

INERA
INC



eXtyle[®]

Schematron: an introduction

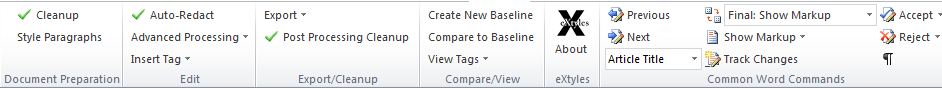
XUG 2012



What is Schematron?

- a rule-based language that reports on an XML document
 - highly customizable
 - easy to use
 - can report on well-formed or valid XML
 - uses XPath syntax (we'll get to that in a minute)
 - can run at multiple stages of a workflow
-
- **Schematron is NOT transformative:**
it reports on the structure or content of an XML document without modifying it in any way.

Report vs. Validation



Publication: Bulletin of the World Health Organization; Type: Research
Article ID: 05-023150

Effects of insecticide-treated bed net protection during early infancy in an African area of intense malaria transmission: randomized controlled trial

Document Information

* Publisher: WHO Volume: 88 Issue: 2

* Publication: BLT: Bulletin of the World Health Organization Day:

* Type: Research Month:

Second Article Type: Year:

Related Article Type: First Page: Last Page:

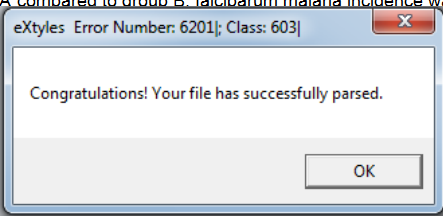
* Article ID: 05-023150 Language: en

* Required Entries Add Document Information to Header

OK Cancel

Methods A total of 3,387 newborns from 41 villages in rural Burkina Faso were individually randomized to ITN protection from birth onwards (group A) versus from month six onwards (group B). Primary outcomes were all-cause mortality in all study children and falciparum malaria incidence in a representative sub-sample of the study population.

Findings After a mean follow-up period of 27 months, there were 129 deaths in group A and 128 deaths in group B (RR 1.0; 95% CI: 0.78–1.27). In group A compared to group B, falciparum malaria incidence was significantly lower (RR 1.3, 95% CI: 1.1–1.6).



```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE article SYSTEM "journalpublishing3.dtd">
<article article-type="research-article" dtd-version="3.0" xml:lang="en">
  <front>
    <journal-meta>
      <journal-id journal-id-type="publisher-id">
        BLT
      </journal-id>
      <journal-title-group>
        <journal-title>
          Bulletin of the World Health Organization
        </journal-title>
        <abbrev-journal-title abbrev-type="pubmed">
          Bull. World Health Organ.
        </abbrev-journal-title>
      </journal-title-group>
      <issn pub-type="ppub">
        0042-9686
      </issn>
      <publisher>
        <publisher-name>
          World Health Organization
        </publisher-name>
      </publisher>
    </journal-meta>
    <article-meta>
      <article-id pub-id-type="publisher-id">
        05-023150
      </article-id>
      <article-id pub-id-type="doi">
        10.2471/05-023150
      </article-id>
      <article-categories>
        <subj-group subj-group-type="heading">
          <subject>
            Research
          </subject>
        </subj-group>
      </article-categories>
    </article-meta>
  </front>
</article>
```

Schematron can tell us...

- whether or not an element or attribute is present
- how many times an element or attribute is present
- about the content of an element or attribute
- about the sequence of elements in the document

And it can tell us these things in
whatever words you want to use
(no opaque parsing errors)

How does Schematron do this?

• XPath!

- XPath is a node-walking language for XML documents
- you tell it where to look in the document (node-walking)
- once you've established where you are in the document, you can either makes assertions (this MUST be true, tell me if it's not) or ask questions by getting reports (tell me how many authors are in the contrib-group)

XML elements and attributes

```
<body>  
  <section section-type="intro">  
    <title>  
      Introduction  
    </title>  
    <para>  
      This is the text of the paragraph.  
    </para>  
  </section>  
</body>
```

XPath and XML document structure

➤ article

- body

- section

- ◆ title

- ◆ para

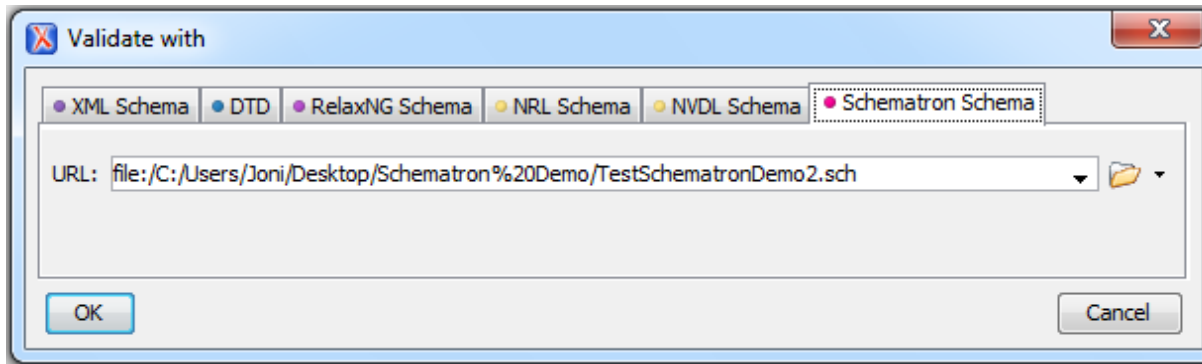
- ◆ para

- ◆ para

in XPath, if you want to find the title element, you could represent it like this:
`article/body/section/title`

Example: report if article-title is longer than 100 characters


```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <schema xmlns="http://purl.oclc.org/dsdl/schematron">
3
4   <title>Schematron demonstration rules</title>
5
6   <pattern id="title-length">
7     <rule context="title-group/article-title">
8       <assert test="string-length(normalize-space(.)) &lt;= 100">title is too long</assert>
9     </rule>
10  </pattern>
11
12 </schema>
```



I... Description - 1 item

! - E [ISO Schematron] title is too long (string-length(normalize-space(.)) <= 100) [assert]

In Oxygen, clicking on the report brings you directly to the item:

I...	Description - 1 item
 -	E [ISO Schematron] title is too long (string-length(normalize-space(.)) <= 100) [assert]

```
<article-id pub-id-type="doi">
  10.2471/05-023150
</article-id>
<article-categories>
  <subj-group subj-group-type="heading">
    <subject>
      Research
    </subject>
  </subj-group>
</article-categories>
<title-group>
  <article-title>
    Effects of insecticide-treated bed net protection during early infancy in an African area of intense malaria transmission: randomized controlled trial
  </article-title>
</title-group>
<contrib-group>
```

Schematron language

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <schema xmlns="http://purl.oclc.org/dsdl/schematron">
3
4     <title>Schematron demonstration rules</title>
5
6     <pattern id="title-length">
7         <rule context="title-group/article-title">
8             <assert test="string-length(normalize-space(.)) &lt;= 100">title is too long</assert>
9         </rule>
10    </pattern>
11
12 </schema>
```

- `<title>` is whatever you want to call your Schematron
- `<pattern>` is where you can specify the group your rules fall into (there can be any number of patterns)
- `<rule>` is where you define the context for what you're asserting or reporting (there can be any number of rules)

Schematron language, cont.

```
<title>Schematron demonstration rules</title>

<pattern id="titles">
  <rule context="sec">
    <assert test="title">The title is missing!</assert>
  </rule>

  <rule context="sec">
    <report test="count(//sec[title=current()/title]) > 1">Section has the same title as another section</report>
  </rule>
</pattern>
```

- <assert> allows you to declare that something must be true, and returns a message if it's not
- <report> asks if something is true, and returns a message if it is

Why Schematron is useful

- there are limitations to DTD validation
- Schematron can look at content
 - Whitespace in an email address
 - Character length of titles
 - Empty elements
- additional level of validation
 - Check that an ISSN has 8 characters with a hyphen in the middle
 - Ensure that DOI syntax matches business rules
- Whether you have well-formed or valid XML, you can use Schematron to report on these things

Additional information

- Schematron was developed by Rick Jelliffe
- ISO Schematron is a freely available standard:
<http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html>
- Schematron can be implemented based on XSLT 1.x using freely available extensions:
<http://code.google.com/p/schematron/downloads/list>
- Many thanks to Deborah Lapeyre for her excellent Schematron tutorial at JATS-Con