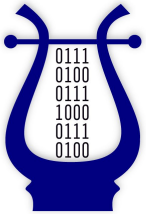


RHAPSODE
CONSULTING

PDF Versioning

A DPC Connect Session

Tim Allison, Rhapsode Consulting, LLC
Peter Wyatt, CTO, PDF Association

July 2023

What is PDF Versioning?

- What a PDF tells us
  - File header
  - Document Catalog's `/Version`

```
%PDF-1.4
%öäüß
1 0 obj
<<
/Type /Catalog
/Version /1.4
```

- What features a PDF might contain
  - e.g. transparent imaging was introduced in 1.4

RHAPSODE CONSULTING

What is PDF Versioning?

- The PDF version is supposed to define what a file is allowed to contain
  - The PDF specification allows a PDF to state a higher version than its features.
  - The PDF spec requires that a PDF's stated version must be ≥ the feature with the highest version:
    - a PDF 1.6 file may contain only features defined in PDF 1.4
    - a PDF 1.4 file should *not* contain features defined in PDF 1.6

RHAPSODE CONSULTING

Analysis of ~1 Million PDFs

- Sampled from **CC-MAIN-2021-31-PDF-UNTRUNCATED**
- Tools:
  - `Poppler's pdfinfo` & `Apache Tika` for what the PDF file tells us
  - `Arlington TestGrammar` for feature identification

RHAPSODE CONSULTING

Analysis of a ~1 Million PDFs: Limitations

- PDFs from the web
  - Not to be confused with “all PDFs”
- `pdfinfo` reported PDF version
  - Combines PDF header (`%PDF-`*`x.y`*) and Document Catalog `/Version` entry
  - Does ***not*** report PDF version in XMP data or features

https://gitlab.freedesktop.org/poppler/poppler/-/blob/master/poppler/PDFDoc.h#L290

```
290     // Return the PDF version specified by the file (either header or catalog).
291     int getPDFMajorVersion() const { return std::max(headerPdfMajorVersion, catalog->getPDFMajorVersion()); }
292     int getPDFMinorVersion() const
293     {
294         const int catalogMajorVersion = catalog->getPDFMajorVersion();
295         if (catalogMajorVersion > headerPdfMajorVersion) {
296             return catalog->getPDFMinorVersion();
297         } else if (headerPdfMajorVersion > catalogMajorVersion) {
298             return headerPdfMinorVersion;
299         } else {
300             return std::max(headerPdfMinorVersion, catalog->getPDFMinorVersion());
301         }
302     }
```

- Arlington performs feature checking of *keys*, but not key *values*
  - e.g. page object `/Tabs` key was introduced in PDF 1.5 with values `R`,`C`,`S`.
    PDF 2.0 added extra values `A`,`W`. that Adobe Extension Level 3 (PDF 1.7) also had.
    Arlington currently reports all `/Tabs` entries as PDF 1.5 *regardless of key value*.

RHAPSODE CONSULTING

Versions Confusion Matrix

Self-Identified Version →

Latest Feature Version ↓

	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.0
1.0	208	218	1,389	3,232	5,578	372	452	279	-
1.1	2	656	1,712	10,787	11,797	278	397	5,007	-
1.2	2	103	1,547	13,599	5,131	4,008	149	14,235	-
1.3	17	169	2,379	22,998	8,022	645	148	3,034	-
1.4	18	72	1,351	31,091	154,162	14,691	10,374	79,704	23
1.5	-	1	276	13,339	59,530	143,205	72,959	90,049	202
1.6	-	-	32	1,244	26,847	62,120	33,504	62,396	22
1.7	-	-	11	394	1,610	907	1,695	10,785	-
2.0	2	10	13	398	829	599	822	1,440	12

RHAPSODE CONSULTING

© 2022 California Institute of Technology. Government sponsorship acknowledged.

Summarized Version Data

Outcome	Count	Percent
PDF's stated version equals latest feature version	367k	37%
PDF's stated version is higher than the latest feature version	421k	42%
PDF's stated version is lower than the latest feature version	207k	21%

At most, 788K (79%) are correctly versioned*

* Arlington currently checks *keys*, not *key values*, so this is the **best case**

RHAPSODE CONSULTING

PDF file format extensibility

- **PDF has always been extensible by design**
  - PDFs can include features unsupported by older viewers
  - PDFs can include private (vendor/product-specific) data
  - But, as with any PDF feature, implementers can choose to support or not support any given extension

```
<< /Type     /ExtGState
   /AAPL:ST << /Type        /Style
               /Subtype     /Shadow
               /Offset      [ 1.5 -1.5 ]
               /Radius      4
               /ColorSpace 4 0 R
               /Color       [ 0 0 0 0.7 ]
            >>
>>
<< /Type /ExtGState /AAPL:AA false >>
<< /Type /ExtGState /AAPL:AA true >>
```

- **PDF is generally(*) backwards compatible**
  - PDFs with unrecognized features can still be processed
- PDF software is not obligated to fully implement every feature
  - And many do not…
- It is *not possible* for software to understand the impact of unknown features
  - Rendering, text extraction, interactivity, metadata, …

“Generally” backwards compatible…

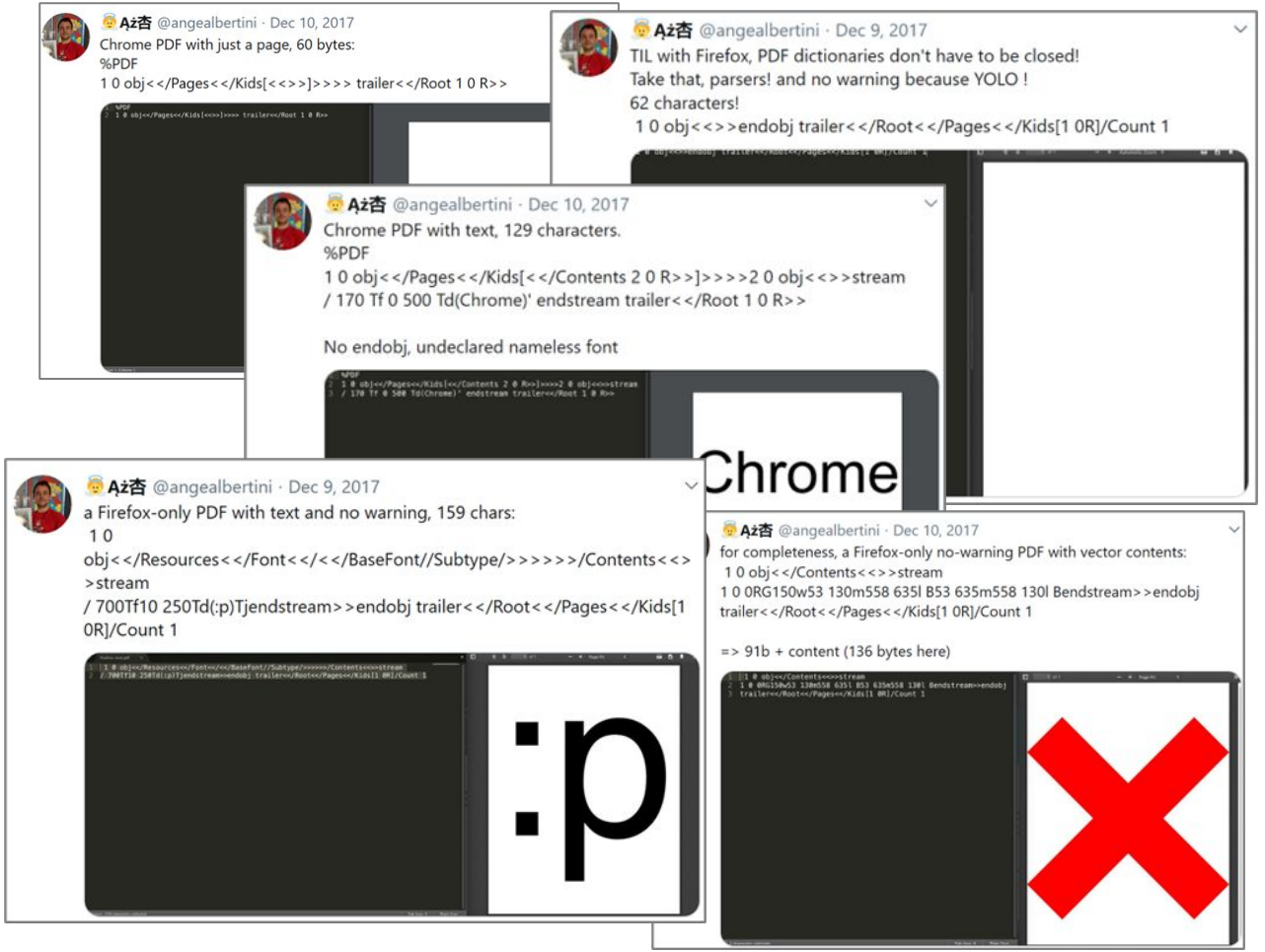
- New filter types → cannot read specific objects
  - If an image format → could leave a “hole” on the page (or fail the page/document)
    - PDF 1.4: JBIG2, PDF 1.5: JPEG 2000
  - If general compression → cannot read specific stream object
    - PDF 1.1: binary data, PDF 1.2: FLATE
    - “Compact PDF”: BZIP2 https://multivalent.sourceforge.net/Research/CompactPDF.html
- New encryption → cannot render as content streams are encrypted
  - *Encryption only affects strings and streams*
  - But can still read PDF file and others kinds of objects
- PDF 1.4 transparent imaging model → possibly wrong appearance
- PDF 1.5 object streams and cross reference streams → cannot read file

PDF Version

- **PDF header line `%PDF-x.y` is the official file identifier magic**
  - Should be correct version for original PDF
- **No PDF software changes behavior/functionality based on PDF version**
  - When software reads a feature it understands, it uses the feature
  - Software ignores features it doesn't support, regardless of version
- **Reported version is only for users, but has no function**
  - Implementations are unclear if reported version is header version, Document Catalog `/Version` key, XMP PDF version, feature-based, implementation capability, some combination, or something else…

When is a file a PDF?

- Most software (*but not all!*) need `%PDF-x.y`
- Most do not need `%%EOF` or `xref`
- Some always ignore `xref`
- Many do not report errors
  - Just show blank pages
  - Ask to save PDF on exit



Is it a PDF?

- **Is it a PDF file?**
- **Is it a functional PDF file?**
  - *By what definition of "functional"?*
  - a.k.a. "compatible"
- **Is it a compliant PDF file?**
  - "Compliant" = as allowed by core PDF specification
  - *Is it malformed in some way?*
- **Is it a conforming PDF/A file?**
  - "Conforming" = self-declared PDF/A (ISO 19005) conformance level
  - *Did file validate accordingly?*
  - *Is it being viewed and used only with conforming software?*

All PDF specifications allow private data and extensions

PDF association

What PDF features matter for #digipres?

- Metadata
- Encryption
- Images (format, dpi, EXIF, etc)
- Embedded fonts (with Unicode mappings!)
- Device independent color
- Attachments and other embedded files
- Sensitive data (PII):
  - Stored by tools unbeknownst to human creator
  - Unsuccessful or shallow redaction
- Proprietary “extensible” data
- Incremental updates
- ???

Contact information

Raf Hens
Chair
E – raf.hens@apryse.com

Dietrich von Seggern
Vice-Chair & ISO Liaison Officer
E –d.seggern@callassoftware.com

Duff Johnson, CEO
Main North America contact
E – duff.johnson@pdfa.org

Peter Wyatt
CTO & ICC Liaison Officer
E – peter.wyatt@pdfa.org

Betsy Fanning
Standards Director
E – betsy.fanning@pdfa.org

Thomas Zellmann, Evangelist
Main Europe contact
E – thomas.zellmann@pdfa.org

Matthias Wagner
Operations Director
E – matthias.wagner@pdfa.org

https://pdfa.org

info@pdfa.org

https://github.com/pdf-association

https://www.facebook.com/PDFassociation/

https://www.youtube.com/user/ThePDFAssociation

https://www.linkedin.com/company/pdf-association/

Tim Allison, Founder

tallison@rhapsodeconsulting.com

https://www.linkedin.com/company/98366844

https://github.com/tballison