

# Harvesting Webpages that contain Mathematical Information

Winfried Neun  
Konrad-Zuse-Zentrum  
für Informationstechnik Berlin  
D-14195 Berlin-Dahlem  
Federal Republic of Germany

Email: neun@zib.de

## Abstract

The aim of the Math-Net project (under the aegis of the International Mathematical Union) is to build up a pool of high quality information on mathematical research and mathematicians worldwide. In the framework of this project we at ZIB are harvesting pages with mathematical contents from the Web. These pages contain, besides simple text information, mathematical formulae or keywords. These formulae are traditionally encoded in LaTeX, but with the emerging new standards like MathML, OpenMath, OMDoc, MathBook, and others we have to encounter more webpages that use the new standards. Our goal is to retrieve as much semantic information as possible independent of the encoding style used for formulae in a mechanized way by providing extensions to the Harvest software. We finally want to classify the mathematical information in the webpage based on the type of formulae included and completed by mathematical keywords. In this talk we discuss some problems with the automatic detection of semantics which are caused by the encoding schemes. One example is the well-known encoding in MathML, where two different encoding types serving different needs of the users as well as mixed types are defined. Some of the attempts we make to overcome these problems are based on heuristics.

## 1 Introduction

Mathematical information on the Web (which is not only administrative) consists of an enormous legacy of (La-)Tex Documents (e.g. preprints, course material, problems and solutions) plus a couple of files in recently developed open formats like MathML, OpenMath, OMDoc or even proprietary formats like pdf, Mathematica notebooks, or Maple worksheets to name a few.

In the Math-Net project we want to collect mathematical information (i.e. including these files) and analyze the mathematical content in order to make the information retrievable in the Math-Net service Sigma, the Math-Net search engine. Queries

for special types of differential equations for example, should be possible. E.g. analyzing this paper should find out informations like: contains some formulae, no theorems, maybe some definitions. We will see shortly that is goal is not achievable given the present state of encoding of mathematical objects in the vast majority of Web files.

## 2 T<sub>E</sub>X and L<sup>A</sup>T<sub>E</sub>X files

There is an enormous wealth of T<sub>E</sub>X and L<sup>A</sup>T<sub>E</sub>X files in the WWW. Nearly two generations of mathematicians worldwide have published most of their research using this format. Unfortunately, the encoding supports the presentation of mathematical expressions. Without further semantic annotation, it is rather difficult to retrieve the information one needs to answer queries about the mathematical content of the paper.

Let us look at a (rather) simple example. The well-known Riemann Zeta function [1] is defined as :

$$\zeta(s) = \sum_{k=1}^{\infty} k^{-s} \quad \text{Re}(s) > 1$$

whereas the less well-known generalized Zeta Function (or Hurwitz Zeta Function) [3] is defined as:

$$\zeta(s, a) = \sum_{k=0}^{\infty} (k + a)^{-s} \quad - a \notin N$$

.

The L<sup>A</sup>T<sub>E</sub>X encoding of the formula above is:

```
\[ \zeta(s,a) = \sum_{k=0}^{\infty} (k + a)^{-s} \quad -a \notin N \]
```

which does not contain semantic information about the Zeta function in mind. Just a remark: s, a, and k are bound variables here. The names do not have a specific meaning. It seems to be an acceptable idea to distinguish on the number of parameter used for the function.

## 3 OpenMath and MathML

Starting in the 1990 years, some members of the mathematical community (most of them stemming from the computer algebra group) realized that the common mathematical notation (e.g. T<sub>E</sub>X) is not able to adapt properly to the requirements of the WWW. Formulae could be included into web documents as pictures only, causing some of trouble e.g. when the user wants to resize the page. And, of course, the contents of the formulae in pictures are not retrievable.

As a consequence, two efforts were started: **OpenMath**, partially funded by the European Union and **MathML** by a working group of the WWW consortium. Both

groups came up with a standard for encoding mathematical objects in XML. We have to take into account today, that there will be an increasing number of mathematical documents which have the formulae encoded in OpenMath or MathML.

We cannot discuss the standards in detail here. We will focus on features or lacks of features which are interesting in our context.

### 3.1 MathML

MathML [2] is a recommendation of the W3C, the latest revision is version 2.0 dated 21. Feb 2001. This recommendation has attracted a lot of interest in the software industry and is used as standard exchange language between modern mathematical processors, e.g. computer algebra systems or browsers. It is supported by a wide range of mathematical software today. It is not intended to be human readable.

MathML consists of two markup versions, namely presentation and content markup. The names of these markup languages explain their proposed usage. Most software systems use the presentation markup, some use content markup and some of them produce both markups. This feature helps a lot, since we can reconstruct the content relatively easy. One can mix both markups in a single expression at the same time.

The example [3] below is an excerpt from a Web page provided by Wolfram Research, Inc. describing (again) the Hurwitz Zeta Function in MathML syntax. The annotation mechanism of MathML allows the author to add the Mathematica encoding and the MathML-content encoding to the presentation markup. On the other hand, this example does not contain the information to identify the symbols via the MathML C-Symbol mechanism (discussed below).

```
<math xmlns='http://www.w3.org/1998/Math/MathML'
mathematica:form='TraditionalForm'
xmlns:mathematica='http://www.wolfram.com/XML/'>

<semantics> <mrow> <mrow> <semantics> <mrow> <mi>&#950;</mi>
<mo>&#8289;</mo> <mo></mo>
<mrow> <mi>s</mi> <mo>,</mo> <mi>a</mi> </mrow> <mo></mo> </mrow>

<annotation encoding='Mathematica'>TagBox[RowBox[List[&quot;\[Zeta]&quot;,
&quot;(&quot;, RowBox[List[TagBox[&quot;s&quot;, Rule[Editable,
True]], &quot;,&quot;, TagBox[&quot;a&quot;, Rule[Editable, True]]]],
&quot;)&quot;]], InterpretTemplate[Function[List[a, b], Zeta[a, b]]]]
</annotation> </semantics>

<mo>&#10869;</mo> <mrow> <munderover> <mo>&#8721;</mo> <mrow> <mi>k</mi>
<mo>=</mo> <mn>0</mn> </mrow> <mi>&#8734;</mi> </munderover>
<mfrac> <mn>1</mn> <msup> <mrow> <mo></mo> <msup> <mrow>
<mo></mo> <mrow> <mi>a</mi> <mo>+</mo> <mi>k</mi> </mrow>
<mo></mo> </mrow> <mn>2</mn> </msup> <mo></mo> </mrow>
<mrow> <mi>s</mi> <mo>/</mo> <mn>2</mn> </mrow> </msup> </mfrac>
</mrow> </mrow> <mo>/</mo> <mrow> <mrow> <mo>-</mo> <mi>a</mi>
</mrow> <mo>&#8713;</mo> <mi>&#8469;</mi> </mrow> </mrow>
```

```

<annotation-xml encoding='MathML-Content'> <apply> <ci>Condition</ci>
<apply> <eq/> <apply> <ci>Zeta</ci> <ci>s</ci> <ci>a</ci> </apply>
<apply> <sum/> <bvar> <ci>k</ci> </bvar> <lowlimit>
<cn type='integer'>0</cn> </lowlimit> <uplimit> <infinity/> </uplimit>
<apply> <times/> <cn type='integer'>1</cn> <apply> <power/> <apply>
<power/> <apply> <power/> <apply> <plus/> <ci>a</ci> <ci>k</ci>
</apply> <cn type='integer'>2</cn> </apply> <apply> <times/>
<ci>s</ci> <apply> <power/> <cn type='integer'>2</cn>
<cn type='integer'>-1</cn> </apply> </apply> </apply>
<cn type='integer'>-1</cn> </apply> </apply> </apply>
</apply> <apply> <notin/> <apply> <times/>
<cn type='integer'>-1</cn> <ci>a</ci> </apply>
<ci>#8469;</ci> </apply> </apply>
</annotation-xml> </semantics>
</math>

```

Please note also, that the special characters like summation sign and greek letters refer to the Unicode standard[4].

MathML encoding can be produced by most computer algebra systems and a couple of formula editors and will be rendered by most of the prominent web browsers.

Content Markup as shown in the last paragraph of the above example is limited to high-school education (Kindergarden to 12th form), e.g. the power/ operator is defined here. If an object from beyond this level should be identified, there is a standard escape mechanism in the form of the **CSYMBOL** element which allows us to point to any web location. Unfortunately, if any two of these pointers may address the same mathematical object at different locations. if the pointers are not identical, there is no way for a machine to find out.

### 3.2 OpenMath

OpenMath is a standard proposed by the OpenMath Society [5]. This standard is semantically rich, i.e. each object in the presentation of a mathematical expression is identified by a pointer to a unique object, a Content Dictionary (CD), where it is described. Then there is no problem to identify the right version of the Zeta function example, unfortunately there is at present no CD which contains the Zeta function.

The OpenMath standard is content markup only, i.e. no elements are available to define the rendering of OpenMath objects.

In the follow example an excerpt from the transc1 CD is shown, were the sin function is described. The encoding in which the mathematical expressions are put here, is exactly the MathML encoding.

sin

This symbol represents the sin function as described in Abramowitz and Stegun, section 4.3. It takes one argument.

Commented Mathematical property (CMP):

$$\sin(x) = (\exp(ix) - \exp(-ix)) / 2i$$

Formal Mathematical property (FMP):

```
<OMOBJ>
  <OMA>
    <OMS cd="relation1" name="eq"/>
    <OMA>
      <OMS name="sin" cd="transc1"/>
      <OMV name="x"/>
    </OMA>
    <OMA>
      <OMS name="divide" cd="arith1"/>
      <OMA>
        <OMS name="minus" cd="arith1"/>
        <OMA>
          <OMS name="exp" cd="transc1"/>
          <OMA>
            <OMS name="times" cd="arith1"/>
            <OMS name="i" cd="nums1"/>
            <OMV name="x"/>
          </OMA>
        </OMA>
      </OMA>
      <OMA>
        <OMS name="exp" cd="transc1"/>
        <OMA>
          <OMS name="times" cd="arith1"/>
          <OMA>
            <OMS name="unary_minus" cd="arith1"/>
            <OMS name="i" cd="nums1"/>
          </OMA>
        </OMA>
      </OMA>
    </OMA>
  </OMA>
  <OMA>
    <OMS name="times" cd="arith1"/>
    <OMI>2</OMI>
    <OMS name="i" cd="nums1"/>
  </OMA>
</OMA>
</OMOBJ>
```

$$\text{eq}(\sin(x), \text{divide}(\text{minus}(\text{exp}(\text{times}(i, x)), \text{exp}(\text{times}(\text{unary\_minus}(i), x))), \text{times}(2, i)))$$

Commented Mathematical property (CMP):

$$\sin(A + B) = \sin A \cos B + \cos A \sin B$$

Formal Mathematical property (FMP):

```
<OMOBJ>
  <OMA>
    <OMS cd="relation1" name="eq"/>
    <OMA>
      <OMS cd="transc1" name="sin"/>
      <OMA>
        <OMS cd="arith1" name="plus"/>
        <OMV name="A"/>
        <OMV name="B"/>
        </OMA>
      </OMA>
    <OMA>
      <OMS cd="arith1" name="plus"/>
    <OMA>
      <OMS cd="arith1" name="times"/>
      <OMA>
        <OMS cd="transc1" name="sin"/>
        <OMV name="A"/>
        </OMA>
      <OMA>
        <OMS cd="transc1" name="cos"/>
        <OMV name="B"/>
        </OMA>
      </OMA>
    <OMA>
      <OMS cd="arith1" name="times"/>
      <OMA>
        <OMS cd="transc1" name="cos"/>
        <OMV name="A"/>
        </OMA>
      <OMA>
        <OMS cd="transc1" name="sin"/>
        <OMV name="B"/>
        </OMA>
      </OMA>
    </OMA>
  </OMA>
</OMOBJ>

eq (sin (plus ( A , B ) ) , plus (times (sin ( A ) , cos ( B ) )
  , times (cos ( A ) , sin ( B ) ) ) )
```

Commented Mathematical property (CMP):

sin A = - sin(-A)

Formal Mathematical property (FMP):

```
<OMOBJ>
  <OMA>
```

```

    <OMS cd="relation1" name="eq"/>
    <OMA>
      <OMS cd="transc1" name="sin"/>
<OMV name="A"/>
    </OMA>
    <OMA>
      <OMS cd="arith1" name="unary_minus"/>
<OMA>
  <OMS cd="transc1" name="sin"/>
  <OMA>
    <OMS cd="arith1" name="unary_minus"/>
    <OMV name="A"/>
  </OMA>
</OMA>
  </OMA>
  </OMA>
</OMOBJ>

eq ( sin ( A ) , unary_minus ( sin ( unary_minus ( A ) ) ) )

```

### 3.3 OMDoc

The Standard Open Mathematical Document (OMDoc) [9] allows the author to preserve the complete mathematical structure of the document in the encoding. It uses OpenMath or MathML to markup the expressions and on top of this, OMDoc defines XML tags for definition, lemma, theorem, and proof. This standard is new compared to OpenMath and MathML and so far not too many OMDoc documents can be found in the web. For the first stages of the project of acquisition of content it will not play an important role, but we hope that this may change in the near future.

## 4 Extracting Mathematical Content

Our aim is to extract mathematical content from all sources of mathematical information which we gather from all nodes of the Math-Net [6] project worldwide. In order to achieve this, we have to add components to the Harvest [7] system which allow us to summarize and afterwards index the documents with mathematical contents which are gathered. In almost all cases the author will not (at least today) have added semantic annotations to the document. For the time being we have to be aware of different formats: (La)TeX, MathML, OpenMath, and soon OMDoc.

Definitely, we have to apply mathematical knowledge in this process. A simple example:  $\int f(x)dx = \int f(y)dy$ , i.e. the names of bound variables are not significant. Of course there is a trade-off between providing complete mathematical knowledge and keeping the classification scheme manageable. Common cases like equations, differential equations, etc. should be doable.

We will start with the extraction and analysis of the Metadata and keywords, MSC classification of the document. But this is not in the main scope of this paper.

Afterwards, we will try to interpret the mathematical content. Handling the easy cases first: Finding OMDoc or OpenMath files, we can directly reason from the markup. Using pattern matching to identify substructures plus recursing into the elements of the operations will provide most of the information needed. Examples: the encoding of an integral (definite or indefinite) in OpenMath should be recognizable by pattern matching, the same applies for function calls with several parameter. If two such blocks are combined by an operator like  $=$ , like in the Zeta Function definition example, this may be very well seen automatically.

If we find MathML content markup, the case is similar to OpenMath. If Csymbols are used, we can apply heuristics.

With (pure) MathML presentation things are much more complicated. We have to try to flatten the structure, eliminate the typesetting information. After that, one can try to identify Unicode symbols and other symbols, e.g. brackets. This allows us to guess the content of the expression. Experience must be obtained here by looking into particular cases from real mathematical papers. Also, experiences from other projects working in similar areas have to be taken into account.

We can apply similar techniques like in the SearchFor [8] project, even though we do not search for formulae, but we want to classify them.

$\text{\TeX}$  documents are probably the most important class of documents. Several projects are concerned with the transformation of  $\text{\TeX}$  and  $\text{\LaTeX}$  to presentation MathML. As a first approach we can use such a tool for transformation and later on apply techniques which we use for presentation markup in general.

## 5 Conclusions

Starting from files found on the Math-Net, we try to recoup parts of the semantic information, which the author had in mind when he wrote the paper. This will not succeed to our full satisfaction, of course. But we hope that we can improve information available on the Math-Net in this respect.

For the vast number of old papers this seems to be worth the effort, unless one is able to find a human source for semantic annotations.

We expect that new papers in mathematics will be written with authoring tools which produce semantic information together with the presentation as a side-effect.

If successful, future considerations may include proprietary formats like Mathematica notebooks or Maple worksheets.



## References

- [1] Milton Abramowitz and Irene A. Stegun: Handbook of Mathematical Functions, Dover Publications, New York, 1972
- [2] Mathematical Markup Language (MathML) Version 2.0, W3C Recommendation, 21 February 2001, available at <http://www.w3c.org/math>
- [3] Webpages at [functions.wolfram.com](http://functions.wolfram.com), Wolfram Research, 2000–2003
- [4] The Unicode Consortium. The Unicode Standard, Version 4.0.0, defined by: The Unicode Standard, Version 4.0 (Reading, MA, Addison-Wesley, 2003)
- [5] The OpenMath Society: Website at <http://www.openmath.org>.
- [6] Math-Net: Website at <http://www.math-net.org>.
- [7] Kang Jin Lee: Development status of Harvest, these proceedings.
- [8] S. Dalmas, M. Gaëtano: Indexing Mathematics with SearchFor, abstract at <http://www.mathmlconference.org/2000/Talks/dalmas>.
- [9] A Standard for Open Mathematical Documents, OMDoc Homepage at [www.mathweb.org/omdoc](http://www.mathweb.org/omdoc).