
XMLmind XML Editor - Support of RELAX NG Schemas

Hussein Shafie, Pixware <xmleditor-support+xmlmind.com>

July 30, 2010

Abstract

This document describes how RELAX NG schemas are supported by XMLmind XML Editor.

| | |
|--|---|
| 1. Implementation of RELAX NG in XMLmind XML Editor | 1 |
| 2. Specifying which RELAX NG schema to use for validating a document | 1 |
| 2.1. The <code>relaxng</code> configuration element | 2 |
| 2.2. The <code><?xex-relaxng-schema></code> processing instruction | 2 |
| 3. XMLmind XML Editor-friendly content models | 3 |
| 3.1. The non-validating, lenient, editing mode | 5 |
| 3.2. Help provided by the "Show Content Model" window | 6 |
| 3.3. Other content models which are not XXE-friendly | 8 |
| 4. Known problems | 9 |

1. Implementation of RELAX NG in XMLmind XML Editor

The implementation of RELAX NG in XMLmind XML Editor (XXE for short) is based on Jing, an *OpenSource*, *industrial strength*, *streaming* validator written by James Clark.

The trimmed version of Jing included in XXE (`relaxng.jar`) is used

- to load and validate RELAX NG schemas associated to XML documents (XML and compact syntaxes are both supported);
- to fully validate XML documents conforming to RELAX NG schemas, each time these documents are opened and saved, and each time a full validation is explicitly requested by the user (command Tools → Check Validity).

Jing is not used to implement guided editing. That is, Jing is not used to determine the content model of the element being edited. A quick, incremental, version of the algorithm that computes the derivative of a pattern is used for that.

The implementation of W3C XML Schema Datatypes used in RELAX NG schemas (e.g. `xsd:int`) is the work of XMLmind. This implementation is very different from the implementation of W3C XML Schema Datatypes included in the original Jing. This implementation is shared by the version of Jing included in XXE and by our own W3C XML Schema validator.

XXE supports attribute default values as specified in RELAX NG DTD Compatibility. The compatibility of the schema with this feature is (*very strictly*) checked by XXE and not by Jing. This means that a schema found valid by Jing but improperly using this feature will be rejected by XXE.

2. Specifying which RELAX NG schema to use for validating a document

This section is just a primer. The reference documentation about this topic is really *XMLmind XML Editor - Configuration and Deployment*.

2.1. The `relaxng` configuration element

A document type declaration (`<!DOCTYPE>`) can be used to associate a DTD to a document. Attributes `xsi:schemaLocation`/`xsi:noNamespaceSchemaLocation` can be used to associate W3C XML Schemas to a document. But there is no standard way to associate a RELAX NG schema to a document. Therefore this association must be made using an external specification such as the Namespace Routing Language (NRL).

In the case of XMLmind XML Editor, this external specification is simply a configuration element called `relaxng`.

XHTML example:

```
<configuration name="XHTML Strict [RELAX NG]"
  xmlns:html="http://www.w3.org/1999/xhtml"
  xmlns="http://www.xmlmind.com/xmleditor/schema/configuration">
  <include location="xxe-config:schema/ns_xhtml.incl" />

  <detect>
    <rootElementNamespace>http://www.w3.org/1999/xhtml</rootElementNamespace>
  </detect>

  <relaxng location="xxe-config:common/rng/xhtml11/xhtml-strict.rng" />❶

  <preserveSpace elements="html:pre html:style html:script" />❷

  <css name="XHTML" location="xhtml_rng.css" />
  <template name="Page" location="page.html" />
</configuration>
```

- ❶ The `relaxng` configuration element specifies the location of the RELAX NG schema (XML syntax or compact syntax) to which conforms the document being opened.
- ❷ Unlike the DTD, `xhtml-strict.rng` does not specify a `preserve` default value for attribute `xml:space` of elements such as `pre`. Therefore, the `preserveSpace` configuration element must be used to specify whitespace-preserving elements.

2.2. The `<?xxe-relaxng-schema>` processing instruction

The following processing instruction is a non standard, proprietary, way to associate a document to a RELAX NG schema. *Its use should be restricted to testing and other quick and dirty experiments.*

DocBook example:

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xxe-relaxng-schema name="-//OASIS//RELAX NG DocBook V4.3//EN"❶
  location="http://www.docbook.org/rng/4.3/docbook.rnc"
  compactSyntax="true" encoding="US-ASCII" ?>

<!DOCTYPE article [❷

  <!ENTITY % sgml.features "IGNORE">
  <!ENTITY % xml.features "INCLUDE">

  <!ENTITY euro "&#x20AC;">

  <!ENTITY % dbcent PUBLIC
    "-//OASIS//ENTITIES DocBook Character Entities V4.3//EN"
    "http://www.oasis-open.org/docbook/xml/4.3/dbcentx.mod">
    %dbcent;

]>

<article>
  <title></title>
  <section>
    <title></title>
    <para></para>
```

```
</section>  
</article>
```

- ❶ The `<?xxe-relaxng-schema>` processing instruction has pseudo-attributes identical to the attributes of the `relaxng` configuration element.
- ❷ Using a RELAX NG schema should not prevent you from specifying a document type declaration for character entities.

3. XMLmind XML Editor-friendly content models

Validating a document against a RELAX NG schema is similar to matching some text against a regular expression. If the document ``matches" the schema, the document is valid, and this, no matter which sub-expressions were used during the match.

Example: string "b" matches regular expression "(a?,b) | (b,c?)" and we don't care if it matches sub-expression "(a?,b)" or sub-expression "(b,c?)". The situation is exactly the same with RELAX NG schemas, simply replace the characters and the character classes used in a regular expression by RELAX NG patterns.

The job of a RELAX NG schema is to validate a document as a whole, and that's it. For XXE, the problem to solve is different. One of the main jobs of XXE is to guide the user when he/she edits an XML document. That is, one of the main jobs of XXE is to identify the content model of the element which is being edited, in order to suggest the right attributes and the right child elements for it.

To do that, XXE needs to know precisely which ``sub-expressions were used during the match". Unfortunately, sometimes, this is impossible to do.

All examples used in this section are found in `XXE_install_dir/doc/rngsupport/samples/`. Note that they are all valid schemas and valid documents.

Example 1. Ambiguous elements

RELAX NG schema, target.rnc:

```
start = build-element

build-element = element build {
    target-element*
}
target-element = element target {
    attribute name { xsd:ID },
    element list { ref-element* }?,
    element list { action-element* }?
}
ref-element = element ref {
    attribute name { xsd:IDREF }
}
action-element = element action { text }
```

Document conforming to the above schema, target_bad.xml:

```
<build>
  <target name="all">
    <list>
    </list>
  </target>

  <target name="compile"/>
  <target name="link"/>
</build>
```

If you open target_bad.xml in XXE and select the list element, *XXE is lost*: is it the list element which contains refs or is the list element which contains actions? Both list content models are fine in the case of an empty list element!

Now, if you open target_good.xml in XXE, there is no problem at all:

```
<build>
  <target name="all">
    <list>
      <ref name="compile"/>
      <ref name="link"/>
    </list>
    <list>
      <action>cc -c *.c</action>
      <action>cc *.o</action>
    </list>
  </target>

  <target name="compile"/>
  <target name="link"/>
</build>
```

However, in the vast majority of realistic cases, XXE knows how to make a difference between two child elements having the same name and having different content models.

RELAX NG schema, sect.rnc:

```
start = doc-element

doc-element = element doc {
    (simple-sect|
    recursive-sect)+
}

simple-sect = element sect {
    attribute class { "simple" }, paragraph-element*
}
```

```
recursive-sect = element sect {  
    attribute class { "recursive" }?, (recursive-sect|simple-sect)*  
}  
  
paragraph-element = element paragraph { text }
```

Document conforming to the above schema, sect.xml:

```
<doc>  
  <sect>  
    <sect></sect>  
  
    <sect class="simple">  
      <paragraph>Paragraph 2.</paragraph>  
    </sect>  
  </sect>  
  
  <sect class="simple"></sect>  
</doc>
```

In the above example, XXE has no problem at all making a difference between the empty `<sect>` element and the empty `<sect class="simple">` element. The reason is obviously because the first kind of `sect` element has a required attribute `class` with `simple` as its fixed value.

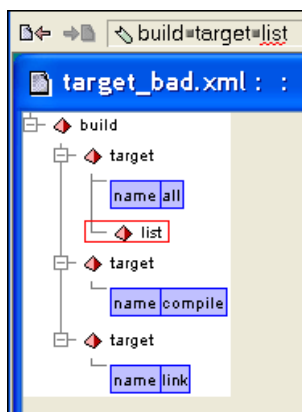
3.1. The non-validating, lenient, editing mode

When XXE is "lost", it automatically enters a *lenient editing mode*. In this mode, XXE can no longer guide the user when he/she edits the element which poses problems.

The node path bar is used to signal elements which are in this non-validating, lenient editing mode:

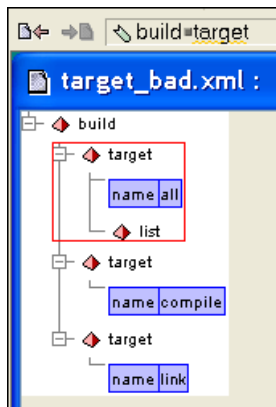
- An element underlined in *red* means that this element is in *non-validating mode 2*. In this mode, XMLmind XML Editor is not able to suggest the right attributes and the right child elements to the user. The user may add and remove any attributes and child elements he/she wants, at any place and in any number.

Figure 1. XXE is completely lost when empty `list` is selected



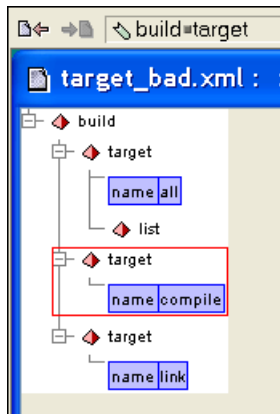
- An element underlined in *orange* means that this element is in *non-validating mode 1*. In this mode, XMLmind XML Editor still suggests the right attributes and child elements to the user. But these are only suggestions: the user may add and remove any attributes and child elements he/she wants, and this, at any place and in any number.

Figure 2. XXE has problems when `target` containing empty `list` is selected



Note that the lenient editing mode is local to an element and its descendants. It is not used for the whole document, but just for the element for which XXE has troubles.

Figure 3. XXE has no problem at all when `target` named `compile` is selected



Also note that, after modifying an element which poses problems to XXE, if these problems are solved, XXE will automatically switch to its normal, strict, validating mode.

3.2. Help provided by the "Show Content Model" window

The node path bar is not the only tool in XXE which can help the user recognize attributes and elements which pose problems to the XML editor. The window opened by command Help → Show Content model also displays very useful information.

Figure 4. `sect.xml` example when first `sect` element is selected

Element sect

Content model

This element can only contain child elements.

```
(@class? ,  
(sect | sect#2)*)
```

Using these attributes or child elements may cause XMLmind XML Editor to switch a lenient editing mode.

Using these attributes or child elements will cause XMLmind XML Editor to switch a lenient editing mode.

Attributes

| Name | Data type | Value |
|-------|-----------------------------------|-------------------|
| class | token enumeration: "recursive" | FIXED "recursive" |

Parents

The following elements contain sect: [doc](#), [sect](#).

Children

The following elements occur in sect: [sect](#), [sect#2](#).

3.3. Other content models which are not XXE-friendly

Example 2. Not specific to RELAX NG

RELAX NG schema, name.rnc:

```
start = names-element
names-element = element names {
    name-element+
}
name-element = element name {
    element fullName { text } |
    (element firstName { text } & element lastName { text })
}
```

Document conforming to the above schema, name.xml:

```
<names>
  <name><fullName>John Smith</fullName></name>

  <name><firstName>John</firstName><lastName>Smith</lastName></name>

  <name><lastName>Smith</lastName><firstName>John</firstName></name>
</names>
```

XXE allows to replace the `firstName`, `lastName` pair by a `fullName`. Simply select both child elements and use command Edit → Replace. But it is impossible to replace a `fullName` by a `firstName`, `lastName` pair.

The only way to do this is to select the `fullName` to be replaced and then, to use command Edit → Force Deletion. This will force XXE to enter the lenient editing mode. Remember that in this mode, the user is allowed to add any child elements he/she wants, including a `firstName`, `lastName` pair¹.

Note that the above example is not specific to RELAX NG. It is possible to model this kind of content with a DTD or a W3C XML Schema.

The example below is very similar but can only be expressed using a RELAX NG schema. This is the case, because, unlike a DTD and a W3C XML Schema, a RELAX NG schema can be used to specify the places within an element where text nodes may occur.

¹The right approach here is to define two *named element templates* for element `name`, one containing a `fullName` child element and the other containing a `firstName`, `lastName` pair.

Example 3. Specific to RELAX NG

RELAX NG schema, name2.rnc:

```
start = names-element
names-element = element names {
    name-element+
}
name-element = element name {
    text |
    (element firstName { text } & element lastName { text })
}
```

Document conforming to the above schema, name2.xml:

```
<names>
  <name>John Smith</name>

  <name><firstName>John</firstName><lastName>Smith</lastName></name>

  <name><lastName>Smith</lastName><firstName>John</firstName></name>
</names>
```

The situation is worse with the name2.rnc example than with the name.rnc example. It is always allowed to delete a text node and this includes the text node containing "John Smith". That is, there is no way to force XXE to enter its lenient mode in order to be able to replace text node "John Smith" by a `firstName`, `lastName` pair.

In such case, using named element templates is the only way to cope with such content models. Simply specify two named element templates for element `name`, one containing a text node with a placeholder string and the other containing a `firstName`, `lastName` pair.

4. Known problems

Known problems:

- `include` and `externalRef` elements in RELAX NG schemas are XML catalog aware, but the equivalent notations in the compact syntax are not XML catalog aware.