



PDF XMP Metadata Validation with Relax NG Schemas and the PoDoFo Library

An Open Source Implementation of ISO 16684-2:2014

[Francesco Pretto](#) | Software Engineer | The PoDoFo Community

PDF Days
Europe
2025

Outline

- Background/context
 - XMP, PoDoFo library, problem definition
- Normative framework
 - Requirements and interesting findings
- The solution development process
- Next Steps/TODOs
- Summary
- Questions

Context - XMP *Extensible Metadata Platform*

- An ISO standardized cross format (PDF, JPEG, etc.) container for storing metadata
- Developed by Adobe in 2001, initially with focus on PDF
- XML based: still good (focus in formal validation) but old



```
1  <?xpacket begin="✪" id="W5M0MpCehiHzreSzNTczkc9d"?>
2  <x:xmpmeta xmlns:x="adobe:ns:meta/" >
3    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4      <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
5        xmlns:xmp="http://ns.adobe.com/xap/1.0/" rdf:about="">
6      <dc:title>
7        <rdf:Alt>
8          <rdf:li xml:lang="x-default">Document title</rdf:li>
9        </rdf:Alt>
10     </dc:title>
11     <xmp:CreatorTool>Adobe Illustrator CS3</xmp:CreatorTool>
12     <xmp:CreateDate>2015-05-21T13:22:16+02:00</xmp:CreateDate>
13     <xmp:ModifyDate>2015-06-12T18:02:02+02:00</xmp:ModifyDate>
14   </rdf:Description>
15 </rdf:RDF>
16 </x:xmpmeta>
17 <?xpacket end="w"?>
```

Context - Issues with XMP and normal validation strategies

- XMP has an ambiguous data model (same content, multiple representations):

```
1 <xmpMM:History>
2   <rdf:Seq>
3     <rdf:li rdf:parseType="Resource">
4       <stEvt:action>converted</stEvt:action>
5       <stEvt:instanceID>uuid:f6227579...</stEvt:instanceID>
6       <stEvt:parameters>converted to PDF/A-1b</stEvt:parameters>
7       <stEvt:softwareAgent>Preflight</stEvt:softwareAgent>
8       <stEvt:when>2022-06-20T10:56:38+02:00</stEvt:when>
9     </rdf:li>
10  </rdf:Seq>
11 </xmpMM:History>
```

```
1 <xmpMM:History>
2   <rdf:Seq>
3     <rdf:li>
4       <rdf:Description>
5         <stEvt:action>converted</stEvt:action>
6         <stEvt:instanceID>uuid:f6227579...</stEvt:instanceID>
7         <stEvt:parameters>converted to PDF/A-1b</stEvt:parameters>
8         <stEvt:softwareAgent>Preflight</stEvt:softwareAgent>
9         <stEvt:when>2022-06-20T10:56:38+02:00</stEvt:when>
10      </rdf:Description>
11    </rdf:li>
12  </rdf:Seq>
13 </xmpMM:History>
```

- Regular schema based validation is not normally possible
- Adobe [XMP Toolkit SDK](#) comes to rescue and is available for C++ (and Java, but it's not open source anymore)
- Validation with XMP data model parsing libraries is possible ([veraPDF](#) does it) but arguably less formal than a schema based validation
- Actually I didn't know about XMP Toolkit SDK until 2025 :-)

Context - What is PoDoFo?

- A **modern C++17** PDF manipulation library
- Multiplatform (Win/Lin/Mac, also working in Android/iOS)
- Open Source, **LGPL2+** (planned: **MPL2**)
- Advanced font/text cleanup and sanitization capabilities, powered by *Adobe Font Development Kit for OpenType* ([AFDKO](#))
- Originally written by German developer Dominik Seichter
- Currently developed/maintained by the author of this presentation
- If you are interested, please reach out at:



<https://github.com/podof/podof>

Context - Original problem and other use cases

- **Use case:** PDF/A *reduction*, aka “*conversion*” (with PoDoFo). Existing metadata packets must be sanitized from invalid properties
- Rigorous XMP packet validation (can we do a better job than XMP data model parsing strategy?)
- XMP extensions pluggable validation: `<pdfaExtension>` in PDF/A-1 to 3, `AFRelationship` in PDF/A-4
- Platform/language independent validation: RELAX NG validators are available in multiple platforms (at least Java, C, C#/.NET)

```
1 <?xpacket begin="⌘" id="W5M0MpCehiHzreSzNTczkc9d"?>
2 <x:xmpmeta xmlns:x="adobe:ns:meta/">
3   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4     <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
5       xmlns:xmp="http://ns.adobe.com/xap/1.0/"
6       xmlns:pdf="http://ns.adobe.com/pdf/1.3/" rdf:about="">
7     <dc:format>application/pdf</dc:format>
8     <xmp:CreatorTool>Apache FOP</xmp:CreatorTool>
9     <xmp:CreateDate>2019-03-23T16:51:05Z</xmp:CreateDate>
10    <xmp:ModifyDate>2022-05-20T17:19:11+02:00</xmp:ModifyDate>
11    <xmp:MetadataDate>2019-03-23T16:51:05Z</xmp:MetadataDate>
12    <pdf:Trapped>True</pdf:Trapped>
13    <CustomProp>Custom property value</CustomProp>
14  </rdf:Description>
15 </rdf:RDF>
16 </x:xmpmeta>
17 <?xpacket end="r"?>
```

Not available
until PDF2.0
(PDF/A-4)

Requires
`<pdfaExtension>`
in PDF/A-1 to 3

Normative framework - The requirements

- [ISO 16684-1:2019](#):

Graphic technology - **Extensible metadata platform (XMP)**, Part 1: Data model, serialization and core properties

- [XMP Specification Part 2](#) [*various revisions*]: **Additional Properties**

- ISO 19005 (aka [PDF/A](#)):

Document management - Electronic document file format for long-term preservation

- [ISO 19005-1:2005](#) Part 1: **Use of PDF 1.4** (PDF/A-1)

- [ISO 19005-2:2011](#) Part 2: **Use of ISO 32000-1** (PDF/A-2)

- [ISO 19005-3:2012](#) Part 3: **Use of ISO 32000-1 with support for embedded files** (PDF/A-3)

- [ISO 19005-4:2020](#) Part 4: **Use of ISO 32000-2** (PDF/A-4)

Normative framework - Standards to the rescue

- [ISO 16684-2:2014](#):
Graphic technology - Extensible
metadata platform (XMP), Part 2:
**Description of XMP schemas
using RELAX NG**

- [ISO/IEC 19757-2:2008](#):
Information technology -
Document Schema Definition
Language (DSDL)
Part 2: **Regular-grammar-based
validation - RELAX NG**

- RELAX NG: simple, intuitive and
precise validation schema syntax
for XML content

The logo for RELAX NG, featuring the text "RELAX NG" in a white, serif, all-caps font, centered within a solid teal rectangular background.

- ISO 16884-2 sketches the solution
for XMP validation but “leaves it
as a trivial exercise to the reader”

Solution - Fully implement ISO 16884-2

1. Normalize the XMP packets with a canonization algorithm (chapter 5 *"Canonical serialization of XMP"* of ISO 16684-2:2014)
2. Write a full schema for the PDF XMP packets. PDF/A compliances are more accurate with regard to XMP specification revision:
<https://github.com/ceztko/XMP-RNG-Schema>
3. [**C/C++**] Implement a better streaming validation in libXML2:
https://gitlab.gnome.org/GNOME/libxml2/-/merge_requests/334
[**Merged!**] *Added xmlRelaxNGValidCtxtClearErrors to recover from errors during streaming validation*

Solution - XMP canonization algorithm 1

- Implemented in [PoDoFo](#) (C++)
- `<rdf:RDF>` outermost XML element, which shall contain a single `<rdf:Description>` element to contain all XMP properties
- Property (and structure fields) attribute notation shall not be used:

 `<rdf:Description rdf:about="" xmp:Rating="3" />`



```
1 <rdf:Description rdf:about="">
2   <xmp:Rating>3</xmp:Rating>
3 </rdf:Description>
```

Solution - XMP canonization algorithm 2

- The `rdf:parseType="resource"` attribute notation shall not be used



```
1 <xmpMM:History>
2   <rdf:Seq>
3     <rdf:li rdf:parseType="Resource">
4       <stEvt:action>converted</stEvt:action>
5       <stEvt:instanceID>uuid:f6227579...</stEvt:instanceID>
6       <stEvt:parameters>converted to PDF/A-1b</stEvt:parameters>
7       <stEvt:softwareAgent>Preflight</stEvt:softwareAgent>
8       <stEvt:when>2022-06-20T10:56:38+02:00</stEvt:when>
9     </rdf:li>
10  </rdf:Seq>
11 </xmpMM:History>
```



```
1 <xmpMM:History>
2   <rdf:Seq>
3     <rdf:li>
4       <rdf:Description>
5         <stEvt:action>converted</stEvt:action>
6         <stEvt:instanceID>uuid:f6227579...</stEvt:instanceID>
7         <stEvt:parameters>converted to PDF/A-1b</stEvt:parameters>
8         <stEvt:softwareAgent>Preflight</stEvt:softwareAgent>
9         <stEvt:when>2022-06-20T10:56:38+02:00</stEvt:when>
10      </rdf:Description>
11    </rdf:li>
12  </rdf:Seq>
13 </xmpMM:History>
```

- The canonical serialization of XMP shall use a nested `rdf:Bag`, `rdf:Seq`, or `rdf:Alt` element for array values [unclear]:
maybe malformed  `<dc:date>2019-03-23T16:51:05Z</dc:date> ?`

Solution - XMP canonization algorithm 3

- All general qualifiers shall be serialized as XML elements within an `rdf:Description` element [...]. The value of that XMP entity shall be within an `rdf:value` element within that `rdf:Description` element.



```
1 <xmp:Identifier>
2   <rdf:Bag>
3     <rdf:li xmpidq:Scheme="ISBN">0-13-110941-3</rdf:li>
4   </rdf:Bag>
5 </xmp:Identifier>
```



```
1 <xmp:Identifier>
2   <rdf:Bag>
3     <rdf:li>
4       <rdf:Description>
5         <rdf:value>0-13-110941-3</rdf:value>
6         <xmpidq:Scheme>ISBN</xmpidq:Scheme>
7       </rdf:Description>
8     </rdf:li>
9   </rdf:Bag>
10 </xmp:Identifier>
```

Solution - PDF XMP packets schema design

- The schema must be different depending on XMP specification revision (or PDF/A compliance)
- XSLT compatible expressions are introduced to allow for easy templatzation of the schema
- Some edge cases present while reproducing XMP complexity with RELAX NG schemas (eg. User defined structures, *open choice* of lists, XPath validation)

```
1 <rng:element name="rdf:RDF">
2   <rng:element name="rdf:Description">
3     <rng:attribute name="rdf:about"/>
4     <rng:interleave>
5       <rng:ref name="XMP_Properties-dc"/>
6       <rng:ref name="XMP_Properties-xmp"/>
7       <rng:ref name="XMP_Properties-xmpMM"/>
8       <!-- ... -->
9       <rng:ref name="XMP_Properties-pdf"/>
10      <rng:ref name="XMP_Properties-photoshop"/>
11      <rng:ref name="XMP_Properties-tiff"/>
12      <rng:ref name="XMP_Properties-crs"
13        condition="$IsPDFA2OrGreater"/>
14      <rng:ref name="XMP_Properties-exif"/>
15      <rng:ref name="XMP_Properties-exif_aux"
16        condition="$IsPDFA2 or $IsPDFA3"/>
17      <rng:ref name="XMP_Properties-exifEX"
18        condition="$IsPDFA4OrGreater"/>
19      <!-- ... -->
20      <rng:ref name="XMP_Properties-pdfaid"/>
21      <rng:ref name="XMP_Properties-pdfExtension"
22        condition="$IsPDFA1 or $IsPDFA2 or $IsPDFA3"/>
23    </rng:interleave>
24  </rng:element>
25 </rng:element>
```

Solution - API Integration (C++, PoDoFo)

```
1 auto packet = PdfXMPPacket::Create(buffer);
2 packet->PruneAndValidate(PdfALevel::L2B, [](const PdfXMPPProperty& prop) {
3     cout << "Invalid property \"" << prop.GetName() << "\"\" << endl;
4 });
5 cout << packet->ToString() << endl;
```

```
1 <?xpacket begin="✪" id="W5M0MpCehiHzreSzNTczkc9d"?>
2 <x:xmpmeta xmlns:x="adobe:ns:meta/">
3   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4     <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
5       xmlns:xmp="http://ns.adobe.com/xap/1.0/"
6       xmlns:pdf="http://ns.adobe.com/pdf/1.3/" rdf:about="">
7     <dc:format>application/pdf</dc:format>
8     <xmp:CreatorTool>Apache FOP</xmp:CreatorTool>
9     <xmp:CreateDate>2019-03-23T16:51:05Z</xmp:CreateDate>
10    <xmp:ModifyDate>2022-05-20T17:19:11+02:00</xmp:ModifyDate>
11    <xmp:MetadataDate>2019-03-23T16:51:05Z</xmp:MetadataDate>
12    <pdf:Trapped>True</pdf:Trapped>
13    <CustomProp>Custom property value</CustomProp>
14  </rdf:Description>
15 </rdf:RDF>
16 </x:xmpmeta>
17 <?xpacket end="r"?>
```



```
1 <?xpacket begin="✪" id="W5M0MpCehiHzreSzNTczkc9d"?>
2 <x:xmpmeta xmlns:x="adobe:ns:meta/">
3   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4     <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
5       xmlns:xmp="http://ns.adobe.com/xap/1.0/"
6       xmlns:pdf="http://ns.adobe.com/pdf/1.3/" rdf:about="">
7     <dc:format>application/pdf</dc:format>
8     <xmp:CreatorTool>Apache FOP</xmp:CreatorTool>
9     <xmp:CreateDate>2019-03-23T16:51:05Z</xmp:CreateDate>
10    <xmp:ModifyDate>2022-05-20T17:19:11+02:00</xmp:ModifyDate>
11    <xmp:MetadataDate>2019-03-23T16:51:05Z</xmp:MetadataDate>
12  </rdf:Description>
13 </rdf:RDF>
14 </x:xmpmeta>
15 <?xpacket end="w"?>
```

Bonus - Validate packet with both PDF/A and PDF/UA compliance

- PDF/UA, PDF/VT, PDF/X properties require extensions in PDF/A-1 to 3
- PoDoFo automatically add the extensions

```
1 <?xpacket begin="⌘" id="W5M0MpCehiHzreSzNTczkc9d"?>
2 <x:xmpmeta xmlns:x="adobe:ns:meta/">
3   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4     <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/"
5       xmlns:xmp="http://ns.adobe.com/xap/1.0/"
6       xmlns:pdf="http://ns.adobe.com/pdf/1.3/"
7       xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/"
8       xmlns:pdfuaid="http://www.aiim.org/pdfua/ns/id/" rdf:about="">
9     <dc:format>application/pdf</dc:format>
10    <xmp:CreatorTool>Apache FOP</xmp:CreatorTool>
11    <xmp:CreateDate>2019-03-23T16:51:05Z</xmp:CreateDate>
12    <xmp:ModifyDate>2022-05-20T17:19:11+02:00</xmp:ModifyDate>
13    <xmp:MetadataDate>2019-03-23T16:51:05Z</xmp:MetadataDate>
14    <pdfaid:part>2</pdfaid:part>
15    <pdfaid:conformance>B</pdfaid:conformance>
16    <pdfuaid:part>1</pdfuaid:part>
17  </rdf:Description>
18 </rdf:RDF>
19 </x:xmpmeta>
20 <?xpacket end="r"?>
```



```
1 <?xpacket begin="⌘" id="W5M0MpCehiHzreSzNTczkc9d"?>
2 <x:xmpmeta xmlns:x="adobe:ns:meta/">
3   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4     <rdf:Description xmlns:dc="http://purl.org/dc/elements/1.1/" ...
5       xmlns:pdfaExtension="http://www.aiim.org/pdfa/ns/extension/"
6       xmlns:pdfaSchema="http://www.aiim.org/pdfa/ns/schema#" ... >
7     <!-- ... -->
8     <pdfaid:part>2</pdfaid:part>
9     <pdfaid:conformance>B</pdfaid:conformance>
10    <pdfuaid:part>1</pdfuaid:part>
11    <pdfaExtension:schemas>
12      <rdf:Bag>
13        <rdf:li>
14          <rdf:Description>
15            <pdfaSchema:namespaceURI>http://www.aiim.org/pdfua/ns/id/
16          </pdfaSchema:namespaceURI>
17          <pdfaSchema:prefix>pdfuaid</pdfaSchema:prefix>
18          <pdfaSchema:schema>PDF/UA ID Schema</pdfaSchema:schema>
19          <!-- ... -->
20        </rdf:Description>
21      </rdf:li>
22    </rdf:Bag>
23  </pdfaExtension:schemas>
24 </rdf:Description>
25 </rdf:RDF>
26 </x:xmpmeta>
27 <?xpacket end="r"?>
```

Next Steps

- Support extensions: convert from `<pdfaExtension:schemas>` to RELAX NG schemas (PDF/A-1 to 3)
- Validate **AFRelationship** (PDF/A-4)
- Handle edge cases: user defined structures (eg. `xmpMM:Pantry`), *open choice* of lists (eg. `tiff:YCbCrSubSampling`), XPath validation
- Feedback welcome in the XMP RNG Schema (and PoDoFo) project:
<https://github.com/ceztko/XMP-RNG-Schema>

Conclusions

- **Future-proof:** XMP validation with RELAX NG is now a practical alternative
- **Open and free:** Schemas released under a liberal license
- **Technical value:** PoDoFo delivers innovation to the PDF technical world
- **Cross community value:** Enhancements related to RELAX NG parsing integrated into libXML2

This presentation was
sponsored by



The technology
vendor for Trust
Solutions

PDF Days
Europe
2025

A black and white photograph of a person's hands signing a document on a tablet. The person is holding a pen and writing on the screen. The image is partially obscured by a large orange triangle on the left side.

euronovategroup.com

Questions (to and from the audience)

- Why there was no public RNG grammar for the PDF XMP packets until now?
- Answering to anything (I try), also on PoDoFo...

Surf now to this
▼ presentation ▼

