

SAA 69th Annual Meeting, New Orleans, August 14-21, 2005

PDF/A

The Development of a Digital Preservation Standard

Stephen Abrams

Harvard University

Betsy Fanning

Association for Information and Image Management

Diana Helander

Adobe Systems, Inc.

Susan Sullivan, CRM

National Archives and Records Administration

Agenda

- The preservation problem
- The ISO standards process
- Benefits of PDF/A
- Technical overview
- Questions

The preservation problem

- What is the best option for preserving electronic documents over archival time spans?
 - TIFF?
 - Widely adopted
 - No access to underlying text without OCR
 - No mechanism for capturing logical structure
 - Difficult to create “born-digital” documents
 - XML?
 - Good for describing logical structure, but not appearance
 - Many incompatible domain-specific schemas
 - Native Format (e.g., MS Word)?
 - Several ubiquitous, but closed proprietary formats
 - PDF?

The preservation problem

- PDF is a ubiquitous open format for electronic documents
 - Proprietary, but with publicly available specification
- Many statutory, regulatory, and institutional policies mandate the retention of PDF-based documents over multiple generations of technology
- The feature-rich nature of PDF can complicate preservation efforts

Desirable properties of a preservation format

- PDF/A objectives
 - 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

Desirable properties of a preservation format

- PDF/A objectives
 - (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

PDF/A usage

- PDF/A standard may be used by vendors to:
 - Develop applications that read and write and otherwise process PDF/A files
- These applications will be used by organizations to:
 - Create and process PDF/A conformant files
 - As part of their business processes
 - In conjunction with necessary adjunct archival and records management policies and procedures

Current support for PDF/A

- There is no “formal” support for PDF/A today
 - Acrobat 7 support for “draft” version
- Nor has PDF/A yet been adopted as a “required” format by any governmental, academic, or commercial body
- However, once ISO 19005-1 is formally published, we can expect tools to be developed quickly
 - Acartus ApertureONE ERM
 - Many other vendors participated in the standards process
 - Appligent, Callas, Global Graphics, PDF Sages
- And we expect that the mandated (or recommended) use of PDF/A will follow

PDF/A caveats

- However...
 - PDF/A *alone* does not guarantee preservation
 - PDF/A *alone* does not guarantee exact replication of source material
 - The intent of PDF/A is *not* to claim that PDF-based solutions are the best way to preserve electronic documents
 - But once you have decided to use a PDF-based approach, PDF/A defines an archival profile of PDF that is *more* amenable to long-term preservation

The PDF/A standard

- “This International Standard specifies how to use the Portable Document Format (PDF) 1.4 for long-term preservation of electronic documents”
 - Applicable to documents containing character, raster, and vector data
 - The standard does not address:
 - Processes for generating PDF/A files
 - Specific implementation details of rendering PDF/A files
 - Methods for storing PDF/A files
 - Hardware and software dependencies

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

Agenda

- ✓ The preservation problem
- **The ISO standards process**
- Benefits of PDF/A
- Technical overview
- Questions

The PDF/A standard

- Multi-part ISO International Standard
 - ISO 19005-1:2005, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*
 - Part 2 (19005-2) intended to bring PDF/A into conformance with PDF 1.6
 - And additional future parts, as necessary

Time Line for Part 1

- 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

Time Line for Part 1

- Submitted to ISO Central Secretariat for publication as International Standard
 - Should be publicly available September 2005
- Throughout the process, PDF/A has been reviewed by technical experts from 15 national standards bodies

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*

Role of AIIM and NPES

- AIIM, Association for Information and Image Management
 - Secretariat to ISO/TC 171 and ISO/TC 171/SC2
 - Secretariat to US Technical Advisory Group (TAG) for ISO/TC 171
- NPES, The Association for Suppliers of Printing, Publishing, and Converting Technologies
 - Secretariat to ANSI Committee for Graphic Arts Technologies Standards (CGATS)
 - Secretariat to US TAG for ISO/TC 130
- Joint sponsors of the initial US PDF/A committee

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 obsoleting previous Parts

Agenda

- ✓ The preservation problem
- ✓ The ISO standards process
- **Benefits of PDF/A**
- Technical overview
- Questions

PDF/A

- 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

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

Intellectual property rights

- PDF/A is a *file format* standard
- Anyone can use the *PDF Reference* and *XMP Specification* in conjunction with ISO 19005-1 to create applications that read, write, or process PDF/A files
- Adobe has granted a general royalty free license to use certain of its patents to create applications that read, write, or process PDF/A files

Supplemental information

- 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
 - Required for compliance with NARA's *PDF Transfer Guidance*
- PDF/A FAQ
 - Under development
 - Will be available on AIIM and NPES web sites

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 archive 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

Agenda

- ✓ The preservation problem
- ✓ The ISO standards process
- ✓ Benefits of PDF/A
- **Technical overview**
- Questions

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

Nevertheless...

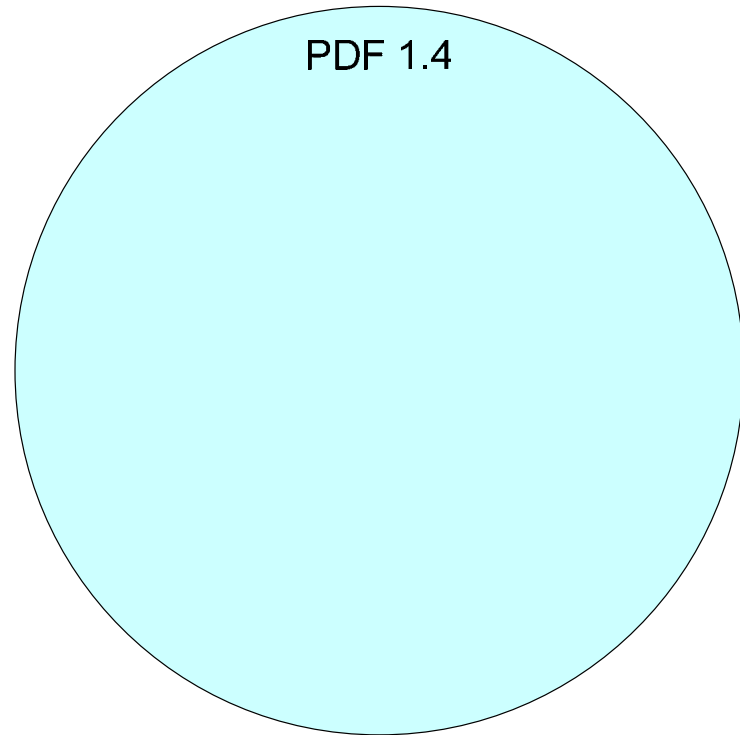
- PDF/A may not be the last preservation format you will need
- However, proper application of PDF/A should result in reliable, predictable, and unambiguous access to the full information content of electronic documents

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

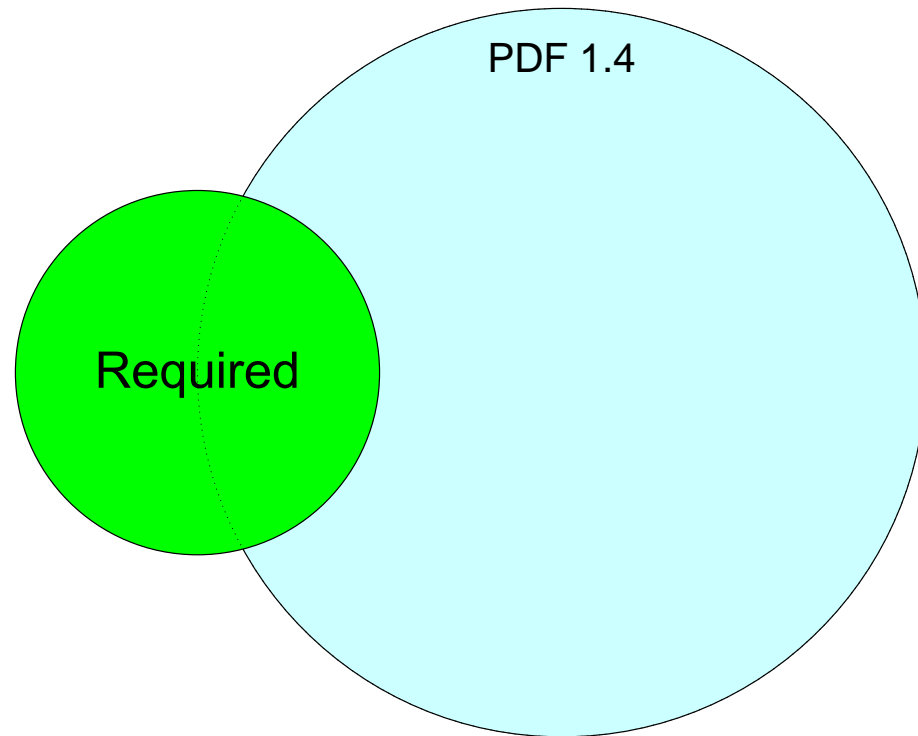
PDF/A requirements

- Conformance to PDF 1.4



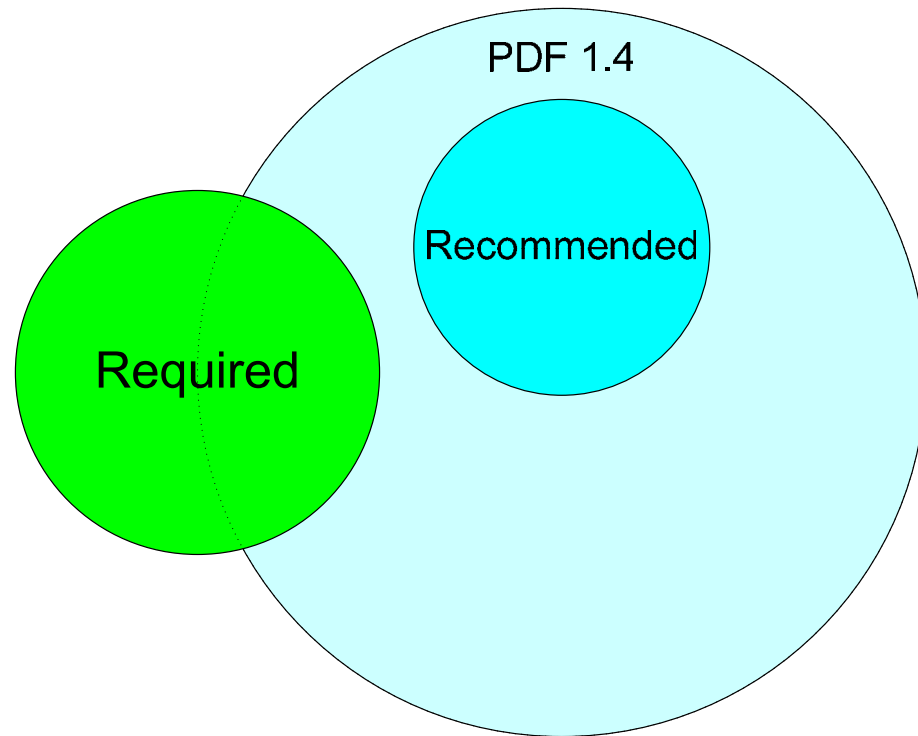
PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required



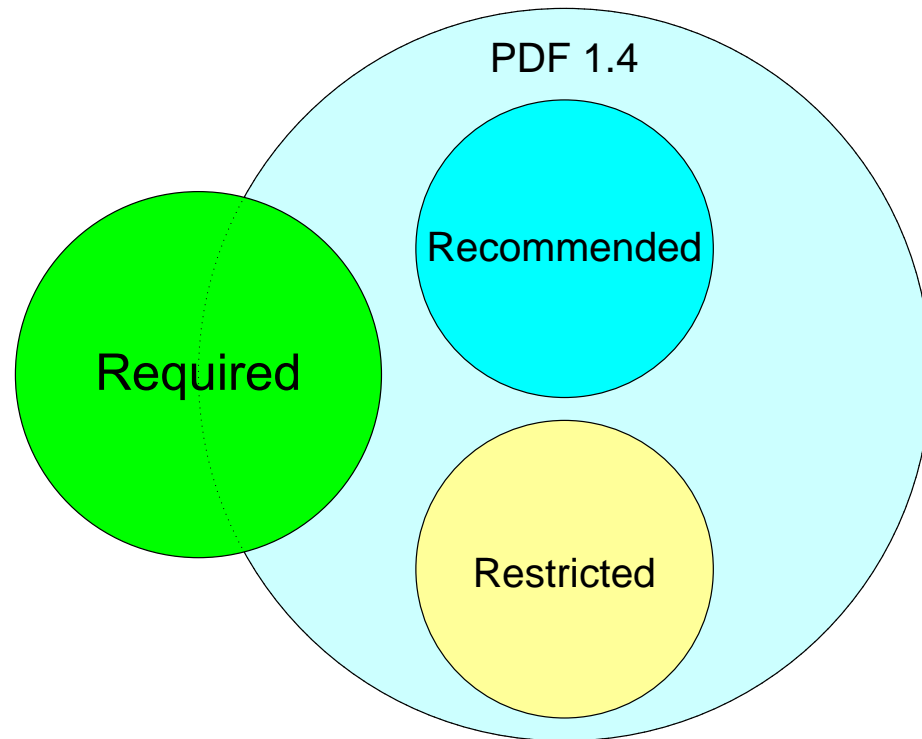
PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required
 - Recommended



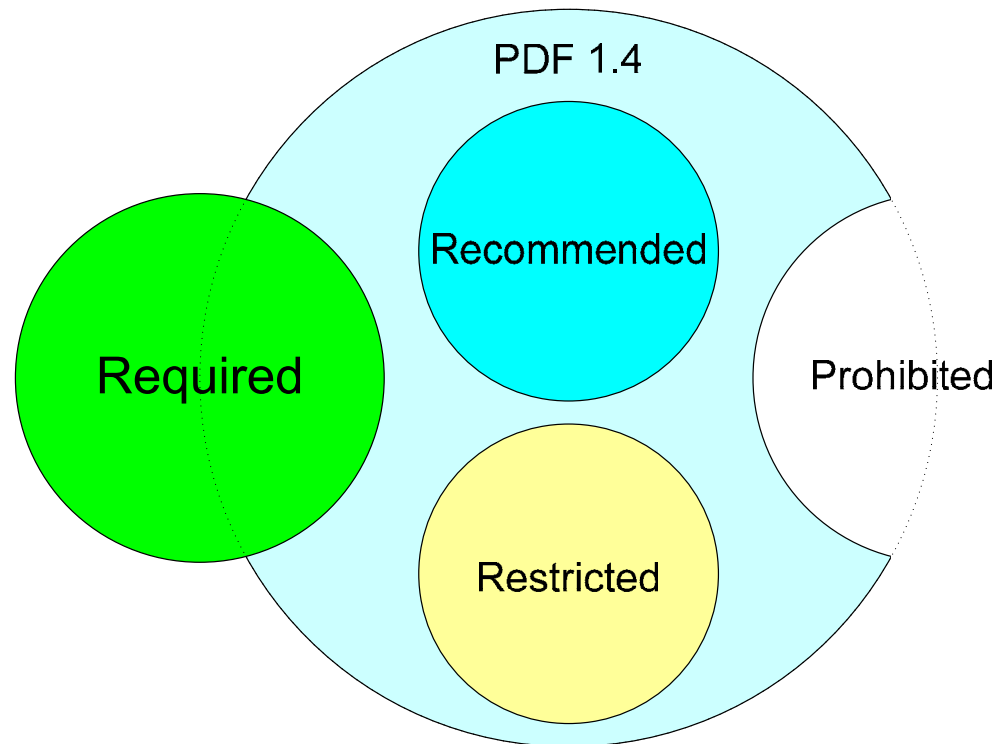
PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required
 - Recommended
 - Restricted



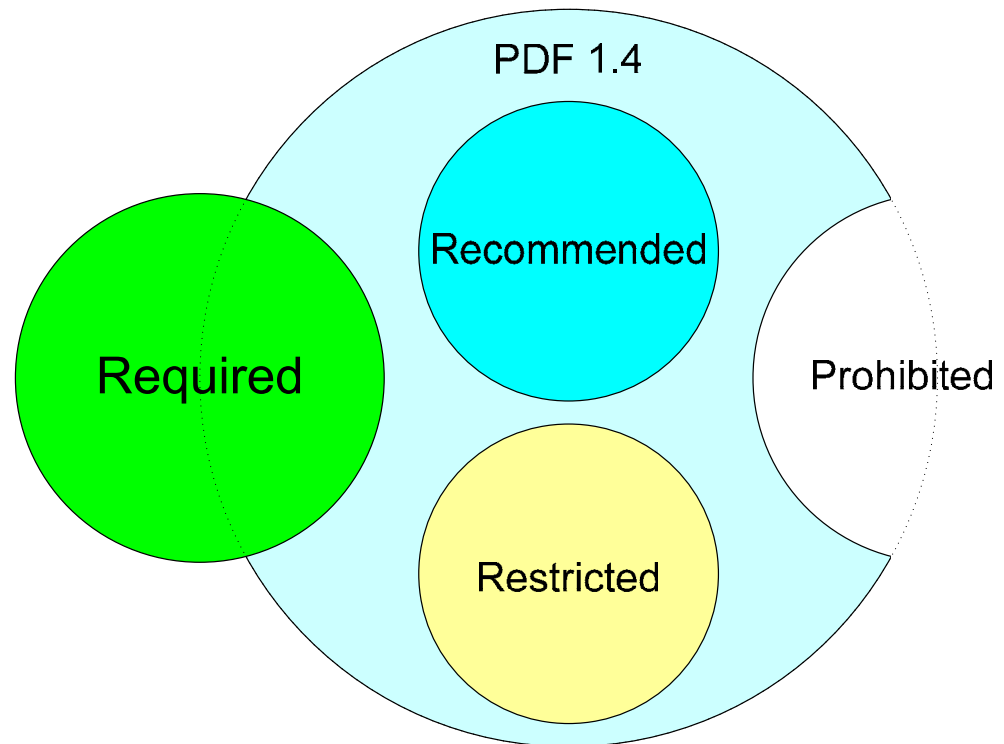
PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required
 - Recommended
 - Restricted
 - Prohibited



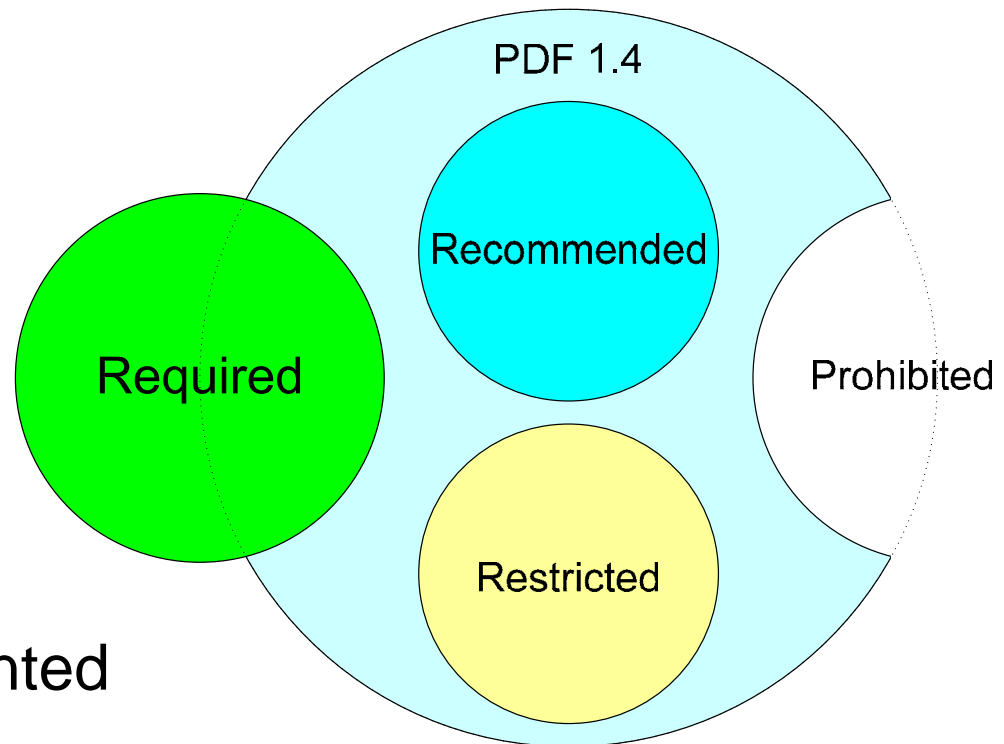
PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required
 - Recommended
 - Restricted
 - Prohibited
- Reader functional requirements



PDF/A requirements

- Conformance to PDF 1.4
- With features that are
 - Required
 - Recommended
 - Restricted
 - Prohibited
- Reader functional requirements
- Features not documented in 1.4 are ignored by PDF/A readers



General

- **Required**
 - Conformance to 1.4 requirements
- **Recommended**
 - Linearization hints should be ignored
- **Restricted**
 - Document information dictionary must be consistent with XMP metadata
- **Prohibited**
 - Encryption
 - LZW compression
 - Embedded files
 - Optional content
 - Sound and movie media types

Graphics

- Required
 - Device independent color
 - Embedded color spaces
- Restricted
 - Image dictionaries
 - **Separation** and **DeviceN** color spaces
 - Form XObjects
 - Extended graphics state
 - Rendering intents
- Prohibited
 - Reference XObjects
 - PostScript XObjects
 - Non-PDF 1.4 defined operators
 - Transparency

Fonts

- Required
 - Fonts legally embeddable for unlimited, universal rendering
 - Embedded font programs
 - Embedded CMaps
 - Consistent font metrics
 - Unicode character map (For Level A conformance only)
- Recommended
 - Font subsets
- Restricted
 - Character encodings

Annotations

- Required
 - Reader mechanism to expose the annotation dictionary
Contents key
- Restricted
 - Annotation dictionaries
- Prohibited
 - Non-PDF 1.4 defined types
 - **FileAttachment**, **Sound**, and **Movie** types

Actions

- Required
 - Behavior for **NextPage**, **PrevPage**, **FirstPage**, and **LastPage** actions as defined in PDF 1.4
 - Reader mechanism to expose **GoToR** dictionary **F** and **D** keys, URI action dictionary **URI** key, and SubmitForm action dictionary **F** key
- Prohibited
 - **Launch**, **Sound**, **Movie**, **ResetForm**, **ImportData**, and **JavaScript** actions
 - Deprecated **set-state** and **no-op** actions
 - Named actions other than the 4 page navigation actions
 - Widget annotation or Field dictionary **AA** key

Metadata

- Requires use of Extensible Metadata Platform (XMP)
 - Proprietary, but open format
 - Used for metadata creation, processing, and interchange
 - Based on Resource Description Framework (RDF)
 - Open World Wide Web Consortium (W3C) standard
 - Cornerstone of Semantic Web
 - Pre-defined schemas
 - Base, DC, DRM, DAM, Workflow, EXIF, PDF, PSD
 - Defined extension mechanism
 - Embedding rules
 - TIFF, JPEG, JPEG 2000, HTML, AI, PSD, PDF, ...

Metadata

- Required
 - Document level XMP metadata
 - Equivalent XMP metadata for all appropriate Document Information Dictionary properties
 - Embedded extension schema
 - Version and conformance self-identification
- Recommended
 - File identifier
 - File provenance
 - Font metadata
- Prohibited
 - XMP packet header **bytes** and **encoding** attributes

Logical structure (Level A conformance only)

- Required
 - Tagged PDF
 - Explicit word breaks
- Recommended
 - Tagging for pagination, layout, and page artifacts
 - “Strongly structured” block-level structural tagging
 - Natural language tagging
 - Alternative description, non-textual annotation, replacement text, and abbreviation/acronym expansion tagging

Interactive forms

- Required
 - Field appearance dictionary
- Restricted
 - **NeedAppearance** flag
 - Explicit word breaks
- Prohibited
 - **A** and **AA** keys in Widget and Field dictionaries
- Note
 - There is *no* restriction on the use of digital signatures, as defined by PDF 1.4

What's under consideration for Part 2?

- Based on PDF 1.6
- The following specific features are under consideration for inclusion in Part 2
 - JPEG 2000 image compression
 - More sophisticated digital signature support
 - OpenType fonts
 - 3D graphics
 - Audio/video content
 - Consistency with PDF/X, PDF/E, PDF/UA

What's under consideration for Part 2?

- If PDF/A-1 does not meet your specific needs, get involved in the process
 - Contact Betsy Fanning, Director, AIIM Standards Program

PDF/A summary

- ISO 19005-1 (should be available September 2005)
- File format standard
- One component of a comprehensive archival strategy
- Based on PDF 1.4
- Two conformance levels
 - Level A for structural/semantic tagging
 - Level B for appearance only
- Emphasis on reliable and predictable rendering of static visual appearance
 - Do's: embed fonts, device-independent color, XMP metadata, tagging
 - Don'ts: encryption, LZW, embedded files, external content references, transparency, multi-media, JavaScript

PDF/A summary

- Consistency with PDF/X
- Work planned for Part 2

Agenda

- ✓ The preservation problem
- ✓ The ISO standards process
- ✓ Benefits of PDF/A
- ✓ Technical overview
- **Questions**

Questions?

<http://www.iso.org/>

<http://www.aiim.org/pdfa/app-notes>

<http://www.npes.org/standards/toolspdfa.html>

stephen_abrams@harvard.edu

bfanning@aiim.org

helander@adobe.com

susan.sullivan@nara.gov