veraPDF: industry supported, open source PDF/A validation for digital preservationists

PREFORMA Experience Workshop, Berlin
23 November

Why veraPDF?

- PDF/A, and the standards on which PDF/A is based, are complex; agreement on meaning isn't always clear
- Need for a **trusted** open source tool
- A single **commercial** entity cannot define conformance with PDF/A
- Industry assistance guarantees the means of **interoperability** (it's PDF's holy grail, after all!)

veraPDF consortium

- **Digital preservationists**
  - **Lead**: Open Preservation Foundation
  - Digital Preservation Coalition
  - KEEP Solutions
- **PDF technology industry**
  - **Lead**: PDF Association
  - Dual Lab (veraPDF lead developer)

Functionality & Quality

- **Functional requirements**
  - PDF/A conformance checker
  - PDF features extraction (characterization)
  - Reliable batch processing
  - Clear reporting
  - Ensure accurate use of PDF/A metadata

- **Quality**
  - Test corpora
  - Open source community
  - Proven quality control practices

Other components

- **Policy Checker**
  - Extract PDF Features with various data from the Document (font names, page count and boundaries, security info, images, etc)
  - Verify the Report against Policy requirements in Schematron syntax
- **Metadata Fixer**
  - Fix any incorrect claims of specific PDF/A validity
  - Add PDF/A identification to an otherwise conforming PDF/A document

veraPDF test corpus stats

- Part 1 Level B: 179 test files complementing Isartor
- Part 2 Level B: 223 test files complementing BFO
- Part 3 Level B: 12 extra tests on embedded files
- Level U: 3 test files on Unicode character map
- Level A: 7 test files on tagging, predefined roles
- XMP 2004: 367 tests on predefined schemas
- XMP 2005: 549 tests on predefined schemas

Industry support & transparency

- **Industry support**
  - PDF Association's Validation Technical Working Group
  - Review test documents
  - Discuss and resolve ambiguities in the specifications

- **Transparency of development practices**
  - Open grammar and validation profiles
  - Dual license scheme: MPLv2+ and GPLv3+
  - Availability of functional and technical specifications
  - All materials available at GitHub open repositoty: github.org/verapdf

Today: beta version 0.26

- Validation of all PDF/A parts and conformance levels
- Java API, REST API, CLI, GUI integration interfaces
- Cross-platform GUI installer
- Validation rules wiki, demo web site
- Test corpus with over 1000+ newly generated files
- Available at: http://downloads.verapdf.org
- All sources at: https://github.com/veraPDF
- **The 1.0 release is planned for mid-December**

Helping to understand PDF/A

- Open format of validation profiles:
  - All 8 validation profiles for 1b,1a,2b,2u,2a,3b,3u,3a
  - Each consists of ~100 atomic rules
- The source code for the PDF model as well as all validation profiles is openly available at github:
  - https://github.com/veraPDF/veraPDF-model
  - https://github.com/veraPDF/veraPDF-validation-profiles
- Documentation: Wiki
  - https://github.com/veraPDF/veraPDF-validation-profiles/wiki

Resolution of ambiguities

- **Identification**:
  - 27 cases formally reported to the mailing list
  - discussed at regular TWG calls
  - formally resolved at joint TWG / ISO committee meetings
  - When applicable to future parts of PDF/A, recommendations are forwarded to the PDF/A Project Leader
- **Distribution**:
  - Resolutions posted on the veraPDF mailing list
- **Documentation**:
  - Included into validation Wiki at https://github.com/veraPDF/veraPDF-validation-profiles/wiki

PDF parser implementation

- veraPDF's proof of concept was implemented using PDFBox, an open-source Java library under Apache
- PREFORMA's licensing requirements mandate dual licensing (GPLv3 and MPLv2); no suitable codebase with such licensing was available
- Accordingly, the veraPDF consortium was obliged to develop a greenfield implementation of a PDF parser
- **The first beta of the new greenfield PDF parser is included into the latest 0.26 release**

Interoperability: PREFORMA suppliers

- Two other suppliers:
  - Easy Innova - TIFF validation
  - MediaArea - video streams validation
- Uniform shell:
  - detects file type and forwards it to the corresponding validation tool.
- Embedded video files:
  - veraPDF provides a sample plug-in for validating embedded video files (AVI, MKV) via MediaArea validator

Extensibility via plug-ins

- PDF/A specifications refer to relevant PDF specifications, which rely on a number of external standards. Validating embedded fonts, ICC profiles, XMP metadata, image compression and more is crucial to establishing archival quality for the whole document
- veraPDF provides a plug-in mechanism to access the embedded ICC profiles, fonts, images, attachments
- Collaboration with experts in relevant technologies is necessary for complete coverage

Today: validation extents update

- **JPEG2000**: plug-in based on Jpylyzer python script
- **Other images**: none, looking for contributors
- **Fonts**: partnering with Compart AG (Germany)
- **XMP**: validated internally by veraPDF
- **ICC profiles**: plug-in based on the official ICC validator
- **Digital signatures**: none, looking for contributors
- **Video attachments:** plug-in based on MediaArea

Development processes

- **Github** is the repository for both code and test corpora
  https://github.com/verapdf
- **Travis** manages the continuous builds
  https://travis-ci.org/veraPDF
- **Jenkins** manages automated tests and deployment
  http://jenkins.opf-labs.org/job/veraPDF-library/
- **Sonar** monitors code quality
  http://sonar.opf-labs.org/dashboard/index/8021

Get involved!

- Join the Open Preservation Foundation http://openpreservation.org/join/
- Download the candidate test suite files from GitHub
- Share your own test files
- Test the code posted on GitHub, and report issues
- Develop plug-ins for validating embedded data or PDF features not related to PDF/A
- Think about how you could use industry supported open source PDF/A validation

Stay in touch


http://verapdf.org/


http://verapdf.org/subscribe/ (news)


users@lists.verapdf.org (Q&A, discussions)


https://twitter.com/_verapdf


https://github.com/veraPDF


info@verapdf.org

Thank you

Questions?