chem-bla-ics

Test File Repository and RelaxNG

Egon Willighagen

Published June 25, 2007

Citation

Willighagen, E. (2007, June 25). Test File Repository and RelaxNG. *Chem-bla-ics*. https://doi.org/10.59350/a1np2-c3x03

Keywords

Blue-obelisk, Openscience

chem-bla-ics

Last week I started the Blue Obelisk Chemical Test File Repository, a repository of OSI-approved-licenced test files (from various sources) to improve interoperability between chemoinformatics software.

Following a discussion on the mailing list earlier, a directory hierarchy has been set up, and each files contains an index.xml to describe the content. In case of a directory with actual test files, it may look like:

```
<dir name="asn/pubchem/valid" xmlns:dc="http://purl.org/dc/elements/1.1/">

  <chemfiles>

    <file name="cid1.asn" valid="yes">
       <dc:format>chemical/x-asn-pubchem</dc:format>
       <dc:source>PubChem</dc:source>
       <dc:creator>Unknown</dc:creator>
       <dc:rights>PublicDomain</dc:rights>
       <test by="CDK"/>
    </file>

  </chemfiles>

</dir>
```

As is clear, Dublin Core is reused for much of the meta data.

To improve and ensure some quality, the XML must be valid in addition to just well-formed, so that I can set up XSLT stylesheets to create XHTML indices and summaries. Therefore, I wanted to setup a schema for the index.xml files. My first thought was to use XML Schema which has XML Namespaces support and has well defined (and extensible) data types. I have hacked in it in the past my the details have slipped me. Already in 1998 I worked with DTDs, around the time that the XML specification was declared a recommendation. Originating from the SGML year, it is not XML based, had no knowledge of namespaces, and only a limited amount of data types.

Then there is RELAX NG. XML based, uses the same data types are XML Schema and has support for namespaces. Since I had to look up the specs for either DTD or XML Schema for the details anyway (e.g. on how to allow the DC namespace in the main namepsace), why not try something new. Well, I was amazed. RELAX NG has a syntax simplicity like that of DTD, but the functionality from XML Schema. So, I hacked up in 30 minutes a XML spec for the test file repository, including a (too short) list of recognized MIME types. Just a combination of some `<element>`, `<attribute>`, `<oneOrMore>`, etc elements. The results is available as schema.relaxng in SVN.