www.pdfa.org

The Future of PDF/A and Validation

Dietrich von Seggern

Managing Director, callas software GmbH
Member of the Board of the PDF Association

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

The Future of PDF/A (and) Validation

www.pdfa.org

- **veraPDF (June 2017)**
- **PDF 2.0 (Fall 2017)**
- **PDF/A-NEXT (2018)**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What is veraPDF?

- Funded by the European Commission's PREFORMA (PREservation FORMAts) project
  - PDF/A
  - TIFF
  - Video
- Aim: Develop an "industry supported" open source validator
- Covering all PDF/A parts and conformance levels
- The veraPDF consortium:
  - Open Preservation Foundation (OPF)
  - PDF Association

http://www.verapdf.org

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Why veraPDF?
Resolve differences between PDF/A validators

www.pdfa.org

- Different interpretations of the standard text
- Bugs in validators
- Has been a major task of the PDF Association's PDF/A Competence Center
  - Important achievement in establishing PDF/A as an international standard
  - Vendors are at the same time competitors
  - Number of discussions down to very, very few per year
- Limitations:
  - No systematic approach but dependent on inquiries from end users or from vendors
  - Resolutions where not summarized in a formal document
  - In a few cases the standard text was misleading and would result e.g. in an unnecessary problems during conversion.
    These cases remained without "official" clarification

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

www.pdfa.org

veraPDF – Deliverables to the industry

Opportunity for the PDF/A industry to further converge and consolidate different PDF/A implementations

- **Test Corpus**
- **Open Source Validator**
- **TechNote 010 to be published by the PDF Association**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

veraPDF
Test Corpus

www.pdfa.org

- **Before: Isartor Testsuite from 2008 developed by PDF Association**
  - **204 fail tests for PDF/A-1b**
- **veraPDF Test Corpus**
  - **179 fail or pass tests for PDF/A-1b (in addition to Isartor) plus 367 tests for XMP (Spec from 2004)**
  - **223 fail or pass tests for PDF/A-2b plus 549 tests for XMP (Spec from 2005)**
  - **13 fail or pass tests for PDF/A-3b**
  - **3 fail or pass tests for conformance level U**
  - **7 fail or pass tests for conformance level A**
  - **Totals to 1541 fail or pass tests**
- **General improvements**
  - **Additional pass tests**
  - **All standard parts and conformance levels**
- **Most important achievement for developers**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

www.pdfa.org

veraPDF – Deliverables to the industry

Opportunity for the PDF/A industry to further converge and consolidate different PDF/A implementations

- **Test Corpus**
- **Open Source Validator**
- **TechNote 010 to be published by the PDF Association**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

PDF association

www.pdfa.org

veraPDF

callas

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

2017-05-15

- **Online: http://demo.veraPDF.org**
- **Downloads: http://downloads.veraPDF.org**
- **Sources: https://github.com/veraPDF**

veraPDF
Open Source Validator

- Currently Version 1.4
- Will be final in June 2017
- Maintenance and bug fixes will continue
- First reference when it comes to differences or uncertainties between vendors or applications
- In validation only environments
- For additional validation in PDF/A conversion

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

veraPDF – Deliverables to the industry

www.pdfa.org

Opportunity for the PDF/A industry to further converge and consolidate different PDF/A implementations

- **Test Corpus**
- **Open Source Validator**
- **TechNote 010 to be published by the PDF Association**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

www.pdfa.org

veraPDF
TechNote 010 published by the PDF Association

- During implementation and creation of Test Corpus double checking with existing validators
- Discussion on remaining differences, 29 ambiguities
- Clarifications were summarized and discussed with the TC 171 at ISO (responsible for PDF/A)

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

TechNote 10 published by the PDF Association II

- **According to the rules of the PDF Association members will have the opportunity to submit comments and to vote in a ballot**
- **Final resolutions will then be available to the public as TechNote 10**

veraPDF
TechNote 010 published by the PDF Association

www.pdfa.org

- None of the ambiguities clarified in Tech Note 10 is related to "real" PDF problems
  - Rendering (conformance level B)
  - Accessibility (conformance level A)
  - Text indexing, search, copy (conformance level U)
- From a technical and internal PDF structure point of view they are all tiny details
- They could become important issues in workflows that build on a mix of PDF/A validators

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

veraPDF
TechNote 010 published by the PDF Association II

www.pdfa.org

- **Examples (3 out of 29)**
  - **Named resources not used for rendering**
    - are relevant for validation
  - **Validation of XMP Metadata value types**
    - only simple value types are relevant
  - **Metadata must not be compressed – even if it is not on document level**
    - Metadata can be attached to images, fonts, pages
    - Provision in PDF/A-1 (and later) is clear: „Metadata object stream dictionaries shall not contain the Filter key."
    - Background is that it should be easy to read metadata, however, that becomes more difficult when no metadata objects are compressed

- **Some of the clarifications have been taken over into a possible next part for PDF/A, based on PDF 2.0…**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

www.pdfa.org

Summary:
Achievements of the veraPDF project

Opportunity for the PDF/A industry to further converge and consolidate different PDF/A implementations

- Test Corpus
- Open Source Validator
- TechNote 010 published by the PDF Association

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

The Future of PDF/A (and) Validation

- veraPDF (June 2017)
- PDF 2.0 (Fall 2017)
- PDF/A-NEXT (2018)

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

PDF 2.0

- **Developed in seven years**
  - **ISO 32000-1 was more or less the same as PDF 1.7**
  - **No fundamental changes to the main purpose of PDF**
    - Many new features and details
    - Major parts rewritten to be more precise and clear

- **ISO is currently working on new versions based on ISO 32000-2 for**
  - **PDF/A**
  - **PDF/E**
  - **PDF/UA**
  - **PDF/X**

- **Next: What is new in PDF 2.0 for PDF archiving? Complements Dov Isaacs' presentation after lunch "PDF 2.0-based PDF/X Standards for Print Workflows"**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What does PDF 2.0 change for PDF/A? (1 of 3) – encryption and signatures

www.pdfa.org

- **Encryption**
  - **Custom security handlers (not defined in PDF 2.0)**
    - May be combined with an unencrypted wrapper document for PDF processors that do not have access to the security handler
  - **Unicode passwords**
  - **More powerfull encryption algorithms (256-bit AES, Advanced Encryption Standard)**
- **Signatures and certificates**
  - **ECC-based certificates (Elliptic Curve Cryptography)**
  - **Signatures based on PAdES (PDF Advanced Electronic Signatures)**
  - **Long-term signature validation (LTV)**
    - Document Security Store
      information needed to verify a signature, e.g. certificates
    - Document Timestamp Dictionary
      information about the expiration of a certificate etc.

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What does PDF 2.0 change for PDF/A? (2 of 3) – metadata and else

- **Support for geospatial coordinate systems for maps or satellite imagery**
- **Interactive Annotations: 3D, Rich Media**
  - **Projections (measurements e.g. in an active 3D model)**
  - **Support for the new ISO standard for 3D 'PRC'**
  - **Extensions to 3D viewing conditions, incl. transparency**
  - **Deprecate "old" sound and movie actions and annotations**
- **Color**
  - **Page-level output intents**
  - **Device independence can be inherited from group color space**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What does PDF 2.0 change for PDF/A? (3 of 3) – conformance levels

www.pdfa.org

- Tagged PDF (for accessibility)
  - Namespaces for better extensibility
  - Additional standard structure tags
  - Pronunciation hints for text to speech
- Embedded Files
  - Associated Files (from PDF/A-3)
  - Thumbnails for embedded files

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What does PDF 2.0 *deprecate* for PDF/A?

- Fonts
  - CharSets and CIDSets
- XFA forms
- Document Information (Metadata has to be XMP)
- Outdated digital signatures
  - adbe.pkcs7.sha1
  - adbe.x509.rsa_sha1

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

The Future of PDF/A (and) Validation

- veraPDF (June 2017)
- PDF 2.0 (Fall 2017)
- PDF/A-NEXT (2018)

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

PDF/A-NEXT – Taken into account

- Lessons learned with PDF/A-1, 2, 3
- Results from veraPDF (Tech Note 10)
- PDF 2.0

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Current state of discussions at ISO (1 of 4) – overall structure

- "PDF/A-4"
- No conformance level U
  Unicode requirements go from conformance level "U" into "main" but with a "should"
  - Reason: glyphs that are not characters (e.g. bullet point indicators)
- PDF/A-3 (allowing for any kind of embedded file) becomes "PDF/A-4f"

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Current state of discussions at ISO (2 of 4) – taken from 2.0

- No equivalence between Document Information metadata and XMP
- Color may be defined in a device independent way using page level output intents
- CIDSet and CharSet not required to be complete or present for font subsets
- Support for new signatures (but not for encryption)
- Improvements to Tagged PDF

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Current state of discussions at ISO (3 of 4) – not or partly taken from 2.0

- **Disallowed: New interactive annotations (3D, Rich Media)**
- **Non-static content, JavaScript, permitted but limited**
  - **Associated with form fields or other actions**
  - **An interactive processor has to provide special treatment for JavaScript actions:**
    - May only be executed when invoked by the user (via a button or outline entry)
    - A non interactive processor must not execute them
- **Viewers must not use thumbnail images for rendering (which may be present in a page's metadata)**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Current state of discussions at ISO (4 of 4) - metadata

www.pdfa.org

- **Requirements XMP Extension Schemas for custom XMP metadata fields**
  - **Embedding XMP Extension Schemas as in PDF/A1 through PDF/A-3 is neither required nor recommended**
  - **It is recommended to embed a RELAX NG schema (ISO 16684-2:2014 Description of XMP schemas using RELAX NG) for automated validation as with an XML schema**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

www.pdfa.org

Summary:
The Future of PDF/A (and) Validation

- veraPDF (June 2017)
  - Test Corpus
  - Open Source Validator
  - TechNote 010 to be published by the PDF Association
- PDF 2.0 (Fall 2017)
- PDF/A-NEXT (2018)

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

Summary: The Future of PDF/A (and) Validation

- veraPDF (June 2017)
  - Test Corpus
  - Open Source Validator
  - TechNote 010 published by the PDF Association
- PDF 2.0 (Fall 2017)
- PDF/A-NEXT (2018)

Thank you!

Any questions?

Dietrich von Seggern

Get in touch: d.seggern@callassoftware.com
Web site: www.callassoftware.com

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What is PDF 2.0?

- **Makes it clear that resources on a page (e.g. a default color space) is not inherited into annotations, type 3 fonts, form XObjects or pattern. That makes sure that e.g. a Type 3 font or an XObject is not different if present on two different pages**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association

What is PDF 2.0?

- **Security**
  The encryption algorithms included in previous versions of PDF have fallen behind current best practices in security, so PDF adds AES-256-bit and states that all passwords used for AES-256 encryption must be encoded in Unicode.
  A PDF 1.7 reader will almost certainly error and refuse to process any PDF files using the new AES-256 encryption.

- ***PDF 2.0 header:*** **It's only a small thing, but a PDF reader must be prepared to encounter a value of 2.0 in the file header and as the value of the Version key in the Catalog.**
  PDF 1.7 readers will probably vary significantly in their handling of files marked as PDF 2.0. Some may error, others may warn that a future version of that product is required, while others may simply ignore the version completely.

- ***UFT-8 text strings:*** **Previous versions of PDF allowed certain strings in the file to be encoded in PDFDocEncoding or in 16-bit Unicode. PDF 2.0 adds**

Dietich von Seggern
Managing Director

Member of the Board
of the PDF Association