Semantics of mathematics on the web

Mika Seppälä and Jouko Väänänen

May 28, 2002

1 Introduction

Almost any scientific information contains mathematical formulae. For may disciplines the mathematics that is needed is rather limited, but even areas like law need to be able to speak of lengths, areas, weights and volumes. Hence they need to be able to express information involving mathematical concepts. Sometimes these concepts have to be translated into equivalent other concepts, say like US units, inches, feet, miles, pounds and gallons, into equivalent metric concepts, or vice versa. This is a task that could well be done automatically given all the computing power that we have in the internet.

But even such a modest task is not being done automatically. Manual conversions are used, and sometimes this may cause troubles if not done diligently. CNN's News item entitled "NASA's metric confusion caused Mars orbiter loss" (September 30, 1999) reported about the accident in which NASA lost a $ 125 million Mars orbiter. According to this CNN's news item,

> "the spacecraft was lost because one engineering team used metric units while another used English units for a key spacecraft operation. For that reason, information failed to transfer between the Mars Climate Orbiter spacecraft team at Lockheed Martin in Colorado and the mission navigation team in California. Lockheed Martin built the spacecraft."

For a layman this appears to be a trivial error in such a complex undertaking. Had the conversion of the units taken place automatically one could have avoided this error.

The Ariane 5 failure in 1996 was another example of a very costly mistake that could have been avoided provided that the conversion of mathematical objects had been done correctly. The destruction of Ariane 5 in June 1996 was caused by a failure to convert a 64-bit floating point number to a 16-bit signed integer. The floating point number which was converted had a value greater than what could be represented by a 16-bit signed integer. This caused an operand error in the navigation system of the rocket, and hence initiated self-destruction of the spacecraft. For details see [1].

These examples underline the need to embed mathematical formulae into documents and programs so that the correct meaning of the object can be automatically understood.

2 Languages for mathematical information

2.1 MathML

The Mathematics Meta Language, MathML, of the W3 Consortium allows the visual representation of mathematical formulae in the web. MathML can also express the content of simple formulae contained typically in the K12 (or, after the last revisions, perhaps K14) curriculum.

MathML has two components: the presentation MathML that allows one to write mathematical formulae that can be rendered correctly in the web browsers, and the content MathML that can be used to understand the meaning of simple formulae automatically.

Presentation MathML is a slight improvement to expressing mathematical information as gif pictures. The improvement is in the fact that formulae behave nicely when increasing or decreasing the font sizes, or when the window in which the formulae are being viewed, is resized. Presentation MathML also results more satisfying printed documents.

One can argue that presentation MathML is not very useful, since other solutions, like pdf files, can be used to produce a similar result: high quality screen documents that yield also high quality printed versions.

The power of MathML lies in the fact that mathematical software packages, like Maple or Mathematica, can now be used to *import* and to *export* MathML encoded formulae automatically.

2.2 OpenMath

OpenMath extends the power of content MathML so that, in principle, any scientific information can be embedded in documents so that the meaning of the formulae can be automatically understood. This has been achieved through extensible Content Dictionaries that are used to define the meaning of formulae. These CD's are publicly available at the OpenMath web site (www.openmath.org, [5]).

2.3 Examples of mathematics embedded in documents

This article has been typeset with LaTeX. Mathematical formulae are embedded in the document using the typesetting commands provided by LaTeX. These typesetting commands will produce very high quality printed documents. The typesetting commands do, however, not contain any semantic information about the formula in question. The mathematical expression $\sin(x^2+1)$ is embedded in the document as `$\sin(x^2+1)$`. These commands define he visual presentation of this mathematical expression, but a human is needed in order to correctly interpret its meaning.

The above mentioned languages OpenMath and MathML will provide means to attach semantics to formulae which semantics can be automatically understood.

MathML can be viewed as OpenMath Lite. The capabilities of MathML to define semantics is limited, but it certainly suffices in situations like the above formula.

The following is self–explanatory.

2.3.1 Presentation MathML encoding of the expression $\sin(x^2+1)$

```
<math xmlns='http://www.w3.org/1998/Math/MathML'>
<mrow>
    <mi>sin</mi>
    <mo>&ApplyFunction;</mo>
    <mfenced>
    <mrow>
        <msup>
            <mi>x</mi><mn>2</mn>
        </msup>
        <mo>+</mo>
        <mn>1</mn>
    </mrow>
    </mfenced>
</mrow>
</math>
```

2.3.2 Content MathML encoding of the expression $\sin(x^2 + 1)$

```
<math xmlns='http://www.w3.org/1998/Math/MathML'>
    <apply id='id7'><sin id='id1'/>
        <apply id='id6'><plus/>
            <apply id='id4'> <power/>
            <ci id='id2'>x</ci>
            <cn id='id3' type='integer'>2</cn>
            </apply>
        <cn id='id5' type='integer'>1</cn>
        </apply>
    </apply>
    </math>
```

2.3.3 OpenMath encoding of the expression $\sin(x^2 + 1)$

```
<OMOBJ>
    <OMA>
        <OMS cd="transc1" name="sin"/>
        <OMATTR>
        <OMATP>
            <OMS cd="presentation" name="left"/>
                <OMSTR>(</OMSTR>
            <OMS cd="presentation" name="right"/>
                <OMSTR>)</OMSTR>
        </OMATP>
        <OMA>
        <OMS cd ="arith1" name="plus"/>
            <OMA> <OMS cd="arith1" name="power"/>
            <OMV name="x"/>
            <OMI>2</OMI>
        </OMA>
        <OMI>1</OMI>
        </OMA>
        </OMATTR>
        </OMA>
</OMOBJ>
```

2.3.4 Comments

Presentation MathML does not define any semantic information of the mathematics it is displaying. That can be done by Content MathML. The above example shows how content MathML embeds the formula in question in an XML document. Several commercial products that allow the creation, editing and rendering of MathML (both Presentation and Content) exists already now (March 2002). Hence the inclusion of mathematical formulae into XML documents with semantics that can automatically understood is already possible. This will make XML preferred mode for scientific publishing very soon.

The semantics supported by MathML is limited, but that can partly be circumvented by the use of annotations which allow one to add a program specific representation of a mathematical object as an annotation to a MathML encoding of a formula.

A better solution has been provided by OpenMath. The core of the OpenMath language for mathematics is the collection of Content Dictionaries (see www.openmath.org) in which mathematical objects are defined using the OpenMath language. The above OpenMath encoding of the formula $\sin(x^2 + 1)$ illustrates the use of the OpenMath CD's. To define the function sin one refers to the respective CD in which this function is defined. In fact, the name sin is overloaded. There are two different kind of sin functions: one with real arguments and values in the interval $[-1, 1]$, and one with complex arguments and complex values.

In the same way the operation '+' is defined by pointing to a particular CD. This may look like an overkill, but one should observe that the symbol '+' is heavily overloaded: the same symbol may mean anything from usual addition to an operation in an vector space or a group. Giving the CD in which the meaning of this symbol is defined, the semantics of this formula can be automatically understood.

One may wonder what is the purpose of Presentation MathML since LaTeX typesetting produces excellent results as such. One reason for Presentation MathML is in the desire to be able to express typesetting information in XML documents using XML itself to do this rather than embedding LaTeX formulae in XML documents. Presentation MathML renders XML such typesetting abilities.

The following is a quote from an editorial in the journal Scientific Computing World [6]:

> Murray-Rust and Rzepa have just published the first ever peer-reviewed journal article - about the CML language - completely in XML (see references). And he has hopes for much greater things.
>
> "I pay great tribute to the W3C," he says. "They have built an infrastructure out of XML that covers the whole of networked computing - metadata, machine communications, etc. - all which is moving towards a number of advanced IT application goals. However, Tim Berners-LeeŠs vision of a 'semantic Webt', a true information infrastructure, is where I think the significance really lies. I think it will be universal, and XML will be the transport mechanism."

This is a strong statement about the role of XML in developing semantic web. Among the community developing tools and content for the semantic web, nobody challenges this statement. For the prominent role of XML in this undertaking it is necessary to develop XML so that it can deal with scientific information. MathML was the first meta-language developed for XML, and there are others too. OpenMath provides a mechanism to harness the power of XML in any area of mathematics.

3 OpenMath architecture

The extensible OpenMath CD's (see [5]) define mathematical objects so that they can be embedded in the documents retaining their semantics. In order to be able to understand these objects, one needs auxiliary programs that translate the OpenMath encoding of mathematical objects to encodings pertinent to the particular program.

For example,

the OpenMath definition for $\int \sin(x)dx$

```
 <OMOBJ>
  <OMA>
    <OMS cd="calculus1" name="int"/>
    <OMS cd="transc1" name="sin"/>
  </OMA>
</OMOBJ>
```

will result to the nicely displayed formula $\int \sin$ (the name of the argument is actually not defined) when placed on a web page.

When glued in an OpenMath compatible mathematics program, the same expression will result into the formula $-\cos$.

So how the mathematical object is being displayed in a program – or rather what a program does to a mathematical object – will depend on the program and, of course, on the object. The phrase books take care of the necessary translations between program specific representations and the OpenMath encoding.

4 Multiple encodings of mathematics

OpenMath provides a flexible way to embed semantic information in web pages and other documents. These methods can be applied to any discipline.

The following is a quote from A. M. Odlyzko's important paper ([2]). He wrote, in 1996:

> ”Standards, or lack of them, can be a significant impediment to the adaption of new technologies.
>
> . . .
>
> . . .
>
> Mathematics, computer science, and physics all seem to have settled on TEX and its various dialects as the de facto typesetting standards. This makes it easier for these disciplines to move into electronic publishing than it is for others that have not converged on a solution. However, while TEX is adequate for almost all current papers, it may not suffice in the future as we move into a multimedia world. It is also possible that commercial packages such as Microsoft Word will be enhanced with addition of modules to handle scientific material, and may become the prevalent tools. (The era when scientists dominated electronic communications is coming to an end, and systems such as TEX, developed by scholars for scholars, might soon be eclipsed by general purpose packages.) We should not become too committed to any particular standard, as it may be transitional. ”

Now 8 years later mathematics still continues to be dominated by TEX and LATEX but the transition to other ways of publishing mathematics is clearly happening. One can cut mathematical formulae rendered by IBM's techexplorer in PowerPoint slides and paste them to Maple so that the semantics will be automatically understood. The techexplorer display quality of the mathematics in PowerPoint is as good as what can be produced by LATEX but the encoding is not anymore LATEX nor TEX.

Almost all of the current mathematics is, however, encoded in LATEX. The new emerging ways of embedding mathematics into documents have to accommodate also TEX and LATEX encodings. The above mentioned IBM's techexplorer program does just that, and there are other ways too. For several years now we are going to see scientific documents where a combination of LATEX and MathML and OpenMath is used. The reasons for this are

- the necessity of being able to use existing materials encoded in TEX or LATEX,
- the production of MathML or OpenMath encoding is still very expensive because of the lack of good editors.

The best way today to create MathML or OpenMath encoded mathematics is to express the formula in question in a mathematics program like Maple or Mathematica and then export the formula into the desired format. There are tools for the LATEX to OpenMath conversion

5 Example of the usages of the OpenMath concepts: the MAMMA project and the Helsinki Learning System

Mathematics with Multimedia (MAMMA) project directed by the authors of this paper has developed an adaptive interactive learning system, the Helsinki Learning System (HLS, see [4]), that currently supports a mixed LaTeX- MathML encoding for mathematics in the system. The first application area of the Helsinki Learning System is calculus, but the system is in no way limited to mathematics only. It can be used in any discipline.

The core of the HLS is formed by the problem and exercise database and the Course Content Dictionaries (CCDs), which define the subject areas of problems. Through the CCDs one can associate the problem database with any text-book.

6 History of the OpenMath project

It was proposed, at the International Conference of Mathematicians in Warsaw, 1983, that a database for mathematical facts should be developed. Consequently the European Mathematical Council, chaired by Sir Michael Atiyah, set up the Database working group.

Principal leaders of that working group were Michael Demazure (former N. Bourbaki) and Flemming Tøpsoe. Practically all West European countries were represented in this project which later became known as the Euromath Project. Mika Seppälä was the Finnish representative.

The Euromath Project started before any of the people involved in the project had personal computers. Most of them had no experience with TeX either. So it was no wonder that the Euromath working groups could not fully understand the difficulties in creating the Database of Mathematical Facts, which was the original charge of the project.

By 1990 it became apparent that the development of the Database of Mathematical Facts was not possible. The project redefined itself and aimed at producing a method to include mathematics in SGML documents. The project was successful in doing this in the sense that it did produce a DTD for mathematical SGML documents, and an editor, the Euromath editor, to create and to manipulate such documents.

In the Euromath DTD mathematics was embedded in SGML documents using LaTeX encoding. The Euromath editor allowed one to write mathematics so that the mathematical expressions are entered using their LaTeX encoding. By pressing a suitable sequence of keys, this encoding was then shown as a traditional mathematical expression in the document. all this was fine, but in 1991 still required too much computing power. Only a handful of mathematicians had, at that time, access to computers powerful enough to run the Euromath editor. Also the Euromath DTD in which mathematics was embedded as TeX expressions in SGML documents was not a satisfactory solution.

The need of developing this standard further became apparent (see the Editorial of the Euromath Bulletin Vol. 1, N. 1 by M. Seppälä [3]). In November 1993 M. Seppälä submitted a proposal, to the European Community, to establish a European network to develop a standard for mathematics in the internet. Shortly thereafter (December 1993) the first OpenMAth workshop was organized by Gaston Gonnet at ETH in Zürich.

This initial OpenMath project was entitled ”Editing and Computing” and it laid foundations for the current MathML and OpenMath languages.

Currently the service mark ”OpenMath” belongs to the international OpenMath Society, which is a registered society in Helsinki.

References

[1] European Space Agency. Ariane 501 - Presentation of Inquiry Board report. Published in the www, http://www.esa.int/export/esaCP/Pr_33_1996_p_EN.html, July 1996.

[2] A.M. Odlyzko. Tragic loss or good riddance? The impending demise of traditional scholarly journals. *Intern. J. Human-Computer Studies*, 42:71 – 122, 1995. Original paper dated Nov. 6, 1994, URL http://www-mathdoc.ujf-grenoble.fr/textes/Odlyzko/amo94/amo94.html.

[3] Mika Seppälä. Standarization of mathematical documents. *Euromath Bulletin*, 1(1):5–6, August 1992.

[4] Mika Seppälä and Jouko Väänänen. The Helsinki Learning System. URL http://mark.math.helsinki.fi/hls/, 2001.

[5] OpenMath Society. Content Dictionaries. Published in the web. URL http://www.openmath.org/cd/.

[6] Vanessa Spedding. XML to take Science by Storm. Available on the wwww. SECIS ONLINE, URL http://www.secis.rl.ac.uk/index.html?rhs=collaboration/scw_xml_science_storm.html.