Representing Mathematical Formulae in Content MathML using Wikidata

Philipp Scharpf[1], Moritz Schubotz[1], and Bela Gipp[1]

Information Science Group
University of Konstanz, Germany
first.last@uni-konstanz.de

Abstract. In this paper, we describe how to represent mathematical formulae in Content MathML referring to the open knowledge-base Wikidata for the grounding of the semantics. By doing so, we link identifiers and symbols in MathML to Wikidata items to annotate mathematical identifiers or operators. In contrast to other mathematical knowledge-bases, which define symbols in a deductive fashion, the terms in Wikidata emerged inductively from Wikipedia articles in different languages. In this context, we discuss the term of a mathematical *formula content* and its relation to the data representation in Content MathML and Wikidata in detail.

1 Introduction

We are confronted with a constantly increasing rate of published documents from the disciplines of Science Technology, Mathematics and Engineering (STEM) that contain a large amount of mathematical formulae [1]. While there are already advancements to process their textual contents using technologies from Natural Language Processing (NLP) [2], there is still a huge potential for development concerning the processing of mathematical content. Pagel and Schubotz coined an analogous term of Mathematical Language Processing (MLP) [3]. Exceeding a mere processing, it will certainly be beneficial if the computer is able to *understand* and *interpret* the processed content. This means being able to include the semantics of the mathematical formulae and their constituents in the computational analysis. To achieve this, approaches will be developed and refined that enable automated semantic enrichment of mathematical formulae, i.e., the computer should be able to automatically infer the meaning of a formula and its constituents from the context (surrounding text, mathematical topic or discipline, etc.). Furthermore, a definition of a *formula content* is needed to formalize a mapping to its digital representation in a markup language and semantic database. To assess the quality of the semantic enrichment by various approaches and tools, we need a benchmark dataset as a reference. So far, there are only little high-quality samples available. In the remainder of this paper, we will develop a definition of a mathematical *formula content*, which can be represented in the Content MathML language making use of the semantic knowledge-base Wikidata as a Content Dictionary as proposed in [4]. One crucial benefit

of linking semantic elements of a formula to Wikidata is to have a language-independent representation that can be used to retrieve information in a large number of languages, which do not provide thorough textbooks. Furthermore, we will describe our annotation tool that we used to construct a Gold Standard *MathMLben* [5], which was recently introduced to facilitate the conversion between different mathematical formats such as LaTeX variations and Computer Algebra Systems (CAS). Finally, we assess the quality of the MathML benchmark and its suitability for the evaluation of automatic semantic enrichment approaches or tools.

2 Background

In this chapter, we will review the currently available encoding standards for the semantics of mathematical formulae to lay the foundations for our contribution of a new semantic annotation in Content MathML linking to Wikidata items.

(La)TeX LaTeX (shortening of Lamport TeX named after its author) is a software package for the typesetting system TeX that was first released in 1983 [6]. TeX was initiated by Knuth in 1977 as both a typesetting for the annotation of document parts (title, sections, etc.) and a markup macro language (font style, math environments, etc.). So aside from being a standard for the document layout, TeX can be used to annotate the semantic parts of a document, but also within a formula. Kohlhase developed a semantic markup format for LaTeX formulae [7]. "sTeX: An Infrastructure for Semantic Preloading of LaTeX Documents" is available as a package[1] collection to bring a mathematical document in a format that can be used for Mathematical Knowledge Management (MKM). Moreover, LaTeX macros were defined for the Digital Repository of Mathematical Formulae (DRMF) [8,9], e.g. to annotate mathematical functions such as the Euler Gamma function $\Gamma(z)$ as `\EulerGamma@{z}`.

OpenMath "OpenMath is an extensible standard for representing the semantics of mathematical objects" [10]. It was developed to enable a representation of the semantics of mathematical formulae and facilitate the exchange between different mathematical software systems and languages. OpenMath encoding is especially also beneficial in document analysis, for purposes of search, recommendation, plagiarism and novelty detection where knowing the semantics of a formula and its parts is required. The encoding of OpenMath objects allows them to be rendered in browsers, exchanged between software systems or transferred between different mathematical documents while preserving the semantics. Moreover, it is an important step towards the goal of automatically checking the *soundness* of mathematical statements by considering their semantic content. While the OpenMath Standard was built to annotate individual formulae or simple mathematical statements, an additional markup language was developed to extend the markup to entire documents. The Open Mathematical Documents

[1] https://github.com/KWARC/sTeX

(OMCDoc) standard [11] is part of the *MathWeb.org* initiative for supporting mathematics on the web, which is maintained by Michael Kohlhase on GitHub at `https://kwarc.github.io/mathweb-org/`. OMDoc enables a markup for mathematical expressions on three levels: The *Object level* (individual formulae with Content MathML markup), the *Statement level* (definitions, theorems, proofs, examples and the relations between them) and the *Theory level* (set of contextually related statements). Each level can possibly include both logical syntax and natural language information. Since physics has its own characteristics, a specific dialect was developed, the PhysML content markup language for physics, which is maintained by Hilf, Kohlhase, and Stamerjohanns on Github at `https://github.com/OMdoc/OMDoc/wiki/PhysML`. It aims to extend the OMDoc standard by "an infrastructure for the principal concepts of physics: observables, physical systems, and experiments" [12]. OpenMath, OMDoc, and PhysML are all important steps towards a Semantic Web for mathematics and physics, consisting of ontologies in OWL format, which contain mathematical or physical markup.

MathML The Mathematical Markup Language (MathML), is "an XML application for encoding mathematics on the Web" [13]. It encodes both the visual structure (Presentation MathML) and the formula content (Content MathML) of mathematical formulae. The original goal of MathML was to represent mathematics on the web as a first XML application, which was promoted by the World Wide Web Consortium (W3C). In 1997, only three years after the W3C was formed, the W3C Math Working Group began to design the MathML standard, which was first released (MathML 1.0) in 1999 and extended and refined (MathML 2.0) in 2001. It was gradually supported by numerous software systems and browsers as well as organizations, prominently the American Mathematical Society (AMS). According to its design by the W3C Math Working Group, MathML is both human-readable for mathematicians with little computer science background and machine-readable for software systems performing calculations and automated reasoning. It is easily extensible by new markup tags, dictionaries, rules, etc. Also, it enables the exchange of mathematical content between various software systems, prominently Computer Algebra Systems and LaTeX editors for academic writing, web pages, etc. A recent assessment of its suitability as an intermediate language for the conversion between a collection of mathematical tools and LaTeX was made by the authors [5]. Since HTML5, MathML is an integral part of HTML. The markup is displayed in many, even interactive browser environments. In comparison to other display methods for online mathematics, such as formulae as images in gif, jpeg, png or pdf format, it is much more flexible, lean, combinable and reusable. Being an application of XML, MathML extents its syntax and rules by additional restrictions on types and values of content. For an extensive description of MathML, we refer to "the W3C MathML standard", which is available online at `https://w3.org/Math/`.

Content MathML A central aspect of the OpenMath standard was to introduce the possibility of Content Dictionaries, collections of symbols or identifiers with

declarations of their semantics - names, descriptions, and rules. According to OM Society, they serve as an agreement on a joint "OpenMath language" [10]. The Content Dictionaries are collected by the MathML CD Group. A list can be found at `http://www.openmath.org/cdgroups/mathml.html`, comprising basic algebraic concepts, symbols for common arithmetic functions and relations, calculus operations, operations on and constructors for complex numbers, linear algebra matrix operations, limits, basic logic functions, basic multiset theory, symbols for creating numbers and constants, roundings etc. The official CD collection is reviewed by the OpenMath Society. Content MathML was introduced to complement Presentation MathML. While the latter focuses on the visual structre of the marked mathematical formulae in documents or websites, the former was built for the description of content elements (identifiers, operators, functions, and bindings) by using the OpenMath specification [10]. Content MathML adopts the prefix notation of OpenMath, which stems from the style of functional programming languages such as Lisp or Scheme, allowing for nested lists or tree-structured data when marking the application of operators. As an example, the expression $x + y$ is represented in prefix notation as $+xy$, meaning that the operation $+$ takes the arguments x and y as bound variables.

Wikidata Wikidata is an open semantic knowledge-base that can be read and edited by humans and machines. It was initially built to provide well-maintained high-quality structured data to other Wikimedia projects, prominently being a source of the lean Wikipedia infoboxes. Wikidata aims to contain the "semantic bones" extracted from Wikipedia, being its central data management platform. It is an internationalized multilingual database for the collaborative adding and editing of the world's knowledge. It allows inconsistent and contradictory facts to coexist to represent the plurality of knowledge about a specific topic [14]. Since its start in 2012, Wikidata was enriched by an enormous amount of structured data such as numbers, coordinates, dates, names, formulae, taxonomies, etc. The data is disposed to be searched, analyzed and reused by direct access through query services or regular (machine-readable) data exports[2]. Users can extend and edit the content of Wikidata items even without having to create an account [16]. References have to be provided and are checked by the community on a regular basis. Wikidata is continuously growing in size: today (May 2018) it contains almost 50 million items In 2014, Google offered the data of Freebase to the Wikidata community who developed the *Primary Sources Tool* [17] to facilitate data migration by an interface in which users can approve or reject alleged and referenced statement [18]. In the last few years, many other datasets were imported, including migration from Wikimedia sister projects. The data model of Wikidata consists of entities or items (referenced by QIDs Qxxx) connected by properties (referenced by Pxxx) as triples, e.g., Albert Einstein (Q937) - native language (P103) - German (Q188). An illustration of the data model is shown in Figure 1. In this case, *London* is the main item with a statement that consists

[2] The RDF data dumps that connect Wikidata to the Linked Data Web are available at `http://tools.wmflabs.org/wikidata-exports/rdf` [15].

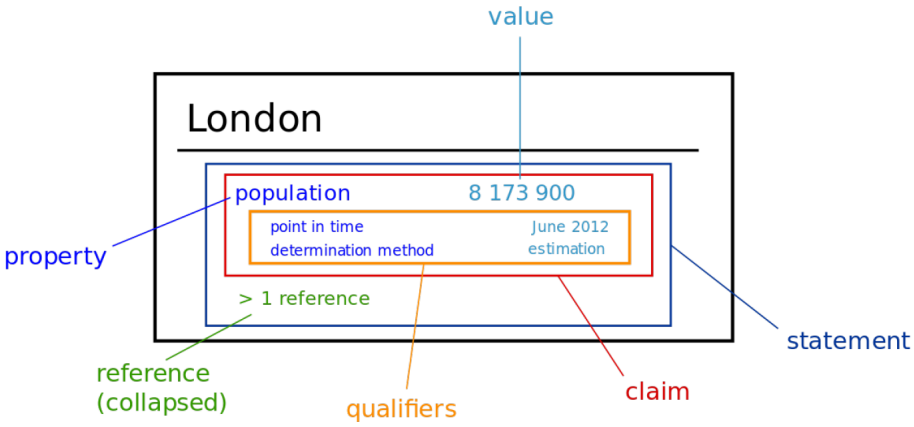

Fig. 1. Wikidata statement terminology illustrated by an example [19].

of a claim and a reference for that claim (statement = claim + reference). The claim is formed by the main property-value pair that represents the main fact, here *population* = value, specified by optional qualifiers with values, the *point in time* and *determination method*. Compared to the amount of non-mathematical knowledge, Wikidata contains only a few mathematical formulae. To address this issue, we recently have seeded around 17 thousand mathematical formulae fetched from Wikipedia articles (first defining formula) into Wikidata. They are now at disposal to be approved or rejected by the community using the *Primary Sources Tool*.

3 Approach

3.1 Formula Content

We define a mathematical *formula content* as the composition of its constituting features, the properties or attributes that are relevant to identify the content. There are mainly three types of formula constituents: numbers, identifiers, and operators. To uniquely define the *formula content*, additional information on the relations between these constituents is required. The Content MathML tree can be visualized by VMEXT, as illustrated in [20]. Two formulae could possibly contain the same identifiers or numbers, but with different conjunctions or bindings. As a simple example, $a = 2 + 3 \cdot b + c$ contains the same numbers (2 and 3), identifiers (a, b, c) and operators $(+, \cdot, =)$ as $c = 2 + 3 \cdot a + b$, but in a different functional composition. Analogous to natural language, mathematical language could be decomposed into a triple representation, where the mathematical numbers or identifiers play the role of the natural language subject or object, e.g., $3 < 6$, but this only applies to binary operators, inequalities, and elementary expressions. For composite equations containing a variety of arbi-

trary n-ary operators (e.g. an unary faculty $x!$), a general functional definition is necessary.

This can be illustrated examining $r = \sqrt{x^2 + y^2}$. This formula contains one number (2), three identifiers (r, x, y) and three operators $(=, \sqrt{}, +, \hat{})$, if we classify the equality sign = as operator. In functional notation, it can be written as a nested function:

```
=(r,\sqrt(+(^(x,2),^(y,2))))
```

Having included the set of numbers, identifiers, and operators as well as their functional relations in our definition of a *formula content*, we still need a disambiguation of the identifiers (and operators - although they are much less ambiguous) in the sense of defining the mathematical data type and/or physical units. In our example, the identifiers r, x and y could be any variables, complex numbers, real numbers or represent radius and coordinates. Besides, also a number could need clarification of its meaning. E.g., 3.14 could be an arbitrary value or a rounded representation of the irrational transcendent number π. So a clarification of the meaning is the last ingredient to grasp the *formula content*.

In short, we summarize our definition of a *formula content* as:

- The set of numbers, identifiers and operators (including equality = or inequality signs $<, >$) the formula contains.
- The (nested) functional relations of the identifiers and numbers, with the operators being the functions and numbers or identifiers being the variables.
- The individual meanings (disambiguation) of the numbers, identifiers, and operators.

or more formally: A *formula content* $\mathcal{F}$

$$\mathcal{F} = \{\mathcal{N}, \mathcal{I}, \mathcal{O}, \mathcal{R}\}$$

is given by the sets of the numbers $\mathcal{N}$, identifiers $\mathcal{I}$ and operators $\mathcal{O}$ the formula contains and the set of functional relations $\mathcal{R}$. In the first three sets, each element is a (symbol, meaning)-tupel while the last set contains the nested functionality of the formula.

The example formula $r = \sqrt{x^2 + y^2}$ is formalized as:

$\mathcal{F}_r = \{\mathcal{N} = \{$`(2, natural number)`$\}$,
$\mathcal{I} = \{$`(r, radius), (x, dimension), (y, dimension)`$\}$,
$\mathcal{O} = \{$`(=, equality sign),(\sqrt(), square root),`
`(^(), power operator)`$\}$,
$\mathcal{R} = \{$`=(r, \sqrt(+(^(x,2),^(y,2))))`$\}$

The disambiguation of 2 as natural number, = as equality sign, `\sqrt()` as square root and `^()` as power operator could arguably be omitted, but for completeness, we fill all tuples (number, disambiguation), (identifier, disambiguation) and (operator, disambiguation).

3.2 Wikidata annotation in Content MathML

Our definition of a *formula content* can be mapped into Content MathML with Wikidata annotation, i.e., using Wikidata as an OpenMath Content Dictionary. Figure 2 shows an example markup of the formula $E = mc^2$ from physics with its identifiers linked to Wikidata items referenced by their QID, e.g., Q11379 for "energy".

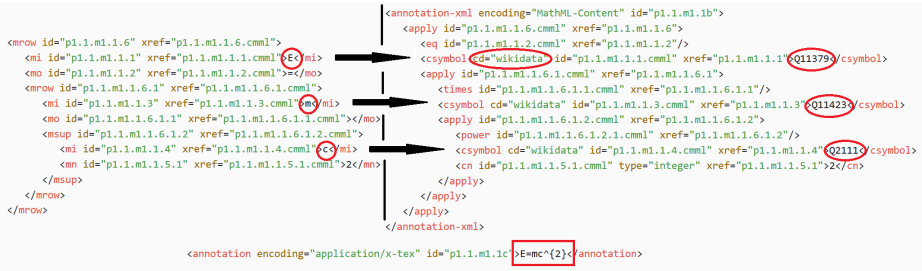

Fig. 2. Wikidata annotation in Content MathML. The identifiers E, m and c within the formula $E = mc^2$ are linked to their corresponding Wikidata QIDs, which are *energy* (Q11379), *mass* (Q11423) and *speed of light* (Q2111). The XML tags $< mi >$ and $< mo >$ annotate mathematical identifiers and operators respectively.

Since the semantics of the identifiers or operators is taken care of by the Wikidata Content Dictionary, we claim that our definition of a *formula content* can fully be covered by a description in Content MathML with Wikidata markup. In the following, we illustrate how the functionality of a formula can be mapped to Strict Content MathML, an XML encoding of OpenMath objects with Content Dictionaries. Table 1 contains the most important tags in Strict Content MathML and OpenMath representation, which will be described below[3].

The tags are intended to use as follows:

- The `<cn>` ("content number") element is built to contain numbers such as integers, real and (double) floating point numbers. Additionally, an *e-notation* as well as *complex-cartesian* and *complex-polar* notations are supported. In the case of complex numbers, the real and imaginary part can be separated using the `<sep>` tag (possibly be rewritten using the `<apply>` element).
- The `<csymbol>` ("content symbol") element is used in `<annotation-xml>` encoding to refer to a Content Dictionary by the *cd* attribute. Identifiers or operators can be cross-referenced by an *id* and *xref* attribute respectively.
- The `<ci>` ("content identifier") element is a markup for mathematical variables who - in contrast to symbols - do not have a fixed value. All variables with the same name in the same scope (see bindings `<bind>` and bound variables `<bvar>` below) are considered equal or identical.

[3] The description follows the W3C Recommendation for Content Markup available online at `https://www.w3.org/TR/MathML3/chapter4.html#contm.cds`.

Table 1. Most important tags in Strict Content MathML and OpenMath to capture the functionality and semantics of a *formula content* excerpted from `https://www.w3.org/TR/MathML3/chapter4.html#contm.cds`.

Strict Content MathML	**OpenMath**
cn	OMI, OMF
csymbol	OMS
ci	OMV
cs	OMSTR
apply	OMA
bind	OMBIND
bvar	OMBVAR
share	OMR
semantics	OMATTR
annotation, annotation-xml	OMATP, OMFOREIGN

- The `<cs>` ("content string") element encodes string literals in the form of text.
- The `<apply>` element is used to fundamentally build compound objects by recursively applying a function or an operator to some arguments (numbers, symbols or variables). E.g. $x + 3$ can be implemented as `<apply><csymbol cd="arith1">plus</csymbol> <ci>x</ci><cn>3</cn></apply>`.
- The `<bind>` element is indented to build mathematical expressions where a variable is bound in the scope of a function, operator or quantifier. The latter can be an integral, a sum, product or logical quantifier such as $\forall$ (*for all*) or $\exists$ (*there exists*) while the former is the bound *dummy variable* (renaming it does not change the meaning of the expression).
- The `<bvar>` element denotes the individual bound variables that occur as children in a nested binding expression.
- The `<share>` can be used to avoid duplicate passages by allowing for copies that are pasted by reference. This can be done using the *href* and *id* attributes.
- The `<semantics>`, `<annotation>`, and `<annotation-xml>` element wrap content elements that provide additional semantic markup or annotations via a Content Dictionary. The tags are regarded as part of both Presentation MathML and Content MathML.

Concluding, we claim that the semantic level of our defined *formula content* is covered by the `<semantics>`, `<annotation>` and `<annotation-xml>` markup while the structure and functional level can be described by utilizing the `<apply>` and `<bind>` environments.

The remainder of this paper describes the construction, maintenance and quality assessment of a MathML benchmark dataset for the evaluation of automated semantic Information Retrieval. We recently introduced it [5] as a Gold Standard for evaluating the conversion between different mathematical formats, especially LaTeX markup and Computer Algebra Systems.

3.3 MathMLben

Our original motivation for MathML-Wikidata annotation was to define semantic relatedness for formulae by counting Wikidata links between them. Having relations with other Wikidata items enables to improve the taxonomic distance measure [21]. The dataset - *MathMLben* - comprises 305 mathematical expressions (ranging from individual symbols up to complex multi-line formulae). Additionally, it contains meta-information such as the source URL or document page it is retrieved from. The expressions were selected from three different sources [5]:

Expressions 1 to 100 are random samples taken from the "National Institute of Informatics Testbeds and Community for Information access Research Project" (NTCIR) 11 Math Wikipedia Task [22].

Expressions 101 to 200 are random samples taken from the "NIST Digital Library of Mathematical Functions" (DLMF) [23] available on the website `https://dlmf.nist.gov/` containing around 10.000 labeled LaTeX formulae with semantic markup classified in 36 categories [8,9]. In case of multiple equations, we randomly chose one and discarded the others.

Expressions 201 to 305 were selected from the NTCIR arXiv and NTCIR-12 Wikipedia dataset retrieval. 70 % of these formulae were taken from the arXiv [24] and 30 % from a Wikipedia dump.

We created a Graphical User Interface (see Figure 3) as a web application with a variety of input fields to easily maintain the gold standard. For each Gold ID entry or formula, you can chose a *Formula Name*, specify a *Formula Type* (definition, equation, relation or general formula) and insert the **Original Input TeX** and manually **Corrected TeX** together with a **Hyperlink** to the source. The **Semantic LaTeX Input** field is used for the semantic annotations, providing the basis for the generation of Content MathML with Wikidata annotations by LaTeXML [25,26]. The corrected TeX is rendered in real time by Mathoid [4] as an SVG image. Moreover, an expression tree is displayed, rendered by our visualization tool VMEXT [20]. For each symbol in the tree, the assigned annotation is shown as a yellow mouse-over infobox containing the Wikidata QID, name and description (if available).

As described in the gold-standard [5], the data is publicly available at `https://mathmlben.wmflabs.org` with a user guide on how to access raw data or contribute by extending or correcting the expression tree or (Wikidata) annotations.

4 Discussion

Finally, we will discuss in detail how our proposed usage of Wikidata both as LaTeX markup and a MathML Content Dictionary will improve grasping the semantics of *formula content*. We will compare our approach of implementing Wikidata annotation in comparison to already existing DLMF LaTeX macros and default OpenMath Content Dictionaries.

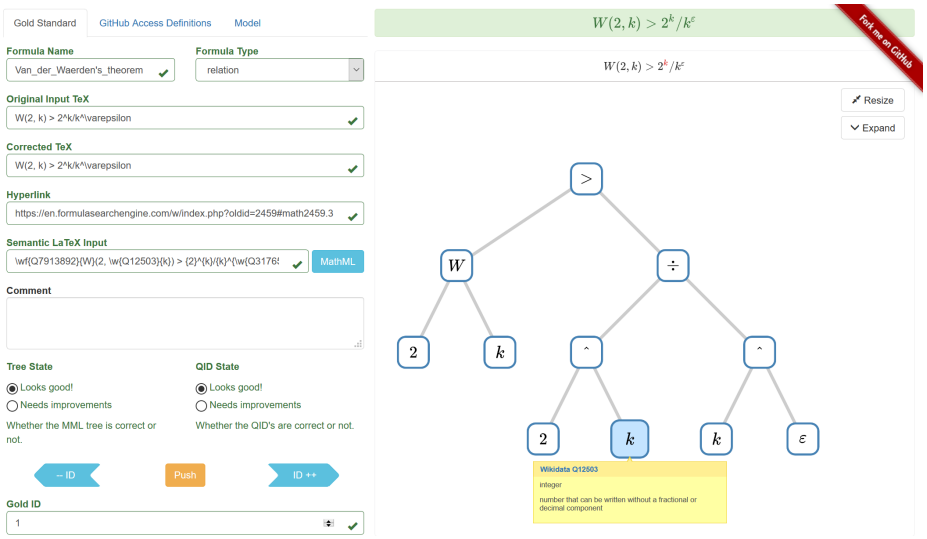

Fig. 3. Graphical User Interface of *MathMLben* providing several TeX input fields (left) and a mathematical expression tree rendered by the VMEXT visualization tool (right) [5].

4.1 DLMF LaTeX macros

LaTeXML [25,26] supports the usage and conversion of a selection of DLMF LaTeX macros, the community agreed upon [27]. The macros are designed to facilitate a human-readable semantic annotation and the compilation continously will be extended. They are a huge benefit for the creation of semantically enriched mathematical knowledge, since the direct editing in MathML can be tedious or confusing and researchers and editors in the mathematical sciences are usually more familiar with LaTeX markup [28]. Some examples are [5]:

- `\EulerGamma@{z}`: $\Gamma(z)$: gamma function,
- `\BesselJ{\nu}@{z}`: $J_\nu(z)$: Bessel function of the first kind,
- `\LegendreQ[\mu]{\nu}@{z}`: $Q_\nu^\mu(z)$:
  associated Legendre function of the second kind or
- `\JacobiP{\alpha}{\beta}{n}@{x}`: $P_n^{(\alpha,\beta)}(x)$:
  Jacobi polynomial.

However, the defined macros are naturally not extensive, and expansion requires community consensus and implementation. In contrast, Wikidata items can be easily created at any time by anyone. This is why we propose using Wikidata markup as a supplement. We are aware of the fact that since the content of Wikidata items is monitored by a much larger community than the DLMF responsibles can be either beneficial or harmful to the quality of the semantic markup.

4.2 OpenMath and Wikidata Content Dictionaries

To assess the quality of Wikidata annotations at the level of MathML markup, we will now compare the Content Dictionaries (CDs) using the simple example of adding two numbers $a + b$.

Using the `plus` tag of the OpenMath CD *arith1* for arithmetic operations[4], Strict Content MathML reads

```
<apply><csymbol cd="arith1">plus</csymbol><ci>a</ci><ci>b</ci></apply>
<apply><csymbol cd="arith1">plus</csymbol><ci>b</ci><ci>a</ci></apply>.
```

In contrast, a proposed Wikidata markup reads in LaTeX

```
\w{Q12916}{a} \w{Q32043}{+} \w{Q12916}{b}
```

and the Content MathML annotation generated by LaTeXML is

```
<semantics id="p1.1.m1.1a">
    <mrow id="p1.1.m1.1.4" xref="p1.1.m1.1.4.cmml">
       <mi id="p1.1.m1.1.1" xref="p1.1.m1.1.1.cmml">a</mi>
       <mo id="p1.1.m1.1.4.1" xref="p1.1.m1.1.4.1.cmml"></mo>
       <mi id="p1.1.m1.1.2" mathvariant="normal" xref="p1.1.m1.1.2.cmml">+</mi>
       <mo id="p1.1.m1.1.4.1a" xref="p1.1.m1.1.4.1.cmml"></mo>
       <mi id="p1.1.m1.1.3" xref="p1.1.m1.1.3.cmml">b</mi>
    </mrow>
    <annotation-xml encoding="MathML-Content" id="p1.1.m1.1b">
       <apply id="p1.1.m1.1.4.cmml" xref="p1.1.m1.1.4">
          <times id="p1.1.m1.1.4.1.cmml" xref="p1.1.m1.1.4.1"/>
          <csymbol cd="wikidata" id="p1.1.m1.1.1.cmml" xref="p1.1.m1.1.1">Q12916</csymbol>
          <csymbol cd="wikidata" id="p1.1.m1.1.2.cmml" xref="p1.1.m1.1.2">Q32043</csymbol>
          <csymbol cd="wikidata" id="p1.1.m1.1.3.cmml" xref="p1.1.m1.1.3">Q12916</csymbol>
       </apply>
    </annotation-xml>
    <annotation encoding="application/x-tex" id="p1.1.m1.1c">a+b</annotation>
</semantics>
```

While the *arith1*-markup allows for a specification of the identifiers a and b as "integer", "rational", "real", "complex" etc., the Wikidata-markup assigns them to the Wikidata item *real number (Q12916)*. Additionally it is possible to specify the meaning, e.g. annotate as a physical *observable* (Q845789). The plus-operation is assigned to the Wikidata item *addition (Q32043)*.

The Wikidata markup is larger and maybe less easy to read, but as already discussed it can be extended at all time by creating or adjusting Wikidata items. Furthermore, it has the advantage that all items are linked to Wikipedia articles that provide extensive human-readable descriptions. Moreoover, MathML is seldomly edited manually. Therefore, we propose to use tools such as VMEXT, which uses human readable labels rather than the Wikidata QIDs in the GUI.

[4] For the specification see `http://www.openmath.org/cd/arith1#sum`.

5 Conclusion and Outlook

With our proposed Wikidata markup, we achieved advantages of both LaTeX and MathML markup at the same time. The Wikidata macros in LaTeX are easy to write and read and thus adequate for editors who are not familiar with the syntax and structure of MathML. It will be an important mission to convince researchers in the mathematical sciences of the benefits of making their published content machine-readable. In their documents, they should mark as many entities as possible:
- named entities, e.g. `\w{Q210546}{Equivalence principle}`
- whole formulae, e.g. `\w{Q210546}{$E=mc^2$}`
- parts of formulae, e.g. `$\w{Q11379}{E} = \w{Q11423}{m} \w{Q2111}{c}^2$`
We are planing to develop a GUI to facilitate the annotation process. This will produce labeled data to use for the future discovery and recognition of formula concepts, our long-term research goal.

Formula Concept Discovery (FCD) will be an approach to develop an understanding of the definition of a mathematical concept by using labeled data from Wikipedia, the arXiv, other resources and self-annotated formulae. We strive to engineer an automatic retrieval of the defining formula of a given mathematical concept within a document. Machine Learning methods such as formula clustering, decision trees, and feature analysis will help to approximate our ambition.

Formula Concept Recognition (FCR) will subsequently be applicable to spot the defined formula concepts in given documents by identifying its semantic components (i.e., the constituting identifiers, operators, and numbers) with Wikidata items. If there is finally a sufficient amount of labeled data, we can explore the possibility of using neural networks to recognize a given formula by matching its features to concept labels, starting with a small set of classifiers. Our future research will eventually aim at a largely representation-independent definition and recognition of formula concepts.

6 Acknowledgements

We thank Wikimedia Foundation and Wikimedia Deutschland providing cloud computing facilities and inviting us for a research visit. This work was supported by the FITWeltweit program of the German Academic Exchange Service (DAAD) as well as the German Research Foundation (DFG grant GI-1259-1).

References

1. Radu Hambasan and Michael Kohlhase. Faceted search for mathematics. In Ralph Bergmann, Sebastian Görg, and Gilbert Müller, editors, *Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, Trier, Germany, October 7-9, 2015.*, volume 1458 of *CEUR Workshop Proceedings*, pages 33–44. CEUR-WS.org, 2015.
2. Mohamed Kurdi. *Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications*, volume 2. John Wiley & Sons, 2017.
3. Robert Pagel and Moritz Schubotz. Mathematical language processing project. In Matthew England, James Davenport, Andrea Kohlhase, Michael Kohlhase, Paul Libbrecht, Walther Neuper, Pedro Quaresma, Alan Sexton, Petr Sojka, Josef Urban, and Stephen Watt, editors, *Joint Proceedings of the MathUI, OpenMath and ThEdu Workshops and Work in Progress track at CICM co-located with Conferences on Intelligent Computer Mathematics (CICM 2014), Coimbra, Portugal, July 7-11, 2014.*, volume 1186 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.
4. Moritz Schubotz and Gabriel Wicke. Mathoid: Robust, scalable, fast and accessible math rendering for wikipedia. In Stephen Watt, James Davenport, Alan Sexton, Petr Sojka, and Josef Urban, editors, *Intelligent Computer Mathematics - International Conference, CICM 2014, Coimbra, Portugal, July 7-11, 2014. Proceedings*, volume 8543 of *Lecture Notes in Computer Science*, pages 224–235. Springer, 2014.
5. Moritz Schubotz, Andre Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard Cohl, and Bela Gipp. Improving the representation and conversion of mathematical formulae by considering their textual context. *arXiv preprint arXiv:1804.04956*, 2018.
6. Leslie Lamport. *LaTeX - A Document Preparation System: User's Guide and Reference Manual, Second Edition*. Pearson / Prentice Hall, 1994.
7. Michael Kohlhase. Using as a semantic markup format. *Mathematics in Computer Science*, 2(2):279–304, 2008.
8. Howard Cohl, Marjorie McClain, Bonita Saunders, Moritz Schubotz, and Janelle Williams. Digital repository of mathematical formulae. In *Conference on Intelligent Computer Mathematics (CICM), Coimbra, Portugal*, pages 419–422, 2014.
9. Howard Cohl, Moritz Schubotz, Marjorie McClain, Bonita Saunders, Cherry Zou, Azeem Mohammed, and Alex Danoff. Growing the digital repository of mathematical formulae with generic sources. volume 9150, pages 280–287, 2015.
10. OpenMath Society. Openmath. `http://www.openmath.org`. Accessed: 2018-05-09.
11. Michael Kohlhase. *OMDoc - An Open Markup Format for Mathematical Documents [version 1.2]*, volume 4180 of *Lecture Notes in Computer Science*. Springer, 2006.
12. EberhardHilf, Michael Kohlhase, and Heinrich Stamerjohanns. Capturing the content of physics: Systems, observables, and experiments. In Jonathan Borwein and William Farmer, editors, *Mathematical Knowledge Management, 5th International Conference, MKM 2006, Wokingham, UK, August 11-12, 2006, Proceedings*, volume 4108 of *Lecture Notes in Computer Science*, pages 165–178. Springer, 2006.
13. Pavi Sandhu. *The MathML Handbook CD-ROM*. Charles River Media, Inc., 2002.
14. Denny Vrandečić. Wikidata: a new platform for collaborative data collection. In Alain Mille, Fabien Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 1063–1064. ACM, 2012.

15. Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. Introducing wikidata to the linked data web. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandecic, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 50–65. Springer, 2014.
16. Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.
17. Sebastian Schaffert (Google). Wikidata:primary sources tool. `https://www.wikidata.org/wiki/Wikidata:Primary_sources_tool`. Accessed: 2018-04-11.
18. Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web*, pages 1419–1428. International World Wide Web Conferences Steering Committee, 2016.
19. Lydia Pintscher. Wikidata statement. `https://commons.wikimedia.org/wiki/File:Wikidata_statement.svg`. Accessed: 2018-05-09.
20. Moritz Schubotz, Norman Meuschke, Thomas Hepp, Howard Cohl, and Bela Gipp. VMEXT: A visualization tool for mathematical expression trees. In *CICM, Edinburgh, UK*, pages 340–355, 2017.
21. Moritz Schubotz, Abdou Youssef, Volker Markl, Howard Cohl, and Jimmy Li. Evaluation of similarity-measure factors for formulae based on the NTCIR-11 math task. In Noriko Kando, Hideo Joho, and Kazuaki Kishida, editors, *11th NTCIR, Tokyo, Japan*, 2014.
22. Moritz Schubotz, Abdou Youssef, Volker Markl, and Howard Cohl. Challenges of mathematical information retrieval in the NTCIR-11 math wikipedia task. In Ricardo Baeza-Yates, Mounia Lalmas, Alistair Moffat, and Berthier Ribeiro-Neto, editors, *Special Interest Group on Information Retrieval (SIGIR), Santiago, Chile*, pages 951–954, 2015.
23. Daniel Lozier. NIST digital library of mathematical functions. *Ann. Math. Artif. Intell.*, 38(1-3):105–119, 2003.
24. Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. NTCIR-11 math-2 task overview. In *Proc. 11th NTCIR Conf. on Evaluation of Information Access Technologies, Tokyo, Japan*, 2014.
25. Bruce Miller. `LaTeXML`: A LaTeX to XML converter. `http://dlmf.nist.gov/LaTeXML/`. Accessed: 2018-05-09.
26. Deyan Ginev, Heinrich Stamerjohanns, and Michael Kohlhase. The latexml daemon: Editable math on the collaborative web. In *LWA 2011, Magdeburg, Germany*, pages 255–256, 2011.
27. Bruce Miller and Abdou Youssef. Technical aspects of the digital library of mathematical functions. *Ann. Math. Artif. Intell.*, 38(1-3):121–136, 2003.
28. Howard Cohl, Marjorie McClain, Bonita Saunders, Moritz Moritz Schubotz, and Janelle Williams. Digital repository of mathematical formulae. In *Intelligent Computer Mathematics*, pages 419–422. Springer, 2014.