veraPDF after PREFORMA

Real world adoption and industry needs for more PDF standards

Boris Doubrov, Martin Wrigley
veraPDF Consortium

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

History of veraPDF / PREFROMA

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

The PREFORMA project

- veraPDF development has been funded by the PREFORMA project
- PREservation FORMAts for culture information/e-archives, is a Pre-Commercial Procurement (PCP) project co-funded by the European Commission under its FP7-ICT Programme.
- The project's main aim is to address the challenge of implementing standardized file formats for preserving digital objects in the long term, giving memory institutions full control over the acceptance and management of preservation files into digital repositories.
- http://www.preforma-project.eu/

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

The veraPDF consortium

Led by the

- Open Preservation Foundation
- PDF Association

With partners

- Dual Lab
- KEEP Solutions
- DPC

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

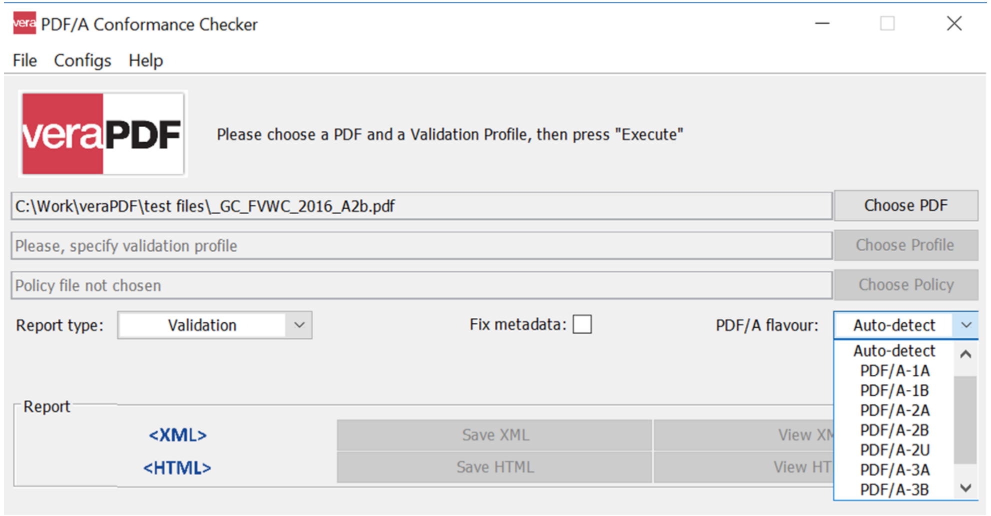
PDF/A validation

- Conformance checker: full support for all PDF/A versions (1,2,3) and levels (A,B,U)
- Test corpus of over 1000+ files
- GUI - desktop version for single file evaluation
- CLI - command line interface targeting large volume batch processing
- Web - Online demo web site: http://demo.verapdf.org
- Java library - Calling a Java API from custom Java-based applications

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Open source for sustainability

- Software licensed under dual MPL v2+ / GPL v3+ allowing anyone to download or integrate the software free of charge
- Test datasets and documentation licensed under CC-BY-4
- Code and test files are available at http://github.com/verapdf



PDF/A Conformance Checker

Please choose a PDF and a Validation Profile, then press "Execute"

C:\Work\veraPDF\test files_GC_FVWC_2016_A2b.pdf

Please, specify validation profile

Policy file not chosen

Report type: Validation

Fix metadata: □

PDF/A flavour: Auto-detect

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Current state of the art

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Recent news and developments

- Latest release: 1.12 (May 9, 2018)
- Multi-threading support
- CLI optimized for processing large collections in several parallel processes
- Support for ISO 32000-2 in PDF feature extraction
- Improved user interface for defining institutional policy profiles
- Multiple fixes for the support issues

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

PDF Association: Industry support

- PDF Association’s Validation Technical Working Group operates since December 2014 and recently conducted its 36th meeting
- Discuss and resolve ambiguities in the specifications
- Published TN0010 “Clarifications of ISO 19005, parts 1-3 for developers of PDF/A creators and validators”, which contains 28 resolutions
- There are 17 more ambiguities under study
- Many these discussions resulted in the comments for the ISO specifications!

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Current questions under discussion

- Are Type1 fonts in PFB format allowed in PDF?
- How to specify identity ToUnicode maps?
- Is non-breaking space the same as SPACE?
- Are custom encodings in 14 standard fonts allowed?
- When transfer functions may / shall be present in the Halftone dictionaries?
- Is /AA entry permitted in non-Widget annotations?

Many of these questions end up to be submitted to the ISO committee as comments against newly developed standards!

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Long-term support

PREFORMA funding ended in July 2017.

veraPDF is transitioning from a funded project to a stand alone open source project.

- Is EU-funded project dead after the end date?

As of now the OPF and Dual Lab continue the support the project:

- continue to address bugs reported on GitHub;
- test and merge small new features submitted as pull requests submitted to GitHub; and
- provide answers and support for all mailing list enquiries.

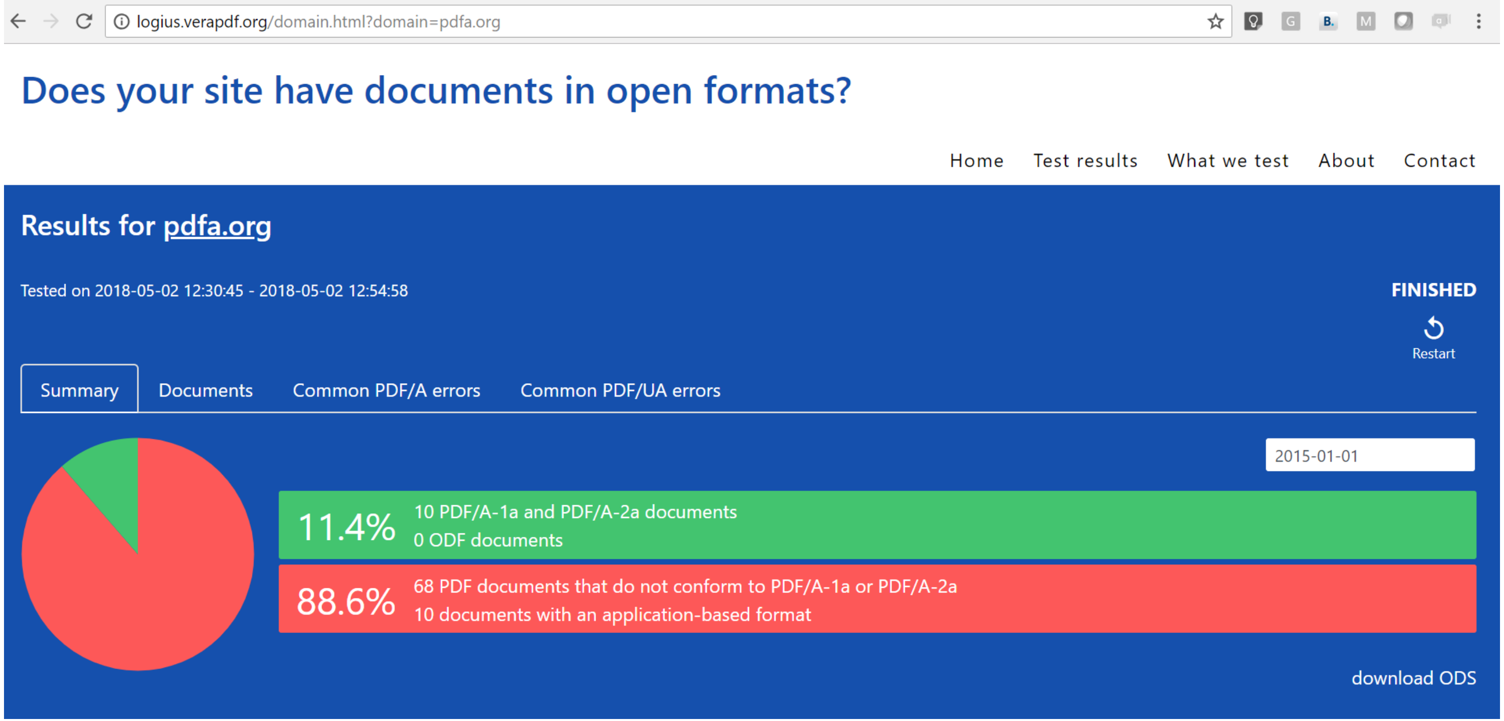
Boris Doubrov,
Martin Wrigley
veraPDF Consortium

veraPDF-based applications

- The veraPDF consortium is in an active development partnership with Logius, the digital government service of the Netherlands Ministry of the Interior and Kingdom Relations (BZK)Web domain(s) crawler for all PDF documents with their validation against PDF/A standards: http://logius.verapdf.org/
- Web-based converter of Office documents to PDF/A-1a with post-conversion validation: http://cnv.verapdf.org/
- ZUGfeRD project has developed an open-source tool for validate PDF/A-3 based ZUGfeRD documents using the veraPDF and a plug-in to check the XML attachments against EN16931 UN/CEFACT SCRDM v16B
- https://github.com/ZUGFeRD

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Web crawler with veraPDF-based PDF/A validation



Does your site have documents in open formats?

Results for pdfa.org

Tested on 2018-05-02 12:30:45 - 2018-05-02 12:54:58

FINISHED

11.4% 10 PDF/A-1a and PDF/A-2a documents
0 ODF documents

88.6% 68 PDF documents that do not conform to PDF/A-1a or PDF/A-2a
10 documents with an application-based format

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

2018-05-14

Future of veraPDF

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Future developments?

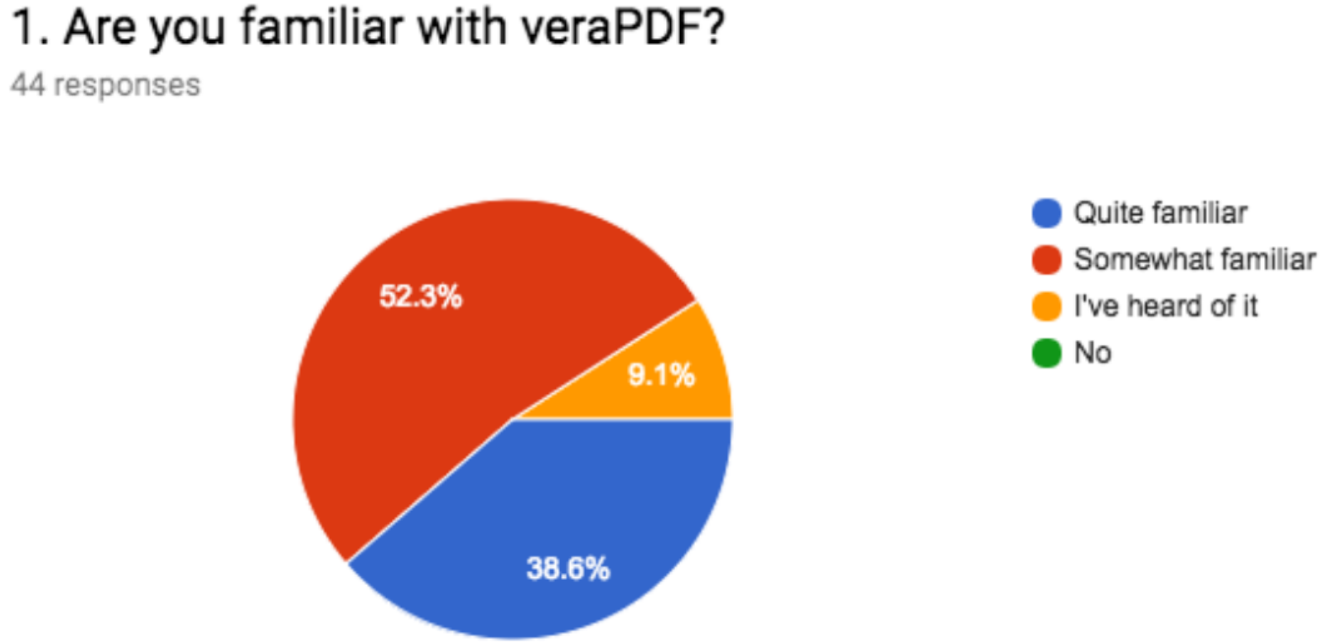
veraPDF is functionally complete for PDF/A but there is still work to be done, for example:

- developing validators for other parts of the PDF standard
- veraPDF supports a plug in mechanism that can be used to validate other formats embedded or attached to PDF/A documents
- integrating veraPDF with other preservation systems and other services
- the creation of more complex institutional policies that reflect collections of real institutional acceptance criteria

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

PDF Association Members’ Survey

- Conducted December 20, 2017 - January 12, 2018
- 45 responses

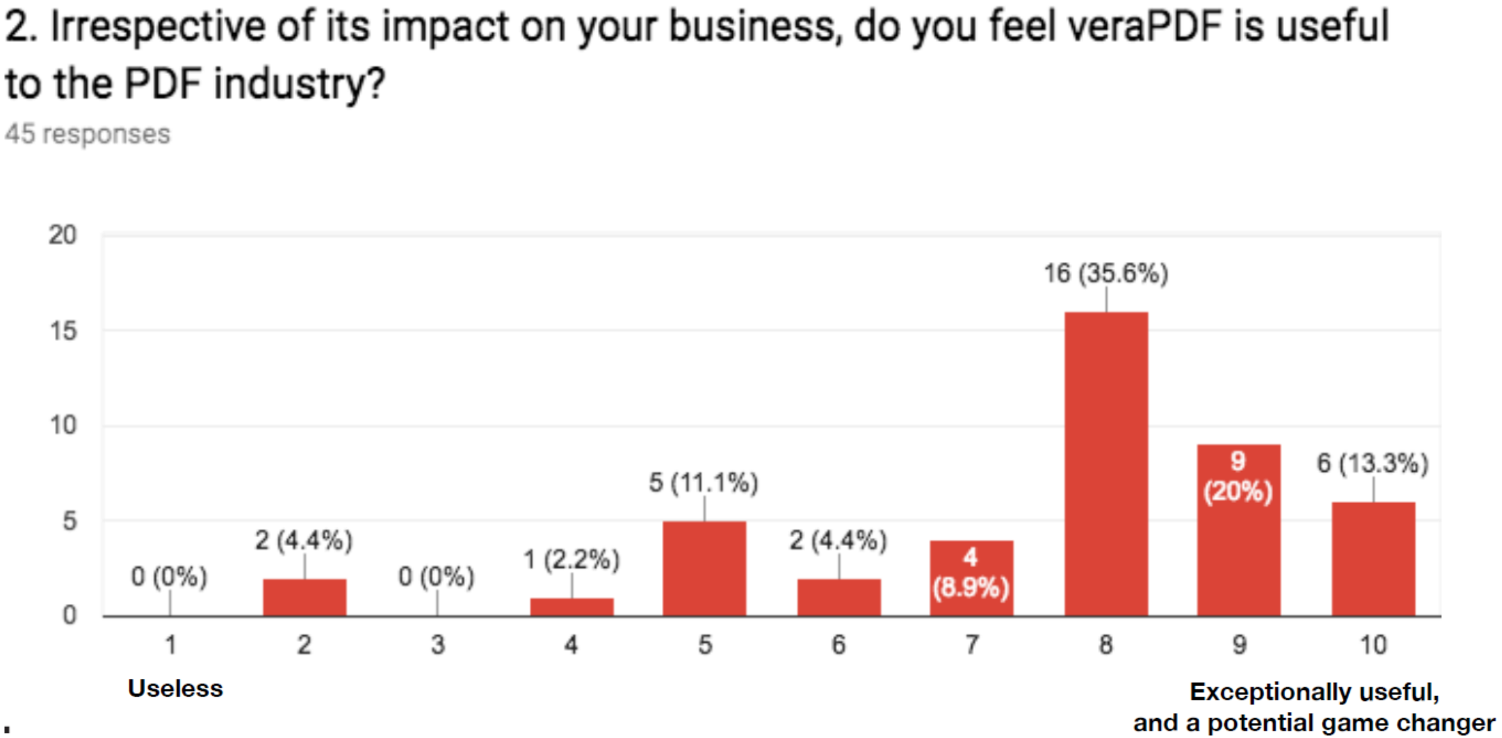


1. Are you familiar with veraPDF?

44 responses

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Is veraPDF useful to the PDF industry?

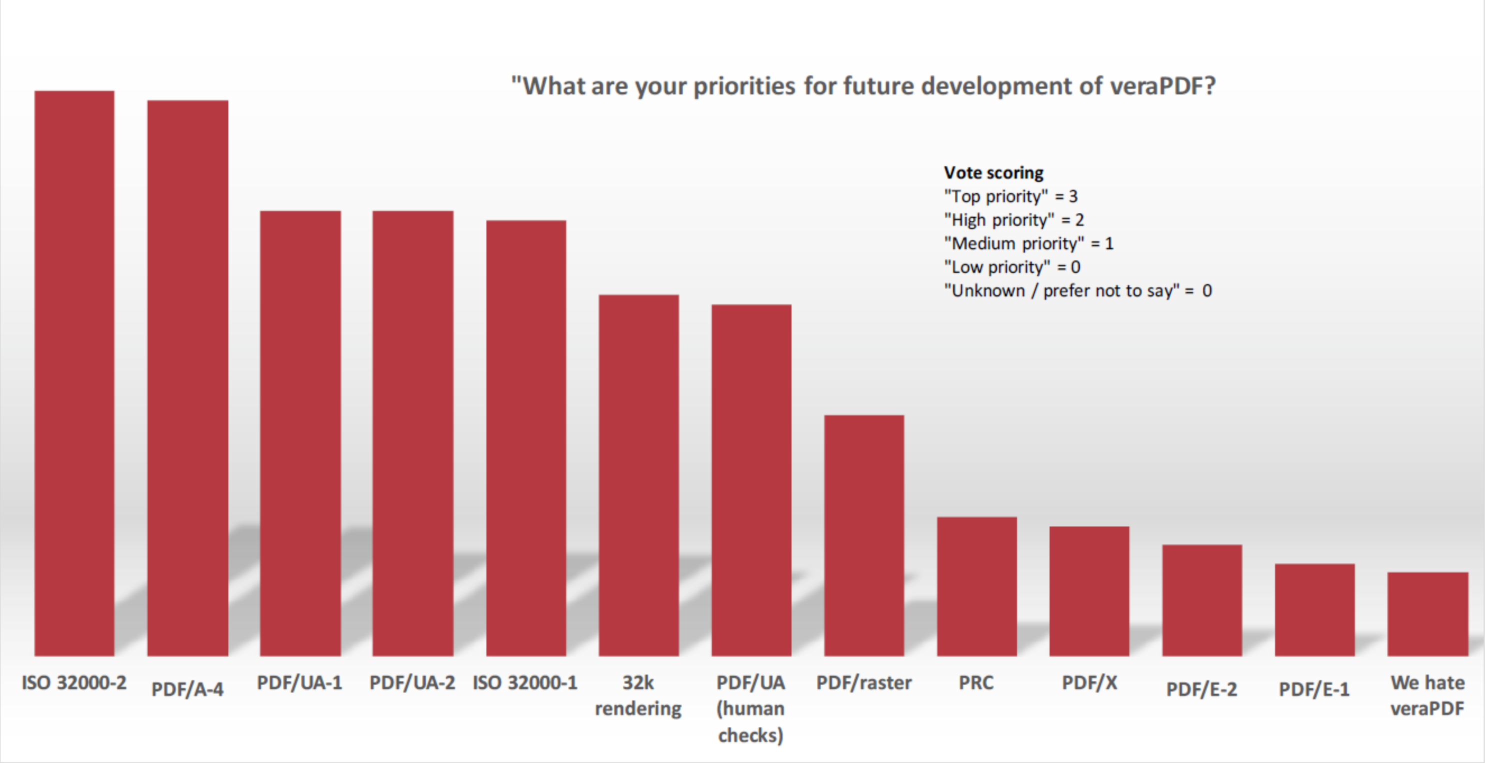


2. Irrespective of its impact on your business, do you feel veraPDF is useful to the PDF industry?

45 responses

www.pdfa.org

What are your priorities for future development?



Boris Doubrov,
Martin Wrigley
veraPDF Consortium

2018-05-14

Feedback from archival institutions

iPRES 2017: need for **PDF/UA**

- Marco Klindt (ZIB), PDF/A considered harmful for digital preservation
- *"Require data producers to implement workflows that adhere to the Matterhorn protocol to assure fully, meaningful tagged PDFs (including MathML formulas, semantically tagged data and so on) and to provide /ActualText for every textual information contained in the PDF that is not easily extractable otherwise."*

JHOVE community: need for **32000k** validation

- ISO 32000-1 validation (syntax level, document and page structure) is badly needed with no tools available today

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Partnership between the industry and memory institutions

- Memory institutions alone cannot have expertise in-house for every format they collect and preserve.
- The PDF family of formats is widely used in digital preservation. PDF/A is just one specification, there are no complete open source validators for the others.
- OPF and PDF Association have bridged a gap between memory institutions and industry to create a successful product.
- Growing community drives us further to cover more gaps!

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Open Preservation Foundation

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Open Preservation Foundation

OPF

Not for Profit, Global membership association tasked with stewarding open-source tools for the digital preservation community.

OPF reference toolset includes veraPDF, JHOVE and more

Martin Wrigley – Executive Director OPF

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

OPF & veraPDF: Sustaining open source tools

OPF Vision:

Open sustainable digital preservation

OPF Mission:

Enabling shared solutions for effective and efficient digital preservation; the Open Preservation Foundation leads a collaborative effort to create, maintain and develop the reference set of sustainable, open source digital preservation tools.

This set of tools (including software and standards) enables organisations to evaluate, validate, document, mitigate risk, and process digital content to be preserved in line with desired policies and community best practice.

Open Preservation Foundation

veraPDF

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

OPF Members

- Austrian Institute of Technology
- British Library
- Bibliotheque Nationale de France
- Goportis
- International Atomic Energy Archives
- Jisc
- Koninklijke Bibliotheek
- Det Kgl. Bibliotek
- Nationaal Archief
- The National Archives UK
- Nasjonalbiblioteket
- Rigsarkivet
- Ex Libris
- Rahvusarhiiv
- Latvijas Nacionala biblioteka
- Österreichische Nationalbibliothek
- Preservica
- Yale University Library
- Albert-Ludwigs Universitat
- University of North Carolina
- Portico
- PSNC (Poznan Supercomputing & Networking Centre)
- Artefectual
- Biblioteca Nacional de Portugal
- Arcsys Software

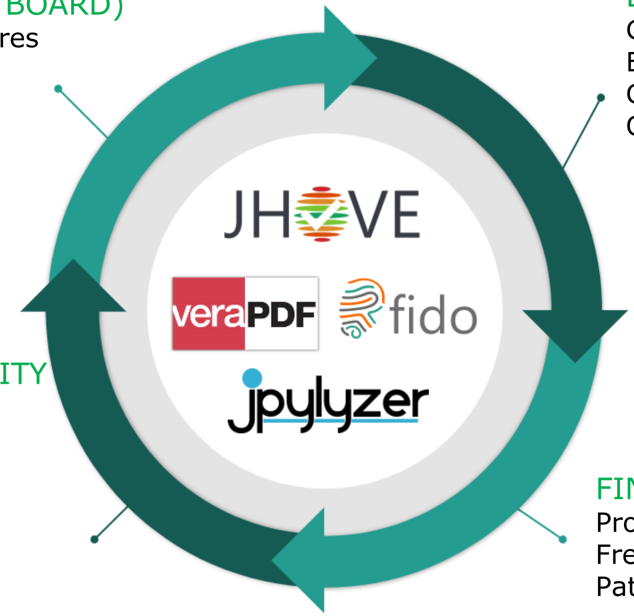
The OPF process

FUNDING

OPF membership
Donations
Project income

PLANNING (PRODUCT BOARD)

Prioritise fixes and features
Define the release
Manage the roadmap



DEVELOPMENT & TESTING

GitHub for OS development
Build a set of test data
Continuous integration
Quality Assurance

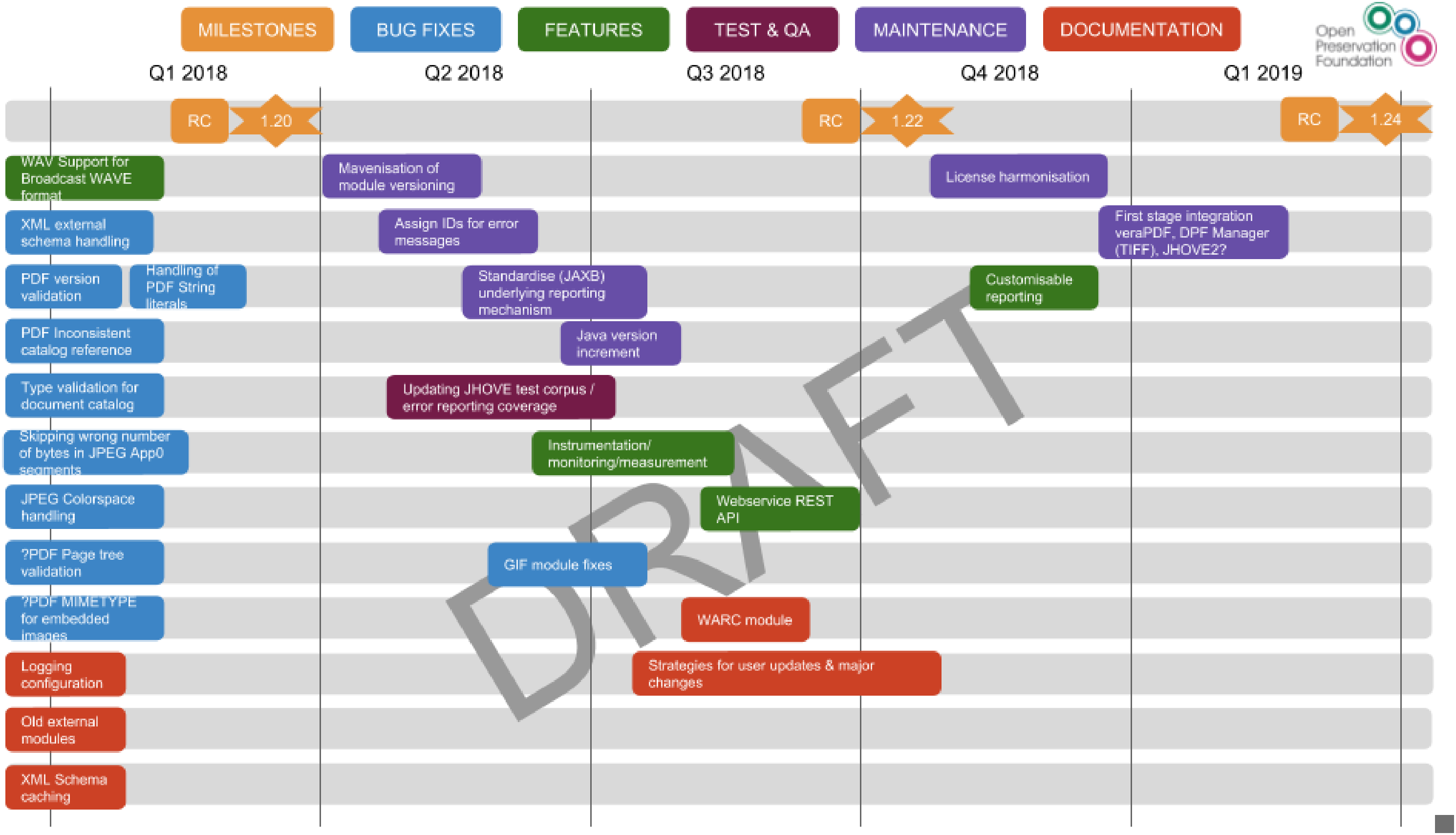
REQUIREMENTS & COMMUNITY FEEDBACK

Bug reports and new feature requests
Hack day activities
Code contributions
Input from OPF interest groups
Contribution of test files
Improvements to documentation

FINAL TEST & RELEASE

Production release
Freely available to community
Patches (essential fixes)

Example (JHOVE) Product Roadmap



MILESTONES · BUG FIXES · FEATURES · TEST & QA · MAINTENANCE · DOCUMENTATION

Q1 2018 · Q2 2018 · Q3 2018 · Q4 2018 · Q1 2019

RC · 1.20 · RC · 1.22 · RC · 1.24

WAV Support for Broadcast WAVE format

XML external schema handling

PDF version validation

Handling of PDF String literals

PDF Inconsistent catalog reference

Type validation for document catalog

Skipping wrong number of bytes in JPEG App0 segments

JPEG Colorspace handling

?PDF Page tree validation

?PDF MIMETYPE for embedded images

Logging configuration

Old external modules

XML Schema caching

Mavenisation of module versioning

Assign IDs for error messages

Standardise (JAXB) underlying reporting mechanism

Java version increment

Updating JHOVE test corpus / error reporting coverage

Instrumentation/ monitoring/measurement

Webservice REST API

GIF module fixes

WARC module

Strategies for user updates & major changes

License harmonisation

First stage integration veraPDF, DPF Manager (TIFF), JHOVE2?

Customisable reporting

Open Preservation Foundation

veraPDF

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

2018-05-14

Participate!

- Web site: http://verapdf.org/
- OPF Web site: www.openpreservation.org
- User mailing list: users@lists.verapdf.org
- Source code: https://github.com/veraPDF
- News: http://verapdf.org/subscribe/
- Twitter: @_verapdf
- Email: info@verapdf.org
- Validation TWG: verapdf-tech@googlegroups.com (for members of PDF Association)

Boris Doubrov,
Martin Wrigley
veraPDF Consortium

Thank you! Any questions?

veraPDF

Get in touch:	matt.kuznicki@pdfa.org
Web site:	www.pdfa.org
Twitter:	PDFassocation

Boris Doubrov,
Martin Wrigley
veraPDF Consortium