PDF/A: An Electronic Document File Format for Long-Term Preservation

Stephen L. Abrams
Harvard University
Cambridge, Massachusetts, USA

Stephen P. Levenson
Administrative Office of the US Courts
Washington, DC, USA

Abstract

PDF has emerged as one of the predominant formats for creating, capturing, storing, and delivering electronic documents. As it is increasingly applied to content deserving or requiring long-term retention, questions arise as to what technical mechanisms and best practices can ensure effective preservation. Long-term preservation entails not merely maintaining the fixity of the byte stream, but also ensuring continuing access to, and interpretation of, the full information content of archived PDF documents. PDF supports many sophisticated features that tend to subvert efforts towards effective preservation. To address these concerns the International Organization for Standardization (ISO) has established a joint working group to develop a standard to specify how to use PDF for long-term preservation of electronic documents. ISO 19005-1, known familiarly as PDF/A, defines a constrained subset of the file format and a set of functional requirements for PDF/A readers. This paper provides a snapshot of the evolving PDF/A standard and addresses its potential impact for digital preservation.

Introduction

PDF is increasingly the format of choice for electronic documents. For many important and sensitive institutional, commercial, and government applications the PDF electronic document is the document of record. Under various statutory, regulatory, and policy regimes the informational and evidentiary integrity of these documents must be preserved. Unfortunately, the very basis of PDF's ubiquity – its ability to encapsulate a wide range of static and dynamic information resources and media types, potentially in device-dependent representations – tends to complicate the preservation process.[1] How can color fidelity be ensured if a document uses uncalibrated colorspaces? How can the visual integrity of a document be maintained if a rendering agent can use substitute fonts? How can any preservation activity be performed on an encrypted file?

In order to address these concerns an open meeting of interested parties was held in Washington, DC, in October 2002, under the joint auspices of the Association for Information and Image Management (AIIM) and NPES, The Association for Suppliers of Printing, Publishing and Converting Technologies, with the goal of fostering the establishment of an international standard for a preservation profile for PDF. The intent of this profile was to define a set of constraints and best-practice guidelines for the PDF format necessary to ensure predictable and reliable behavior of future rendering of PDF documents. Additionally, the profile enables the future retrieval of the full information content of conforming documents at syntactic, structural, and semantic levels.

The PDF/A Standards Process

Through a series of subsequent meetings, the ad hoc AIIM/NPES group, consisting of representatives from government, industry, library and archival institutions, and software developers including Adobe Systems Incorporated, produced a draft specification for an archival form of PDF, known informally as PDF/A. This draft was submitted to the International Organization for Standardization (ISO) for consideration as a new work item (NWI) in April 2003. The proposal was accepted in August 2003 and ISO subsequently established a new Joint Working Group (JWG), ISO/TC 171/SC 2/WG 5 (*Document management applications – Application issues – PDF/A*), to continue work on the PDF/A standard, ISO 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF (PDF/A)*.

The diverse JWG membership reflects the widespread interest in PDF/A among the standards community and includes representatives of the following ISO Technical Committees and Subcommittees:

- TC 42 *Photography*
- TC 46/SC 11 *Information and documentation – Archives / records management*
- TC 130 *Graphics technology*
- TC 171/SC 2 *Document management applications – Application issues*

US interests in the PDF/A standards process are represented by the ANSI-accredited US Technical Advisory Group (TAG) to the JWG.

The JWG work is on schedule to lead to a promulgated international standard as early as January 2005:

- October 2002 Formation of AIIM/NPES committee
- August 2003 Approved ISO NWI
- October 2003 JWG New Orleans meeting
- December 2003 First Committee Draft (CD)
- March 2004 JWG New York meeting

. . .

- January 2005 International Standard

The current work on PDF/A will result in a document that is part one of a multi-part standard. Additional parts may be added in the future to address other considerations of PDF-based preservation, such as multi-media content or data conversion practices.

Scope of the Standard

As expressed in the first committee draft of December 2003, PDF/A is scoped to "[specifying] how to use the Portable Document Format (PDF) for long-term preservation for electronic documents. It is applicable to documents containing combinations of character, raster, and vector data."[2] The PDF/A document model explicitly does not encompass multi-media documents containing audio or video content. Additionally, PDF/A specifically excludes considerations of methods of document capture or conversion; specific technical design, user interface, or implementation details of rendering; and physical methods of document storage. While the standard does include some functional requirements for conforming software, its main emphasis is on defining a preservation file format, leaving specific implementation details to vendors and workflow considerations to local policy.

The definition of "long-term" is taken from the Open Archival Information System (OAIS) reference model: a "period of time long enough for there to be concern about the impacts of changing technology…. [extending] into the indefinite future."[3] Thus, many of the requirements specified by PDF/A are designed explicitly to de-couple the interpretation of document content from any specific hardware or software environment. Effective long-term preservation involves more than just maintenance of the fixity, or bit-stream integrity of an electronic document; more important is the preservation of the usability of the document. An exact copy of a 10 year old bit-stream is of no use if it cannot be reliably rendered. PDF/A is intended to preserve the usability of electronic documents at syntactic, structural, and semantic levels.

Syntactic preservation maintains only the visual integrity of rendered PDF/A documents. For example, a textual document could be preserved syntactically as an image. However, higher-level understanding of such a document may require a human interpreter. Structural preservation maintains the logical organizational integrity of a document (its pages, chapters, sections, paragraphs, etc.) in a manner that is recognizable by automated processes. Similarly, semantic preservation, which is also amenable to automatic processing, maintains the integrity of higher-order information content such as the document text stream in its native script system and natural reading order and descriptive and administrative metadata. An effective preservation format for electronic documents should support the explicit internal representation of document content at all three levels in order to support anticipated preservation activities such as emulation or migration.

Just as bit-level preservation is not sufficient to guaranty long-term usability of document content, the mere fact that PDF/A is used as a representation format also may not be sufficient in itself to guaranty usability. However, incorporating the use of PDF/A into a comprehensive program for digital preservation with well thought-out policies and practices for data capture, managed storage, and pro-active intervention against obsolescence does represent the best strategy for long-term electronic document preservation.

What's Allowed, What's Prohibited

The base-line criterion for PDF/A is adherence to PDF 1.4 as amended by its public errata.[4] Beyond this, the standard defines two conformance levels: (1) minimal conformance, which includes requirements applicable to ensuring the integrity of the visual appearance of documents; and (2) full conformance, which includes requirements applicable to ensuring the recoverability of higher-order structural and semantic properties. Minimal conformance is a necessary prerequisite for full conformance.

The specific requirements defined by PDF/A can be separated into three categories: (1) features that are mandatory; (2) features that are allowed; and (3) features that are prohibited. Decisions regarding the categorization of PDF 1.4 features were based upon the evaluation of those features with respect to the following preservation criteria:[5]

- Device independence: the degree to which a PDF/A file is independent of the software and hardware platform on which it is interpreted and rendered.
- Self-containment: the degree to which a PDF/A file contains all resources necessary for its reliable and predictable interpretation and rendering.
- Self-documentation: the degree to which a PDF/A file documents itself in terms of descriptive, administrative, structural, and technical metadata.

- Transparency: the degree to which a PDF/A file is amenable to direct analysis with basic tools, including human readability.

The majority of PDF 1.4 features fall into the category of being allowed in PDF/A.

The following summary of PDF/A requirements is based on the December 2003 Committee Draft (CD). The requirements are subject to revision during subsequent ISO activities.

General File Format

No data can precede the file header, follow the last file trailer, or appear between indirect objects. The header must be followed by a comment containing at least four binary byte values to facilitate 8-bit safe transmission over transport protocols.

The LZWDecode filter is prohibited due to its legal encumbrances. The ASCII filters are prohibited as being of little practical value. Encryption is prohibited to ensure the transparency of PDF/A documents and to avoid intellectual property restrictions and the necessity for external password management.

Stream data cannot be defined by external files in order to ensure that PDF/A files are self-contained. Content streams cannot contain any operators not documented in PDF 1.4 even if they are bracketed by the BX/EX compatibility operators. This is to ensure the known semantics of all operators found in PDF/A content streams.

A PDF/A file must conform to the limits on quantities defined in Table C.1 of the PDF 1.4 reference in order to provide predictable rendering behavior across all software and hardware platforms.

Graphics

All colors must be specified in a device-independent colorspace or must be used in conjunction with an OutputIntent to provide guidance to facilitate color accuracy and reproducibility in rendering.

Alternative images are prohibited in order to ensure predictability in rendering. Interpolation is prohibited as being implementation dependent. Reference XObjects are prohibited in order to ensure that PDF/A files are self-contained. PostScript XObjects are prohibited as unnecessary and unevenly implemented by rendering software. OPI proxies are prohibited from images and form XObjects in order to ensure that PDF/A files are self-contained.

Fonts

All font subsets referenced within a PDF/A file must be embedded within that file. Furthermore, these fonts must be legally embeddable for unlimited, universal rendering. For full conformance all fonts shall have a means to map their characters to the equivalent Unicode characters, either using a well-defined encoding scheme or Cmap or by specifying a ToUnicode dictionary, allowing the retrieval of character semantic properties.

Transparency

Consistent with the decision made in PDF/X (ISO 15930), the pre-press data exchange standard, transparency is not allowed due to the device-dependent manner in which it is often implemented. The visual effect of transparency can be simulated by other techniques such as pre-rendered data or flattened vector objects.

Annotations

The FileAttachment, Movie, and Sound annotation types, as well as any type not defined in PDF 1.4 are prohibited. All non-textual annotations should provide a human-readable alternative description in their Contents key. Conforming PDF/A rendering software must provide a means to expose the contents of annotations independent of an annotation's appearance stream.

Actions

The Sound and Movie actions are prohibited as multi-media content is out of scope for the initial version of PDF/A. The Launch and JavaScript actions are prohibited since arbitrary executable code could affect the visual appearance of a document and raised security concerns. The ResetForm and ImportData actions are similarly prohibited. The GoToR, SubmitForm, and URI actions are allowed but conforming PDF/A rendering software may choose to make hyperlinks non-actionable, but shall provide a means to expose the destination of the links. Additional-actions are prohibited from form fields and the document catalog since JavaScript execution could affect the visual appearance of a document.

Metadata

The document information dictionary is prohibited in favor of a mandatory XMP metadata stream accessed through the document catalog Metadata key. (This decision remains controversial among the JWG and may be rescinded in a subsequent Committee Draft.) XMP (Extensible Metadata Platform) is an RDF-based metadata standard embeddable in PDF documents as XMP packets.[6] Most document information dictionary entries can and should be expressed in XMP form. At the instigation of the JWG Adobe is revising the XMP core schema to include a repeatable Identifier property for providing unambiguous identification of a document. XMP file provenance metadata should be used to describe all steps taken to create, modify, and transform the PDF/A document. Furthermore, all modifications to XMP metadata values as a document moves through its life-cycle should also be described.

The JWG is exploring ways in which to specify finer-grained metadata relative to specific structures within a PDF/A document such as individual images.

Logical Structure

PDF/A requirements concerning document logical structure are applicable only for full conformance to the standard. These requirements relate to preserving the higher-order structural and semantic properties such as document

logical structure, text stream in natural reading order, and descriptive and administrative metadata. Meeting these requirements may place greater burdens on PDF/A writers but access to these properties does expose the full information content of PDF/A documents and may facilitate future preservation migrations or other activities.

A fully conforming PDF/A document shall meet all of the requirements for Tagged PDF as defined in PDF 1.4. Pagination, layout, and page artifacts should be specified to the fullest extent possible. Similarly, the logical structure of a document should be captured in a structure hierarchy whose block-level elements should follow the PDF 1.4 strongly structured paradigm to the fullest extent possible.

A document's default natural language must be specified in the document catalog. Textual content that differs from the default language should be identified using the Lang property. Alternative descriptions of non-textual structure elements, replacement text for non-standard text representations, and abbreviation and acronym expansion should be provided to the fullest extent possible.

Forms

All form fields must have an appearance dictionary associated with the field's data. Conforming PDF/A rendering software must render a field according to that dictionary without regard to the underlying form data. Rendering software is prohibited from implementing any feature that would allow a document's appearance to change.

Use of PDF/A in Preservation Workflows

PDF/A can be introduced easily into existing document management workflows. Since PDF/A is a constrained subset of PDF 1.4, a PDF/A document can be substituted freely for a "generic" PDF document without adverse consequences. Based on the example of PDF/X it can be anticipated that the promulgation of PDF/A as an approved ISO standard will lead to a variety of commercial and open source tools with native PDF/A support for file creation, conversion, validation, and rendering.

The US National Archives and Records Administration (NARA) has recently issued guidance for the transfer of permanent electronic records.[7] PDF/A meets all of NARA's general requirements with regard to security features and fonts. The PDF/A standard also provides an informative annex on best practices that is consistent with the NARA guidance on document conversion. In the future, additional parts may be added to the 19005 standard defining normative standards for records management policies and practices.

The Administrative Office of the US Courts (AOUSC) has been implementing automation solutions for document management in the federal judiciary since 1993. In 1995 PDF was selected as the preservation format for both paper and born-digital documents. PDF/A should facilitate the ability of AOUSC to carry out its statutory and regulatory obligations to maintain the integrity of court documents into the future. Similar projects are underway in other branches of government as well as by regulated industry.

The PDF/A standards process was instigated by the widespread use of PDF as a representation format for electronic documents. Although PDF/A represents a best effort at defining a preservation profile for PDF, it is not necessarily certain that a PDF-based strategy is indeed the best one for document preservation. Should future preservation activity require the migration of PDF-based content to another substantive format, for example XML, the features of PDF/A will permit the recovery of the full information content of PDF/A documents at a syntactic, structural, and semantic level, mitigating potential loss incurred during the transformation process.

Conclusion

PDF has emerged as one of the predominant formats for creating, capturing, storing, and delivering electronic documents. To meet the needs of librarians, archivists, and others interested in the future usability of electronic documents, ISO has initiated a process leading to an International Standard, ISO 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF* (PDF/A), defining a constrained profile of PDF suitable for the long-term preservation of document content at syntactic, structural, and semantic levels. PDF/A defines the set of PDF 1.4 features that are mandatory, those that are allowed, and those that are prohibited. The use of PDF/A in the context of a managed digital preservation program should ensure the continued integrity of the visual appearance, logical organization, and semantic content of electronic documents for future use.

References

1. John Mark Ockerbloom, Archiving and Preserving PDF Files, *RLG DigiNews* **5:1** (2001).
2. ISO/CD 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF (PDF/A)*, 2003.
3. ISO 14721, *Space data and information transfer systems – Open archival information system – Reference model*, 2002.
4. Adobe Systems Incorporated, *PDF Reference: Adobe Portable Document Format, Version 1.4*, Addison-Wesley, Boston, 3rd ed., 2001.
5. Caroline R. Arms and Carl Fleischhauer, Digital Formats: Factors for Sustainability, Functionality, and Quality, *DLF Fall Forum* (2003).
6. Adobe Systems Incorporated, *XMP Specification* (2004).
7. National Archives and Records Administration, *Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: Records In Portable Document Format (PDF)* (2003).

Biographies

Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, providing technical leadership for strategic planning, design, and coordination of

the Library's digital systems, projects, and assets. He is currently engaged in research and implementation of effective methods for archival preservation of digital objects. Mr. Abrams is the ISO project leader and document editor for ISO/TC 171/SC 2/WG 5, the joint working group developing the PDF/A standard. He is a member of ACM, ALA, ASIS&T, and the IEEE Computer Society.

Stephen Levenson is the Electronic Records Policy Officer for the Administrative Office of the US Courts. Mr. Levenson is the ISO project convener for ISO/TC 171/SC 2/WG 5, the joint working group developing the PDF/A standard. He is currently engaged in both research and implementation issues involved in the conversion of case files to digital formats