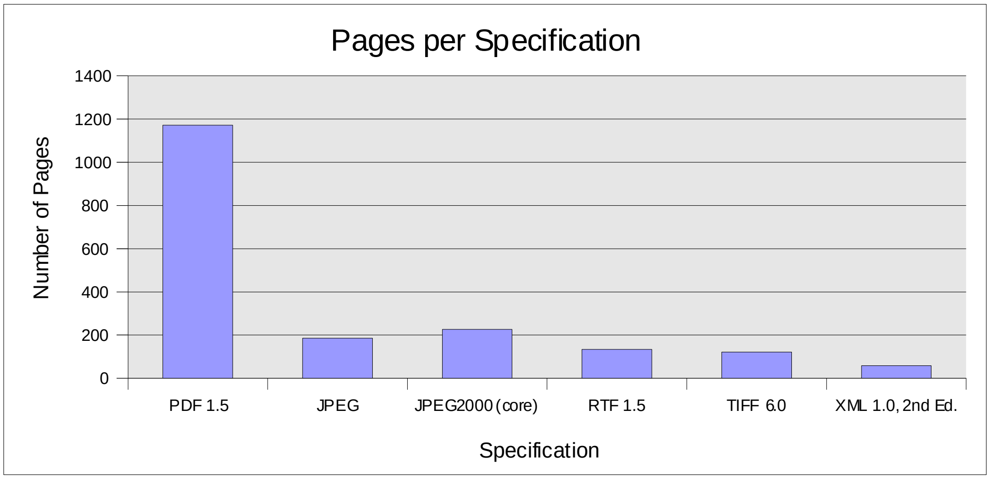


Figure 3: Size of PDF 1.5 specification as compared to other specifications.

## 6 Discussion

PDF has been around for over a decade. Even though PDF is a proprietary format, it is a format that cannot be ignored for archiving due to its thriving popularity and sustainability. However, data formats are heavily dependent on the software that creates and supports them. The dynamic nature of software inherently guarantees the limited timespan of any data format. Just like the once popular PostScript file, there will come a time when PDF become obsolete.

The increasingly popular XML poses as a possible successor format to PDF. Unlike PDF which is a proprietary format, XML is open standard, non-proprietary and royalty-free. This has led to a wide adoption of XML in many commercial and open source software. For example, OpenOffice uses XML as the foundation of its Open Office XML file format (OO.o XML). It consists of a zip file containing XML files and stand-alone images. The most recent release of Microsoft Office, Office 2003, also uses XML as its native format. Though the original release of InfoPath did not include support for binary data, the most recent service pack release of InfoPath, SP1, allows attaching binary data into InfoPath form by storing it as a BLOB (binary data encoded as base64 in the XML). To improve interoperability and collaboration with other vendors, Microsoft has published its XML-based office schema used by InfoPath and Excel on a royalty-free basis. XML has become one of the most popular formats used by many office software.

PDF is more complex than word processing formats and provides a richer set of functionalities than XML. Though they are both container formats, PDF provides additional security and access control features that are currently not available in XML. Realizing the important of XML, Adobe has provided a comprehensive XML architecture which allows XML data to be embedded in PDFs and makes PDF better integrated with XML.

One possible solution for normalizing PDF files is to use XML (and other formats) to store the intellectual content extracted from PDF files. This will served as a normalized version for long-term archival purposes. The benefit of 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[2] [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.

## 7 Acknowledgment

This background report is based on PDF 1.4 Action Plan Background written by Andrea Goethals while she was on staff at FCLA. The author cordially appreciates her previous research on PDF format.

## 8 References

[Adobe] Adobe Systems Incorporated, “Adobe Company fact sheet”, (http://www.adobe.com/aboutadobe/pressroom/companyprofile.html

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

[Adobe, 2000] 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,

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

2001.

[Adobe, 2001c] Adobe Systems Incorporated, "Minimum and Maximum Parameters for Acrobat Capture 2.0x Products and PDF Files", Document 323083, Support Knowledge Base, Adobe website, September 24, 2001. (http://www.adobe.com/support/techdocs/27fae.htm)

[Adobe,2003] Adobe Systems Incorporated, “PDF Reference: Adobe Portable Document Format Version 1.5”, August, 2003.

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

[Adobe, 2005] Adobe Systems Incorporated, “Annual Report (10-K)”, February 2005.

[Adobe, 2005a] Adobe Systems Incorporated, “Adobe to acquire Macromedia”, press release, April 18, 2005

[AIIM 2004] AIIM, “New Standard Address Document Collaboration and Exchange Within Engineering Process”, 2004

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

[Enfocus, 2002] Enfocus Software. “The Ultimate PDF/X Guide”, Enfocus Software, 2002.

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

[Fluckinger] Don Fluckinger, “Adobe + Macromedia = Boost”, PDFZone, April 18, 2005,

[Gammon 2004] Gammon, Ralph, “Electronic Archiving Standard Taking Shape”, Document Imaging Report, May 21, 2004. http://www.aiim.org/article-aiim.asp?ID=28406

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

[Harvard, 2002] Harvard University Library Mellon Project Technical Team and Harvard University Library Mellon Project Steering Committee, " Report on the Planning Year Grant for the Design of an E-journal Archive", April 1, 2002. (http://www.diglib.org/preserve/harvardfinal.html#_Toc4915258)

[IDRSolutions, 2003] IDRSolutions. JPedal website, 2003.

http://www.jpedal.org/

[ISO 2003] ISO/CD 19005-1, "Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF (PDF/A)", October 31, 2003.

[Karr, 2003] Karr, Ron. Email Correspondence, Senior Technical Writer, Developer Resources, Adobe Systems Incorporated; January 27, 2003.

[King 2004] James C. King, "A Format Design Case Study: PDF", Proceedings of the fifteenth ACM conference on Hypertext & hypermedia, August 2004

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

[Millard] Elizabeth Millard, "Adobe Releases Reader 7.0 for Linux", PDF Zone, April 12, 2005, http://www.pdfzone.com/article2/0,1759,1784911,00.asp

[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