Generating OpenMath Content Dictionaries from Wikidata

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

Abstract

OpenMath content dictionaries are collections of mathematical symbols. Traditionally, content dictionaries are handcrafted by experts. The OpenMath specification requires a name and a textual description in English for each symbol in a dictionary. In our recently published MathML benchmark (MathMLBen), we represent mathematical formulae in Content MathML referring to Wikidata as the knowledge base for the grounding of the semantics. Based on this benchmark, we present an OpenMath content dictionary, which we generated automatically from Wikidata. Our Wikidata content dictionary consists of 330 entries. We used the 280 entries of the benchmark MathMLBen, as well as 50 entries that correspond to already existing items in the official OpenMath content dictionary entries. To create these items, we proposed the Wikidata property P5610. With this property, everyone can link OpenMath symbols and Wikidata items. By linking Wikidata and OpenMath data, the multilingual community maintained textual descriptions, references to Wikipedia articles, external links to other knowledge bases (such as the Wolfram Functions Site) are connected to the expert crafted OpenMath content dictionaries. Ultimately, these connections form a new content dictionary base. This provides multilingual background information for symbols in MathML formulae.

1 Introduction and Prior Works

Traditionally, mathematical formulae occur in a textual or situational context. Human readers infer the meaning of formulae from their layout and the context. An essential task in mathematical information retrieval (MathIR) is to mimic parts of this process to automate MathIR tasks. There are different approaches to evaluate the effectiveness of MathIR tasks. The naïve approach is to

1. identify typical information needs,
2. use existing corpora that embed the formula layout in textual content, and
3. evaluate the effectiveness of IR system for the defined information needs with human domain experts.

The advantage of this approach is that it evaluates the practical relevance of real-world applications. The disadvantages are the following:

1. it is expensive, due to the considerable evaluation effort,
2. it is tailored to the information needs of the task, and

In: O. Hasan, J. Davenport, M. Kohlhase (eds.): Proceedings of the 29th OpenMath Workshop, Hagenberg, Austria, 13-Aug-2018, published at http://ceur-ws.org

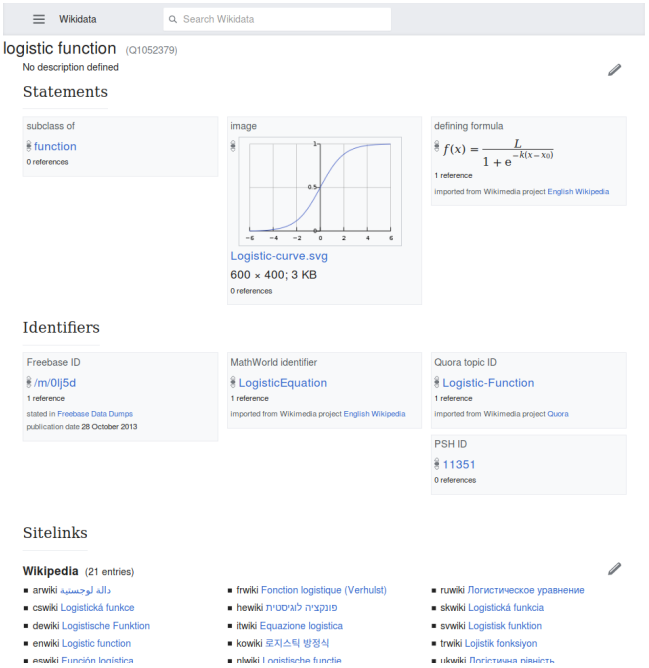


Figure 1: Wikidata entry for the logistic function.

3. it is difficult to identify the reasons for the limited performance of a system.

As an alternative, we suggest breaking down complex MathIR tasks into several subtasks. The first subtask is to convert formulae from their source representation to a machine-readable format which describes the semantics. With our MathMLBen project [11, 13], we created an open gold standard for this task. MathMLBen provides a list of over 300 formulae and their textual contexts. The example formulae include formulae that were used in the NTCIR search tasks [1, 2, 3], as well as formulae from the DLMF [5] and DRMF [4]. For all formulae, the original LaTeX representation, a corrected LaTeX form (that corrects typographic errors in the layout), and a semantic LaTeX interpretation is given. From the semantic LaTeX representation, we generate parallel content and presentation MathML markup using LaTeXML [8]. We consider the generated MathML as a first version of a machine-readable format which describes the semantics and use it to measure the effectiveness of different systems for the task described above [13]. However, this machine readable format is not yet compatible with the OpenMath standard. In Section 2, we describe how we generated a first version of the Wikidata content dictionary `wikidata.ocd` contains all symbols occurring in our gold standard. However, not all symbols of our gold standard were associated with Wikidata items, but with standard OpenMath symbols. In Section 3 we experiment with exchanging those standard symbols with Wikidata items and analyse the effects. Finally, Section 4 concludes the paper and points out future works.

2 The first version of a Wikidata content dictionary

As introduced earlier, a key feature of MathMLBen are special LaTeX(ML) macros [11]. Those macros link `csymbol`-elements in the MathML to entries in Wikidata. For example, one element of our gold standard includes the logistic function $f(x) = \frac{L}{1+e^{-k(x-x_0)}}$. We did not find a symbol for the logistic function in the OpenMath content dictionaries[1]. However, articles regarding the logistic function can be found in many different languages in Wikipedia. Moreover, the Wikidata item Q1052379 connects all these Wikipedia articles. This item was manually improved by 19 users (excluding bots). Through these improvements, additional data such as properties, relations to other items and external identifiers were added. Figure 1 shows a screenshot of the item, including statements, external identifiers, and links to Wikipedia articles as well as other Wiki projects such as Wikisource and Wikiversity. In our gold standard, we used the semantic LaTeX macro[2] `\wf{Q1052379}{f}`

[1] A list of all OpenMath symbols is available from `https://www.openmath.org/symbols/`.

[2] The difference between the macros `\w` and `\wf` is that `\wf` associates the role function with a symbol. For example, the first invisible operator in $f(a+b)$ is interpreted as function application rather than multiplication, if `\wf` is used [11].

to encode that LaTeXML should treat the symbol f as `<csymbol cd="wikidata">Q1052379</csymbol>` [11]. However, the content dictionary `wikidata` that the `csymbol` element refers to with its `cd` attribute did not exist.

Currently, there are more than 49 million items in Wikidata. Not all items are part of mathematical formulae. Thus, creating a Wikidata content dictionary based on all items would be impractical. In this paper, we present a Wikidata content dictionary that includes all the 280 entries used in MathMLBen. We call our content dictionary `wikidata.ocd`. It can be downloaded from `https://cd.formulasearchengine.com/wikidata.ocd`.

```
1362  <CDDefinition>
1363    <Name>Q1052379</Name>
1364    <Role>application</Role>
1365    <Description>
1366      logistic function
1367      https://en.wikipedia.org/w/index.php?title=Logistic+function
1368
1369      This description was generated from
1370      https://www.wikidata.org/w/index.php?oldid=648452086
1371    </Description>
1372  </CDDefinition>
```

Listing 1: Definition for the logistic function in wikidata.ocd.

Listing 1 shows the content dictionary entry for the symbol logistic function (Q1052379). All 280 symbols follow the same pattern. This is because they generated them with MathTools [6] using the Wikidata item numbers given in the `csymbol`-elements in MathMLBen. The description includes the English label of the Wikidata entry (line 1366), the English description (not available in Listing 1), a link to the English Wikipedia article (line 1367) and finally a static link to the version of the Wikidata item that was used to create the content dictionary entry (line 1370). While this description is currently in English, the language is only a configuration parameter in our tool. Theoretically, any other language could be used. However, the success depends on the number of available community-maintained texts in that language (cf. Section 3).

Our content dictionary `wikidata.ocd` improves the standard compliance of the MathMLBen gold standard. It provides a content dictionary for third party MathML processing software that can read user-contributed content dictionaries without requiring a special implementation to fetch data from Wikidata. This improves the standard compliance the MathMLBen gold standard.

3 Using Wikidata as cdbase

After having discussed how to automatically derive a content dictionary from a set of Wikidata items, we discuss how to create a cdbase that contains the standard OpenMath symbols from Wikidata in this chapter.

As Figure 2 shows, the nature of content dictionaries and Wikipedia pages (here in English) is different. While the CD description is brief in human-readable content, the Wikipedia page shows a lot of human-readable information. On the other hand, there are formal mathematical properties that are hard to extract from the Wikipedia page. Consequently, we analyse the strength and weaknesses of both approaches in the following. However, before doing so, we need to map entries in Wikidata to OpenMath. We therefore proposed a new property in Wikidata (P5610) of type external identifier which was approved on August 9th 2018. This identifier, labeled OpenMath ID, allows one to refer to OpenMath symbols from within Wikidata. That way, we connect both communities, Wikidata and OpenMath. Everyone (even without a Wikidata account) is now able to create new mappings.

The new property P5610 is of type string and has three constraints: a single-value, a unique value, and a regexp filter `([a-z]+[0-9]*)\#([a-z_]+)` constraint. These constraints prevent common mistakes. Table 1 lists the 50 standard OpenMath symbols that occur in the MathMLBen project. We uploaded these manually created mappings on August 10th, 2018 to Wikidata. Consequently, we now have the opportunity to describe all symbols that occur in the gold standard in terms of Wikidata without referring to any OpenMath definitions. One can now create a *redirect service* which redirects traditional MathML and OpenMath IDs such as `<plus>` which correspond to `arith1#plus` to the associated Wikidata entry (e.g., Q32043 for plus). According to the MatML standard (Section 4.2.3.2) the URI of a definition can be given as `URI = cdbase + '/' + cd-name + '#' +`

symbol-name. The SPARQL interface query.wikidata.org allows one to find an item that is associated with the last part of the url (cd-name#symbol-name). For instance, the query for arith1#plus reads SELECT ?x WHERE { ?x wdt:P5610 'arith1#plus'.} which returns the URL http://www.wikidata.org/entity/Q32043, the Wikidata item for plus. To materialise the results, we used the method described in Section 2 to generate CDDefintions for the 50 standard symbols.

```
264  <CDDefinition>
265    <Name>Q32043</Name>
266    <Role>application</Role>
267    <Description>
268      addition
269      arithmetic operation of adding (augend+addend=summand+summand=sum,
           total). (Add, Sum, Plus, Increase, Total)
270      https://en.wikipedia.org/w/index.php?title=Addition
271      See also
272      https://www.openmath.org/cd/arith1#plus
273
274      This description was generated from
275      https://www.wikidata.org/w/index.php?oldid=720254931
276    </Description>
277  </CDDefinition>
```

Listing 2: Definition for the + symbol in wikidata.ocd.

Listing 2 shows an entry in wikidata.ocd that corresponds to the Wikidata item plus. The description section contains more information than Listing 1. Line 269 is the English description from the item that represents the addition in Wikidata. Moreover, line 272 links to the definition form of arith1#plus from the OpenMath content dictionary.

For the remainder of the section, we discuss the differences between traditional OpenMath symbol definition entries and Wikidata generated symbol definitions. Our Wikidata content dictionary contains 330 symbols in a single content dictionary. In contrast, there are 289 official OpenMath symbols divided into 38 content dictionaries. 247 OpenMath symbols have a role attribute (application 193, constant 39, binder 3, semantic-attribution 2, error 3). While we did research on identifying Wikidata items as numerical constants [14], this information is not included in the current version of wikidata.ocd. Moreover, the official OpenMath content dictionaries contain 149 examples, 180 formal mathematical properties (FMP), and 179 commented mathematical properties (CMP). Currently, wikidata.ocd does not contain any of the aforementioned features. Due to the lack of time, the MathMLBen data has not been converted to the OpenMath XML format, which would be required to create examples. Deriving reasonable CMP or FMP from Wikidata requires semantic enhancement of the defining formula statement which are currently only available in presentation form. The description field in the official OpenMath content dictionaries is on average 131 words, as compared to 212 words (including 14 words for the reference to the source) in wikidata.ocd. While Wikidata items are not divided into a structure comparable to content dictionaries, they have hierarchical relations such as the instance of (P31) relation. As displayed in Table 1, the instance of relation is not modelled consistently. We hypothesise that this is typical for corpora which emerged from community interactions. Finally, the symbol names in the standard OpenMath dictionaries are easier to remember for English speakers. Therefore, using IDE or smart editors is a prerequisite to work with Wikidata items conveniently. Otherwise, the long numeric item identifiers are hard to read and to remember. To support this purpose, we released the node module codemirror-wikidata [12], which provides autocompletion based on the description rather than on the numeric values.

4 Conclusions and Future Works

In this paper, we released a first version of the Wikidata content dictionary wikidata.ocd to the public. It contains all the symbols used in the MathMLBen open gold standard. Moreover, it contains 50 entities that correspond to standard OpenMath symbols. Furthermore, we introduced the new Wikidata property P5610 and described how it can be used to create an alternative cdbase. Also, we compared symbol descriptions

that were generated automatically from Wikidata to the manually crafted OpenMath symbol definitions. While multilingualism and links to Wikipedia might be considered as an advantage of the Wikidata cdbase, many other formal aspects such as structure and type information are currently better modeled in the traditional OpenMath content dictionaries. On the other hand, Wikidata has far more items that could be possibly used as symbol definitions.

Future research should investigate how the missing formal information in `wikidata.ocd` can be automatically extracted. If there was a mechanism to generate content dictionaries from Wikidata that have the same formal quality as the current OpenMath content dictionaries, a good foundation for CD extension, based on Wikidata, would ease the expansion of the OpenMath standard. Another promising research direction is to better understand how information given in distributed data sources can be connected using alignments [7, 9, 10].

Acknowledgments

We thank the Wikimedia Foundation and Wikimedia Deutschland for providing cloud computing facilities and for providing office space for us. 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). The author would like to thank Howard Cohl for constructive criticism of the manuscript.

References

[1] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. "NTCIR-10 Math Pilot Task Overview". In: *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies* (2013).

[2] Akiko Aizawa et al. "NTCIR-11 Math-2 Task Overview". In: *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014*. National Institute of Informatics (NII), 2014.

[3] Akiko Aizawa et al. "NTCIR-12 Math-3 Task Overview". In: *NTCIR*. National Institute of Informatics (NII), 2016.

[4] Howard S. Cohl et al. "Growing the Digital Repository of Mathematical Formulae with Generic LaTeX Sources". In: *Proc. CICM*. Ed. by Manfred Kerber et al. Vol. 9150. Springer, 2015.

[5] F.W.J. Olver et al., eds. *NIST Digital Library of Mathematical Functions*. `http://dlmf.nist.gov/`, Release 1.0.14 of 2017-12-21. F.W.J. Olver, A.B. Olde Daalhuis, D.W. Lozier, B.I. Schneider, R.F. Boisvert, C.W. Clark, B.R. Miller and B.V. Saunders, eds. 2017.

[6] André Greiner-Petter et al. "MathTools: An Open API for Convenient MathML Handling". In: *11th Conference on Intelligent Computer Mathematics CICM, RISC, Hagenberg, Austria*. RISC, Hagenberg, Austria, Aug. 2018.

[7] Cezary Kaliszyk et al. "A Standard for Aligning Mathematical Concepts". In: *Joint Proceedings of the FM4M, MathUI, and ThEdu Workshops, Doctoral Program, and Work in Progress at the Conference on Intelligent Computer Mathematics 2016 co-located with the 9th Conference on Intelligent Computer Mathematics (CICM 2016), Bialystok, Poland, July 25-29, 2016*. Ed. by Andrea Kohlhase et al. Vol. 1785. CEUR-WS.org, 2016.

[8] Bruce Miller. *LaTeXML: A LaTeX to XML/HTML/MathML Converter*. Web Manual at `http://dlmf.nist.gov/LaTeXML/`. Seen 2018.

[9] Dennis Müller et al. "Alignment-based Translations Across Formal Systems using Interface Theories". In: *Proceedings of the Fifth Workshop on Proof eXchange for Theorem Proving, PxTP 2017, Brasilia, Brazil, 23-24 September 2017*. Ed. by Catherine Dubois and Bruno Woltzenlogel Paleo. Vol. 262. 2017.

[10] Dennis Müller et al. "Classification of Alignments Between Concepts of Formal Mathematical Systems". In: *Proc. CICM*. Ed. by Herman Geuvers et al. Vol. 10383. Springer, 2017.

[11] Philipp Scharpf, Moritz Schubotz, and Bela Gipp. "Representing Mathematical Formulae in Content MathML using Wikidata". In: *Proceedings of the 3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2018) co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018*. Ed. by Philipp Mayr, Muthu Kumar Chandrasekaran, and Kokil Jaidka. Vol. 2132. CEUR-WS.org, 2018.

Table 1: Standard OpenMath IDs, their corresponding Wikidata labels, and the Wikidata 'instance' of' relation.

Wikidata Label	**OpenMath ID**	**Instance of**
absolute value	arith1#abs	piecewise function, even function, idempotent function
division	arith1#divide	binary operation
greatest common divisor	arith1#gcd	function
subtraction	arith1#minus	binary operation, operation
addition	arith1#plus	binary operation
exponentiation	arith1#power	operation
nth root	arith1#root	type of mathematical function, algebraic function
sum	arith1#sum	mathematical expression
multiplication	arith1#times	binary operation
opposite number	arith1#unary_minus	
derivative	calculus1#diff	unary operation, mathematical concept
Lambda expression	fns1#lambda	Wikimedia disambiguation page
function composition	fns1#left_compose	operator, operation
gamma function	hypergeo0#gamma	function
factorial	integer1#factorial	function
range	interval1#interval_oo	part
limit	limit1#limit	mathematical concept
0	limit1#null	integer, Fibonacci number, triangular number, automorphic number, even number, non-negative integer, 0 number class, non-positive integer
determinant	linalg1#determinant	invariant
matrix	linalg2#matrix	array data structure, tensor
row vector	linalg2#matrixrow	row and column vectors
vector	linalg2#vector	tensor
list	list1#list	creative work
logical conjunction	logic1#and	logical connective, boolean function
equivalence relation	logic1#equivalent	transitive relation, symmetric relation, reflexive relation
e	nums1#e	real number, transcendental number, irrational number
imaginary unit	nums1#i	square root, mathematical constant, Gaussian integer, imaginary number
infinity	nums1#infinity	mathematical concept
pi	nums1#pi	real number, transcendental number, mathematical constant, irrational number
approximation	relation1#approx	relation, estimation
equality	relation1#eq	equivalence relation, partial order
greater or equal to	relation1#geq	inequation
greater than	relation1#gt	inequation
less or equal to	relation1#leq	inequation
less than	relation1#lt	inequation
not equals to sign	relation1#neq	inequality sign
subset	set1#in	binary relation, subclass
intersection	set1#intersect	binary operation, set operation
set	set1#set	Wikidata metaclass, Wikidata metaclass, Wikidata metaclass, class (set theory), formalization, collection
arccosine	transc1#arccos	inverse trigonometric function, decreasing function
arctangent	transc1#arctan	inverse trigonometric function, increasing function
cosine	transc1#cos	trigonometric function, even function
hyperbolic cosine	transc1#cosh	hyperbolic function, even function
natural exponential function	transc1#exp	exponential function, type of mathematical function
natural logarithm	transc1#ln	type of mathematical function, logarithm
logarithm	transc1#log	type of mathematical function, type of mathematical function, multivalued function, elementary transcendental function
sine	transc1#sin	trigonometric function, odd function
hyperbolic sine	transc1#sinh	hyperbolic function, odd function
tangent	transc1#tan	trigonometric function
hyperbolic tangent	transc1#tanh	hyperbolic function

[12] Moritz Schubotz. "VMEXT2: A Visual Wikidata aware Content MathML Editor". In: *Joint Proceedings of the CME-EI, FMM, CAAT, FVPS, M3SRD, OpenMath Workshops, Doctoral Program and Work in Progress at the Conference on Intelligent Computer Mathematics 2018 co-located with the 11th Conference on Intelligent Computer Mathematics (CICM 2018)*. Ed. by Osman Hasan et al. 2018.

[13] Moritz Schubotz et al. "Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context". In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*. Ed. by Jiangping Chen et al. ACM, 2018.

[14] Moritz Schubotz et al. "Introducing MathQA - A Math-Aware Question Answering System". In: *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Workshop on Knowledge Discovery*. Fort Worth, USA, June 2018.

gamma

Description:

Euler's gamma function

Commented Mathematical property (CMP):
gamma(z)=\int_0^{+\infty} t^{z-1} e^{-z} dt (Re(z)>0)

Formal Mathematical property (FMP):

OpenMath XML (source)
Strict Content MathML
Prefix
Popcorn
Rendered Presentation MathML

$$\text{real}(z) > 0 \Rightarrow \text{gamma}(z) = \int_0^{\infty} t^{(z-1)} e^{(-z)} dt$$

Example:
gamma(n) = (n-1)! (n \in N)

OpenMath XML (source)
Strict Content MathML
Prefix
Popcorn
Rendered Presentation MathML

$$n \in \mathbb{N} \Rightarrow \text{gamma}(n) = (n-1)!$$

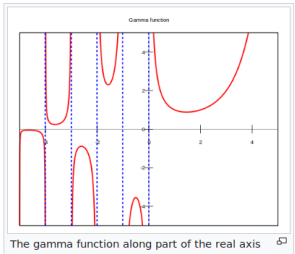
Signatures:
sts

Gamma function

From Wikipedia, the free encyclopedia

In mathematics, the **gamma function** (represented by Γ, the capital Greek alphabet letter gamma) is an extension of the factorial function, with its argument shifted down by 1, to real and complex numbers. If n is a positive integer,

$$\Gamma(n) = (n-1)!$$



The gamma function along part of the real axis

The gamma function is defined for all complex numbers except the non-positive integers. For complex numbers with a positive real part, it is defined via a convergent improper integral:

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$$

This integral function is extended by analytic continuation to all complex numbers except the non-positive integers (where the function has simple poles), yielding the meromorphic function we call the gamma function. It has no zeroes, so the reciprocal gamma function $1/\Gamma(z)$ is a holomorphic function. In fact the gamma function corresponds to the Mellin transform of the negative exponential function:

$$\Gamma(z) = \{\mathcal{M} e^{-x}\}(z)$$

The gamma function is a component in various probability-distribution functions, and as such it is applicable in the fields of probability and statistics, as well as combinatorics.

Figure 2: OpenMath content dictionary entry for hypergeo0#gamma (left) and the corresponding Wikipedia article (right).