High-Level Semantic Correspondences for Mathematics

Stephen M. Watt

Ontario Research Center for Computer Algebra
University of Western Ontario

10 Years of OpenMath, May 21-22 2004, Helsinki

Translations Among Mathematical Data Formats

- Compose mathematical tools

Translations Among Mathematical Data Formats

- Compose mathematical tools
- Context: OpenMath, MathML, TeX, Monet, Maple, Aldor

Translations Among Mathematical Data Formats

- Compose mathematical tools
- Context: OpenMath, MathML, TeX, Monet, Maple, Aldor
- People:
  Elena Smirnova, Sandy Huerter, Clare So,
  Yuzhen Xie, Igor Rodionov, Xuehong Li

Translations Among Mathematical Data Formats

- Content MathML ⇒ Presentation MathML

Content MathML

$$(a + b)^2$$

```
<apply>
       <power/>
       <apply> <plus/> <ci>a</ci> <ci>b</ci> </apply>
       <cn>2</cn>
</apply>
```

Presentation MathML

$$(a+b)^2$$

```
<msup>
      <mfenced open="(" close=")">
            <mi>a</mi>
            <mo>+</mo>
            <mi>b</mi>
      </mfenced>
      <mn>2</mn>
</msup>
```

Translations Among Mathematical Data Formats

- Content MathML ⇒ Presentation MathML

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML

Parallel Markup

```
<semantics>
    <mrow>
        <mrow>
            <mo>(</mo> <mi>a</mi> <mo>+</mo>
            <mi>b</mi> <mo>)</mo>
        </mrow>
        <mo>&InvisibleTimes;</mo>
        <mrow>
            <mo>(</mo> <mi>c</mi> <mo>+</mo>
            <mi>d</mi> <mo>)</mo>
        </mrow>
    </mrow>
    <annotation-xml encoding="MathML-Content">
        <apply>
          <and/>
          <apply>
              <xor/> <ci>a</ci> <ci>b</ci>
          </apply>
          <apply>
              <xor/> <ci>c</ci> <ci>d</ci>
          </apply>
        </apply>
    </annotation-xml>
</semantics>
```

Parallel Markup

```
<semantics>
    <mrow id="G0">
        <mrow id="G1">
            <mo id="G2">(</mo> <mi id="G3">a</mi> <mo id="G4">+</mo>
            <mi id="G5">b</mi> <mo id="G6">)</mo>
        </mrow>
        <mo id="G7">&InvisibleTimes;</mo>
        <mrow id="G8">
            <mo id="G9" >(</mo> <mi id="G10">c</mi> <mo id="G11">+</mo>
            <mi id="G12">d</mi> <mo id="G13">)</mo>
        </mrow>
    </mrow>
    <annotation-xml encoding="MathML-Content">
        <apply xref="G0">
          <and xref="G7"/>
          <apply xref="G1">
              <xor xref="G4"/> <ci xref="G3">a</ci> <ci xref="G5">b</ci>
          </apply>
          <apply xref="G8">
              <xor xref="G11"/> <ci xref="G10">c</ci> <ci xref="G12">d</ci>
          </apply>
        </apply>
    </annotation-xml>
</semantics>
```

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML
  Notation selection

Translations Among Mathematical Data Formats

- Content MathML ⇒ Presentation MathML $\otimes$ Content MathML
  Notation selection

- Content MathML ⇒ OpenMath + Core CDs
- OpenMath + Core CDs ⇒ Content MathML

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML
  Notation selection

- Content MathML $\Rightarrow$ OpenMath $+$ Core CDs
- OpenMath $+$ Core CDs $\Rightarrow$ Content MathML
- Stylized Content MathML $\Rightarrow$ General OpenMath
- General OpenMath $\Rightarrow$ Stylized Content MathML

Translations Among Mathematical Data Formats

- Content MathML ⇒ Presentation MathML ⊗ Content MathML
  Notation selection

- Content MathML ⇒ OpenMath + Core CDs
- OpenMath + Core CDs ⇒ Content MathML
- Stylized Content MathML ⇒ General OpenMath
- General OpenMath ⇒ Stylized Content MathML

- MathML ⇒ TeX
- TeX⇒ MathML

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML
  Notation selection

- Content MathML $\Rightarrow$ OpenMath + Core CDs
- OpenMath + Core CDs $\Rightarrow$ Content MathML
- Stylized Content MathML $\Rightarrow$ General OpenMath
- General OpenMath $\Rightarrow$ Stylized Content MathML

- Presentation MathML $\Rightarrow$ $\TeX$
- $\TeX$ $\Rightarrow$ Presentation MathML

Conversion must know about macros

$$J_3(z) = \left(\frac{8}{z^2} - 1 \right) J_1(z) - 4J_0(z)/z$$

```
\newcommand{\J}[2]{J_{#1}(#2)}
$$
\J3z = \left( \frac8{z^2} - 1 \right ) \J1z - 4\J0z/z
$$
```

Similarly with XSLT for MathML to TeX.

Naïve Approach to TEX Translation

TEX ⇒ MathML

- Macro expansion:
  High-level TEX ⇒ Low-level TEX
- Translate:
  Low-level TEX ⇒ Low-level Presentation MathML

Resulting MathML has correct visual structure, BUT...

This boils off all implicit semantics from the input TEX and gives MathML with insufficient structure.

Implicit Semantics

What is that J?

- An angular momentum operator?
- A current?
- A special function?

Implicit semantics have been seen to be quite common in practice, *e.g.* Boeing corpus of mathematical TEX.

Each XSLT stylesheet/TEX style file induces semantics on the markup.

Expanding macros loses information.

How much to expand?

- We *must* expand macros, but expanding macros loses information!
- How do we decide which macros to expand away, and which to take note of?

Better Approach

- Mapping file: associates TEX macros with XSLT templates

  E.g. `\J{u}{z}` ↔ `<apply> <mmlx:J/> ... ... </apply>`

- Converter uses mapping files rules to short-circuit detailed translation.

The Mapping File

- The mapping file describes the correspondence between MathML and TEX patterns.

  The mapping file has an XML-form and consists of templates, representing MathML - TEX patterns.

  Each template is of the form

```
<map:template>
    <map:tex op="\[TeX name]" parameters="TeX expression"/>
    <map:mml op="mml-element" mode="math|text|spec"
        . . .
        [MathML expression]
        . . .
    </map:mml>
</map:template>
```

- Suppose a user has defined 2 style sheets, one defining `<mmlx:binom>` for XSLT and one defining `\binom` for TeX.

combinatorics.xsl:

```
<!-- Template for an element <mmlx:binom> -->
<xsl:template match = "apply/mmlx:binom[position()=1]">
   <mfenced>
      <mfrac thikness="0ex">
         <xsl:apply-templates select="*[2]"/>
         <xsl:apply-templates select="*[3]"/>
      </mfrac>
   </mfenced>
   <annotation-xml encoding="Content MathML">
     <!-- encoding of expression n!/(n-m)!m! -->
   </annotation-xml>
   <annotation-xml encoding="OpenMath">
     <!-- encoding of expression n!/(n-m)!m! -->
   </annotation-xml>
</xsl:template>
```

combinatorics.cls:

```
\newcommand{\binom}[2]{left({#1} \atop {#2}\right)}
```

Our mapping file should give the bidirectional conversion between a use of `\binom` and a use of `<mmlx:binom>`:

```
<map:template>
   <map:tex op="\binom"params="\patVAR!{a}\patVAR!{b}"/>
   <map:mml op="apply-mmlx:binomial">
      <apply>
         <mmlx:binomial/>
         <map:variable name="a"/>
         <map:variable name="b"/>
      </apply>
   </map:mml>
</map:template>
```

Then we would translate

```
<apply>
  <mmlx:binomial/>
  <apply> <plus/> <ci> a </ci> <ci> b </ci> </apply>
  <apply> <plus/> <ci> c </ci> <ci> d </ci> </apply>
</apply>
```

to/from

```
\binom{a+b}{c+d}
```

instead of the lower-level expression

```
\left(\atop{a+b}{c+d}\right)
```

General Process

- Macro expansion:

  High-level $\TeX \Rightarrow$ Low-level $\TeX$

  except macros mentioned in mapping files, i.e.

  High-level $\TeX \Rightarrow$ Low-level $\TeX$ + XML

- Translate:

  Low-level $\TeX$ + XML $\Rightarrow$ Presentation MathML + XML

- Refine:

  Coarse MathML tree $\Rightarrow$ (re) associated MathML tree

... and Vice Versa

- MathML ⇒ TEX from same mapping files
- Non-trivial problem to render as high-quality TEX
- E.g. line-breaking
  - deciding where to break
  - notational conventions at break (where to put operator, etc)
  - jumbo sub-expressions (e.g. in radicals, fractions, scripts, etc)

Vocabularies of functions

- Mapping files associated with standard TEX class files.
- Mapping files associated with OpenMath CDs
- E.g. NIST Special functions (math handbook 2ed)
- Each mini-ontology gives a `.cls` and `.xsl` file.

Anatomy of a pre-processor

A typical use of one of our translation modules:

- Read TEX document
- Read mapping files
- Expand macros in TEX document to find math modes (undo the ones without math mode)
- Convert math islands to MathML
- Save TEX document with each math mode replaced with `\begin{verbatimXML}...\end{verbatimXML}` or `verbXML=...=`
- Run, e.g. Latex2HTML (or one of the others)

Translations Among Mathematical Data Formats

- Content MathML ⇒ Presentation MathML ⊗ Content MathML
  Notation selection

- Content MathML ⇔ OpenMath + Core CDs
- Stylized Content MathML ⇔ General OpenMath

- Extended MathML ⇔ $\TeX$

Translations Among Mathematical Data Formats

- Content MathML $\Rightarrow$ Presentation MathML $\otimes$ Content MathML
  Notation selection

- Content MathML $\Leftrightarrow$ OpenMath + Core CDs
- Stylized Content MathML $\Leftrightarrow$ General OpenMath

- Extended MathML $\Leftrightarrow$ $\TeX$

- OpenMath $\Leftrightarrow$ Maple

OpenMath to Maple

- Relatively straightforward
- Phrasebook written in Maple
- Use per-cd mapping files

Maple to OpenMath

- More complicated
- Phrasebook written in Maple
- Use same per-cd mapping files

- Build translator dynamically out of list of cd mappings

Maple to OpenMath

- More complicated
- Phrasebook written in Maple
- Use same per-cd mapping files

- Build translator dynamically out of list of cd mappings

- Detect clashes and require disambiguation hooks

Coming back to TEX
What you see isn't what you get

- $(a+b)^2$ `(a + b)^2`
- $a \cdot b + c \cdot d$ `a \cdot b + c \cdot d`

- Remove non-semantic line-breaks
- Re-group (insert `<mrow>`s, re-associate, ...)
- Impossible in complete generality.
- Can guide process with knowledge of domain restrictions.

Pen-Based Mathematics

- Common issues with general TEX⇒ MathML, Presentation ⇒ Content
- Lift Ink to math objects, participating with other tools