
# Technology Watch Report

# Preserving the Data Explosion: Using PDF

Betsy A. Fanning

AIIM

1100 Wayne Avenue, Suite 1100, Silver Spring, MD 20910 USA

bfanning@aiim.org

DPC Technology Watch Series Report 08-02

April 2008

**Executive Summary**

The introduction of the Internet and desktop computing has transformed the way information is handled. The Internet placed a new level of urgency on the business world in expecting that information would be made available immediately – much faster than what the paper-based information world could handle. This need for information now led to the transition from a paper-based document centric environment to an environment of electronic documents and electronic email. This transition from paper to electronic documents led the way for document file formats like PDF, Portable Document Format to be introduced. A PDF document is essentially an electronic document that is equivalent to the paper document.

This report reviews the use of PDF, Portable Document Format, more specifically, PDF/Archive as an archival file format to preserve an organization's knowledge. It should be noted that the use of PDF or PDF/Archive alone will not ensure the long-term preservation of electronic documents. When PDF/Archive is combined with a comprehensive records management program and formally established records policies and procedures, an organization can be sure that their electronic documents will be preserved.

Electronic documents just like paper-based documents are preserved for reasons beyond why they were created. Organizations preserve documents based on their historical value, to preserve the organization's knowledge, for research, and the uniqueness of the information. Not all information needs to be preserved. The archival value of the information must be assessed based upon the content of the documents being preserved.

While focused on PDF and PDF/Archive, this report will address some of the other file formats which may be used or considered for archiving electronic documents. It will explore the PDF standards efforts and their suitability to long-term preservation including:

- PDF/Engineering,
- PDF/Exchange,
- PDF/Universal Access, and
- PDF Healthcare.

PDF/Archive was developed as a result of numerous organisations needing to be able to preserve their electronic documents and be able to access and view the documents at a future time with the document appearing as it had when it was produced. In order to ensure that the documents could be accessed, it was important to ensure that the standard addressed the goals of device independence, files being self-contained and self-documenting, and that there would be no restrictions like encryption that would impede the access to the documents.

Given the wide acceptance of PDF, the development of PDF/Archive for long-term preservation of electronic documents is a logical use of the file format. Through the use of PDF/A, organizations can be sure that their documents will be preserved for the long term. While PDF/A may be a suitable file format today for long-term preservation of electronic documents, it should be noted that there may be other file formats introduced in the future that may better serve the needs of an organization. Therefore, organizations should be continually reviewing the available file formats to ensure they have selected the best format for their purposes.

**Keywords:** PDF, Archive, Preservation, Engineering, File Format, Healthcare, Portable Document Format

## 1.0 Introduction

The Internet has transformed the workplace and changed the way paper is used. The use of electronic documents has exploded and become the normal business practice. It is estimated that over 90% of all business records are electronic[1]. This transition from paper to electronic documents has led the way for document file formats such as PDF, Portable Document Format, to be introduced. PDF was made popular by the fact that it enabled users to see electronic documents as they did their paper counterparts. With a large amount of information being based in electronic formats, it is critical to be able to preserve or archive this knowledge for future generations. This report will describe the PDF standards activities currently being developed and their relevance to digital preservation.

## 2.0 Background

Records, paper- or electronic-based, are preserved for specific reasons beyond the reason they were created. The records may have historical value, need to be preserved to maintain corporate knowledge, provide a glimpse of the direction the organisation was taking at a specific point in time, or have some need for future research or value as new products are being introduced. An organisation cannot archive all the information it creates. Therefore, it is important to assess the uniqueness of the information and the requirement for archiving. Assessing the archival value of the information requires a review of the content of the information or documents. This requires judgments to be made as to the value of the content for reference, research, validation or reinforcement for decisions; preserving the organisation's history; maintaining relevant information for legal, administrative, or fiscal purposes; or celebrating the history of the organisation at notable anniversaries.

Not many years ago, predictions were made that information was exploding at astronomical rates and that this information was to be in paper format at that time, the electronic document was only used for non-business purposes and not the essence of business that it is today. An electronic document is an "electronic representation of a page-oriented aggregation of text and graphic data, and metadata useful to identify, understand and render that data, that can be reproduced on paper or optical microform without significant loss of its information content[2]." Shortly after these predictions were made, organisations made the distinct decision that they had to better manage their paper or be overcome by the tons of paper that were being generated in their organisations on a daily basis. At approximately the same time, the concept of the "paperless office" was born, which was to be a utopia for organisations. Take a moment to picture in your mind the vision that these early crusaders had. In this "paperless" society, organisations would not use paper for any business transactions. Instead they would use electronic documents, electronic forms, electronic mail (e-mail), and telephone calls to conduct

[1] http://www.arma.org/erecords/index.cfm?view=publications
[2] ISO 19005-1: 2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)

business. Imagine what a typical office in this "paperless" world would look like…the desks would be bare of paper and have a computer monitor and a telephone on them. There would be no clutter, no chaos, no records filing rooms, no copiers, etc. Soon after the concept was introduced, it was realised that a "paperless" office was nearly impossible to achieve and the concept of a "less paper" office was born which was thought to be easy to achieve. The paperless office concept could not be fully achieved due to our human dependence on paper to do our work and the lack of tools available for the electronic documents. The "less paper" concept contributed to the overall acceptance of electronic documents.

The explosion of information has been the result of improving digital technologies including more powerful computation resources, faster data streams, and more storage resources. Data needs to be shared across both organisational and geographical boundaries, and where storage tends to be independently managed. The ease and speed with which this data is generated and changed also makes it increasingly difficult to ensure its quality.

With all of this electronic information, the challenge is in finding the nuggets of information and knowledge buried under this avalanche of bits and bytes. Reliable tools to access and integrate legacy, raw and derived data, and manage its transformation into knowledge are in high demand. The process of transforming data into knowledge requires access and integration.

PDF is a file format originated by Adobe Systems in the early 1990s created for the primary purpose of exchanging documents. PDF was developed because Adobe Systems needed to exchange document electronically amongst their employees and the technologies at the time did not provide that capability. It was intended to make electronic documents essentially similar to their paper equivalents by being authentic, reliable and easy to use. The *PDF Reference* is an open specification that defines the features and functions for the PDF file format. An open specification is one that is publicly available. Adobe made all the *PDF Reference* specifications freely available on their web site and allows any software developer to use the specification in designing their products. When PDF was first introduced, it was used primarily by graphics artists, designers, and publishers for producing and exchanging proofs. Now, PDF is used to exchange all types of data including vector graphics (illustrations and designs), text, and raster graphics (photographs and other types of images). Through the evolution of the Internet and use of PDF, PDF has become a de facto standard for exchanging documents.

While the "paperless" office did not come to be, businesses learned a number of valuable lessons which have been implemented today. The experiment of the paperless office taught organisations that their employees were more comfortable with paper. Organisations learned that paper was easier to navigate through because it facilitated cross-referencing. It was easy to apply notes to the paper. Paper was found to facilitate collaboration efforts and made coordinating activities easier. It was not too long ago that when attending a meeting, the attendees used paper to follow the meeting. Now, meeting attendees rely on electronic documents and leave the paper back in the office. These characteristics provided a challenge to the software developers who were developing electronic document applications as they

needed to make it easy to navigate through the document as well as make the electronic documents as easy to use as possible. As PDF evolved, it became the electronic equivalent to paper through the functionality and features added to the specification. This led the various PDF developers to introduce products that made the user experience of working with an electronic document seem very similar to the way they were accustomed to working with paper.

PDF as a document format is feature-rich. This richness can be a hindrance when developing applications to fit specific needs such as exchanging or archiving documents. As users became more proficient with electronic documents, they began to request functionality and features that used the existing features of PDF and added new features and functionality to the PDF specification. In 1994, PDF 1.1 introduced support for:

- External links
- Article threads
- Security features
- Device independent colour
- Notes

PDF 1.2 introduced in 1996 provided new features that enabled PDF to be beneficial in the prepress industry including:

- Support for OPI (Open Prepress Interface) 1.3 specifications
- Support for the CMYK (colour model short for cyan, magenta, yellow and key (black)) colour space
- Spot colours could be maintained in PDF
- Halftone functions could be included as well as overprint instructions

In 1999, PDF 1.3 provided support for:

- 2-byte CID fonts
- OPI 2.0 specifications
- New colour space called DeviceN to improve support for spot colours
- Smooth shading, a technology that allows for efficient and very smooth blends (transitions from one colour or tint to another)
- Annotations

May 2001, PDF 1.4 was released providing:

- Transparency support that allows text or images to be seen through
- Improved security
- Improved support for JavaScript

PDF 1.5 released in May 2003, introduced:

- Improved compression techniques including object streams and JPEG2000 compression

- Support for layers
- Improved support for tagged PDF

November 2006 marks the date that PDF 1.7 was introduced providing:

- Improved support for commenting and security
- 3D support was improved

In 2000, Adobe Systems initiated the first of what would become several efforts to standardise subsets of the *PDF Reference* for specific purposes. The first subset to be introduced became known as PDF/X after which came numerous other ISO PDF standards. The current PDF standards portfolio consists of:

- PDF – Portable Document Format
- PDF/A – Portable Document Format/Archive
- PDF/E – Portable Document Format/Engineering
- PDF/UA – Portable Document Format/Universal Access
- PDF Healthcare – Portable Document Format Healthcare
- PDF/X – Portable Document Format eXchange

## 2.1 Graphics Exchange

The PDF/X, PDF/eXchange, is the family of standards that was and is continuing to be developed by ISO TC 130, Graphic Technology. ISO TC 130, Graphic Technology, is responsible for developing standards for the printing and graphic technology fields. The PDF/X standard provides an efficient vehicle for exchanging files representing print ready material. PDF/X is predominantly used by the graphic industry.[3]

The family of PDF/X standards currently consists of:

- ISO 15929, *Graphic technology – Prepress digital data exchange – Guidelines and principles for development of PDF/X standards*
- ISO 15930-1, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 1: Complete exchange using CMYK data (PDF/X-1 and PDF/X-1a)*
- ISO 15930-2, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 2: Partial exchange (PDF/X-2)*
- ISO 15930-3, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 3: Complete exchange suitable for colour managed workflows (PDF/X-3)*
- ISO 15930-4, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 4: Complete exchange of CMYK and spot colour printing data using PDF 1.4 (PDF/X-1a)*
- ISO 15930-5, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 5: Partial exchange of printing data using PDF 1.4 (PDF/X-2)*

[3] PDF/X Application Notes

– ISO 15930-6, *Graphic technology – Prepress digital data exchange – Use of PDF – Part 6: Complete exchange of printing data suitable for colour-managed workflows using PDF 1.4 (PDF/X-3)*

ISO 15930 specifies the use of PDF for the dissemination of complete digital data in a single exchange that contains all elements ready for final print production. This standard defines the data format and how it is to be used to ensure the file transmitted contains all the content information necessary to process and render the document as it had been intended to be. This means that the colour is exchanged in the exact way that the designer had intended and that the hues or tones of the colour are not altered resulting in a higher quality image since the fonts and colours are embedded in the file.

2.2 Archival

The family of PDF/Archive (PDF/A) standards was and is being developed by an ISO Joint Working Group under the auspices of ISO TC 171 SC2, Document Management Applications, Application Issues in cooperation with representatives from ISO TC130, Graphics Technology, ISO TC46 SC11, Information and documentation – Archives/records management and ISO TC 42, Photography. This joint working group brought together experts from the library, archival, document management, records management, graphics, government agencies, industry, and software developer communities to develop this standard.

The family of PDF/A standards consists of:

– ISO 19005-1:2005, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*

This standardisation effort began in the United States as a joint project of AIIM, the Enterprise Content Management Association and NPES, The Association for Suppliers of Printing, Publishing, and Converting Technologies as a result of a need raised by numerous organizations to be able to reliably preserve electronic documents. It quickly gained a great deal of visibility with over 400 individuals requesting to be on the committee listserv to follow the committee activities. The reason PDF/Archive began was because several organisations were being faced with large collections of electronic documents that they needed to manage, preserve, and make searchable and available for generations to come. These groups began by formalising the need for a PDF archival format and identifying the business requirements. In addition to archiving the electronic documents, it was necessary to be able to reliably render the documents a hundred years from now which meant that a format supporting long-term preservation was necessary.

In the initial stages of this effort, the committee discussed the various file formats that could be used for archiving electronic documents. These formats included TIFF (Tagged Image File Format), XML (eXtensible Markup Language), native file formats and PDF. PDF was chosen as the file format best

suited for long-term preservation due to its wide adoption in numerous applications and ease of creating PDF files from digitally born documents.

PDF is an open file format for electronic documents. While the format is considered proprietary, the specification for the file format is publicly available. Adobe Systems owns patents on the format but allows developers to use the specification to develop products that produce and render PDF files royalty free. Regardless of the operating system or tool used to create the PDF file, the PDF file will display exactly as it is intended to be displayed on any device using any operating system.

Due to the de facto adoption of PDF, many organisations already mandate the retention of PDF documents. PDF is recognised as being feature rich which can cause difficulties for specific uses such as long-term preservation. The committee recognised that PDF documents are not self-contained but rely on system fonts and other content stored external to the file. These external links can change or get broken over time which would allow information to be lost causing a preservation problem.

The objectives of the PDF/A working group that guided the technical development of the standard, included:

- Device-independence to ensure files did not require a specific platform or operating system to render
- Files needed to be self-containing and have all the resources necessary for rendering
- Self-documenting containing their own description
- Lack of restrictive elements like encryption that would hamper the rendering of the document years from when it was originally created
- Disclosure
- Widespread use and adoption of the format [4]

The PDF/A standard defines long-term as "the period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository, which may extend into the indefinite future.[5]"

PDF/A was intended to address[6]:

- Defining a file format that preserves static visual appearance of electronic documents over time
- Provides a framework for recording metadata about electronic documents

[4] PDF/A Application Notes
[5] ISO 19005-1: 2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)
[6] PDF/A Frequently Asked Questions (FAQ)

- Provides a framework for defining the logical structure and semantic properties of electronic documents

As the working group developed PDF/Archive, they identified the following as issues that needed to be resolved through the development process:

- Long-term preservation of electronic documents – PDF is feature-rich which imposes problems or issues for the archival nature of electronic documents
- PDF is not necessarily self-contained
- Lack of standardisation of PDF tools which leads to inconsistent results
- Exchange barriers including content such as external links which may be inaccessible
- In the engineering world, there are multiple proprietary formats which require individual and often expensive viewers
- Making PDFs fully accessible to individuals with disabilities
- Need for a secure, electronic container that can store and transmit relevant healthcare information without limitations on the type of information being transmitted while keeping the costs low
- Need to expand PDF
- Easy to use
- Reliability

The PDF/Archive file format is based on and includes the functionality included in the *PDF Reference 1.4*. However, in order to ensure long-term preservation, it was necessary to limit the specific functionality of PDF by establishing specific requirements. Therefore, the PDF/A standard specified features that are allowed and not allowed.

PDF/A-1 (ISO 19005-1) files allow[7]:

- Embedded fonts which includes only those fonts which may be legally embedded without a royalty fee
- Device-independent colour
- XMP Metadata

PDF/A-1 files do not allow[8]:

- Encryption as the method of encryption may not be supported when the files are opened at a later date
- LZW Compression due to intellectual property constraints
- Embedded files
- External content references because the references may change or be broken

[7] PDF/A Frequently Asked Questions (FAQ)
[8] PDF/A Frequently Asked Questions (FAQ)

- PDF Transparency – a file can use other techniques than the use of transparency keys to provide the visual effect of partially transparent graphics
- Multimedia
- JavaScript

PDF/A establishes two conformance levels. Level A denoted as "PDF/A-1a" and Level B denoted as "PDF/A-1b". A Level A conforming file or reader must meet all the requirements of the *PDF Reference 1.4* as modified by ISO 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1).* Conforming Level A files may include any feature in *PDF Reference 1.4* or an earlier version of *PDF Reference* that is not forbidden in ISO 19005-1. Level B conforming files must meet all the requirements in ISO 19005-1 except those identified in clauses 6.3.8 and 6.8. The requirements of a Level B conforming file are those that are necessary to ensure the file and the visual integrity can be preserved over time. In addition to the requirements for specifying a PDF/A file, the standard includes best practices which are intended to help end users and developers implement PDF/A in their organisations.

As the title of ISO 19005-1: 2005 suggests, PDF/A is based upon the *PDF Reference 1.4* and therefore complies with the requirements established by that version of the PDF specification. The committee was careful to state that PDF readers will ignore the features that are not documented in *PDF Reference 1.4*. ISO 19005-1 requires the document information dictionary to be consistent with XMP metadata. Some of the other key requirements established in ISO 19005-1 include:

- Embedded colour spaces
- Device independent colour
- Fonts must be able to be legally embedded for unlimited, universal rendering
- Unicode character map is required for Level A conformance
- Behaviour for NextPage, PrevPage FirstPage, and LastPage actions as defined in *PDF Reference 1.4*
- Requires the use of Extensible Metadata Platform (XMP)
- Tagged PDF

The XMP metadata required by the standard is an open but proprietary metadata methodology that is used for metadata creation, processing, and interchange. It is based on the Resource Description Framework (RDF) that was developed by the World Wide Web Consortium (W3C) and is the cornerstone to the semantic web. Through the use of XMP, applications are able to access and understand the metadata about the documents that they manipulate.

### 2.3 Engineering

The PDF/Engineering family of standards is being developed by TC 171 SC2, Document Management Applications, Application Issues with the assistance of TC 184, Industrial Automation Systems and

Integration. The standard defines a file format for representing electronic documents that preserves the visual appearance and content for the exchange of engineering documentation. PDF/E addresses the need to produce consistent results even if multiple creation tools are used. It breaks down the exchange barrier and addresses external links. Since PDF/E is based on the PDF specification, PDF/E documents may be read on the many readily available PDF readers.

PDF/E was originally developed through a joint effort between AIIM, the ECM Association and NPES, The Association for Suppliers of Printing, Publishing, and Converting Technologies bringing together graphics, construction, manufacturing, geospatial, cartographers, and others. The standard describes how to reliably create, exchange, and review engineering documentation even large format drawings. ISO 24517, *Document management – Engineering document format using PDF – Part 1: Use of PDF 1.6 (PDF/E-1)* establishes only one conformance level by which files, readers, and creators must comply which means PDF/E compliant readers must comply with all requirements in the standard. This standard addresses and allows 3D images to be included in the file as well as digital signatures.

The PDF/Engineering family of standards includes:

- ISO 24517-1, *Document management – Engineering document format using PDF – Part 1: Use of PDF 1.6 (PDF/E-1)*

To ensure the reliable rendering and exchange of engineering documentation, the standard had to establish limits.

PDF/E files must include[9]:

- Embedded fonts
- Device-independent colour
- XMP for metadata

PDF/E files may not include[10]:

- External content references (links may be dropped or changed resulting in the loss of data)
- JavaScript not associated with 3D
- Dynamic (XFA based) forms

PDF/E may include[11]:

- JavaScript associated with 3D
- Embedded files
- Encryption
- Digital Rights

[9] PDF/E Frequently Asked Questions (FAQ)
[10] PDF/E Frequently Asked Questions (FAQ)
[11] PDF/E Frequently Asked Questions (FAQ)

- Digital Signatures
- Transparency
- Layers

It is important to note that PDF/E was created to exchange engineering documentation and not for the purposes of archiving. There is nothing to prevent using PDF/E for archiving engineering documentation as PDF/E files. The fact that the 3D standard specified in the PDF/E standard is maintained by acceptable standards organisations is not a deterrent to preserving PDF/E files. The long-term preservation of engineering documents should be a part of a formal records management program. It should be noted that PDF/A was designed for static documents or documents that do not contain dynamic content. PDF/E documents may contain interactive content, 3D images, and/or JavaScript which is outside the scope of ISO 19005-1. It is anticipated that dynamic content will be addressed in future versions of ISO 19005.

PDF/E should be used to standardise the use of PDF for the creation and exchange of engineering documentation in engineering workflows. PDF/E was not developed as an archiving format but for active file use. PDF/E and U3D (Universal 3D) may be used in files that are intended to be retained for long-term preservation purposes because the standards are maintained by accredited standards organisations. U3D was developed by Intel and the 3D Industry Forum (3DIF) to enable sharing 3D images on the web and in office applications. The use of multimedia in PDF/E files provides limitations on the long-term preservation of PDF/E files. This is because the long-term availability of multimedia players is not assured. It is helpful to note that the PDF/A working group is defining the requirements for preserving multimedia and 3D.

### 2.4 Healthcare

Unlike the PDF/Archive and PDF/Engineering efforts, the PDF Healthcare effort is not producing a proposed standard initially. The PDF Healthcare committee is developing a best practices guide that describes the use of the existing Portable Document Format (PDF) functionality as a secure and portable container for personal and electronic health record information to promote interoperability. The guide is based upon the existing ASTM CCR (Continuity of Care Record) standard. Given the extensible nature of PDF, PDF Healthcare will be able to contain any well-formed XML documents.

PDF Healthcare describes the attributes of PDF necessary to facilitate the capture, exchange, preservation, and protection of healthcare information. The best practices guide will allow healthcare providers and consumers to develop a secure, electronic container that can store and transmit relevant healthcare information, including but not limited to personal documents, clinical notes, lab reports,

electronic forms, scanned images, photographs, digital X-rays, and ECGs that are important to maintaining and improving one's health.[12]

This effort was jointly lead by AIIM, the ECM Association and ASTM, American Society for Testing and Materials. The *Portable Document Format in Healthcare Best Practices Guide* is being finalised for publication and should be available in early 2008. It describes the features and functions of a proposed, voluntary use of the PDF for the healthcare industry. PDF Healthcare does not address long-term preservation as its intent is to describe the use of PDF with regard to healthcare information.

### 2.5 Accessibility

In 2003, development of PDF/UA, or PDF Universal Access, which specifies how to use PDF to produce electronic documents containing character, raster, and vector data which are accessible to users with disabilities began. The committee's goal is to set standards for PDF authoring such that conforming files are accessible and usable to all, including those who use assistive technology. This effort brings together experts from organizations devoted to those with disabilities, education experts, text book publishers, and others developing tools for disabled people including representatives from standards committees working on complementary efforts. This work is unusually challenging as it is important to be able to describe to a person with disabilities the intricacies of the content and structure of the document to provide the context that a person without disabilities can gain in reading the document. This means that a document containing a table must be able to describe the table, its elements, format, and content to the disabled reader. For a sighted person, a table contains information that can be readily interpreted but this is not the case for a person with disabilities. The interpretation must be carried out from the file format as the assistive tool "reads" the document to the user.

While the final direction of the PDF/UA standard has not been established, it is anticipated that the completed standard will be proposed as an ISO standard and eventually adopted as a national standard for ANSI. Prior to being proposed as an ISO standard, the U.S. TAG (Technical Advisory Group) will be polled to achieve consensus on the proposal of the standard to ISO. By submitting the standard to ISO, the development of the standard will be enriched by the introduction of experts from other countries and representatives from other specialized ISO committees producing a more comprehensive standard for the industry.

### 2.6 PDF Reference

In January 2007, Adobe Systems announced its intent to release the full *Portable Document Format (PDF) 1.7* specification to AIIM for the purpose of publication by the International Organisation for Standardisation (ISO). AIIM represents ANSI (American National Standards Institute) at the ISO level. With the relationship developed between AIIM and Adobe Systems and the previous

[12] PDF Healthcare Frequently Asked Questions (FAQ)

PDF/Archive and PDF/Engineering efforts, the release of the PDF Reference to AIIM and ANSI was a logical step in the standardisation of PDF.

When one thinks of PDF tools and applications, the first product that comes to mind is Adobe's Acrobat. It may be hard to believe but there are over 700 PDF tools and applications available on the market and the numbers of tools are increasing at a rapid pace. The PDF Reference enables developers to produce software to ensure reliable, consistent viewing and printing of electronic documents created as PDF documents. By further standardising the PDF Reference as an ISO standard, the developers and users of PDF will now be stakeholders in the features and functionality of the PDF file format.

Adobe Systems has always made the PDF Reference freely available for developers to download. While Adobe holds the copyright on the specification and owns several patents in connection with the technology, they allow developers to use the specification to develop software and tools without the hindrance of royalties.

On behalf of ANSI, the U.S. TAG approved the acceptance of the PDF Reference from Adobe Systems to facilitate the standard through the ISO standards development process. The initial intent for the standard is to publish the PDF Reference 1.7 as an ISO standard without change. In July 2007, the document was submitted to ISO for fast track approval processing so that the standard could be published rapidly allowing the committee to then begin work on the next version of the standard which would include new features. Future versions of the standard will be developed by the ISO committee. Adobe Systems will be a stakeholder on the committee along with other organizations and will be able to recommend modifications for the standard. According to the press release announcing the release of the PDF Reference FAQ provided by Adobe Systems, "Adobe's products will remain conformant to the ISO PDF standard, once approved. Adobe will also continue to innovate and offer PDF extensions to enable new technologies and practices within its own implementation[13]."

**3.0 Issues**

Implementing these PDF file formats is not enough to ensure accurate preservation of electronic documents; organisations must also have the appropriate policies and procedures in place to provide the guidance for maintaining electronic documents and ensure the integrity of their documents. Organisations may obtain guidance for establishing records management policies and procedures from other standards such as *ISO 15489-1, Information and documentation – Records management – Part 1: General* which provides guidance to ensure that records are adequately managed. ISO 15489, part 1, provides guidance on the establishment of records management programme and the processes necessary for a comprehensive records management programme. *ISO 15489-2, Information and documentation – Records management – Part 2: Guidelines*, provides guidance to records managers as to how to implement ISO 15489-1 in their organization and provides an overview of the processes and

[13] http://www.adobe.com/pdf/release_pdf_faq.html

factors to consider when working toward making the organization's records management programme be compliant with ISO 15489-1. *ISO 14721, Space data and information transfer systems – Open archival information system – Reference model*, defines a reference model for a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access[14]. Through the use of these standards to implement a records management program, the use of the PDF/A file format becomes more effective to ensure the long-term preservation of an organization's electronic documents.

The availability of products to easily produce PDF/A, PDF/X, PDF/E, etc. files is an issue that needs consideration. Organizations can support the need for these products by requesting that the products be able to create files using the file formats. One way that these requests can be made is through incorporating the request as a requirement in RFPs (Request for Proposals). When the products are available, it is important to have a method to validate that the files created comply with the standards so that you will know that a PDF/A file is truly a compliant PDF/A file. The development of validating tools is emerging and will become increasingly more important as the file format is adopted. PDF validation is complex and often the validation tools do not agree on whether a file is valid or not.

While PDF/A will assist an organisation in the long-term preservation of their electronic documents, it is important to realise that the PDF/A file format may not be the only format for preservation that will be needed. New file formats are being developed all the time and there may be a newer, much better format developed that will better suit the business needs. Organisations should remain aware of the file format standards development efforts by joining industry associations that will help to keep them informed of these development efforts and by reviewing industry publications on a regular basis to determine if it may be appropriate to change to another file format. It is safe to say that correctly implementing the PDF/A file format should result in reliable, predictable, and unambiguous access to the full information content of electronic documents long-term.

**4.0 Technology**

The competing file formats to PDF and the variations include TIFF, XML, ODF, OOXML, and XPS. It must be understood that due to the specific nature of long-term preservation of electronic documents, the field of available file formats that can be used for preservation purposes is very small. It has only been recently that some competition in these file formats has been created with the proposed release of XPS to be developed as an ISO standard.

TIFF is widely acknowledged and adopted in the industry, but makes indexing the content in the TIFF documents more difficult and takes more processing time because the text is not readily available. TIFF does not provide a mechanism for capturing the logical structure of the document which makes it suitable for use with images. Creating TIFF files from 'digitally born' documents is difficult. While

[14] http://public.ccsds.org/publications/archive/650x0b1.pdf

TIFF is an excellent file format for the transfer of electronic information and good for converting paper to digital format, it is not an appropriate candidate for use in an archival setting for digitally born documents.

XML is good for describing the logical structure of a document. The appearance of the document is maintained through specifications such as XSLT. The XML format provides much capability for working with the text of the document and making the retrieval of XML documents easy. XML documents are easily modified and can maintain the history of the revisions in the XML file. As vendors integrate XML into their office applications and use the lessons learned from implementing the ISO Open Archival Information System (OAIS) Reference Model standard alternate archival capabilities will be acknowledged. When PDF/A is combined with these XML applications, organisations will have an archival alternative that will enable them to better respond to their regulatory and business archival needs while enhancing the access to their documents.

The Open Document Format standard defines an XML schema for office applications including the associated metadata. The schema covers all types of office documents, including text documents, spreadsheets, charts and graphical documents like drawings or presentations, but is not restricted to just these documents. It defines XML structures for office documents that will easily translate to other applications by using XSLT or similar XML-based tools. The OASIS developed OpenDocument v1.1 provides an introduction to the format, describes the metadata that is contained in the documents, and describes the text and paragraph content, tables, text fields, etc. OpenDocument makes use of existing standards such as HTML, SVG, XSL, SMIL, XLink, XForms, MathML and Dublin Core wherever possible to promote interoperability.

Office Open XML or OOXML is an XML based file format for all types of electronic office documents that will promote interoperability and further enable workflow or business process management in the office. It is intended to simplify the exchange of information between Microsoft Office products and enterprise applications. Office Open XML is based on the industry accepted XML standard and ZIP technologies. It was designed for the unstructured content created on products developed by the Microsoft Office products. It is a container format with specialise XML-based mark-up languages that correspond to the individual applications within the Microsoft Office product line. This file format may be less valuable for archival purposes as it may be considered a native file format. This does not mean that as the standard develops that archival functionality may not be included.

XML Paper Specification or XPS is a document storage and viewing specification developed by Microsoft that describes electronic paper in a way that it can be read by hardware, software, and the human eye. XPS provides a page view of the way the document will print. It describes the appearance of fixed format documents by using an XML based format so that the layout will not change. XPS is viewed as a potential competitor to PDF (Portable Document Format) but will not replace PDF in the instances where dynamic content capabilities are required which XPS cannot handle. Dynamic content

is the type of content that may be contained in a drop-down on a form. Microsoft released XPS with a royalty-free patent license to encourage wide adoption by the industry. The XPS document format is included with Windows Vista and 2007 Microsoft Office System products. The XPS viewer is included with Windows Vista and is available for Windows XP and Windows Server 2003. The viewer allows users to open, read, and apply digital signatures to XPS documents without needing the full XPS generating software. XPS allows electronic documents to print better, be shared easier, be archived, and better maintain the security of the information in the document. XPS may be a good archival alternate file format to PDF/A, however, until it is implemented more widely its validity as an archival file format is unknown.

**5.0 Implications**

The PDF/A file format is relatively new with the ISO standard being published only in 2005. Therefore, it is difficult to accurately assess the implications of using the file format. It should be noted that there is a wider acceptance for or adoption of the file format in Europe than is the case in the United States. This is largely due to the efforts of the PDF/A Competence Center (http://www.pdfa.org) that has been promoting the use of the standard since it was published.

PDF/Archive and the other variations of the PDF files are based upon the Adobe Systems, Inc. published *PDF Reference* which makes it easier for the available PDF readers to be able to open and display the various PDF formats leading to wider adoption of the formats. If the digital document life cycle is considered briefly, the PDF readers take care of the end of the process. For wide acceptance of the various PDF formats, tools are needed to be able to easily capture and create documents either digitally born or originally paper-based into the various file formats. PDF/A will become more widely adopted as more products introduce the capability to produce PDF/A compliant files as features in their products. Currently, there are over 20 products that produce PDF/A files. The use of a conformance testing process or file validator is needed to assess the level of compliance to the standard for the products and files created by the products. Both of these items are being discussed by the ISO Technical Committee developing the standard.

The PDF/Archive file format may be used in virtually any industry where electronic documents are created and are needed to be preserved. One example of the use of PDF/Archive is an organisation using PDF/A for the purposes of ensuring the organisation complies with the regulations for legal compliance as established by the government where their business is operating. The file format also ensures that the electronic documents are easily accessible by multiple individuals from various locations. An additional benefit from the use of the file format that this organisation is realising is the reduction of the electronic storage needed to house millions of documents.

When an organisation adopts the PDF/A file format for preserving their electronic documents, it is important that they realise that the file format is not sufficient to guarantee the long-term preservation

of their documents but that they also need to have the appropriate records management policies and procedures implemented as well. Additionally, the organisation needs to have established quality assurance processes and procedures in place that will evaluate the quality of the documents being preserved.

Education and training on the PDF/A file format and how to best implement it in an organisation is needed. The PDF/A Competence Center is leading the way on this effort. AIIM, the ECM Association, is in the process of introducing a web-based and in person training program for PDF/A that will assist organisations in understanding the standard and how it can be implemented in their organisation. Additionally, the PDF/A Competence Center and AIIM are collaborating on providing PDF/A conferences where case studies on the use of PDF/A are highlighted.

We are only at the beginning of the implementation and adoption of PDF/A as an electronic preservation file format, however, it is hoped that PDF/A will be widely adopted as the long-term preservation file format for the future.

**6.0 Activities**

The development work on these PDF standards has only begun. Much work remains to be completed to provide additional functionality to meet the continually developing requirements of users and their organisations to work more effectively with electronic documents. The ISO committees working on the ISO versions of the PDF standards have all adopted a position of using the ISO part numbers for updates and revisions of the standards which bring the base standard current to the latest PDF Reference specification without obsolescing previous parts. This action makes locating the standards easier as PDF/X is the ISO 15930, PDF/A is ISO 19005, and PDF/E is ISO 24517.

The PDF working groups recognised the need for all PDF committees to hold simultaneous meetings in order to better share issues and requirements for the PDF work. As the groups developed the standards they realised the reliance that each standard has on the other standards. Holding the joint meetings will foster better communications as the committees share requirements.

ISO TC 130, Graphic Technology will be publishing the following documents soon:

- ISO 15930-7, *Graphic technology – Prepress digital data exchange using PDF – Part 7: Complete exchange of printing data (PDF/X-4) and partial exchange of printing data with external profile reference (PDF/X-4p) using PDF 1.6*
- ISO 15930-8, *Graphic technology – Prepress digital data exchange using PDF – Part 8: Partial exchange of printing data using PDF 1.6 (PDF/X-5)*

The United States committee, CGATS, Committee for Graphics Arts Technologies Standards, represents the U.S. interest in the work of PDF/X. At the time this report was being written, there were

no plans for any additional work to be performed on PDF/X for the immediate future to allow the standard to be adopted and products to be developed.

Shortly after being published, PDF/X was adopted early by Time Magazine as the required method for advertisers to submit artwork for advertisements. Early adoptions like this, lead to other magazine printers requesting and accepting PDF/X files.

The ISO TC 171 SC2 working group is drafting ISO 19005-2, *Document management – Electronic document file format for long-term preservation – Part 2: Use of PDF 1.6 (PDF/A-2)*. The new features of this version including JPEG 2000 image compression, more sophisticated digital signature support, OpenType fonts, 3D graphics, audio/video content, and consistency with other PDF-based standards, along with being based on PDF Reference 1.6. The draft for part 2 is being reviewed by the member countries to accept it as a new work item and to approve the working draft to make it available for consideration as a committee draft. There will be several more ballots before the standard is published which is not expected until sometime in early 2009.

The working group recently began identifying items to be included in part 3 of this series. It is important to note that parts 1 and 2 dealt with static documents. With the introduction of part 3, the working group began discussions on how to deal with documents that have dynamic content and how that content should be preserved for the long-term. Part 3 will be based on ISO 19005-2 and will enable the archiving of PDF/E (ISO 24517-1) compliant documents. This part of the document is proposed to remain platform-neutral and may loosen some of the PDF/A-2 restrictions such as the embedding of non-PDF/A compliant documents and encryption. Additionally, consideration will be given to including 3D in the PDF/A file format and providing more structure as well as additional metadata options. There is no expected date for a new work item and working draft for PDF/A-3.

In the case of PDF/Archive, which was developed with the need for implementation flexibility to promote its wide adoption, there are numerous products being introduced whose developers claim to comply with the ISO 19005-1 standard, including but not limited to the following:

- Adobe Acrobat 7.0 (compliant to ISO/DIS 19005-1)
- Adobe Acrobat 8.0
- Amylini Technologies PDF Converter
- Apago PDF Appraiser
- Aquaforest TIFF Junction
- BFO PDF Library
- Callas Software Pdfa Pilot
- Callas Software pdfInspekto3 CLI PDF Preflight Check
- Crawford Technologies Pro PDF/A Product line
- CRT Europe Pro PDF/A Products

- Detec myPDFconvert
- eDOCUMAN
- Elpro PDF Data Logger (USB Device)
- Groupware e:PDF Server
- Gumbo Software SpoolMail
- LuraTech PDF Compressor
- LuraTech PDF/A Printer
- LuraTech PDF/A Validator
- Maas AFP2Web MultiConverter
- Nuance ScanSoft PDF Converter 4
- PageTech PCL Works
- PageTech PCL Tool SDK
- PDFlib GmbH PDFlib 7
- PDF-Tools PDF/A Compliance Product
- Prime Recognition PrimeOCR
- Software602 Print2PDF
- Tracker Software PDF XChange
- Visioneer Scanning and document imaging solutions (Support PDF/A)
- Xerox-brand Scanning and document imaging solutions (Support)
- Xenos d2e

Additional products are needed and integration of these products into the traditional Enterprise Content Management suites is a requirement. Organisations that are adopting the PDF/A standard need to be able to easily create PDF/A files whether it is from a native file or through the scanning process.

ISO 24517, *Document management – Engineering document format using PDF – Part 1: Use of PDF 1.6 (PDF/E-1)* developed by ISO TC 171 SC2 WG7, Document Management Applications, Application Issues, PDF/Engineering is expected to be published before the end of 2007. The working group is beginning discussions of the features to be included in ISO 24517-2 or PDF/E-2. In addition to determining what new features should be included in ISO 24517-2, the working group is also finalising application notes which provide background to the standard and serve as an aid for the PDF tool developers and a set of FAQ (Frequently Asked Questions) which help users understand the standard and its benefits.

ISO/DIS 32000, *Document management – Portable Document Format (PDF)* began the ISO fast track approval process in July 2007. The DIS ballot which will close on December 2007 is expected to approve the *PDF Reference* as an ISO standard. Upon publication of the ISO 32000, the ISO working group will begin work on the next version of PDF in which end users, PDF developers, and other software developers will be able to for the first time help specify the way PDF will look and operate.

The *Portable Document Format Healthcare Best Practices Guide* and the *Implementation Guide for the Portable Document Format Healthcare* are expected to be approved and published in early 2008. These companion documents will describe the features and functions for the voluntary use of Portable Document Format (PDF) in the healthcare industry. The Best Practice will establish how PDF can be used to provide secure exchange of personal healthcare information. PDF Healthcare will allow a high level of accessibility and interoperability through its use of XML.

The PDF Healthcare Implementation Guide was designed to facilitate the implementation of the technical items described in the Best Practice Guide so that healthcare providers would be able to use PDF as an electronic container for information. The committee is discussing the next steps it will take in developing guidance for the use of PDF in the healthcare industry. This work in the healthcare industry is more critical now than it has ever been. Given the amount of travel people do and the need to move to electronic healthcare data in light of natural disasters such as what the United States faced with Hurricane Katrina and what Indonesia faced with the tsunami, these best practices and implementation guides are necessary to establish how the use of PDF can provide the control patients need to have over their health information. Through providing accurate clinical and administrative healthcare data to the point where patients are being treated, the quality of healthcare will be improved and the cost associated with it will be decreased.

As medical records migrate to a universal digital format, the adoption of a document encapsulation practice would contain specifications for portability, interoperability, and security will promote the exchange of healthcare information and in turn would promote the adoption of PDF Healthcare.

The PDF Universal Access or PDF/UA committee is targeting being able to conduct an initial approval ballot on this standard by mid-2008. The approval process is expected to extend through the end of 2008 with publication expected in 2009. After approval and publication, the committee will determine if a recommendation to submit the standard to ISO should be made to the U.S. TAG to ISO TC 171.

The work on these committees is done by subject experts who come from many different countries such as Australia, Bulgaria, China, France, Japan, Poland, Russian Federation, South Africa, Spain, Sweden, Switzerland, Ukraine, United Kingdom, and United States. The experts represent all areas of the industry from federal government to local government and private industry representing healthcare and pharmaceuticals, manufacturing, financial institutions, and software developers.

In addition to the standards developed, the committees are also drafting Application Notes, FAQs (Frequently Asked Questions), and other informational materials that explain the standards. The application notes add explanation to the information presented in the standard to assist developers and users of the standard. In order to promote the adoption of the PDF standards, AIIM is in the process of introducing white papers, blog postings, podcasts, and other informational events promoting the

standards. The PDF/A Competence Center offers events on PDF/A and has published a book, *PDF/A in a Nutshell: Long Term Archiving with PDF*. As versions of the standard are nearing publication, the committee's focus shifts to education and awareness in the industry to promote the adoption of the standards as they are published.

Adoption of these standards by organisations also requires education and promotion of the standards. This is one area that most standards organisations do not apply enough attention. Training is a key element to ensuring adoption of the standard. Records managers, IT, and archivists need to know how to implement the standards in their organisation. This requires gaining an understanding of the requirements for the digital documents in the organisation. One international organisation that is championing the PDF/A work in Europe is the PDF/A Competence Center. The aim of this group is to promote the exchange of information and experience in the area of long-term archiving in accordance with PDF/A. The organisation has members from over forty software developers and solution providers. Through the efforts of the PDF/A Competence Center, PDF/A is being adopted throughout Europe at a rapid pace.

**7.0 Recommendations**

PDF has become a widely used file format that is integrated into many desktop applications. Before adopting any of the PDF subset file formats, organisations must consider the alternative file formats that are available, understand their content (documents and records), and how they use electronic information. It is critical to understand the purpose of the electronic information as that will be a determining factor in choosing the file format to best suit the organisation's needs.

Organisations that choose to adopt PDF/A need to implement additional policies, procedures, and requirements on the processes used to generate conforming files to ensure the reliable rendering of the documents. This means organisations need to implement records management policies and retention requirements for the organisation based on ISO 15489. When creating electronic archives, the organisation should also consider the recommendations established in the OAIS Reference Model. Quality assurance processes are needed to ensure the files are created in conformance with the standards and that the archives are being maintained in accordance with the established policies and procedures. The need for these general records management best practices is not reduced by selecting to adopt PDF/A to archive electronic documents.

It is important to note that one file format may not fit all needs in an organisation. As the organisation's needs change, the file formats chosen need to be reviewed to ensure that they fill the needs and meet the ever changing regulatory compliance requirements. The development of file formats such as XPS and OOXML should be followed and implemented where appropriate in an organisation. Increasingly, file formats are being based on XML which provides a level of interoperability amongst the file formats and aids in the preservation of electronic documents created using the formats.

**Glossary**

| | |
|---|---|
| AIIM | Association for Information and Image Management |
| ANSI | American National Standards Institute |
| ASTM | Formerly, American Society for Testing and Materials |
| CCR | Continuity of Care Record |
| CGATS | Committee for Graphics Arts Technology Standards |
| DIS | Draft International Standard (ISO) (Next to the last approval stage prior to publication) |
| ECM | Enterprise Content Management |
| ISO | International Organisation for Standardisation |
| JWG | ISO Joint Working Group |
| HTML | Hypertext Markup Language |
| MathML | Mathematical Markup Language |
| OASIS | Organisation for the Advancement of Structured Information Standards |
| OCR | Optical Character Recognition |
| OOXML | Open Office XML |
| PDF | Portable Document Format |
| PDF/A | Portable Document Format/Archive |
| PDF/A-1 | Portable Document Format/Archive (ISO 19005-1) |
| PDF/E | Portable Document Format/Engineering |
| PDF/UA | Portable Document Format/Universal Access |
| PDF/X | Portable Document Format/Exchange |
| RDF | Resource Description Framework |
| SC | ISO Subcommittee |
| SMIL | Synchronized Multimedia Integration Language |
| SVG | Scalable Vector Graphics |
| TAG | Technical Advisory Group (Represents a country at International Standards Meetings) |
| TC | ISO Technical Committee |
| TIFF | Tagged Image File Format |
| U3D | Universal 3D |
| W3C | World Wide Web Consortium |
| WG | ISO Working Group |
| XFORMS | XML Forms Format |
| XLINK | XML Linking Language |
| XML | eXtensible Markup Language |
| XMP | Extensible Metadata Platform |

| | |
|---|---|
| XPS | XML Paper Specification |
| XSL | Extensible Stylesheet Language |
| XSLT | XSL Transformations |

**References/Bibliography**


http://www.aiim.org/pdfa PDF/A Committee Web Site

http://www.aiim.org/documents/standards/19005-1_FAQ.pdf PDF/A Frequently Asked Questions

http://www.aiim.org/standards.asp?ID=29510 PDF/A Application Notes

http://www.aiim.org/pdfe PDF/E Committee Web Page

http://www.aiim.org/pdfua PDF/UA Committee Web Page

http://www.aiim.org/standards.asp?ID=33736 PDF Expert Corner

http://pdf.editme.com PDF Wiki Collaboration Site

http://www.npes.org/standards/toolspdfa.html PDF/X and PDF/A Resources at NPES

http://www.aiim.org/standards.asp?ID=31832 PDF Healthcare

http://www.aiim.org/standards.asp?ID=31979 PDF Healthcare Frequently Asked Questions

PDF/E Frequently Asked Questions (FAQ) (To be published)

ISO 19005-1:2005, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*

ISO 24517-1:2007, *Document management – Engineering document format using PDF – Part 1: Use of PDF 1.6 (PDF/E-1)*

ISO15489-1:2001, *Information and documentation – Records management – Part 1: General*

ISO 15489-2:2001, *Information and documentation – Records management – Part 2: Guidelines*

Druemmer, Olaf, Alexandra Oettler, and Dietrich von Seggern. *PDF/A in a Nutshell: Long Term Archiving with PDF*. Berlin, Germany: Die Deutsche Bibliothek, 2007.

Khoshafian, Setrag, A. Brad Baker, Razmik Abnous and Kevin Shepherd. *Intelligent Office – Object-oriented Multi-media Information Management in Client Server Architectures*. New York: Wiley, 1992.

Sellen, Abigail J. and Richard H. R. Harper, *The Myth of the Paperless Office*. Cambridge, MA: The MIT Press, 2003.

**About the Author**

Betsy Fanning is the director of Standards for AIIM. In this position, she is responsible for the standards and technical reports produced by AIIM as ANSI (American National Standards), ISO, and AIIM Recommended Practices as well Best Practices. At the international level, she is the secretary for ISO TC 171, Document Management Applications and ISO TC 171 Subcommittee 2, Application Issues and is the administrator for the U. S. Technical Advisory Group to TC 171 that represents the United States at the international meetings. She is also responsible for building liaison relationships with other standards development organisations such as the Workflow Management Coalition, the Object Management Group, the Organisation for the Advancement of Structured Information Standards, and others that interface with AIIM's technologies and is a member of a number of ANSI committees.

Prior to coming to AIIM, Betsy held positions with DynSolutions, a Correspondence, Document and Records Management Company, and Westinghouse Electric. In both of these companies, she has implemented imaging, workflow and document management systems. In 1994, Betsy was awarded AIIM's Distinguished Service Citation. Betsy has a Masters in Library Science from the University of Pittsburgh and a Bachelor of Science in Education specialising in Library and Information Sciences from Clarion State University.

**About AIIM – The Enterprise Content Management Association**

AIIM is the international authority on Enterprise Content Management (ECM). ECM is the technologies used to capture, manage, store, preserve, and deliver content and documents related to organisational processes. ECM tools and technologies provide solutions to help users with the four C's of business: Continuity, Collaboration, Compliance, and Costs.

For over 60 years, AIIM has been the leading non-profit organisation focused on helping users to understand the challenges associated with managing documents, content, records, and business processes. Today, AIIM is international in scope, independent, implementation-focused, and, as the representative of the entire ECM industry - including users, suppliers, and the channel - acts as the industry's intermediary.

As a neutral and unbiased source of information, AIIM serves the needs of its members and the industry by providing educational opportunities, professional development, reference and knowledge resources, networking events, and industry advocacy.

Information about AIIM can be found at www.aiim.org.