Sofia University “St. Kliment Ohridski”
Faculty of Mathematics and Informatics

MathML Statistics

Markup Languages – XML Coursework

Stanislav Peshterliev, 61096

13 January 2011

Table of contents

1. Task

Processing a MathML XML by a DOM application Consider the MathML standard. Objective: to parse a valid XML describing given MathML (and valid for the MathML DTD) with two DOM applications and to extract some part of the meta-information inside a new XML valid for a custom DTD and to another XML valid for a custom XSchema. Both new DTD and XSchema are to be proposed by the students. Students have to demonstrate the XML documents and their DTD validation, and theirs processing by the application. More, it is required a MS Word document describing a MathML overview and task realization (4-7 pages).
Resources: http://www.w3.org/Math/, http://www.w3.org/TR/REC-MathML/,
http://www.mozilla.org/projects/mathml/

2. MathML document

The MathML document that I have chosen to use describes the following rational integral. The reasons to chose that integral are because I know it very well from Calculus classes and it is complex enough to be processed in interesting ways. The document can be found in the project folder src/resources/RationalIntegral.xml.

$$I_n = \int \frac{\mathrm{d}x}{\left(a^2+x^2\right)^{\mathrm{n}}} = \frac{1}{a^2}\int \frac{a^2}{\left(a^2+x^2\right)^{\mathrm{n}}}\mathrm{d}\,x = \frac{1}{a^2}\int \frac{a^2+x^2-x^2}{\left(a^2+x^2\right)^{\mathrm{n}}}\mathrm{d}\,x =$$

$$\frac{1}{a^2}\int \frac{\mathrm{d}x}{\left(a^2+x^2\right)^{\mathrm{n-1}}} - \frac{1}{a^2}\int \frac{x^2}{\left(a^2+x^2\right)^{\mathrm{n}}}\mathrm{d}\,x = \frac{1}{a^2} I_{\mathrm{n-1}} - \frac{1}{2\,a^2}\int x\left(a^2+x^2\right)^{\mathrm{n}}\mathrm{d}\,x^2 =$$

$$\frac{1}{a^2} I_{\mathrm{n-1}} - \frac{1}{2\,a^2}\int x\left(a^2+x^2\right)^{\mathrm{n}}\mathrm{d}\left(a^2+x^2\right) = \frac{1}{a^2} I_{\mathrm{n-1}} - \frac{1}{2\left(\mathrm{-n}+1\right)a^2}\int x\mathrm{d}\left(a^2+x^2\right)^{\mathrm{-n}+1} =$$

$$\frac{1}{a^2} I_{\mathrm{n-1}} - \frac{1}{2\left(\mathrm{-n}+1\right)a^2}\left(\frac{x}{\left(a^2+x^2\right)^{\mathrm{n-1}}} - \int \frac{\mathrm{d}x}{\left(a^2+x^2\right)^{\mathrm{n-1}}}\right) = \frac{1}{a^2}\left(1-\frac{1}{2n-2}\right)I_{\mathrm{n-1}} + \frac{1}{\left(2n-2\right)a^2}\,\frac{x}{\left(a^2+x^2\right)^{\mathrm{n-1}}}$$

3. Generated XML and DTD/XScheme validation

The DOM application generates, from any valid MathML document, statistics for the number of different structures and data types that one document contains. Generated document structure is as follows

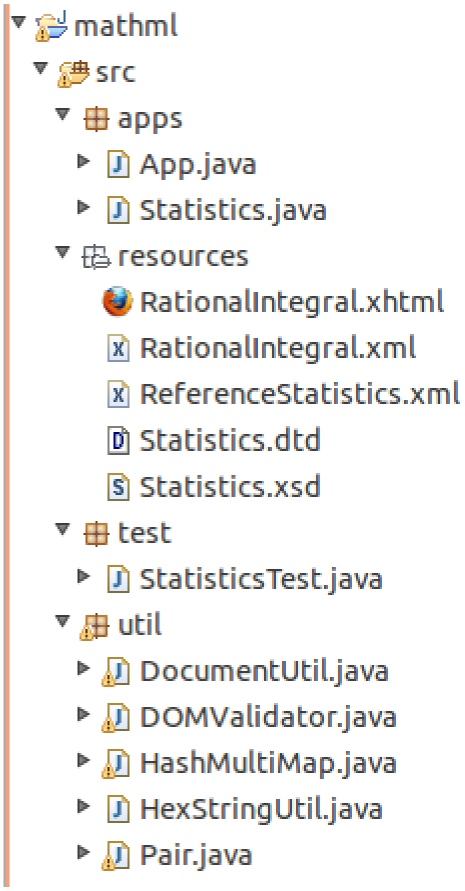
```
<?xml version="1.0" encoding="UTF-8"?>
<statistics>
        <url>path to file</url>
        <rows total="some positive number" />
        <fractions total="some positive number" />
        <subscripts total="some positive number">
                <subscript total="some positive number">
                        <first type="string">l</first>
                        <second type="string">n</second>
                </subscript>
                ....
        </subscripts>
        <supscripts total="some positive number">
                <supscript total="some positive number">
                        <first type="string">a</first>
                        <second type="number">2</second>
                </supscript>
                ....
        </supscripts>
        <identifiers total="some positive number">
                <identifier total="some positive number">n</identifier>
                ....
        </identifiers>
        <operators total="some positive number">
                <operator total="some positive number">+</operator>
                ....
        </operators>
        <numbers total="some positive number">
                <number total="some positive number">1</number>
                ....
        </numbers>
</statistics>
```

The generated document is validated using two separate validation schemas DTD and XScheme, named Statistics.dtd and Statistics.xsd located in project folder src/resources.

4. DOM application

The application that read any valid MathML document and generates statistics is implemented in Java with Eclipse as a development environment. I use only standard Java APIs for working with XML documents. The folder structure is as follows

Folder structure



```
mathml
  src
    apps
      App.java
      Statistics.java
    resources
      RationalIntegral.xhtml
      RationalIntegral.xml
      ReferenceStatistics.xml
      Statistics.dtd
      Statistics.xsd
    test
      StatisticsTest.java
    util
      DocumentUtil.java
      DOMValidator.java
      HashMultiMap.java
      HexStringUtil.java
      Pair.java
```

There are three main packages

- apps – contains main classes of the application. Statistics.java implements the algorithm for generating statistics XML document. App.java is a console application that reads all MathML files from given directory, generates statistics for them and then validates generated files using Statistics.dtd and Statistics.xsd schemas.
- resources – contains all validation schemas and reference XML documents used for testing
- test – contains testing code

util – contains utility classes

5. Abount MathML

Mathematical Markup Language (MathML) is an application of XML for describing mathematical notations and capturing both its structure and content. It aims at integrating mathematical formulae into World Wide Web pages and other documents. It is a recommendation of the W3C math working group. [7]

MathML 1 was released as a W3C recommendation in April 1998 as the first XML language to be recommended by the W3C. Version 1.01 of the format was released in July 1999 and version 2.0 appeared in February 2001. In October 2003, the second edition of MathML Version 2.0 was published as the final release by the W3C math working group. In June 2006 the W3C has rechartered the MathML Working Group to produce a MathML 3 Recommendation until February 2008 and in November 2008 extended the charter to April 2010. A sixth Working Draft of the MathML 3 revision was published in June 2009. On 10 August 2010 version 3 has now become a "Proposed Recommendation" rather than a draft.MathML 3.0 was officially released as a W3C Recommendation on 21 October 2010, as a revision of MathML 2.0. MathML 3 is backward compatible with MathML 2.MathML was originally designed before the finalization of XML namespaces. As such, MathML markup is often not namespaced, and applications that deal with MathML, such as the Mozilla browsers, do not require a namespace. For applications that wish to namespace MathML, the recommended namespace URI is http://www.w3.org/1998/Math/MathML.[7]

There are two MathML standards Presentation MathML and Content MathML. Because the meaning of the equation is preserved separate from the presentation, how the content is communicated can be left up to the user. For example, web pages with MathML embedded in them can be viewed as normal web pages with many browsers but visually impaired users can also have the same MathML read to them through the use of screen readers.[7]

- **Presentation MathML** - Presentation MathML focuses on the display of an equation, and has about 30 elements, and 50 attributes. The elements all begin with m and include token element: <mi>x</mi> - identifiers; <mo>+</mo> - operators; <mn>2</mn> - number. Tokens are combined using layout elements that include: <mrow> - a row; <msup> - superscripts; <mfrac> - fractions. The attributes mainly control fine details of the presentation. A large number of entities are available that represent letters (&pi;), symbols (&RightArrow;) and some non-visible characters such as &InvisibleTimes; representing multiplication.[7]

Quadratic equation using Presentation MathML[7]

```
<?xml version="1.0" encoding="UTF-8"?>
 <!DOCTYPE math PUBLIC "-//W3C//DTD MathML 2.0//EN"
      "http://www.w3.org/Math/DTD/mathml2/mathml2.dtd">
 <math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
   <mi>a</mi>
   <mo>&#x2062;<!-- &InvisibleTimes; --></mo>
   <msup>
    <mi>x</mi>
    <mn>2</mn>
   </msup>
   <mo>+</mo>
   <mi>b</mi>
```

```
    <mo>&#x2062;<!-- &InvisibleTimes; --></mo>
    <mi>x</mi>
    <mo>+</mo>
    <mi>c</mi>
  </mrow>
</math>
```

- **Content MathML** - Content MathML focuses on the semantic meaning of the expression. Central to Content MathML is the <apply> element that represents a function or operator, given in the first child, applied to the remaining child elements. For example <apply><sin/><ci>x</ci></apply> represents sin(x) and <apply><plus/><ci>x</ci><cn>5</cn></apply> represents x+5. The <ci> element represents an identifier, <cn> a number, and there are over a hundred different elements for different functions and operators. Content MathML uses only a few attributes.[7]

Quadratic equation using Content MathML[7]

```
<?xml version="1.0" encoding="UTF-8"?>
 <!DOCTYPE math PUBLIC "-//W3C//DTD MathML 2.0//EN"
          "http://www.w3.org/Math/DTD/mathml2/mathml2.dtd">
 <math xmlns="http://www.w3.org/1998/Math/MathML">
  <apply>
     <plus/>
     <apply>
       <times/>
       <ci>a</ci>
       <apply>
         <power/>
         <ci>x</ci>
         <cn>2</cn>
       </apply>
     </apply>
     <apply>
       <times/>
       <ci>b</ci>
       <ci>x</ci>
     </apply>
     <ci>c</ci>
   </apply>
 </math>
```

6. Tools

RationalIntegral.xml is written using Amaya – editor based on Mozilla Firefox developed by W3C.org. Recommended web browser for viewing MathML documents is Mozilla Firefox 3.5+, as it offers the best support for the MathML standard. DOM application is implemented in Java using Eclipse as development environment.

7. References

[1] http://www.w3.org/TR/xmlschema-0
[2] http://www.w3.org/TR/xmlschema-1
[3] http://www.w3.org/Math
[4] http://www.w3.org/TR/REC-MathML
[5] http://www.mozilla.org/projects/mathml
[6] http://www.w3.org/Amaya
[7] http://wikipedia.org/wiki/MathML