

veraPDF

Industry supported PDF/A validation

About this webinar

What we'll be showing you:

- our current development status;
- the Consortium's development plans for 2016;
- how we've been testing the software so far;
- demonstrations of how to use and test the veraPDF applications; and
- how you can help to improve the pre-release software;
- our ideas and plans for 2017.

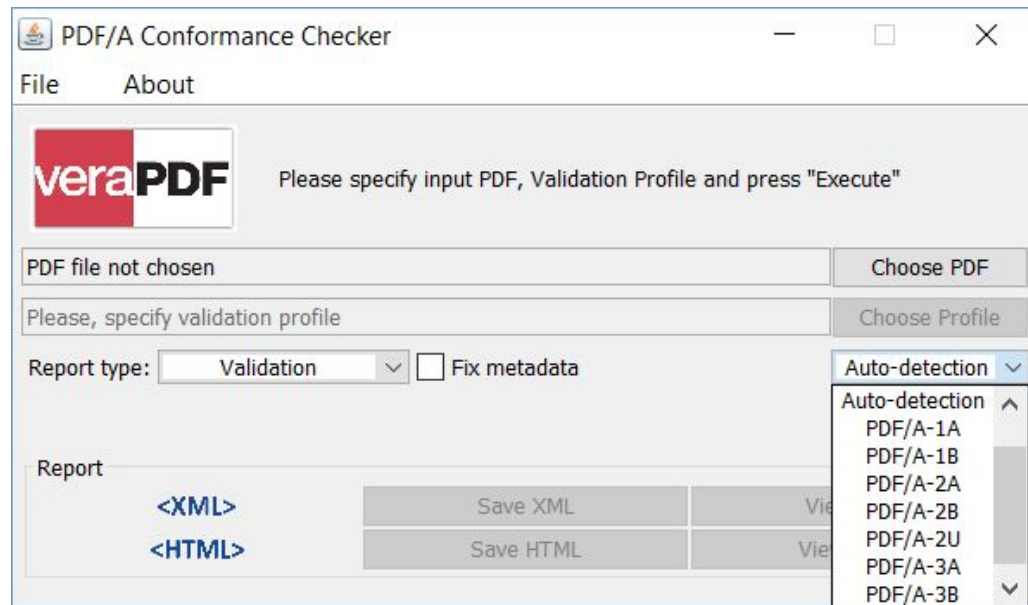
Development status - PDF/A conformance

Latest release:

- Version 0.22 - Sep 9, 2016

Conformance checker:

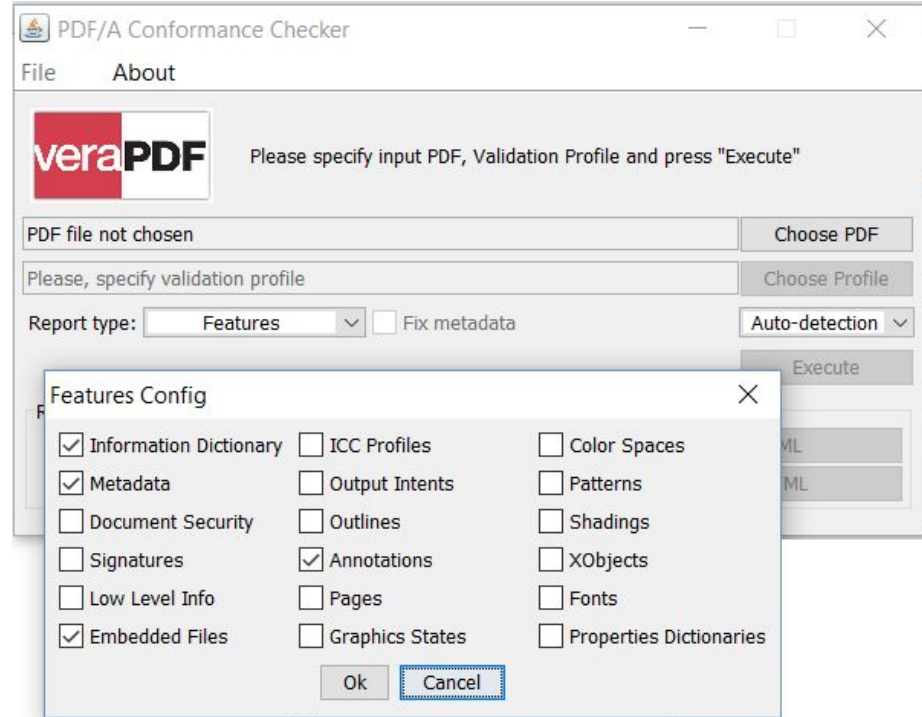
- Full support for all PDF/A versions (1,2,3) and levels (A,B,U)



Development status - custom policies

PDF feature extraction for custom policy checks beyond PDF/A:

- XML report on all metadata, resources, embedded files, pages, annotations, document security, etc



Development status - integration interfaces

- Ready for integration:
 - GUI - Desktop version for single file evaluation
 - CLI - Command line interface targeting large volume batch processing
 - Web - Online demo web site
 - Java library - Calling a Java API from custom Java-based applications
- Plug-in framework for custom validation of embedded formats:
 - XMP metadata, Images, File attachments, Fonts, Digital signatures, ICC profiles
 - Targets integration with third-party tools for verification of other standards inside PDF

Usage: veraPDF [options] FILES

Options:

-c, --config

Loads config form default file, all config flags are ignored.

Default: false

-x, --extract

Extract and report PDF features.

Default: false

--fixmetadata

Performs metadata fix.

Default: false

-f, --flavour

Choose built in Validation Profile flavour, e.g. 1b. Alternatively supply

0 to turn off PDF/A validation or supply auto to automatic flavour
detection from file's metadata.

Default: 1b

Possible Values: [0, auto, 1a, 1b, 2a, 2b, 2u, 3a, 3b, 3u]

--format

Choose output format:

Default: mrr

Possible Values: [xml, mrr, html, text]


-h, --help

Shows this message and exits.

Default: false

VeraPDF | REST Client x

demo.verapdf.org



PDF/A Validation

Prototype PDF/A validation REST service with client.

- 1 Choose
- 2 Configure
- 3 Validate

Configure PDF/A Validator

Validation Profile

PDF/A-1b

Extract features: yes no

[Previous](#) [Next](#)

Development status - test corpus

1500+ atomic self-documented test files:

- Part 1 Level B: 179 test files complementing Isartor
- Part 2 Level B: 223 test files complementing BFO
- Part 3 Level B: 5 extra tests on embedded files
- Level U: 3 test files on Unicode character map
- Level A: 7 test files on logical structure
- XMP 2004 Metadata (PDF/A-1): 367 tests on predefined schemas
- XMP 2005 Metadata (PDF/A-2,3): 549 tests on predefined schemas

<http://tests.verapdf.org> - every new veraPDF build is checked against all test files to guarantee the stability

veraPDF Test Summaries : Tue Sep 20 12:41:42 UTC 2016

Dependencies

core v0.23.4-Tue Sep 20 12:24:00 UTC 2016

pdfbox-validation-model v0.23.7-Tue Sep 20 12:33:00 UTC 2016

BFO-corpus

Valid Cases: 10 | passed: 9 | failed: 1
Invalid: 24 | passed: 24 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

veraPDF-1a-corpus

Valid Cases: 2 | passed: 2 | failed: 0
Invalid: 5 | passed: 5 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

veraPDF-1b-corpus

Valid Cases: 254 | passed: 254 | failed: 0
Invalid: 289 | passed: 289 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

veraPDF-2b-corpus

Valid Cases: 365 | passed: 365 | failed: 0
Invalid: 583 | passed: 583 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

veraPDF-2a-corpus

Valid Cases: 2 | passed: 2 | failed: 0
Invalid: 1 | passed: 1 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

veraPDF-3b-corpus

Valid Cases: 4 | passed: 4 | failed: 0
Invalid: 1 | passed: 1 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

karin-corpus

Valid Cases: 0 | passed: 0 | failed: 0
Invalid: 204 | passed: 204 | failed: 0
Exceptions: 0
Undefined: 0 | Not Applicable: 0

Roadmap till the end of 2016

- Coming releases: ~October 1, ~November 1, ~December 10 (**1.0**)
- Reporting improvements:
 - combined batch processing report,
 - nicer HTML templates,
 - PDF reports
- Policy checks:
 - Impose extra requirements on the document based on its features
 - Risk scoring for PDF/A validation errors
- Common shell for all PREFORMA conformance checkers
- Greenfield PDF parser:
 - current implementation is based on Apache PDFBox and is not compliant with PREFORMA licensing requirements (MPL2+, GPL3+)

Testing veraPDF : Current status

While developing validation functionality we've focussed on:

- Expanding the test corpus to cover the complete PDF/A specifications;
- Testing the validator functionality against the corpus
- Resolving specification ambiguities and corpus creation issues with the Technical Working Group

Application testing and real-world testing has been a secondary concern, although we have tried to address issues raised on GitHub and from external feedback.

Testing veraPDF : External feedback

The DPC have been encouraging organisations to test veraPDF and give us their feedback. We've been listening to this and other external opinions. Our immediate priorities from our initial findings:

- memory usage is an issue;
- reliability of batch processing has held back external testing efforts;
and
- better batch reporting would make testing easier.

Testing veraPDF : Our response

Our end of September v0.24 release will include:

- improved batch processing reliability through more robust exception handling;
- improved reporting of processing errors;
- dedicated batch reporting formats that summarise validation results; and
- memory optimisations where possible, though this is ongoing.

We'll take a look at some of these features now....

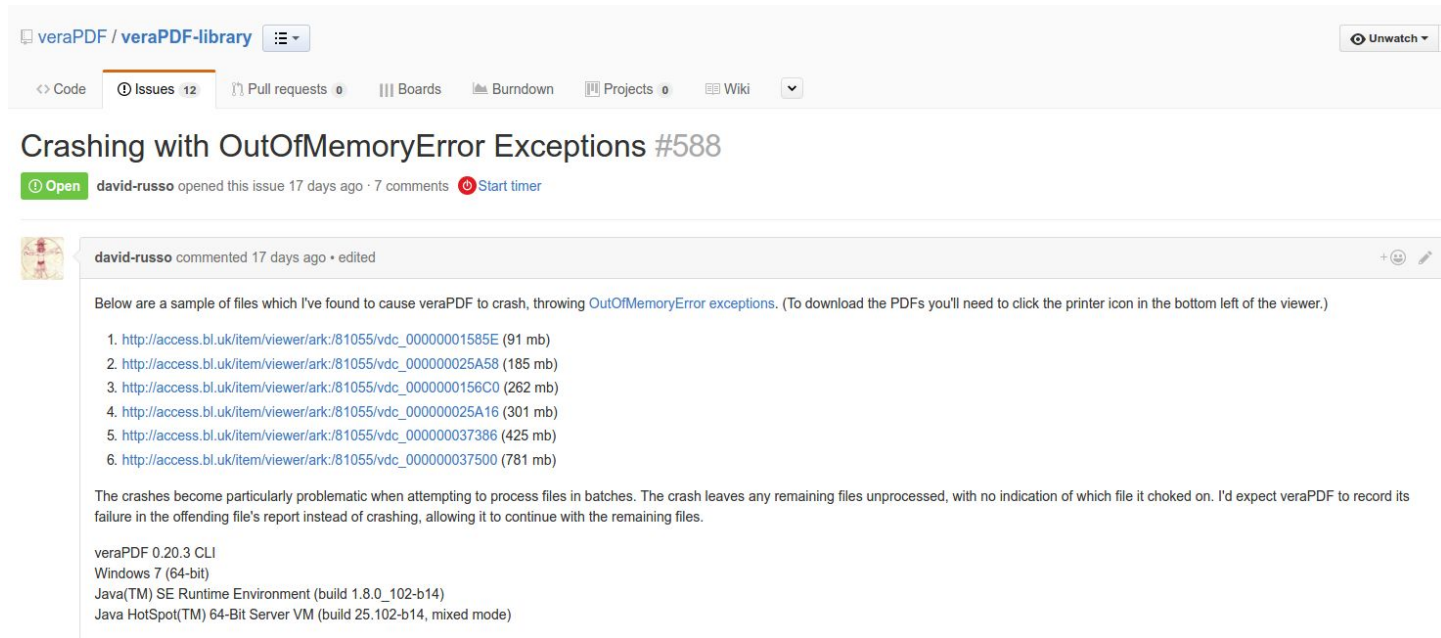
veraPDF Demos

Carl to do demos here



How you can help : GitHub issues

<https://github.com/veraPDF/veraPDF-library/issues/588>



veraPDF / veraPDF-library Unwatch

<> Code **Issues 12** Pull requests 0 Boards Burndown Projects 0 Wiki

Crashing with OutOfMemoryError Exceptions #588

Open david-russo opened this issue 17 days ago · 7 comments **Start timer**

david-russo commented 17 days ago · edited

Below are a sample of files which I've found to cause veraPDF to crash, throwing `OutOfMemoryError` exceptions. (To download the PDFs you'll need to click the printer icon in the bottom left of the viewer.)

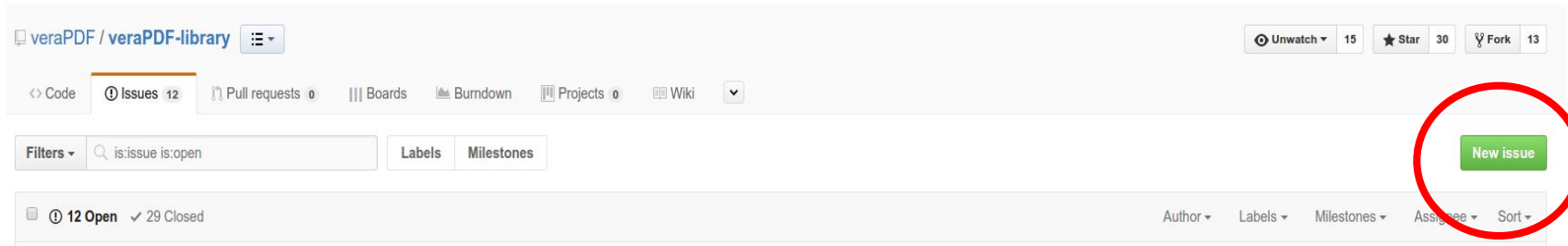
1. http://access.bl.uk/item/viewer/ark:/81055/vdc_00000001585E (91 mb)
2. http://access.bl.uk/item/viewer/ark:/81055/vdc_000000025A58 (185 mb)
3. http://access.bl.uk/item/viewer/ark:/81055/vdc_0000000156C0 (262 mb)
4. http://access.bl.uk/item/viewer/ark:/81055/vdc_000000025A16 (301 mb)
5. http://access.bl.uk/item/viewer/ark:/81055/vdc_000000037386 (425 mb)
6. http://access.bl.uk/item/viewer/ark:/81055/vdc_000000037500 (781 mb)

The crashes become particularly problematic when attempting to process files in batches. The crash leaves any remaining files unprocessed, with no indication of which file it choked on. I'd expect veraPDF to record its failure in the offending file's report instead of crashing, allowing it to continue with the remaining files.

veraPDF 0.20.3 CLI
Windows 7 (64-bit)
Java(TM) SE Runtime Environment (build 1.8.0_102-b14)
Java HotSpot(TM) 64-Bit Server VM (build 25.102-b14, mixed mode)

How you can help : Creating an issue

<https://github.com/veraPDF/veraPDF-library/issues/>



The screenshot shows the GitHub interface for the repository veraPDF/veraPDF-library. At the top, there are navigation options: Code, Issues (12), Pull requests (0), Boards, Burndown, Projects (0), and Wiki. On the right, there are buttons for Unwatch (15), Star (30), and Fork (13). Below the navigation, there is a search bar with the filter 'is:issue is:open' and buttons for Labels and Milestones. A 'New issue' button is highlighted with a red circle. At the bottom, there is a summary bar showing '12 Open' and '29 Closed' issues, along with dropdown menus for Author, Labels, Milestones, Assignee, and Sort.

How you can help : Creating an issue

<https://github.com/veraPDF/veraPDF-library/issues/new>

veraPDF / veraPDF-library

Unwatch 15 Star 30 Fork 13

<> Code Issues 12 Pull requests 0 Boards Burndown Projects 0 Wiki

Title

Write Preview AA B i “ <> ↻ ⋮ ⋮ ⋮ ↶ @ 📎

Leave a comment

Attach files by dragging & dropping, selecting them, or pasting from the clipboard.

Styling with Markdown is supported

Submit new issue

Labels: None yet

Milestone: No milestone

Assignees: No one—assign yourself

How you can help : Feedback

<https://github.com/veraPDF/veraPDF-library/issues/588>



bdoubrov commented 13 days ago

veraPDF member + 😊 ✎ ✕

The latest release veraPDF 0.22.1 improves memory management and provides more user-friendly OutOfMemoryError handling along with instructions how to increase the JVM memory.

In particular, increasing the JVM memory to 4Gb allows to process all above test files.

We still keep this issue open and will try to implement for robust batch processing.



david-russo commented 13 days ago • edited

+ 😊 ✎ ✕

Hi Boris, thanks for looking into this.

Solving the problem by increasing the memory doesn't sound like a complete solution to me, since it's predicated on the user still having to experience the problem, and even outside of batch processing I don't think crashing is an expected exit condition for a user.

Since there's no limit to how large the contents of a PDF can be, it's easy to imagine a scenario where the resources necessary to complete validation might not exist. That being the case, it would seem prudent to include some heuristics which could allow veraPDF to recognize when it doesn't have the resources to finish processing a file and bow out gracefully, leaving a report.

How you can help : Mailing List

We're pleased to announce the new veraPDF Mailing List :

- <http://lists.verapdf.org/listinfo/users> is for:
 - reporting issues,
 - asking questions,
 - suggesting improvements, and
 - talking to other veraPDF users and the development team.

Understanding the results - PDF/A Conformance



Validation Report

Validation Profile: PDF/A-1B validation profile
Compliance status: Failed

Statistics

Version: 0.23.2
Build Date: 2016-09-15T09:17:00-03:00
Processing time: 00:00:06.676
Total rules in Profile: 102
Passed Checks: 306
Failed Checks: 2

Validation information

Rule	Status
Specification: ISO 19005-1:2005, Clause: 6.3.6, Test number: 1	
For every font embedded in a conforming file and used for rendering, the glyph width information in the font dictionary and in the embedded font program shall be consistent.	Failed
2 occurrences	Show

Wiki on error details available at:
<http://docs.verapdf.org/validation/>

Rule 6.3.6-1

Requirement

For every font embedded in a conforming file, the glyph width information stored in the Widths entry of the font dictionary and in the embedded font program shall be consistent.

Error details

Glyph width information in the embedded font program is not consistent with the Widths entry of the font dictionary.

This requirement is necessary to ensure predictable font rendering, regardless of whether a given reader uses the metrics in the Widths entry or those in the font program.

- Object type: Glyph
- Test condition: `isWidthConsistent == true`
- Specification: ISO 19005-1:2005
- Levels: A, B



Understanding the results - Feature report

XML report with all features, HTML report coming soon

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<report xmlns="http://www.verapdf.org/MachineReadableReport" creationDate="2016-09-15T10:08:32.708-03:00" processingTime="00:00:00.503" version="0.23.2" buildDate="2016-09-15T09:17:00-03:00" itemDetails size="71251">
  <pdfFeaturesReport>
    <metadata>
      <xmpPackage>
        <x:xmpmeta xmlns:ns2="http://www.verapdf.org/MachineReadableReport" xmlns="" xmlns:x="adobe:ng:meta/">
          <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
            <rdf:Description rdf:about="" xmlns:dc="http://purl.org/dc/elements/1.1/">
              <dc:creator>
                <rdf:Seq>
                  <rdf:li>veraPDF Consortium</rdf:li>
                </rdf:Seq>
              </dc:creator>
            </rdf:Description>
            <rdf:Description xmp:ModifyDate="2016-02-20T18:07:32+03:00" xmlns:xmp="http://ns.adobe.com/xap/1.0/" rdf:about="" xmp:CreatorTool="veraPDF Test Builder" xmp:CreateDate="2016-06-14T12:48:37.000Z" xmlns:xmp:pdf="http://ns.adobe.com/pdf/1.3/" pdf:Producer="veraPDF Test Builder 1.0" rdf:about="">
            <rdf:Description pdfaid:part="3" pdfaid:conformance="B" rdf:about="" xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/">
            </rdf:Description>
          </rdf:RDF>
        </x:xmpmeta>
      </xmpPackage>
    </metadata>
    <embeddedFiles>
      <embeddedFile xmlns:ns2="http://www.verapdf.org/MachineReadableReport" ns2:id="file1">
        <ns2:fileName>ChartDiagram.csv</ns2:fileName>
        <ns2:description></ns2:description>
        <ns2:subtype>text/csv</ns2:subtype>
        <ns2:filter>FlateDecode</ns2:filter>
        <ns2:creationDate>2016-06-14T12:57:20.000+03:00</ns2:creationDate>
        <ns2:modDate>2016-06-14T12:48:37.000Z</ns2:modDate>
        <ns2:checksum>1128E177280423C33BE5E96C66303EB1</ns2:checksum>
        <ns2:size>33</ns2:size>
      </embeddedFile>
    </embeddedFiles>
  </pdfFeaturesReport>
</report>
```

Future Plans : PREFORMA Testing Phase

- PREFORMA's Prototyping Phase ends 2016 with veraPDF 1.0 release.
- PREFORMA's Testing Phase runs January 2017 - June 2017
 - Acceptance testing against test data sets created by PREFORMA
 - veraPDF consortium will respond to testing feedback from PREFORMA
- During the testing phase the veraPDF consortium will continue to:
 - address Issues and Pull Requests on GitHub;
 - update website;
 - provide support to enquiries on the mailing list; and
 - improve documentation.

We'll also be working to transition from funded to open source project.



Future Plans: June 2017 onwards

PREFORMA funding ends with the testing phase in June 2017. The veraPDF consortium will:

- continue to address bugs reported on GitHub;
- test and merge small Pull Requests submitted to GitHub; and
- provide answers and support for Mailing List enquiries.

Priority/paid support.

Available for paid development or integration with local systems.

New features and functionality will require funding or external development resource.