VMEXT2: A Visual Wikidata aware Content MathML Editor

Moritz Schubotz

Dept. of Computer and Information Science,
University of Konstanz, Box 76, 78464 Konstanz, Germany,
moritz.schubotz@uni-konstanz.de

Abstract

VMEXT is a **V**isualization of the content **M**athML **EX**pression **T**ree. In VMEXT2, we add the functionality to edit the content MathML expressions visually. In addition to the standard OpenMath based content dictionaries, Wikidata items can be used as content symbols. Moreover, we added a MathML source code editor to allow for seamless integration of a visual and nonvisual editing workflow. We regard VMEXT2 as a step forward to a semantic formula editor for Wikipedia.

1 Introduction

Publications in science, technology, engineering, and mathematics as well as science-related content in other formats, e.g., Wikipedia, contain a high density of mathematical formulae. Therefore, computer-related processing of this data requires specific approaches. Editing mathematical formulae using standard LaTeX macros is efficient for the production of print media. Nevertheless, semantically-augmented mathematical expressions, e.g., in content MathML, have significant advantages over presentation-only formulae [4], as they are better findable and computable. Furthermore, semantically augmented mathematical expressions can interactively support the reader in disambiguating terms or identifiers and give additional explanations to the reader. One implementation is active documents for mathematical content [1] that use MathML to implement semantically rich formulae, generated from the LaTeX-dialect sTeX. To edit MathML directly (without LaTeX), a few visual editors exist. Some of them, most notable 'formulator' [2] and the WIRIS editor [3], support the generation of content MathML. From the available tools, only the WIRIS editor seems to be maintained and widely used. However, the editor is not open source, and it supports content MathML only as an alternative to presentation MathML. For most use cases, determining the presentation based on the content form is highly desirable. Nevertheless, for large existing presentation-based digital libraries, such as Wikipedia, the backward compatibility is a sine qua non. Thus, the content editor must provide semantics as an add-on, without changing the presentation form.

Wikimedia made a strong commitment to become more inclusive and allow everyone to create and edit content in Wikipedia. Notably, the development of the VisualEditor eased the process for many contributors, since knowledge of the syntax of wikitext markup is no longer needed. Nonetheless, in contrast to other WYSIWYG editors, the VisualEditor still enables users to conveniently edit the wikitext markup, as it produces very clean and minimalistic markup. This way, the huge existing editor group is not offended. However, the VisualEditor is still unsatisfactory for content with mathematical expressions, as it provides a list of LaTeX templates rather than allowing one to edit the formula layout visually (see Figure 1).

Our goal is to extend the VisualEditor to allow contributors to enrich mathematical expressions semantically. The semantic enrichments should be grounded on the knowledgebase Wikidata. Wikidata, Wikimedia's language-independent central knowledgebase, contains a significant fraction of the world knowledge in machine-readable

form. It is used for many other semantic enrichments in Wikipedia, and in particular in interlanguage links, authority control for persons and other bibliographic data, as well as the management of social, economic and geographic fact data.

In the following, we will describe our approach in section 2, present the current state of our implementation in section 3 and describe the next steps in section 4.

2 Our MathML Interface Approach

As motivated in section 1, we need a math editing tool that fulfills the following requirements:
- The tool must allow the user to augment the formulae with Wikidata items.
- The tool should allow editing semantics without changing the layout.
- The tool and its source code must be available under a MediaWiki compatible license.
- The tool must not disturb current editing workflows.

To take advantage of the related works regarding math editing, we suggest that MathML should be used as the communication standard to store and retrieve formula data from Wikidata. In addition to the existing Wikidata endpoints, which allow accessing, storing and retrieving of data from Wikidata in XML, RDF, and JSON(p) forms, we will establish a new interface for MathML. Therefore, we introduced an implicit content dictionary (called Wikidata) that allows using any[1] Wikidata item as content MathML symbol. Based on this interface, we designed a content MathML editor that uses standard conform MathML extended by the special Wikidata content dictionary.

For the content MathML editor, we extend our content MathML visualization (VMEXT) [5] so that the tree structure of the expression can not only be displayed but also modified. However, we do not intend to hide the MathML format, and thus use four interlinked visualizations for a single formula. VMEXT2 has two visual representations, the traditional presentation rendering and the expression tree representation of the content MathML. Furthermore, we added interlinked source code editors for presentation and content MathML in this new version VMEXT2. To make this representation more explicit, we made the apply-free form, which replaces the explicit display of apply nodes by the first child, optional. In particular, by clicking on the apply node, the first child is expanded. By doing so, the user experiences a one-to-one mapping between content MathML elements. Any change in one representation automatically updates the other representations. As visualized in Figure 2, hovering over elements, e.g., probability P in the screenshot, results in the highlighting of these elements in all three remaining representations. Also, popups display additional semantics. For csymbol MathML elements, this is the combination label and description from the Wikidata element. Especially for the Wikidata items, whose unique identifiers do not carry humanly readable semantics, the popup, and the auto-completion feature in the visual and source code view prevent the obfuscation of the semantics by the Wikidata item IDs.

3 Implementation

For the implementation, we followed the approach to take advantage of well established open source technology and use the APIs in order to extend the functionality according to our needs. In the following, we will describe the individual modules of VMEXT2 and describe the connection to existing technology. VMEXT2 consists of the following modules:
- **vmext** The parent Node.js module that bundles the components and presents the demo. We strive to implement everything in isomorphic JavaScript and use brunch to bring node modules to the browser.
- **mathml** MathML is the core module that allows to load modify and render MathML input. It is based on the XML processing node module xtraverse.
- **cytoscape-mathml** Cytoscape MathML is a plugin to the graph visualization cytoscape with the dagre-plugin and provides the interactivity with the MathML tree.
- **codemirror-mathml** a plugin for codemirror. It extends the CodeMirror XML autocompletion plugin and provides auto completion for MathML
- **codemirror-wikidata** a plugin for codemirror. It provides autocompletion for Wikidata entities and tooltips for Wikidata elements.

Please refer to vmext.formulasearchengine.com for detailed information on the used versions and up-to-date install instructions.

---

[1] In practice, only items describing mathematical symbols seem reasonable.

4 Conclusion and Outlook

With VMEXT2, we transformed a tree visualization for the content MathML expression tree, that was only capable of visualizing the MathML expression tree, into a content MathML editor. Moreover, we linked presentation and content more closely and created direct mappings between the visualization and the code representation. By doing so, a step towards semantic formulae in the Wikimedia projects has been completed. However many steps are still open:

1. We are currently developing a MathML API to Wikidata that allows one to read and write MathML expressions to Wikidata.
2. We are looking for a real visual editing component for presentation MathML that can be integrated in Wikimedia's VisualEditor under the given OpenSource license and backward compatibility constraints.
3. After the completion of Step 1, we are performing user studies to investigate the acceptance of our editor by the community.

Besides the main use case in Wikipedia and other MediaWikis, such as the DRMF, the components described in section 3 can be used for various purposes in the web. Moreover, VMEXT2 can also be used with another knowledgebase or for editing MathML without a dedicated knowledgebase.
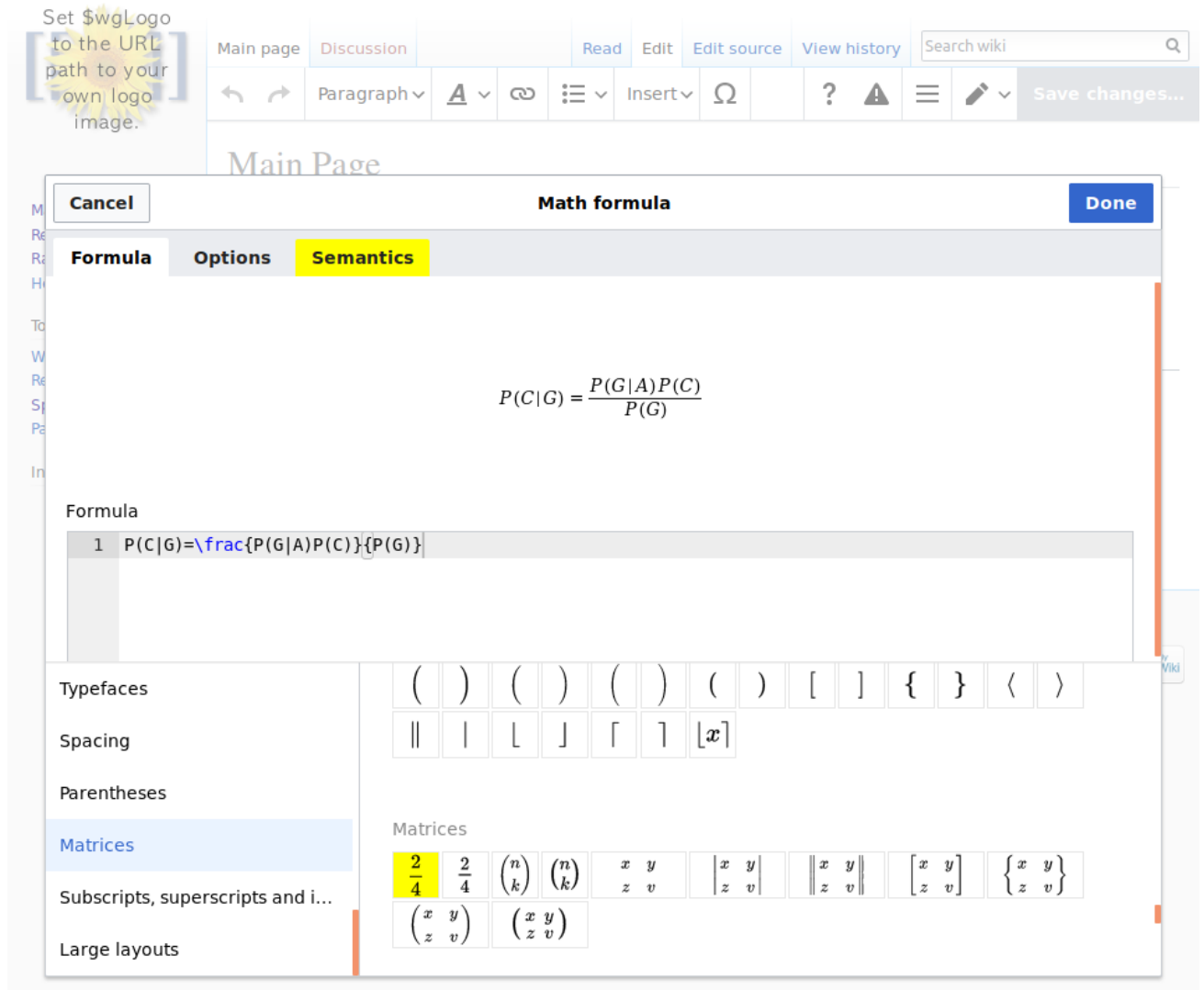
Figure 1: Math editing component of the VisualEditor. As of April 2018, users either need to know the LaTeX-command for 'fraction' or use the fraction template ($\frac{2}{4}$, highlighted in yellow) and replace 2 by the numerator and 4 by the denominator of the formula. We will improve this dialog by adding semantic input as an alternative in the top panel (semantics, highlighted in yellow).

Figure 2: MathML source editor (left) and visual content tree editor (right)