The PDF/A Value Proposition

Stephen Levenson| PDF/A ISO International Convener

Why PDF/A?

Why a "standard" version of PDF

- PDF is powerful and flexible, …
- But too flexible for some applications
- Need higher degree of reliability than required by published spec
- Also, may want standard in hands of neutral non-commercial body (e.g. ISO)

The idea of PDF/A

- A restricted subset of PDF
- Published by recognized standards body such as ISO
- Focused on Archive needs of governments, corporations, libraries

Issues

- Protection against accidental vs. malicious manipulation of format
- Preserving visual appearance vs. content extraction
- Fonts
  - Embed vs. not
  - Restrict the fonts used?

Some interesting things from a 2004 presentation…

PDF/A

- A = Archive (but say it softly)
- Markets: government, regulated industry, libraries
- Focus on statutory requirements for documents to be retrievable after long periods of time
- Minimize risk of not being able to render page content

PDF/A hot issues

- Lingering objections to the whole idea
  - PDF as “proprietary”
  - Too complicated, want black&white standard first
- Desire for restrictions on creation process
  - No downsampling
  - No JPEG
- Metadata: InfoDict vs. XMP
- OutputIntent mandatory?
- Removed syntactic restrictions

Long Term Document Requirements vs. Data Requirements

PDF/A Issues

- Assure predictable appearance – main goal, well understood
- Role of structure, content extraction?
- Should standard address security, signatures?

PDF/A: Reliable Appearance

- Prohibit external content
  - External files, OPI, non-embedded fonts, Reference XObjects
- Restrict programming
  - No JavaScript
- Restrict annotations
- Restrict fonts? (Hazards of TrueType)
- Restrict device-dependent operators (transfer functions, overprint, halftones).

PDF/A Objectives (2005)

- Device independence
  - Can be reliably and consistently rendered without regard to the hardware/software platform
- Self-contained
  - Contains all resources necessary for rendering
- Self-documenting
  - Contains its own description
- Transparency
- Amenable to direct analysis with basic tools
- (Lack of) technical protection mechanisms
  - No encryption, passwords, etc.
- Disclosure
  - Authoritative specification publicly available
- Adoption
  - Widespread use may be the best deterrent against preservation risk

Benefits of PDF/A (2005)

- Non-proprietary standard
  - Based on a proprietary, but open format
- Developed by inclusive set of stakeholders
- Subject to rigorous technical review
- Minimal restrictions necessary to facilitate long-term preservation
- Not reliant on the existence of any particular reader

PDF/A Objectives (Adobe 2005)

- A document format that
  - Can convey critical information
  - Can be rendered accurately
    - “consistent and predictable”
  - Can incorporate associated “marginalia”
    - Notes scrawled in the margins on documents
    - Today perhaps as version history, general metadata, etc.
- A definition of how retrieval devices (Readers) will behave

Issues Not Addressed in a File Format

- Hardware migration issues
- OS and application migration issues
- Document management system issues
  - Document access tracking
  - Digital Rights Management

Agency Archive

- What are the basic points to remember when considering PDF/A-1 for permanent records?
  - PDF/A-1 is **one file format option for electronic records**.
  - PDF/A-1 should **allow PDF records to be maintained longer as PDF** (e.g., within agencies) **because it has fewer "bells and whistles" than traditional PDF** and should minimize future migration requirements.
  - PDF/A-1 is **more open than traditional PDF because PDF/A-1 is maintained by ISO** and not one specific vendor.
  - To ensure the **quality, integrity, and authenticity of information maintained as PDF, agencies should implement PDF/A-1** in conjunction with records management and quality assurance policies and procedures such as ISO 15489-1 Information and documentation - Records management - Part 1: General.
- *http://www.archives.gov/records-mgmt/initiatives/pdf-faq.html*

EPA Example (Environment Protection Agency)

- Full-Text Search
  - PDF/A stores objects (e.g. text, graphics), **allowing for an efficient full-text search in an entire archive**. Files stored as Tagged Image File Format (TIFF) cannot be searched. TIFF is a raster format and must first be scanned with an OCR (optical character recognition) engine.
- File Size
  - **PDF/A files require only a fraction of the memory space of original or TIFF files, without loss of quality**. The smaller file size is especially advantageous for electronic file transfers (FTP, e-mail attachment, etc.)
- Optimization
  - **PDF/A format can be optimized**. The optimization can be focused on images (e.g. scanned checks) or extracting structured data (e.g. voucher information). TIFF treats all file information the same.
- Metadata
  - **Metadata** like title, author, creation date, modification date, subject, keywords, etc., **can be stored in a PDF/A file**. PDF/A files can be **automatically classified based on the metadata**, without requiring human intervention.

- http://www.epa.gov/records/faqs/pdfa.htm

- The self-contained nature of pdf/a also provides an excellent format for providing on-line access to reports, having far greater longevity than standard pdf files, some of which were starting to produce error messages ten years after deposit – a problem I was grappling with in my last couple of months working at the ADS (not the cause of my leaving I might add!)
- *http://digital-archiving.blogspot.com/2013/03/some-thoughts-on-pdf-version-3.html*

WHY PDF/A

- Why has a special PDF standard now been defined for archiving documents? Are traditional PDF documents not "good enough" for long-term archiving? PDF has some excellent characteristics that lend themselves to the creation of archived documents. Like a container, a **PDF can incorporate completely different elements such as text, images, and fonts. In addition, it reproduces layouts that are true to the original and it is cross-platform capable**.
- **PDF in its native form cannot guarantee long-term reproducibility though**. Certain requirements must be met in order to enable the exact reproduction of content in say 10, 20 30 years or more. **For example, it is essential that fonts must be embedded** as PDF/A is; a link to the font in question is not sufficient as is the case in a PDF file. If a font is not embedded in a PDF document it means that if, in that 10, 20, 30 years plus time-frame, a user who tries to open a document does not have a required font on his or her computer, special characters or symbols will not be displayed correctly. Fonts that are common on today's computers may not be available in the future. Imagine the problems this could create - a critical piece of information may be lost from a case file (legal field) simply because the font used to display the missing characters is no longer available.
- **The Bottom Line: By using PDF/A files, your fonts will look the same 30 years from now as they did the day they were created**.

- *http://www.thechoice4biz.com/pdfa-the-digital-document-archiving-standard.html*

WHY PDF/A (continued)

- **This need for permanence is a key essential for sectors dealing with collections of data such as newspaper, library, government and legal industries**. All deal with textual data that is integral to the general public. These industries rely on the long term availability and searchability of information.
- The **reasons for implementing an archiving strategy with the PDF format lie in the advantages of the PDF format itself**. Recall that the **PDF contains device independent content; compatibility and accessibility; reproduction and compression abilities**. Combining all these features means that numerous documents and files can be capable of being both archived and accessed efficiently and faithfully in the future. With the search for a format that provides consistent and predictable rendering, PDF is the best choice.
- **Yet, PDF/A is only one step in the archiving strategy**. The standard doesn't guarantee a permanent and complete solution in itself.
- Document requirements are individually determined by an organization's needs. These generally include concerns regarding authentication, management, preservation policies and the creation of records.

- *http://www.investintech.com/resources/articles/pdfa/*

Why PDF/A (continued)

- "The **feature-rich nature of PDF can create difficulties in preserving information** over the long term, and **some useful features of the PDF file format are incompatible with the demands of long-term preservation**.
- For example, PDF documents are **not necessarily self-contained**, drawing on system fonts and other content stored externally to the original file. As time passes, and especially as technology changes, these external connections can be broken, and the dependencies cause information to be lost. Additionally, **because of the lack of standardisation among the many PDF development tools on the market, there is inconsistency in the implementation of the file format**.
- **This lack of standardisation could be chaotic for the information managers of the future**, especially as it would be difficult (if not impossible) for them to "get under the hood" of the PDF files unless a format specification were put in place that specifically addressed long-term preservation needs.
- All over the world, tremendous quantities of valuable information are currently being created and saved as PDF, and a specification solution is needed to ensure that digital PDF documents can be rendered, readable, and accessible for the long-term. PDF/A is designed to be that specification."

- *http://blog.accountspayable.net.au/blog/bid/241086/What-is-PDF-A-and-What-Are-The-Benefits*

PDF/A IS IN A FAMILY OF STANDARDS AND GOVERNMENT REQUIREMENTS

- Archiving in PDF/A format to satisfy your corporate, HIPAA, Sarbanes-Oxley or other legal requirements **is the best way to ensure that you will be able to search, retrieve, view and print your documents in the future with the full legal fidelity of the original**. **PDF/A format will not become a "legacy" or "incompatible" file format as PDF technology evolves**.
- This is why the International Standards Organization in conjunction with the US Courts, Library of Congress, ARMA, AIIM, NARA and many more corporations, governments, and other major institutions around the world have approved this new standard for PDF files for long-term archiving of documents. This new standard defines a subset of PDF 1.4 format and composition rules that will withstand the test of time.
- **PDF/A is device-independent, self-contained, self-documenting (with XMP metatags) and transparent to direct analysis with basic tools for indexing, etc., it inhibits the use of PDF functionality that could cause problems in future releases**. If you generate PDF documents from a product that complies with this new PDF/A standard, your files will be safe and viable for the years to come.
- *http://www.pagetech.com/docs/why-PDFA.pdf*

OpenOffice

- Aren't all PDFs already designed to look the same way on any computer? No, not necessarily**. In the distant future, technology will change so much that so archiving cannot make assumptions we can today**. Future readers of today's documents may not have access to Times New Roman, DivX video encoding, or AES encryption algorithms. The **PDF/A format archives documents by embedding all the pieces necessary for faithful reproduction (such as fonts) while forbidding other elements (including encryption)**
- http://www.oooninja.com/2008/01/generating-pdfa-for-long-term-*archiving.html*