

The Making of PDF/A

Stephen P. Levenson
United States Federal Judiciary
Washington DC USA

Stephen P. Levenson

April 10, 2008

© 2008 PDF/A Competence Center, www.pdfa.org



PDF/A – for all Eternity?

- ❖ **A file format is a critical part of a preservation model**
- ❖ **Business need and a consumer market are often out of synchronization**
- ❖ **PDF/A offers a market advantage**



Hall of Fame for PDF/A

- ❖ **John Brinkema**
 - ❖ **Judiciary Architect**
- ❖ **Dr. Richard Fennell**
 - ❖ **Judiciary Chief Technology Officer**
- ❖ **Melonie Warfel**
 - ❖ **Adobe Systems**
- ❖ **Betsy Fanning**
 - ❖ **AIIM International**
- ❖ **Leonard Rosenthal**
 - ❖ **Adobe Systems, Inc. ISO Editor**
- ❖ **Stephen Abrams**
 - ❖ **ISO Editor- Harvard Digital Library**



History and Background

Business Case for the Judiciary

- ❖ **Court documents protect citizen's rights**
- ❖ **Access is assured in trial courts for 20 to 40 years for the Judiciary**
- ❖ **Accessions are often time sensitive**
- ❖ **On-site courthouse storage not cost effective**
- ❖ **Court decisions are permanent records held "until the end of the republic" by the National Archives**
- ❖ **Document format conveys critical information, which must be rendered accurately**



History and Background (continued)

- ❖ **Federal courts began imaging in TIFF for Central Violations Bureau in 1993**
- ❖ **Short retention schedule on tickets and citations (5 years)**
- ❖ **National system allowed for electronic shipment of calendars and tickets**
- ❖ **Operational savings for file pulls and shipping was a 30% savings in first year**
- ❖ **Maritime Asbestos cases, 1995**
- ❖ **Format important and retention long-term**
- ❖ **PDF chosen**
- ❖ **Born electronic**
- ❖ **Born paper**



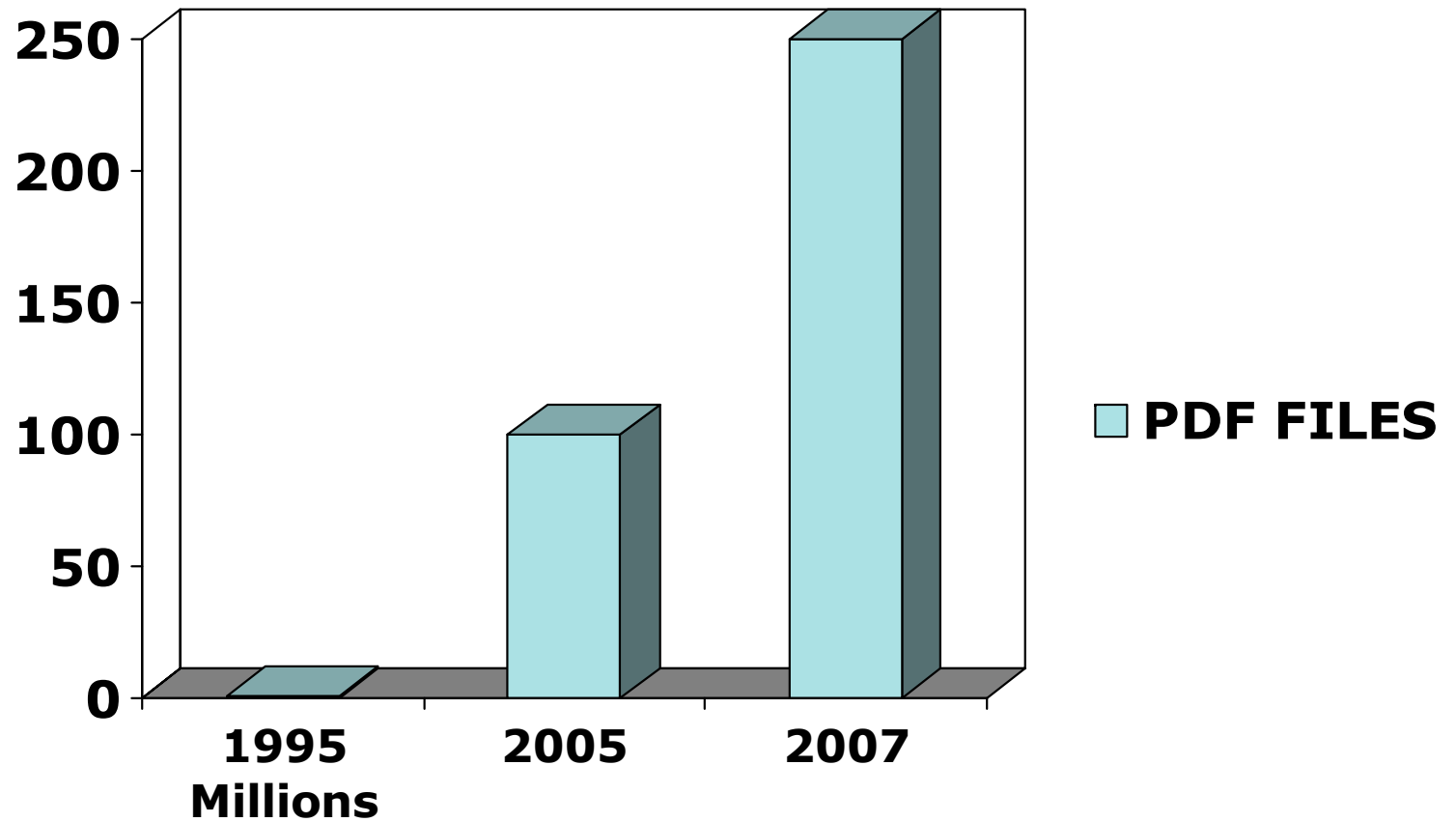
History and Background (continued)

- ❖ **New York Southern Bankruptcy pilot, 1996**
- ❖ **Multi-state filings**
- ❖ **Sophisticated Court and Bar**
- ❖ **Twelve-plus years of filings now in PDF**



Trends and Predictions

File growth



How Many PDF Files Are There?

❖ **466,000,000 pdf pages on Google**



Why PDF/A

- ❖ **Business need and a consumer market are often out of synchronization**
- ❖ **Current workstation application model is in 3 year cycles**
- ❖ **Electronic paper emulation**
- ❖ **Minimal migration**
- ❖ **Accurate, consistent and predictable rendering**
- ❖ **No single solution for all problems**
- ❖ **One answer is PDF/A...**
- ❖ **Free reader, multi-platform and minimum training**



Have You Budgeted for the Migration of Electronic Records?*

- ❖ **Software updates every 18 to 24 months**
- ❖ **Three updates = inability to precisely replicate the original electronic file (Authenticity, Accessibility, and Validity issues)**
- ❖ **Legal case files are kept 6.25 years after the close of a case = 1 migration in the life of the record**
- ❖ **One cubic foot of records = 3,000 pages**
- ❖ **\$4 to \$5 per page to migrate plain text**
- ❖ **Six cubic feet of contract records per year = 18,000 pages**
- ❖ **18,000 pages = \$72,000 to \$90,000 per year in migration costs**

❖ ***Actual US Government Estimate (unpublished)**



Why PDF/A?

- ❖ **PDF is increasingly the format of choice for electronic documents**
- ❖ **Many of the features that drive PDF's ubiquity tend to complicate preservation efforts**
 - ❖ **Encryption, multi-media, external resources, device dependence, etc.**
- ❖ **Effectiveness of preservation repositories will depend upon automation; automation is more successful with homogeneity of content**



Preservation Criteria for These Requirements

- ❖ **PDF/A attempts to maximize:**
- ❖ **Device independence –**
 - ❖ **The degree to which a PDF/A file is independent of the 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**



Design Characteristics

❖ PDF/A

❖ PDF/A is intended to address three primary issues:

- ❖ Define a file format that preserves the static visual appearance of electronic documents over time
- ❖ Provide a framework for recording metadata about electronic documents
- ❖ Provide a framework for defining the logical structure and semantic properties of electronic documents



The Preservation Problem (The Options?)

- ❖ **The Preservation Problem**
 - ❖ **Popular options for preserving electronic documents over archival time spans**
 - ❖ **TIFF?**
 - ❖ **Widely adopted**
 - ❖ **No access to underlying text without OCR**
 - ❖ **Difficult to create “born-digital” documents**
 - ❖ **XML?**
 - ❖ **Good for describing logical structure, but not appearance**
 - ❖ **Native Format (e.g., MS Word)?**
 - ❖ **Several ubiquitous, but closed proprietary formats**
 - ❖ **PDF?**
 - ❖ **PDF/A**



The Preservation Problem (continued)

- ❖ **The Preservation Problem**
 - ❖ **PDF is a ubiquitous open format for electronic documents**
 - ❖ **ISO 32000**
 - ❖ **Long history**
 - ❖ **Early 1980's for Postscript**
 - ❖ **The feature-rich nature of PDF can complicate preservation efforts**
 - ❖ **All PDFs are not created equal**
- ❖ **Much important information maintained in PDF**
 - ❖ **Permanent archival records, in some cases.**



PDF/A Objectives

- ❖ **PDF/A Objectives**
- ❖ **Desirable properties for a preservation format**
 - ❖ **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**
- ❖ **Metadata capabilities**
- ❖ **DRM or IP free**



PDF/A Objectives (continued)

- ❖ **(Lack of) technical protection mechanisms**
 - ❖ **No encryption, passwords, etc.**
- ❖ **Disclosure**
- ❖ **Ubiquitous reader installed on desktops worldwide**
- ❖ **Adoption**
 - ❖ **Widespread use may be the best deterrent against preservation risk**



The PDF/A Standard

- ❖ **The PDF/A Standard**
 - ❖ **PDF/A is a file format standard**
 - ❖ **PDF/A is just one component of a comprehensive preservation strategy**
 - ❖ **Successful implementation depends upon:**
 - ❖ **Records management policies and procedures**
 - ❖ **Additional requirements and conditions**
 - ❖ **Quality assurance processes**

- ❖ **Final form**



Time Line for Part 1 and 2

- ❖ **Time Line for Part 1 (1.4) and 2 (1.7/ISO 32000)**
 - ❖ **October 2002** Initial meeting of AIIM/NPES PDF/A committee
 - ❖ **April 2003** Initial Working Draft (WD)
 - ❖ **August 2003** New Work Item (NWI) approved and Joint Working Group (JWG) formed
 - ❖ **December 2003** First Committee Draft (CD) approved
 - ❖ **September 2004** Second CD approved
 - ❖ **June 2005** Draft International Standard (DIS) unanimously approved
- ❖ **Throughout the process, PDF/A has been reviewed by technical experts from 15 national standards bodies**
- ❖ **Part 2 Committee Draft (CD) out now for comment**
- ❖ **Publication Early 2009**



ISO/TC 171/SC 2/WG 5

- ❖ **ISO/TC 171/SC 2/WG 5**
 - ❖ **ISO Joint Working Group (JWG) for PDF/A**
 - ❖ **ISO/TC 171/SC 2, Document management applications Application issues**
 - ❖ **ISO/TC 130, Graphic technology**
 - ❖ **ISO/TC 46/SC 11, Information and documentation Archives/records management**
 - ❖ **ISO/TC 42, Photography**
- ❖ **AIIM and NPES (PDF/X) Secretariat**



PDF/A Conformance

- ❖ **PDF/A Conformance**
 - ❖ **Two conformance levels**
 - ❖ **PDF/A-1a**
 - ❖ **Compliance with all requirements of 19005-1**
 - ❖ **Including those regarding structural and semantic tagging**
 - ❖ **PDF/A-1b**
 - ❖ **Compliance with all requirements of 19005-1 minimally necessary to preserve the visual appearance of a PDF/A file**



PDF/A Terminology

❖ PDF/A Terminology

- ❖ **PDF/A-1 refers to the format defined by Part 1 (ISO 19005-1) of the standard**
- ❖ **Part 2 (ISO 19005-2) will define PDF/A-2**
- ❖ **New Parts can be added to the PDF/A family of standards without obsolescing previous Parts**



PDF/A- A Foundation To Build On

- ❖ **PDF/A**
 - ❖ **Open standard**
 - ❖ **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**
 - ❖ **Self Contained**
 - ❖ **Free of Digital Rights Management (DRM) or Intellectual Property Rights Claims, giving Archivists Full Control in the future**



Relationship to Other Standards

- ❖ **Relationship to Other Standards**
 - ❖ **PDF/X for pre-press data exchange**
 - ❖ **ISO 15390 parts 4 (PDF/X-1a), 5 (PDF/X-2), and 6 (PDF/X-3)**
 - ❖ **Currently based on PDF 1.4; work underway to extend to PDF 1.6**
 - ❖ **It is possible for a file to be both PDF/A and PDF/X compliant**
- ❖ **PDF/E for engineering, architectural, and GIS documents**
 - ❖ **Provisionally based on PDF 1.6**
- ❖ **PDF/UA for accessibility**
 - ❖ **Intended to address Section 508 concerns**



Supplemental Information

- ❖ **Supplemental Information**
 - ❖ **PDF/A Competence Center**
 - ❖ **Informative annexes to ISO 19005-1**
 - ❖ **PDF/A-1 conformance summary**
 - ❖ **Best practices**
 - ❖ **Guidelines for capturing or converting electronic documents to PDF/A**
 - ❖ **For documents created according to specific institutional rules**
 - ❖ **Replicates the exact quality and content of source documents within the PDF/A file**
- ❖ **PDF/A FAQ**
 - ❖ **Under development**
- ❖ **Will be available on AIIM and NPES web sites**
 - ❖ **PDF Expert Corner**
 - ❖ **<http://www.aiim.org/standards.asp?ID=33736>**



Supplemental Information (continued)

- ❖ **Supplemental Information**
 - ❖ **Application notes**
 - ❖ **Will provide specific guidance on the use of PDF/A**
 - ❖ **Similar in intent to those produced for PDF/X**
 - ❖ **Under development**
 - ❖ **Will be available on AIIM and NPES web sites**
 - ❖ **AIIM and NPES will maintain copies of, and maintain public access to, the PDF Reference and XMP Specification**
 - ❖ **As well as other freely available, non-ISO normative references of ISO 19005-1**



What Is Next?

❖ Collections

❖ Further Contact

- ❖ **Stephen P. Levenson**
- ❖ **Stephen_Levenson@ao.uscourts.gov**
- ❖ **202-502-2625**

