Metadata

Table of contents

1. Overview

Document metadata is an important tool for categorizing and finding documents. Various formats support different kinds of metadata representation and to different levels. One of the more popular and flexible means of representing document or object metadata is XMP (eXtensible Metadata Platform, specified by Adobe). PDF 1.4 introduced the use of XMP. The XMP specification lists recommendation for embedding XMP metdata in other document and image formats. Given its flexibility it makes sense to make use this approach in the XSL-FO context. Unfortunately, unlike SVG which also refers to XMP, XSL-FO doesn't recommend a preferred way of specifying document and object metadata. Therefore, there's no portable way to represent metadata in XSL-FO documents. Each implementation does it differently.

2. Embedding XMP in an XSL-FO document

As noted above, there's no officially recommended way to embed metadata in XSL-FO. Apache FOP supports embedding XMP in XSL-FO. Currently, only support for document-level metadata is implemented. Object-level metadata will be implemented when there's interest.

Document-level metadata can be specified in the `fo:declarations` element. XMP specification recommends to use `x:xmpmeta`, `rdf:RDF`, and `rdf:Description` elements as shown in example below. Both `x:xmpmeta` and `rdf:RDF` elements are recognized as the top-level element introducing an XMP fragment (as per the XMP specification).

2.1. Example

```
[..]
</fo:layout-master-set>
<fo:declarations>
  <x:xmpmeta xmlns:x="adobe:ns: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/">
        <!-- Dublin Core properties go here -->
        <dc:title>Document title</dc:title>
        <dc:creator>Document author</dc:creator>
        <dc:description>Document subject</dc:description>
      </rdf:Description>
      <rdf:Description rdf:about=""
          xmlns:xmp="http://ns.adobe.com/xap/1.0/">
        <!-- XMP properties go here -->
        <xmp:CreatorTool>Tool used to make the PDF</xmp:CreatorTool>
      </rdf:Description>
    </rdf:RDF>
  </x:xmpmeta>
</fo:declarations>
<fo:page-sequence ...
[..]
```

Note:

`fo:declarations` **must** be declared after `fo:layout-master-set` and before the first `page-sequence`.

3. Implementation in Apache FOP

Currently, XMP support is only available for PDF output.

Originally, you could set some metadata information through FOP's FOUserAgent by using its set*() methods (like setTitle(String) or setAuthor(String). These values are directly used to set value in the PDF Info object. Since PDF 1.4, adding metadata as an XMP document to a PDF is possible. That means that there are now two mechanisms in PDF that hold metadata.

Apache FOP now synchronizes the Info and the Metadata object in PDF, i.e. when you set the title and the author through the FOUserAgent, the two values will end up in the (old) Info object and in the new Metadata object as XMP content. If instead of FOUserAgent, you embed XMP metadata in the XSL-FO document (as shown above), the XMP metadata will be used as-is in the PDF Metadata object and some values from the XMP metadata will be copied to the Info object to maintain backwards-compatibility for PDF readers that don't support XMP metadata.

The mapping between the Info and the Metadata object used by Apache FOP comes from the PDF/A-1 specification. For convenience, here's the mapping table:

Document information dictionary		XMP		
Entry	PDF type	Property	XMP type	Category
Title	text string	dc:title	Text	External
Author	text string	dc:creator	seq Text	External
Subject	text string	dc:description["x-de	Text	External
Keywords	text string	pdf:Keywords	Text	External
Creator	text string	xmp:CreatorTool	Text	External
Producer	text string	pdf:Producer	Text	Internal
CreationDate	date	xmp:CreationDate	Date	Internal
ModDate	date	xmp:ModifyDate	Date	Internal

Note:
"Internal" in the Category column means that the user should not set this value. It is set by the application.

Note:
The "Subject" used to be mapped to `dc:subject` in the initial publication of PDF/A-1 (ISO 19005-1). In the Technical Corrigendum 1 this was changed to map to `dc:description["x-default"]`.

3.1. Namespaces

Metadata is made of property sets where each property set uses a different namespace URI.

The following is a listing of namespaces that Apache FOP recognizes and acts upon, mostly to synchronize the XMP metadata with the PDF Info dictionary:

Set/Schema	Namespace Prefix	Namespace URI
Dublin Core	dc	http://purl.org/dc/elements/1.1/
XMP Basic	xmp	http://ns.adobe.com/xap/1.0/
Adobe PDF Schema	pdf	http://ns.adobe.com/pdf/1.3/

Please refer to the XMP Specification for information on other metadata namespaces.

Property sets (Namespaces) not listed here are simply passed through to the final document (if supported). That is useful if you want to specify a custom metadata schema.

4. Links

- Adobe's Extensible Metadata Platform (XMP) website
- Adobe XMP Specification
- Adobe XMP Specification
- http://dublincore.org/