

MASARYK UNIVERSITY
FACULTY OF INFORMATICS



Normative Textual Representation of Mathematical Formulae

MASTER'S THESIS

Maroš Kucbel

Brno, 2013

Declaration

Hereby I declare, that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Maroš Kucbel

Advisor: doc. RNDr. Petr Sojka, Ph.D.

Acknowledgement

@todo Thanks!

Abstract

@todo Abstract

Keywords

MathML, StAX @todo more

Contents

1	Introduction	3
2	Motivation	4
3	Mathematical Markup Language - MathML	5
3.1	<i>Structure</i>	5
3.1.1	Presentation MathML	6
3.1.2	Content MathML	7
3.2	<i>Creating MathML Markup</i>	9
3.2.1	L ^A T _E X XML	10
3.2.2	Tralics	10
3.2.3	MaxTract	11
3.2.4	INFTY Reader	11
3.2.5	T _E X4ht	11
3.3	<i>Canonicalization</i>	12
3.4	<i>Status Quo of Current Tools</i>	12
3.4.1	DAISY	12
3.4.2	MathPlayer	12
3.4.3	AsTeR	13
3.4.4	MathTalk	13
3.4.5	Web Browser Support	14
4	Analysis and Solution Proposal	15
4.1	<i>Input Processing</i>	15
4.1.1	Multithreading	15
4.1.2	Document Object Model	16
4.1.3	XSL	16
4.1.4	Streaming XML	17
4.1.5	XML Data Binding	19
4.1.6	Proposed Solution	19
4.2	<i>Conversion</i>	20
4.2.1	Presentation MathML	21
4.2.2	Content MathML	23
4.2.3	Operators and Symbols	24
5	Implementation	25

5.1	<i>Input Processing</i>	25
5.1.1	XML Parser	27
6	Results	29
6.1	<i>MREC Corpus</i>	29
6.2	<i>Changing StAX Implementation</i>	29
7	Conclusion	30

Chapter 1

Introduction

1 page Estimated length is 30-40 pages of text, which will be supplemented with samples (possibly images) and maybe some source code snippets.

Chapter 2

Motivation

Chapter 3

Mathematical Markup Language - MathML

World Wide Web pages and their main publishing language, HTML¹, provide many ways to present desired information to the user. However, presenting mathematics is not so easy and straightforward. There aren't any special tags in the HTML specification for including mathematics. In most cases mathematical equations are presented in the form of an image. On one hand this approach renders the equation the same way in every web browser, on the other hand there is no way to copy the equation for further use, not to mention editing the equation.

MathML² was specifically created to circumvent this obstacle. It was therefore designed to be used in web pages alongside HTML. It follows that MathML is an application of XML with special set of tags used to capture the content, structure and even presentation details of mathematical equations. We will take closer look at how this is achieved in the following sections. In April 1998 MathML became the recommendation of the W3C working group³ for including mathematics into web pages.

3.1 Structure

MathML as an application of XML consists of a tree of nodes. Each node is either empty, has textual value or has a list of descendant nodes. To provide additional information, a node can have an arbitrary amount of attributes, key – value pairs. Each MathML tree has to have a root node named `math` that belongs to the MathML namespace. Also it is important not to forget to include XML declaration and MathML doctype declaration.

MathML comes two distinct sets of tags. For the visual form of equations there are Presentation MathML tags, Content MathML tags focus on the

1. HyperText Markup Language

2. <http://www.w3.org/Math/>

3. The World Wide Web Consortium (W3C) is the main international standards organization for the World Wide Web

semantics and meaning of formulas. Presentation tags are being used primarily by web browsers to display mathematical expressions to the users, Content tags are important for further processing of expressions by specialized mathematical programs.

3.1.1 Presentation MathML

The main focus of the Presentation MathML is displaying the equation. For this purpose there are around thirty elements, all starting with the prefix “m”. Elements are divided into two classes, first of which is called tokens. It consist of elements that represent individual content and do not contain other nested elements. These include:

- `<mi>x</mi>` - identifiers,
- `<mn>2</mn>` - numbers,
- `<mo>+</mo>` - operators, fences, separators,
- `<mtext>free text</mtext>` - text.

The content of the token can of course be expressed in more than one character (`<mo>sin</mo>`) or with an XML and HTML character entities, for example `>` and `>` have the same meaning as `>`, greater than. It is completely up to the user which notation he will choose.

Even though nesting other elements in tokens is not allowed, there are exceptions. For example HTML5 allows almost any HTML inline tags inside the `mtext` element. So `<mtext>free text</mtext>` would be rendered with the bold word free. The other class of elements is layout schemata. This collection of elements is further divided into following groups:

- General Layout
 - `<mrow>` - general horizontal grouping,
 - `<mfrac>` - fractions and binomial numbers,
 - `<msqrt>`, `<mroot>` - radicals
- Script and Limit - superscripts, subscripts,
- Tabular Math - tables and matrices,
- Elementary Math - notation for lower grades mathematics.

We can thought of the layout schemata as a form of expression constructors, that specify the way in which sub-expression are constructed into larger expressions, starting with tokens and ending with the math element. Therefore elements belonging to layout schemata class do not contain any characters, only other nested elements (layout schemata or tokens).

```
<math>
  <mrow>
    <mi>e</mi>
    <mo>=</mo>
    <mi>m</mi>
    <mo>*</mo>
    <msup>
      <mi>c</mi>
      <mn>2</mn>
    </msup>
  </mrow>
</math>
```

Figure 3.1: Expression $e = m \cdot c^2$ in Presentation MathML

Presentation MathML also provides over fifty attributes for fine tuning of expressions, like setting colors, dimensions, alignments and many others.

3.1.2 Content MathML

Mathematical notation, rigorous as it is, is still not standardized around the world and there are many different ways of writing mathematical expressions based on cultural customs. Even simple multiplication operation can be written as $x * y$, xy , $x \text{ times } y$, $x \times y$. In many situations there's no need to actually render the expression, only the underlying meaning is important. For this reason Content MathML provides a framework and a markup language that capture the semantics of mathematical expressions.

The structure of Content MathML tree is based on a very simple principle – applying an operator to sub-expressions. For example the quotient x/y can be thought of as applying the division operator to two arguments x and y . It follows that the cornerstone of Content MathML is the function application element `<apply>`. The first child of the `<apply>` element signifies the operator, which can be an `<apply>` element again, and the rest of the `<apply>` element direct descendants are arguments to which the oper-

ator is applied. Token elements `<ci>` and `<cn>` are available to represent numbers and variables respectively.

Content MathML provides two ways of defining the operator. The first one is a set of more than 100 elements, each corresponding to some mathematical operator. For example for the addition operator there is `<plus>`, for logical xor `<xor>` and so on. Then there is the second way, the element `<csymbol>`. This element contains a textual representation of the operator, that is bound to a definition in a content dictionary referenced by either the attribute `cd` or `definitionURL`. External content dictionaries are important for communication between agents and there exist several public repositories of mathematical definitions. Most notable is the OpenMath Society repository.

```
<math>
  <apply>
    <eq/>
    <ci>e</ci>
    <apply>
      <times/>
      <ci>m</ci>
      <apply>
        <power/>
        <ci>c</ci>
        <cn>2</cn>
      </apply>
    </apply>
  </apply>
</math>
```

Figure 3.2: Expression $e = m \cdot c^2$ in Content MathML

In MathML 3, a subset of Content MathML elements is defined: Strict Content MathML. This uses only minimal, but sufficient, amount of elements, most importantly allows only the usage of `<csymbol>` element for defining operators. Strict Content MathML is designed to be compatible with OpenMath standard.

```
<math>
  <apply>
    <csymbol cd="dict">equals</csymbol>
    <ci type="real">e</ci>
    <apply>
      <csymbol cd="dict">times</csymbol>
      <ci type="real">m</ci>
      <apply>
        <csymbol cd="dict">power</csymbol>
        <ci type="real">c</ci>
        <cn type="integer">2</cn>
      </apply>
    </apply>
  </apply>
</math>
```

Figure 3.3: Expression $e = m \cdot c^2$ in Strict Content MathML

3.2 Creating MathML Markup

Creating MathML documents is very simple. All that is needed is a word processor available on every operating system. However, this approach is prone to errors, be it the wrong letter case or unclosed tags. Since MathML is an XML language, the usage of some sophisticated word processor that supports tags highlighting, code completion and XML validation will greatly improve the efficiency of creating MathML documents. But the whole process is still very time consuming and impractical. As is the case with most XML documents, even writing simple mathematical expressions takes up a lot of space and time. Fortunately there are many dedicated MathML editors, that provide some degree of abstraction from the actual MathML markup.

Multiple software products designated for working with mathematical expressions provide an option to output the results of calculations in MathML format. These include popular web service Wolfram Alpha, Mathematica, Maple or Matlab. MathML can also be used as a data exchange format or input format for aforementioned programs.

Another way of creating MathML documents is a conversion from different formats. Among scientist, mathematicians particularly, \TeX is the format of choice. It then comes as no surprise that there are many sources of

mathematical texts written in \TeX . However, not every publication comes with the source files. Often only print-ready PDF files are released[1]. Also many academic writings, especially older ones, are available only in printed form. These have to be scanned using an Optical Character Recognition (OCR) software.

3.2.1 \LaTeX XML

The lack of a suitable tool for converting \LaTeX to XML prompted the participants of the Digital Library of Mathematical Functions project to develop their own solution. Main goals of \LaTeX XML[7] design contain the aspiration to faithfully emulate \TeX behavior, the ease of extensibility and the preservation of both semantic and presentation aspects of original documents. To this end \LaTeX XML provides two main commands, `latexml` and `latexmlpost`. `latexml` converts the initial \TeX document to XML based on a set of \LaTeX XML-bindings files, that define the mapping of \TeX macros to XML. `latexmlpost` then processes resulting XML output by converting mathematics and graphics, cross-referencing and applying an XSLT stylesheet.

Both commands come with a multitude of parameters that allow users to customize the whole process, such as loading user-defined bindings or choosing the output format. Based on the requested output format, mathematics is converted to graphics (png images for HTML) or Presentation MathML in case of XHTML or HTML5.

\LaTeX XML is freely available online as an installation package for Linux systems as well as Windows and MacOS. An extensive and detailed manual is available online or as a PDF document.

3.2.2 Tralics

Tralics[3][4] is a freeware software designed to translate \LaTeX sources into XML documents, that can be further converted into either PDF or HTML. The generated XML is conforming to the local ad-hoc DTD (a simplification of the TEI) with mathematical formulas conforming to the Presentation MathML 2.0 recommendations.

Similarly to \LaTeX XML, Tralics provides many ways for customization of resulting documents, among them the possibility to change element and attribute names in the XML file. Besides Presentation MathML, mathematical formulae can be translated to \LaTeX -like elements as well.

Tralics is readily available online in the form of source files or binaries

for Linux, Windows and MacOS operating systems. An extensive documentation regarding customization and usage is also available online.

3.2.3 MaxTract

MaxTract[2] is a tool that through spacial analysis of symbols and fonts in PDF document reverse-engineers its source files in the form of \LaTeX or XHTML + MathML documents. The conversion process requires valid PDF files to work correctly as it needs the information about symbols, font encoding and width of objects contained in the PDF file. PDF documents created via \LaTeX fulfill these requirements, therefore MaxTract is able to process most of the scientific and mathematical material.

3.2.4 INFTY Reader

INFTY[11] is an Optical Character Recognition (OCR) software especially created for mathematical documents. INFTY reads scanned pages (images) and yields their character recognition results in multiple formats, including \LaTeX and MathML. It does so in four steps: layout analysis, character recognition, structure analysis of mathematical expressions, and manual error correction (optional).

The most important phase, character recognition, is responsible for distinguishing mathematical expressions and running a character recognition engine originally developed for mathematical symbols together with non-specialized commercial OCR engine.

INFTY Reader is available free of charge for limited amount of time, then it is necessary to purchase a license key.

3.2.5 \TeX 4ht

\TeX 4ht[5] is a system for producing structured output in a markup language from sources written in the \TeX -based family of languages. It is highly extensible and configurable, most common configurations include the conversion from \LaTeX to HTML with MathML, braille or DocBook targets.

The conversion is invoked by running the `htlatex` command. To the user the process seems similar to producing standard DVI or PDF outputs. \TeX 4ht uses hooks within the \LaTeX constructs and associates configurations to them. By modifying default configuration files, the user can change the resulting HTML document to his liking. \TeX 4ht comes with the support of

Cascading style sheets⁴, which provides further possibilities of customizing the output files.

3.3 Canonicalization

3.4 Status Quo of Current Tools

3.4.1 DAISY

DAISY (Digital Accessible Information System) is a technical standard for digital audio books, periodicals and computerized text. It is specifically intended for people that have problems reading a printed text including blindness, dyslexia or impaired vision. DAISY is based on combination of XML and MP3 and has various functions that traditional audio books do not provide. These include line by line navigation, searching in the text, adding bookmarks or adjusting the speed of the speech. It also provides support for embedded content such as images, graphics and MathML.

DAISY books (books conforming to the DAISY standard) can be listened to on designated DAISY readers, computers with installed DAISY software, mobile phones or MP3 players. There are many implementation of software players, commercial, free or even open source. They range from standalone applications to addons to internet browsers (especially Mozilla Firefox). Some software players are capable of displaying HTML pages with embedded MathML markup and subsequently read the mathematical expressions (this functionality was tested on a commercial product Dolphin EasyReader that supports Czech language as well).

3.4.2 MathPlayer

MathPlayer is an application developed by Design Science⁵ that enables users of Microsoft Internet Explorer web browser to display mathematical equations written in MathML markup. It uses Microsoft's internal HTML engine (MSHTML) on which Internet Explorer is based. Also for any application that makes use of MSHTML to display formatted content MathPlayer is able to display MathML content. This may include email clients, alternative browsers or RSS readers.

Besides displaying MathML in the browser MathPlayer comes equipped with a wide range of MathML-related functions. It enables equations to be

4. <http://www.w3schools.com/css/>

5. <http://www.dessci.com/en/products/mathplayer/>

copied to the clipboard as MathML markup and pasted to any MathML-aware software, be it simple text editor or more sophisticated application. Drag-and-drop functionality is also supported. Another important feature of the MathPlayer is the ability to speak expressions on the web page.

MathPlayer is shipped as a free to use software with English localization only, but is still a close source product with a license that limits its use to only computers owned by the user and also prohibits translation and reverse engineering of the application.

3.4.3 AsTeR

AsTeR (Audio System for Technical Readings)[8] is an exceptional piece of software developed by T.V. Raman as a part of his dissertation at Cornell University. AsTeR is capable of producing audio renderings of technical documents, even those containing higher mathematical expressions, written in $\text{T}_{\text{E}}\text{X}$ or $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$. However, processing different markup languages does not pose a problem for AsTeR. All that is required is a recognizer for given markup language.

The logical structure of the document is transformed to internal representation, which is then rendered in audio using a collection of rendering rules. This enables AsTeR to provide different views of the document; a user can listen to the whole document or select a portion of the document for listening. An important feature of AsTeR makes use of voice intonation to read long and complicated formulae. Such formulae are divided into logical parts (like a single summation or a numerator of a fraction) and read with different intonation than the surrounding part of the expression.

3.4.4 MathTalk

MathTalk⁶[9] is a commercial text-to-software that enables visually impaired people to read algebra notation in a quick manner. MathTalk was designed and created as a part of doctoral dissertation of R. D. Stevens[10] and employs the use of prosody in the synthetic voice to decrease the mental workload of the listener. MathTalk enables the users to browse the document and change the speed of reading which gives them comfortable control of the reading process.

6. <http://www.mathtalk.com/index.htm>

3.4.5 Web Browser Support

MathML was designed as a standard for including mathematical formulae in web pages. It stands to reason that at the most well-known and widespread web browsers should have good support of MathML, at least the Presentation MathML. Unfortunately the situation is not that good, developers of these browsers are mainly focused on different areas and MathML stand very low on the ladder of priorities.

Mozilla Firefox Firefox has the best native support for MathML rendering out of the most popular web browsers and is capable of displaying most elements of the Presentation MathML. Firefox can be enhanced by an extension called FireMath that adds a rich MathML editor capable of generating complicated mathematical expressions quite easily and right into MathML markup.

Google Chrome and Safari Both Google Chrome as well as Safari are based on the WebKit layout engine that has a development version of MathML. The support for MathML is available in latest versions of Safari (since version 5.1). Unfortunately Chrome does not have native support for MathML, but uses the combination HTML and CSS to render mathematical expression entered in MathML.

Internet Explorer Internet Explorer does not have any native support for MathML, but MathML rendering capability and some additional functionality can be added by the MathPlayer extension, that is described in more detail in section 3.4.2.

Opera Opera was one of the first browsers to include native support for MathML. The rendering of MathML is not as good as in Firefox, Opera has issues with positioning of elements in more complicated constructs.

Browser for Hand-held Devices At the time of writing this thesis none of the browsers for hand-held devices has a satisfactory native support for MathML. However they are capable of rendering expressions using CSS.

Chapter 4

Analysis and Solution Proposal

The goal of this thesis is the development of a conversion tool, that will be able to process XML document or documents and transform each occurrence of MathML markup into the appropriate textual representation. For example $5 \times \alpha = x + 3$ should be transformed into "five times alpha equals x plus three". In case of a visual (not semantic) change of the operator the tool should produce the same result. In our example $5 \times \alpha = x + 3$, $5 \cdot \alpha = x + 3$ and $5\alpha = x + 3$ all have the same meaning, so all should be transformed into identical strings.

Since the tool will be used on a corpus composed of thousands of mathematical documents, the computation speed and spacial requirements must be within reasonable limits.

4.1 Input Processing

The batch processing of a substantial amount of documents in a reasonably low time is one of the most important requirements of our conversion tool. Therefore the tool must be able to utilize the ability of modern processors to run applications in multiple threads at the same time.

4.1.1 Multithreading

Performing conversion one document at a time, especially on a batch of a substantial amount of files, is utterly impractical. The application would run for hours or even days or weeks. Therefore it is necessary to run the conversions in parallel, in multiple threads. The number of disposable threads depends on the hardware equipment of the user's computer. It would be desirable to allow the user to set the number of threads available in the thread pool of the application.

Another important factor of speeding up the conversion process represents the sharing of settings and resources among the threads. It lowers the time that each thread spends waiting for external resources, especially

those residing on the hard-drive. This can be achieved by using an in memory cache.

It is also essential to ensure that all threads work with the same settings. This can be accomplished by providing a single point of entry to the settings instance. A globally visible class (object) that implements the Singleton design pattern is the best solution for accessing conversion settings across the whole application.

4.1.2 Document Object Model

Document Object Model is one of the most wide-spread methods of representing and working with XML and (X)HTML documents. It retains the tree structure of the XML document and is composed of nodes. Node represent elements in the XML structure and form a hierarchical structure.

Each node, except the top one called root, has a parent. Nodes can have arbitrary amount of children or siblings (nodes with the same parent). This makes traversing the document structure very easy and logical.

A major drawback of using DOM lies in creating the node tree for the whole document. The input document may be very extensive, but contains only a few MathML expressions (pieces of MathML markup). We are only interested in these pieces and therefore a lot of information provided by DOM is redundant and will be discarded. Also the memory and time requirements needed for creating DOM in computing memory increase with the size of input documents.

4.1.3 XSL

Extensible Stylesheet Language (XSL) is a family of languages used to transform and render XML documents. It was created by the W3C consortium and consists of three languages:

- XSL Transformation,
- XSL Formatting Objects,
- XML Path Language.

XSL Transformation (XSLT) XSLT is a language designed for transformation of XML documents into other XML documents, but can also be used to create different output formats, such as HTML, plain text XSL Formatting

Objects. During this process the input documents stays unchanged, rather a new document is created.

XSLT employs special stylesheets to transform document. XSLT stylesheets consist of a set of templates. Each template defines a rule that describes how to process a node that matches the template selector (most commonly an XPath expression).

The transformation process follows a simple algorithm: load stylesheets and create *source tree* from the input document; starting at the root node apply the best matching template; continue till there are nodes to process. This method of processing makes XSLT a declarative language.

XSL Formatting Objects (XSL-FO) XSL-FO document describes what the resulting pages look like and where the various components go, but does not specify the layout. XSL-FO documents are often created by transforming ordinary XML documents using XSL Transformation. An application called FO processor then takes these documents and converts them into some readable form, most often a PDF or PS file. Several FO processor can be employed to achieve different results.

XML Path Language (XPath) XPath is a query language capable of selecting nodes in an XML documents. It also comes with a multitude of functions, such as string operations, date functions or operations with numbers.

XPath expression consists of three components:

- an axis - navigation within the tree (parent, child, sibling, . . .);
- a node test - match the node name to a pattern;
- zero or more predicates - restricting a node set based on conditions.

A disadvantage of XSL Transformations is the high time complexity. However modern XSLT processor employ optimization techniques and different tree representations which makes the transformation process perform better than general-purpose DOM implementations.

4.1.4 Streaming XML

Streaming approach of processing XML documents is event-based and sequential. It follows that there is no tree representation of the document created in the memory. In fact the memory requirements of streaming processing are minimal and often considered negligible. The streaming parser, after

reporting an event, usually discards almost all of the information. However it keeps some of the data, for example a list of unclosed nodes in order to catch possible errors.

There is more than a dozen available events that can be reported by the parser, such as:

- document start and end,
- element start and end,
- element text content,
- comments.

The parsing is unidirectional, the parser makes only one run through the document. Therefore every event is reported just once and there is no possibility of traversing the document structure. Hence the streaming processing comes into play when a single pass through the document is sufficient.

Based on the method of reporting events, streaming parsers can be divided into two groups:

- push principle,
- pull principle.

Push Principle When utilizing the push method of streaming XML documents all events are pushed (reported) to the application as soon as they are encountered in the document. The application needs to implement event handlers that will react to individual events and process them accordingly.

The industry standard for parsing based on push model is Simple API for XML (SAX) and has many implementations in multiple programming languages.

Pull Principle Unlike push method, with pull method the application requests the next event when it is ready to process it. The structure of the code that uses pull method resembles structure of XML documents, making it more readable and understandable than applications that use push methods.

There are various APIs that use pull method, among them XML Pull Parser (XPP)¹ and Streaming API for XML (StAX).

1. <http://www.extreme.indiana.edu/xgws/xsoap/xpp/mxpl/index.html>

4.1.5 XML Data Binding

XML data binding refers to a process of mapping XML documents to objects in computer memory (deserialization). Objects in the application represent the types defined in the XML document (e.g. by XML Schema) and can be generated by various tools from XML Schema, minimizing the required for creating such objects (classes).

The process of mapping XML types onto application objects is called *unmarshalling*. The reverse process, serialization of objects, is called *marshalling*.

The obvious advantage of this method lies in the fact, that on application level we work only with objects and don't have to concern ourselves with XML specific issues. It makes the resulting code easily readable and comprehensive.

Unfortunately the disadvantages of XML data binding are numerous. The time and spacial requirements for marshalling and unmarshalling grow substantially with the size of input document. Also the round-trip (unmarshalling the document and then marshalling it back) may not preserve the sibling order of elements, physical structure of documents (CDATA sections), comments and entity references or DTDs.

Since MathML consists of hundreds of elements and each one has to be mapped to a class when using XML data binding method, writing the bindings by hand is out of a question. There is however an XML Schema of MathML available that could be used to generate said bindings. Unfortunately attempts of code generation from this schema failed. Moreover if MathML markup was embedded inside another XML markup, such as XHTML, the bindings for MathML only would not be sufficient.

4.1.6 Proposed Solution

As we can see every method comes with many advantages as well as disadvantages making it the right choice in various situations. We will try to select one method, or a combination of two methods, best suited for our use case and requirements.

From the developer's point of view, the best solution would be to use XML data binding. However the above outlined disadvantages make it impossible to actually implement this method.

The next best approach is using the DOM with its easy to traverse structure. But there remains the issue of memory and time requirements needed to create the model. Going by these requirements the streaming method

wins as it is the fastest and most memory-efficient method.

The optimal solution would be to pass through the document using the streaming method. But streaming method does not provide any means to look ahead or behind on elements that might be needed to process the current element. On the other hand the DOM method is well suited for accessing arbitrary elements in the document tree. So combining these two methods should lead to a satisfactory solution.

Solution The streaming processor traverses the document and every read event that is not a part of the MathML markup automatically writes to the output. As we can see, also the output document is being created dynamically. When the processor encounters a `math` node it creates a in-memory DOM-like representation of MathML tree. We use a customized DOM-like model, because we do not actually need all the information that full DOM implementations tend to provide (such as namespaces), a simpler implementation suffices. After the MathML model is created (`math` end element event gets registered) the conversion needs to take place immediately, so that the conversion result can be written to the output in the right place. This way all non-MathML related information is retained and written to the output and for every occurrence of MathML markup in the document a separate DOM-like representation is created and directly converted. See algorithm 1 for illustration.

4.2 Conversion

We are starting the conversion process with our customized DOM representation of the MathML markup. At the beginning of the conversion we need to determine whether the tree consist of only the Presentation markup, Content markup or both, since each requires a slightly different approach to the conversion. In most cases the Presentation MathML resides directly in the `math` tree (is a direct descendant of `math` element) or in the element `semantics`, while the Content MathML is often found enclosed in the `annotation-xml` element with an attribute `encoding` set to the value `MathML-Content`. Or we can simply traverse the tree till we find an element from either the Presentation of Content MathML markup and continue based on our finding.

The conversion process starts at the top level element, `math`, and then recursively continues to traverse the tree, converting the Presentation and Content elements based on their own specifics. Each converted node is

Algorithm 1 Process input algorithm

```

1: procedure PROCESSINPUT
2:   clear tree ▷ a DOM-like MathML tree
3:   while next event exists do
4:     event  $\leftarrow$  next event
5:     if event is a start of math element then
6:       tree  $\leftarrow$  create a new tree
7:     else if event is an end of math element then
8:       convert tree and write the result to the output
9:       clear tree
10:    else if event is inside math element then
11:      insert a new node, value or attribute in the tree
12:    else
13:      write event to the output
14:    end if
15:  end while
16: end procedure

```

marked as processed to prevent it to be processed twice, since in some cases the conversion requires to look ahead and jump out of the recursion pattern to process a sibling node (or any other node).

4.2.1 Presentation MathML

Every element of the Presentation MathML requires individual approach to processing. But there are still groups of elements with similar characteristics.

The first group consists of token elements. The conversion is straightforward and depends only on the value of the element:

- `mn` - convert number to string or take the original value,
- `mi` - take the original value if it is in Latin script, otherwise convert the value if possible (e.g. Greek alphabet),
- `mo` - determine the operation defined by the `mo` element value ($\Sigma \rightarrow$ sum, $\int \rightarrow$ integral, ...).

It is important to remember that one mathematical operation can be expressed with various symbols and the conversion has to unify them into the same string representation.

The second group contains element with specific purpose. In other words these elements clearly state (with their names) what is their intended function. The conversion uses this fact and is therefore very simple. For example:

- `mfrac` - fractions or binomial numbers,
- `msqrt` - square root,
- `mfenced` - subexpression is enclosed in parenthesis.

The third group is composed of elements that specify only the layout of the subexpression. To determine the exact function the author wanted to convey we need additional information provided by child elements and may also depend on sibling or parent elements.

```
<math>
  <munder>
    <mo>lim</mo>
    <mrow>
      <mi>x</mi>
      <mo>&rarr;</mo>
      <mn>0</mn>
    </mrow>
  </munder>
  <mrow>
    <msqrt>
      <mi>x</mi>
    </msqrt>
  </mrow>
</math>
```

Figure 4.1: Expression $\lim_{x \rightarrow 0} \sqrt{x}$ in Presentation MathML

As we can see in figure 4.1 the `munder` element has to be interpreted in the context of its child elements (especially the first child) and its first sibling element. Similar rules apply to elements `mover`, `munderover`, `msub`, `msup` and `msubsup`, that describe various mathematical constructs such as summations, integrals, products, logarithms and many more.

The last group is formed by elements that have purely presentation function: spacing, padding, and so on. These can be ignored altogether

since they do not provide any information about the meaning of presented expressions.

4.2.2 Content MathML

The cornerstone of the Content MathML is without a doubt the element `apply`, function application. It describes the application of its first child element on the rest of child elements, all of which can be `apply` elements themselves. The `apply` element will therefore serve as a hub, assigning the processing of expression or subexpression based on the operator (the first child element).

To be able to convert the expression correctly we need a different approach than for the Presentation MathML, where the elements are ordered for display purposes and can be converted basically from top to bottom (with a few exceptions). In the Content MathML a function can be applied to multiple elements, but it will be declared only once as seen in figure 4.2 (in the Presentation MathML the plus symbol would be declared twice, between x and y , y and z).

```
<math>
  <apply>
    <plus/>
    <ci>x</ci>
    <ci>y</ci>
    <ci>z</ci>
  </apply>
</math>
```

Figure 4.2: Expression $x + y + z$ in Content MathML

Therefore we can not determine word order of the expression from the position of elements in the document (as is the case in the Presentation MathML), but we have to specify the desired word order for each operation separately. Fortunately operations can be sorted into logical groups based on the word order we use when presenting them.

- Infix form - the operator is used between pairs of inputs, starting with the first and the second input, then the second and the third and so on. These include operators for division ($x/y/z$), multiplication ($x \cdot y \cdot z$), comparison ($x = y = z, \geq, <$) and many more.

- Prefix form - the operator is used at the beginning, preceding inputs. In this case the operator is used only once at the beginning, followed by converted inputs. Examples include absolute value ($|x|$), negation (\neg) or floor ($\lfloor x \rfloor$).
- Prefix form with multiple inputs - a special case of the prefix form, where there are multiple inputs. In this instance the operator is also used only once, but inputs have to be divided by commas, with the last two inputs divided by the word "and". Function $\min(x, y, z)$ should be converted to string: minimum of x , y and z . Another examples might be sets, lists, functions greatest common divisor, maximum or lowest common multiple.
- Quotient and remainder - $\text{rem}(x, y)$ should be converted to: reminder of x divided by y . Similarly the quotient operator.
- Others - operators that do not belong to any of abovementioned categories or belong to more than one (like plus or minus, which can be used in both the prefix and infix form) have to be treated in a separate way.

4.2.3 Operators and Symbols

As we mentioned before many mathematical operations can be expressed using more than one operator or symbol. Our task lies in identifying operators and symbols belonging to each operation, creating a storage structure for this data and developing a method or methods for searching in this structure, i.e. finding an operation for a given operator.

This way all operators that denote the same operation are grouped together and we can easily unify the way each operation is converted.

Chapter 5

Implementation

This chapter describes various implementation aspects that occurred during the creation of the conversion application. The application is written in Java programming language, mainly because of personal preference of the author. Also the existence and availability of many frameworks and tools for Java language (for working with XML documents among others) are a big advantage of using this language.

The application uses Apache Maven for build automation, distribution management and dependency management.

The structure of the application can be seen on *UML component diagram*.

5.1 Input Processing

As we outlined in section 4.1.6, the streaming approach with custom DOM-like internal representation of MathML tree is used.

In Java two major APIs¹ for streaming processing of XML documents are popular among developers:

- SAX (Simple API for XML) - SAX implements the push principle, reporting events as it encounters them. From the point of view of this application a better approach would be
- StAX (Streaming API for XML) - pull principle; the application requests events from the StAX processor.

This application uses default StAX implementation that is shipped with Java Standard Edition 6 runtime. There are also other implementations available, such as Woodstox² or Aalto³, and they might provide better performance. These alternative implementations have a possibility to set up the

1. Application programming interface

2. <http://woodstox.codehaus.org>

3. <http://wiki.fasterxml.com/AaltoHome>

processor for specific use (i.e., for fast processing, low memory usage).

@todo Vymenit implementáciu a porovnať výsledky v poslednej kapitole

StAX offers two ways of traversing documents: Cursor API and Iterator API.

Cursor API As the name suggests, the Cursor API provides a cursor with which we can run through the document one event (a more appropriate name might be an info set element) at a time. The cursor always moves forwards, never backwards. Two main interfaces are available: `XMLStreamReader` for accessing all possible information retrievable from XML documents and `XMLStreamWriter` which in turn provides methods for outputting this information.

```
public interface XMLStreamReader {
    public int next();
    public boolean hasNext();
    public String getText();
    public String getLocalName();
    public String getNamespaceURI();
    ...
}

public interface XMLStreamWriter {
    public void writeStartElement(String localName)
        ;
    public void writeEndElement();
    public void writeCharacters(String text);
    ...
}
```

Figure 5.1: Example of methods in `XMLStreamReader` and `XMLStreamWriter` interfaces of the Cursor API

Iterator API The Iterator API represents XML stream as a set of discrete event objects. Much like the Cursor API, the Iterator API provides two interfaces for reading and writing these event objects: `XMLEventReader` and `XMLEventWriter`. The `XMLEventReader` extends `java.util.Iterator` interface, which means, that values returned by `XMLEventReader` can be cached and work with standard Java Iterator.

```

public interface XMLEventReader extends Iterator {
    public XMLEvent nextEvent();
    public boolean hasNext();
    public XMLEvent peek();
    ...
}

public interface XMLEventWriter {
    public void flush();
    public void close();
    public void add(XMLEvent e);
    public void add(Attribute attribute);
    ...
}

```

Figure 5.2: Example of methods in `XMLEventReader` and `XMLEventWriter` interfaces of the Iterator API

Comparison The Cursor API is generally faster and more memory efficient than the Iterator API. On the other hand subclasses of the `XMLEvent` of the Iterator API can be passed into collections or otherwise used even after the parser has moved on to subsequent events. Our application has very little use of this functionality and does not need to modify the input stream (which the Iterator API enables). However, it benefits from the Cursor API's memory and time efficiency.

5.1.1 XML Parser

The class `XmlParserStAX` is responsible for processing input XML documents and at the same time creating output documents in the same run through the file. It also provides a method for processing string input and producing string output. In this case the parser processes the document defined by this string the same way as any other input document, however the output contains only converted strings, without any of original documents elements outside the MathML markup.

The parser is capable of accepting a folder as an input (instead of a document) and processing all files in this folder, preserving the file structure on the path entered by the user.

- Streaming pull vs push, specifically in Java, **StAX**, SAX


```
public class XmlParserStAX {  
    public String parse(final String inputString,  
        final Locale language);  
    public File parse(final File file, final Locale  
        language);  
    public List<File> parse(final List<File> files,  
        final Locale language);  
}
```

Figure 5.3: Public methods of XmlParserStAX class

- <http://www.jcp.org/en/jsr/detail?id=173>
- <http://javapapers.com/jee/java-stax/>

Chapter 6

Results

6.1 MREC Corpus

`https://mir.fi.muni.cz/MREC/index.html`

[6]

6.2 Changing StAX Implementation

Chapter 7

Conclusion

1 page

Bibliography

- [1] Josef B Baker, Alan P Sexton, and Volker Sorge. Towards reverse engineering of PDF documents. *Towards a Digital Mathematics Library. Bertinoro, Italy, July 20-21st, 2011*, pages 65–75, 2011. <http://dml.cz/dmlcz/702603>.
- [2] Josef B Baker, Alan P Sexton, and Volker Sorge. MaxTract: Converting PDF to \LaTeX , MathML and Text. In *Intelligent Computer Mathematics*, pages 422–426. Springer, 2012.
- [3] José Grimm. Tralics, a \LaTeX to XML Translator. *TUGboat*, 24(3):377–388, 2003.
- [4] José Grimm. Producing MathML with Tralics. In Petr Sojka, editor, *Proceedings of DML 2010*, pages 105–117, Paris, France, July 2010. Masaryk University. <http://dml.cz/dmlcz/702579>.
- [5] Eitan M Gurari. \TeX 4ht: HTML Production. *TUG-boat*, 25(1):39–47, 2004.
- [6] Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec. Web Interface and Collection for Mathematical Retrieval: WebMIaS and MREC. *Towards a Digital Mathematics Library. Bertinoro, Italy, July 20-21st, 2011*, pages 77–84, 2011.
- [7] Bruce Miller. LaTeXXML: A \LaTeX to xml converter. *Web Manual at <http://dlmf.nist.gov/LaTeXXML/>, seen April 2013*, 2013.
- [8] TV Raman. *Audio System for Technical Readings*. PhD thesis, Cornell University, 1994.
- [9] Robert Stevens and Alistair Edwards. Mathtalk: The design of an interface for reading algebra using speech. In *Computers for Handicapped Persons*, pages 313–320. Springer, 1994.
- [10] Robert David Stevens. *Principles for the design of auditory interfaces to present complex information to blind people*. PhD thesis, University of York, 1996.

- [11] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori. INFTY: an integrated OCR system for mathematical documents. In *Proceedings of the 2003 ACM symposium on Document engineering*, pages 95–104. ACM, 2003.