

Portable Document Format Archive (PDF/A)

ISO Standard Development



Overview of the Draft PDF/A Standard

May 11, 2004

Introduction

- ▶ Today's presentation will provide an archival/records management perspective on the Draft PDF/A Standard (PDF/A):
 - Background of PDF/A
 - ▶ Why and how PDF/A was initiated, and its status in the ISO process
 - What is covered in PDF/A
 - The approach used in developing PDF/A
 - An overview of each clause of PDF/A (as currently drafted)
 - ▶ What is specified and why
 - ▶ Examples of topics covered and their use in PDF/A
 - What NARA's expectations are for PDF/A
 - Where you can get more information on NARA and PDF/A

NARA's Involvement in PDF/A

- ▶ U.S. National Archives and Records Administration (NARA) is actively participating in PDF/A development with the following goals:
 - To influence the process so that PDF/A compliant records can be preserved by NARA over the long term, and
 - To provide information used in developing NARA guidance for transferring future permanent records in PDF.
 - ▶ Current NARA transfer guidance for records in PDF issued March 31, 2003
 - http://www.archives.gov/records_management/policy_and_guidance/nwm11_2003.html

PDF/A Background – Wide Use of PDF

- ▶ PDF is a digital format that electronically reproduces the visual appearance of documents whether they are:
 - Created natively in PDF,
 - Converted from other electronic formats, or
 - Digitized from paper or microform
- ▶ Businesses, governments, libraries, archives, and other institutions and individuals around the world use PDF to:
 - Collect and disseminate information over the Internet,
 - Store electronic records, and/or
 - Make scanned images searchable by embedding OCR'd text.
- ▶ As a result, large bodies of important information are maintained in PDF.

PDF/A Background – PDF Not a Suitable Archival Format

- ▶ PDF itself is not suitable as an archival format.
 - Adobe® is under no obligation to continue publishing the specification for future versions.
 - Can include features incompatible with current archival requirements
 - ▶ Encryption
 - ▶ Embedded files
 - PDF documents not necessarily self-contained
 - ▶ Can depend on system fonts and other content drawn from outside the file
 - Multiple PDF development tools on the market
 - ▶ Inconsistency in the file format (all PDFs are not created equal)
- ▶ Long-term solution needed to ensure that digital PDF documents remain accessible for long periods of time
 - Permanent archival records, in some cases

PDF/A Background – Example Business Case for PDF/A

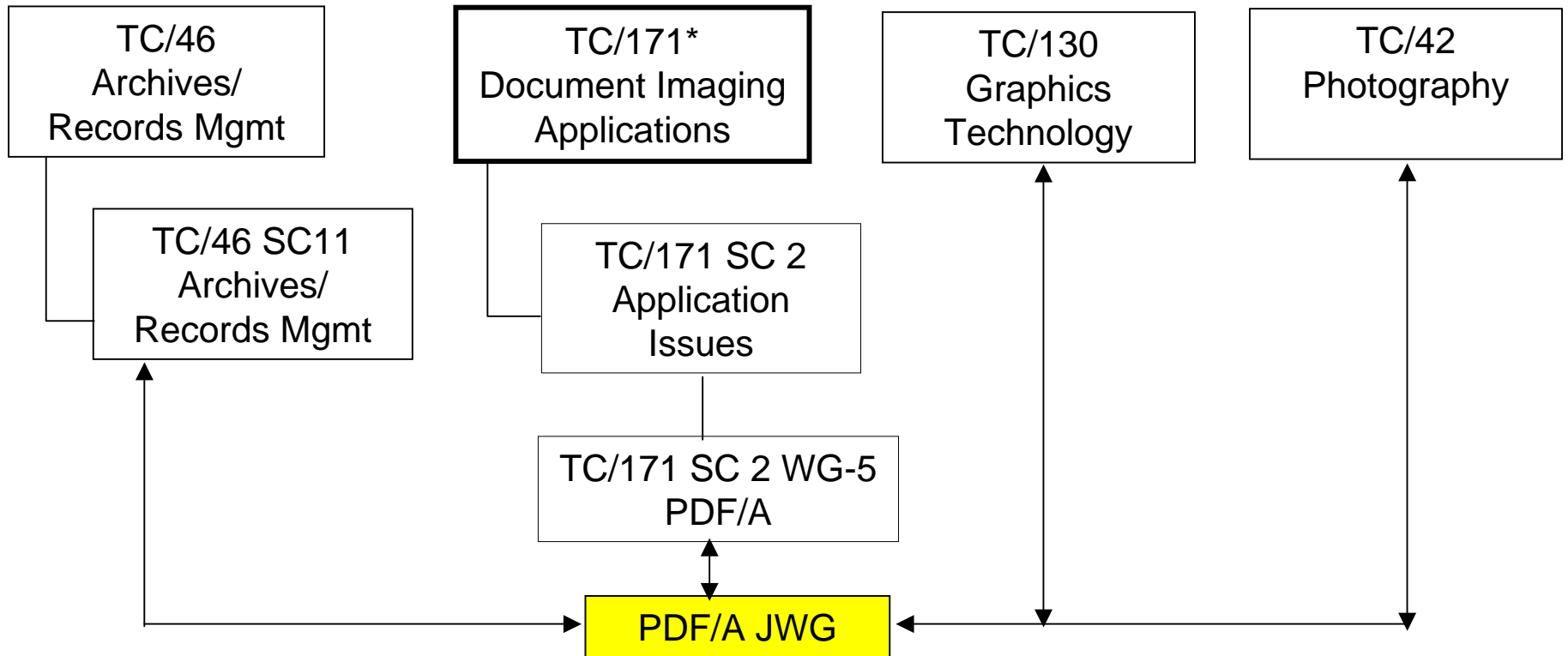
- ▶ Administrative Office of the U.S. Courts (AOUSC)
 - Uses PDF as the electronic format for Electronic Case Filing System
 - ▶ Business need for court documents to maintain their “visual appearance” (i.e., pagination, layout)
 - System accepts filings and provides access to filed PDF documents over the Internet
 - ▶ Electronic court documents are submitted in PDF
 - ▶ Hard copy submissions are scanned to PDF
 - Many AOUSC files must be maintained for long periods of time (i.e., 40 years)
 - ▶ Some will be transferred to the National Archives for permanent retention
 - Future use of and access to the AOUSC’s PDF documents depends on maintaining the ability to reproduce their visual appearance and other properties over the long term (i.e., across multiple generations of technology).

PDF/A Background – Addressing the Long-term Use of PDF

- ▶ AOUSC collaborated with other the U.S. government agencies, libraries, academia, private industry (e.g., imaging product vendors, drug companies, consultants), Adobe®, and other PDF vendors.
- ▶ Formed a U.S. Committee to initiate an ISO Standard to:
 - Ensure preservation of PDF documents over extended periods of time, and
 - Further ensure that PDF documents would be rendered with consistent and predictable results in the future.
- ▶ U.S. Committee working under two Standards Organizations
 - AIIM International (the Association for Information and Image Management, International)
 - NPES (The Association for Suppliers of Printing, Publishing and Converting Technologies)

PDF/A ISO Process – International Joint Working Group

ISO Joint Working Group (JWG) - PDF/A



* JWG formed under the auspices of TC/171

PDF/A ISO Process – Progress and Next Steps

- ▶ Early 2002 PDF/A development initiated
- ▶ September 2003 Approval of ISO New Work Item (NWI)
- ▶ October 2003 TC-171 Meeting - JWG prepared Committee Draft (CD)
- ▶ November 2003 - CD ballot circulated to National Bodies (NBs)
- ▶ February 2004 - CD ballot closed
- ▶ March 2004 - JWG reviewed NB comments on CD
- ▶ June 2004 - Second CD ballot to be circulated to NBs
- ▶ September 2004 - Second CD ballot closes
- ▶ October 2004 - Next JWG Meeting
- ▶ Spring 2005? - Draft International Standard
- ▶ Winter 2005? - International Standard

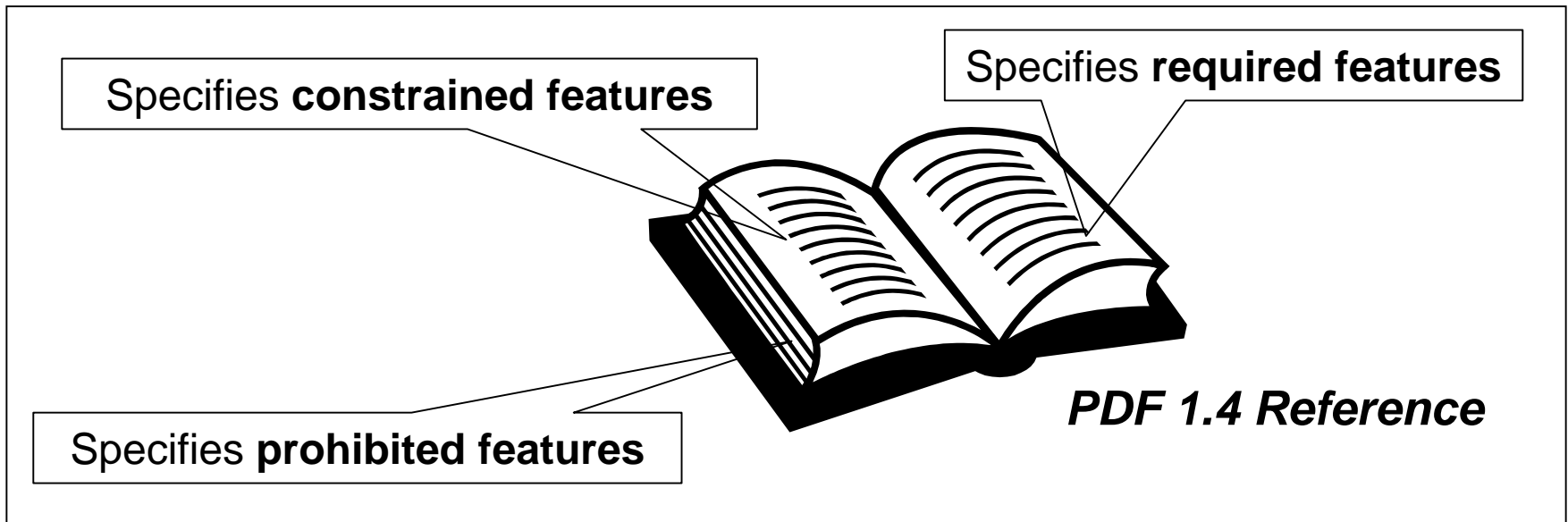
National Bodies Voting/Providing Comments on PDF/A CD

- ▶ Australia
- ▶ France
- ▶ Germany
- ▶ Japan
- ▶ United Kingdom
- ▶ United States
- ▶ Sweden

PDF/A Approach

▶ PDF/A specifies:

- The subset of PDF components, from the Adobe® published specification for Version 1.4 (i.e., PDF 1.4 Reference), that are either mandatory, recommended, or prohibited, **and**
- How these components may be used by software to render the file.



PDF/A

Preservation Criteria for PDF/A 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
 - The degree to which a PDF/A file is amenable to direct analysis with basic tools, including human readability

- ▶ 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

PDF/A Introduction

- ▶ Explains PDF/A background and approach, and further explains that:
 - PDF/A should be used as one component of an organization’s electronic archival environment. Implementation depends on:
 - ▶ Records management policies and procedures and any additional requirements and conditions necessary to ensure the persistence of electronic documents over time
 - ▶ Quality assurance processes necessary to very conformance with requirements

- ▶ Examples of other topics addressed in the Introduction include:
 - Use of “Part 1” in title
 - Intellectual property rights
 - Location of application notes

Structure of the Draft PDF/A Standard

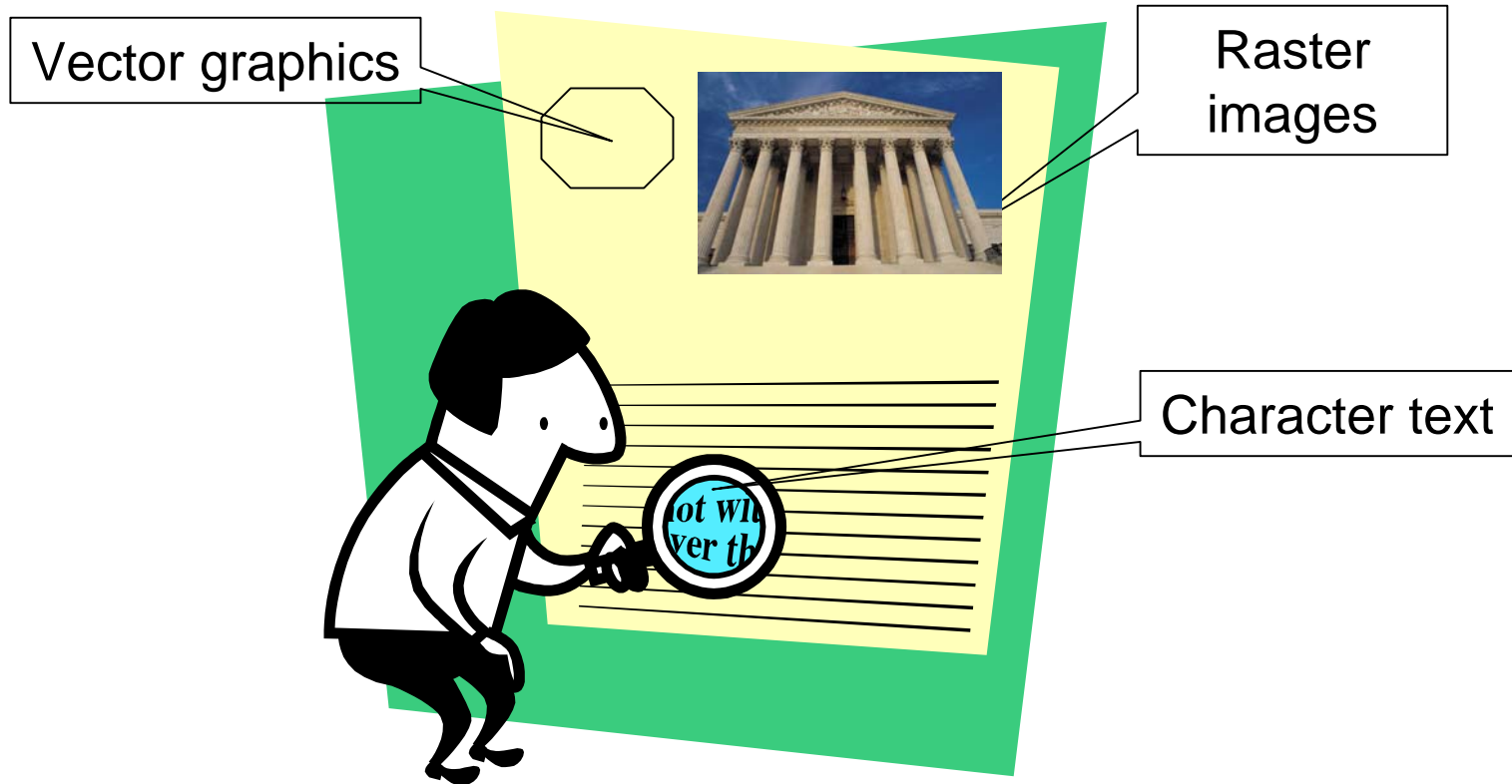
- ▶ 1 Scope
- ▶ 2 Normative References
- ▶ 3 Terms and Definitions
- ▶ 4 Notation
- ▶ 5 Conformance Levels and Identification
- ▶ 6 Technical Requirements
 - 6.1 File Structure
 - 6.2 Graphics
 - 6.3 Fonts
 - 6.4 Transparency
 - 6.5 Annotations
 - 6.6 Actions
 - 6.7 Metadata
 - 6.8 Logical Structure
 - 6.9 Forms
- ▶ Informative annexes
 - Annex A - Summary of Prohibited PDF Features
 - Annex B - Best Practices for PDF/A
- ▶ Bibliography

Clause 1 of the Draft PDF/A Standard – Scope

- ▶ “This International Standard specifies the use of the Portable Document Format (PDF) 1.4 for the long-term preservation of electronic documents.”
(Clause 1, PDF/A Scope)

Clause 1 of the Draft PDF/A Standard – Scope

- ▶ “It is applicable to documents containing combinations of character, raster, and vector data.”

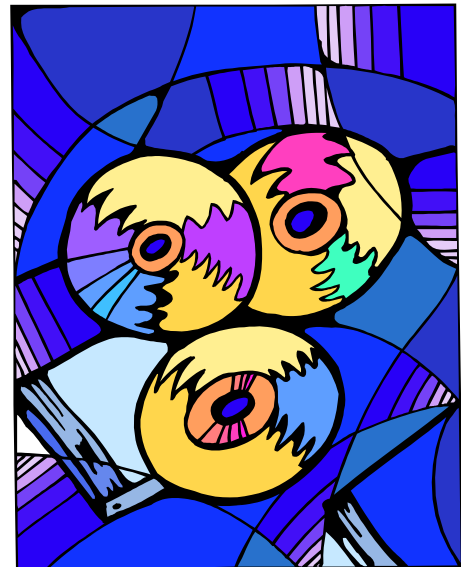


Clause 1 of the Draft PDF/A Standard – Scope

▶ “PDF/A does not apply to:

- “Converting paper or electronic documents to the PDF/A format
- Specific technical design, user interface, implementation, or operational details of rendering
- Specific physical methods of storing these documents such as the media and storage conditions
- Required computer hardware and operating systems”

PDF/A \neq



Clause 2 of the PDF/A Standard – Normative References

- ▶ The **Normative Reference** clause lists documents which, through reference in PDF/A, constitute provisions of PDF/A.

- ▶ Examples of normative references include:
 - Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, 6 October 2000
 - ICC.1:1998-09, File Format for Color Profiles, International Color Consortium
 - ISO/IEC 10646-1, Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane
 - PDF Reference: Adobe Portable Document Format, Version 1.4, Adobe Systems Incorporated – 3rd ed. (ISBN 0-201-75839-3)
 - XMP Specification, January_2004, Adobe Systems Incorporated

Clause 3 of the PDF/A Standard – Terms and Definitions

- ▶ Lists **Terms and Definitions** as they apply to PDF/A
- ▶ Examples of terms defined in this clause include:
 - **Electronic document** - 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
 - **Interactive reader**- reader that requires or allows human interaction during the software's processing phase
 - **Long-term** - 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 [ISO 14721]
 - **Writer**- software application that is able to write files [ISO 15930-1]

Clause 4 of the Draft PDF/A Standard - Notation

- ▶ The **Notations** clause explains conventions used for identifying PDF specific elements in PDF/A.
- ▶ Examples of topics specified in this clause include:
 - Notation for PDF operator names, PDF keywords, the names of keys in PDF dictionaries, and other predefined names
 - Notation for operator values or values of dictionary keys
 - Identification of individual characters (e.g., CARRIAGE RETURN (U+000D)).
 - References to the "PDF Reference"

Clause 5 of the Draft PDF/A Standard – Conformance

- ▶ The **Conformance** clause specifies two levels of conformance for PDF files to allow the exclusion of requirements that could be burdensome.
 - Full conformance - meeting all requirements
 - Minimal conformance - meeting requirements minimally necessary to ensure predictable and reliable rendering

- ▶ Examples of topics specified in this clause:
 - Full conformance (Meets all requirements of PDF Reference, as modified by PDF/A)
 - Minimal conformance - Excludes requirements for:
 - ▶ Unicode character mapping
 - ▶ Logical Structure (Tagged PDF)

Clause 6.1 of the Draft PDF/A Standard – File Structure

- ▶ The **File Format** clause specifies attributes of the PDF format to produce uniform PDF/A files.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Header begins at byte offset 0
- Header followed by comment with at least four bytes < 127
- Conform to limits on quantities defined in Table C.1 of PDF Reference

Recommended

- Reader: ignore any linearization information

Prohibited

- Encryption
- Embedded files
- External content
- Optional content (i.e., OC Properties)

Clause 6.2 of the Draft PDF/A Standard - Graphics

- ▶ The **Graphics** clause specifies attributes of the PDF format to ensure that all the information needed to appropriately render graphics is contained within the file.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Device independent color space (either directly or by Output Intent)

Recommended

- For color-critical applications, follow additional requirements of PDF/X-3

Prohibited

- Form XObjects
- Reference XObjects
- PostScript XObjects
- Content stream operators not documented in PDF 1.4

Clause 6.3 of the Draft PDF/A Standard - Fonts

- ▶ The **Fonts** clause specifies attributes of PDF to ensure that future rendering of the textual content of a PDF file matches the static appearance as originally created.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Embed all referenced fonts
- Reader: only use embedded fonts
- Map to Unicode (only necessary for full conformance)

Recommended

- Font subsets

Prohibited

- Fonts not legally embeddable for unlimited, universal rendering

Clause 6.4 of the Draft PDF/A Standard - Transparency

- ▶ The **Transparency** clause explicitly prohibits the use of transparency within the PDF/A format for overlapping or three dimensional graphics.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Not applicable

Recommended

- Alternatives for displaying “transparent” graphics (e.g., flattened vector objects)

Prohibited

- Transparency Keys

Clause 6.5 of the Draft PDF/A Standard - Annotations

- ▶ The **Annotations** clause specifies the use of annotations to ensure that:
 - The visual presentation of the actual page is rendered exactly as the author intended, and
 - Hidden content is available for viewing.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Reader Behavior
 - ▶ Must display the actual contents of all annotations in human-readable form

Recommended

- Not Applicable

Prohibited

- Hidden Annotations
- File Attachments
- Sound Annotations
- Movie Annotations
- Annotations not defined in the *PDF Reference*

Clause 6.6 of the Draft PDF/A Standard - Actions

- ▶ The **Actions** clause specifies attributes of PDF to ensure that PDF/A files are self-contained and do not rely on external sources for content.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Reader behavior
 - ▶ Interactive readers must display destination of links (address)

Recommended

- Reader Behavior
 - ▶ Not required to act on links, but may
 - Security risk

Prohibited

- Actions external to the document
 - ▶ Launch
 - ▶ Sound
 - ▶ Movie
 - ▶ ResetForm
 - ▶ ImportData
- JavaScript

Clause 6.7 of the Draft PDF/A Standard – Metadata

- ▶ The **Metadata** clause specifies the use of the Adobe® eXtensible Metadata Platform (XMP) for embedding metadata within PDF/A files.
 - Outlines a structured, consistent process to support a broad variety of metadata requirements.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- XMP stream defined by catalog Metadata key
- Info dictionary entries must have equivalent XMP properties
- Embed any extension schemas
- PDF/A version and conformance level

Recommended

- File Provenance
- File Identifiers
- Font Metadata

Prohibited

- Deprecated bytes attribute in XMP packet header

Clause 6.8 of the Draft PDF/A Standard – Logical Structure

- ▶ The **Logical Structure** clause applies only to fully conforming PDF/A files and specifies the use of Tagged PDF to ensure:
 - The recovery of a PDF file’s textual content,
 - The recovery of individual characters that make up each word, and
 - The recovery of information about the logical structure of the document.

- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Tagged PDF (meets all requirements in Clause 9.7 of the *PDF Reference*)
- Explicit white space to indicate word breaks (if appropriate for language)
- Default language

Recommended

- Pagination, layout, and page artifacts
- Structural hierarchy
- Language
- Alternate descriptions
- Replacement text
- Expansion of abbreviations and acronyms

Prohibited

- No current prohibitions

Clause 6.9 of the Draft PDF/A Standard - Forms

- ▶ The **Forms** clause specifies attributes of PDF to ensure that:
 - There is no ambiguity about the rendering of form fields
 - The rendered representation of the page or the content of the document is not changed at any time, and
 - Form fields do not perform actions of any type.
- ▶ Examples of topics specified in this clause and their use:

Mandatory

- Appearance dictionary (required for each form field containing data)
- Reader Behavior (render the field according to the appearance dictionary without regard to the form data)

Recommended

- No current recommendations

Prohibited

- Any feature that would allow the document's appearance to change

Annexes of the Draft PDF/A Standard – Informative Annexes

- ▶ **Informative Annexes** will be included in the Draft PDF/A Standard to provide supplemental information including:
 - Summary of Prohibited PDF Features
 - Best Practices for PDF/A
 - ▶ Recommended software requirements 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

Summary - PDF Background

- ▶ PDF reproduces the visual appearance of electronic documents and is widely used to represent large bodies of important information, some of which must be maintained for long periods of time.
- ▶ PDF, itself, is not a suitable archival format.
- ▶ Long-term solution is needed to ensure that PDF documents can remain accessible over long periods of time (e.g. multiple generations of technology).
- ▶ Draft PDF/A Standard initiated by AOUSC in collaboration with U.S. government agencies, libraries, academia, private industry.
- ▶ PDF/A has been circulated as a Committee Draft and will be circulated again as a Committee Draft
 - Technical changes and additions

Summary - PDF/A Overview

- ▶ Goals - to ensure preservation of PDF documents over extended periods of time, and further ensure that PDF documents will be rendered with consistent and predictable results in the future.
- ▶ Approach
 - PDF/A identifies the subset of PDF components from the *PDF 1.4 Reference* that are either mandatory, recommended, or prohibited, and how these components may be used by software to render the file.

Summary - PDF/A Requirements

- ▶ Two levels of conformance
 - Full
 - Minimal (e.g. No Tagged PDF, UNICODE Mapping)
- ▶ Uniform file format (header, trailer, no encryption)
- ▶ Device-independent rendering of graphics
- ▶ Embedded fonts, character encoding
- ▶ Transparency prohibited
- ▶ Annotations restricted, content should be displayed by readers
- ▶ External actions restricted, no dependence on external content
- ▶ Readers not required to act on hyperlinks, but may
- ▶ XMP metadata “Adobe XML Metadata Framework”
- ▶ Forms based on appearance, not data

Conclusion

- ▶ NARA's expectation for PDF/A
 - Should address some existing archival issues with PDF and enable records in PDF to be maintained for longer periods of time in that format
 - ▶ Standard maintained by external International organization, not just vendors
 - ▶ Increased degree of format reliability
 - ▶ Enhanced future migration capabilities (embedded XMP metadata)

More Information is Available

- ▶ PDF/A standard is projected to be issued as an ISO Standard in 2004
- ▶ More information on PDF/A on AIIM Web Site
 - <http://www.aiim.org/standards.asp?ID=25013>
 - Information about the U.S. National Archives and Records Administration
 - <http://www.archives.gov>
 - Contact Susan Sullivan at susan.sullivan@nara.gov

Questions/Discussion

