Content dictionary description: select symbols from Chapter 9 of the KLS dataset in the DRMF

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

Howard S. Cohl
Applied and Computational Mathematics Division,
National Institute of Standards and Technology,
Gaithersburg, Maryland, U.S.A.,
howard.cohl@nist.gov

Abstract

The NIST Digital Repository of Mathematical Formulae (DRMF) uses parallel MathML markup to express mathematical formulae. However currently, the content MathML symbols do not point to a content dictionary. In this paper, we describe the naming convention and creation process for mathematical symbols in the DRMF. Moreover, we publish a first, experimental version of a content dictionary that contains only a few manually selected symbols from Chapter 9 of the KLS dataset.

1 Introduction

The NIST Digital Repository of Mathematical Formulae [Coh+14] is an online compendium of mathematical formulae data for orthogonal polynomials and special functions. The initial formula data was obtained from the NIST Digital Library of Mathematical Functions [Olv+17]. DLMF uses semantic macros [Mil16], which are converted by LaTeXML [Mil10] to parallel markup including content MathML. We refer to the symbols generated by LaTeXML as implicit content dictionaries [Sch16]. However, the DRMF does not only contain formula-data from the DLMF. In 2015, we started to extend our library with sources that do not use the DLMF macro schema [Coh+15]. In this process, we introduced new symbols and associated LaTeX macros. In this content dictionary, we release an initial version of the symbols introduced while processing Chapter 9 of the KLS dataset [KLS10]. Currently, we are extending our work on Mathematical Language Processing [Sch+16; Coh+17]. A content dictionary is beneficial for this task, since it specifies the semantics in a standard conforming way.

2 Symbol Naming Conventions

We use the same naming convention for LaTeX macro names and the content dictionary symbols. The symbols follow simple naming conventions that we will discuss in this section. The goal for our naming convention is to have short, memorizable and unambiguous names. Currently, the DRMF website uses 689 custom LaTeX macros.

In: H. Geuvers, M. England, O. Hasan, F. Rabe, O. Teschke (eds.): Proceedings of the OpenMath Workshop, Edinburgh, UK, 17-July-2017, published at http://ceur-ws.org

Table 1: Abbreviations used in the macros for Chapter 9 in the KLS dataset.

Abbreviation	Normal form
cts	continuous
norm	normalized

Their length varies between 1 and 26 characters and the median length is 8 characters. The name is derived from a short symbol description using the following conventions:

concatenation Individual elements are concatenated without spaces. Camel case is not used.

capitalization Individual's names are capitalized. Mathematical identifiers such as q and standard macros such as log, Log are used as normal capitalization. Properties such as continuous, monic, generalized, normalized are written in lower case.

abbreviations Only a limited set of abbreviations is used. See Table 1.

3 New Symbols

Extensibility is one of the key goals of the DRMF project. Consequently, we developed a mechanism to update our custom macros or add new macros. Adding a macro follows this procedure:

1. Identify a symbol and its definition in the literature.
2. Describe the symbol with a short name but without ambiguities. This description can include properties, individuals names (cf. Section 2) as well as presentational features.
3. Derive the symbol name based on the description:
   - Remove whitespace.
   - Replace presentational features with their verbalization, e.g., `$\tilde p$` $\mapsto$ ptilde.
   - Replace abbreviations.
   - Remove superflous words like "polynomial". (Our convention to drop the word "polynomial" for this dataset is because these are the most abundant objects.)
4. Add a definition or a longer description of the symbol.
5. Add references to the literature.
6. Update the DRMF LaTeX style file locally and define calling sequences, that is, the different numbers of parameters which can be used for the same macro name. For example the macro `normWilsonWtilde` can be called with 5 or with only one parameters/arguments.
7. Update the macro definitions for the DRMF LaTeXML server.
8. Add a Wikipage in the namespace "Definition" with the name of the symbol that:
   - describes the calling sequences;
   - lists the definition;
   - shows the rendering; and
   - gives references to the literature.
9. Link the new wikipage on the symbol definition overview.

To rename a symbol, the DRMF style file has to be adjusted and to be uploaded again to the server. The definition page has to be moved manually. In contrast, all occurences of the macro can be replaced using the search and replace MediaWiki plugin. This will trigger automated updates for the MathSearch index [Sch17].

4 Outlook and Future Works

Currently, the content dictionary is a proof of concept. For the future, we are planning to automate the creation of content dictionaries for other chapters and sources. Chapter 14 of KLS is already in preparation.

References

[Coh+14] H. S. Cohl, M. A. McClain, et al. "Digital Repository of Mathematical Formulae". In: *Intelligent Computer Mathematics, Lecture Notes in Artificial Intelligence 8543*. Ed. by S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka, J. Urban. Vol. 8543. LNCS. Springer, 2014, pp. 419–422. DOI: 10.1007/978-3-319-08434-3_30.

[Coh+15] H. S. Cohl, M. Schubotz, et al. "Growing the Digital Repository of Mathematical Formulae with Generic LaTeX Sources". In: *Intelligent Computer Mathematics, Lecture Notes in Artificial Intelligence 9150*. Ed. by M. Kerber, J. Carette, et al. Vol. 9150. LNCS. Springer, 2015, pp. 280–287. DOI: 10.1007/978-3-319-20615-8_18.

[Coh+17] H. S. Cohl, M. Schubotz, et al. "Semantic Preserving Bijective Mappings of Mathematical Formulae between Document Preparation Systems and Computer Algebra Systems". In: *10th Conference on Intelligent Computer Mathematics*. 2017.

[KLS10] R. Koekoek, P. A. Lesky, and R. F. Swarttouw. *Hypergeometric Orthogonal Polynomials and Their q-Analogues*. Ed. by T. H. Koornwinder. Springer Monographs in Mathematics. With a foreword by Tom H. Koornwinder. Berlin [u.a.]: Springer-Verlag, 2010, pp. xx+578. XIX, 578. DOI: 10.1007/978-3-642-05014-5.

[Mil10] B. Miller. *LaTeXML: A LaTeX to XML Converter*. 2010. URL: http://dlmf.nist.gov/LaTeXML/.

[Mil16] B. Miller. "Drafting DLMF Content Dictionaries". English. Talk presented at the OpenMath workshop of the 9th Conference on Intelligent Computer Mathematics CICM 2016. Bialystok, Poland, July 25, 2016. URL: http://cicm-conference.org/2016/cicm.php?event=&menu=talks#02 (visited on 10/03/2016).

[Olv+17] F. Olver, A. Olde Daalhuis, 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. URL: http://dlmf.nist.gov.

[Sch+16] M. Schubotz, A. Grigorev, et al. "Semantification of Identifiers in Mathematics for Better Math Information Retrieval". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: ACM, 2016, pp. 135–144. DOI: 10.1145/2911451.2911503.

[Sch16] M. Schubotz. "Implicit Content Dictionaries in the NIST Digital Repository of Mathematical Formulae". English. Talk presented at the OpenMath workshop of the 9th Conference on Intelligent Computer Mathematics CICM 2016. Bialystok, Poland, July 25, 2016. URL: http://cicm-conference.org/2016/cicm.php?event=&menu=talks#03 (visited on 10/03/2016).

[Sch17] M. Schubotz. *Augmenting Mathematical Formulae for More Effective Querying & Efficient Presentation*. Epubli Verlag, Berlin, 2017. URL: https://www.epubli.de/preview/publication/64471.

Listing of the Content Dictionary

```
<CD xmlns="http://www.openmath.org/OpenMathCD">
    <CDComment>

        This document is distributed in the hope that it will be useful,
        but WITHOUT ANY WARRANTY; without even the implied warranty of
        MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

        The copyright holder grants you permission to redistribute this
        document freely as a verbatim copy. Furthermore, the copyright
        holder permits you to develop any derived work from this document
        provided that the following conditions are met.
        a) The derived work acknowledges the fact that it is derived from
        this document, and maintains a prominent reference in the
        work to the original source.
        b) The fact that the derived work is not the original OpenMath
        document is stated prominently in the derived work.  Moreover if
        both this document and the derived work are Content Dictionaries
        then the derived work must include a different CDName element,
        chosen so that it cannot be confused with any works adopted by
        the OpenMath Society.  In particular, if there is a Content
        Dictionary Group whose name is, for example, 'math' containing
        Content Dictionaries named 'math1', 'math2' etc., then you should
        not name a derived Content Dictionary 'mathN' where N is an integer.
        However you are free to name it 'private_mathN' or some such.  This
        is because the names 'mathN' may be used by the OpenMath Society
        for future extensions.
        c) The derived work is distributed under terms that allow the
        compilation of derived works, but keep paragraphs a) and b)
        intact.  The simplest way to do this is to distribute the derived
        work under the OpenMath license, but this is not a requirement.
        If you have questions about this license please contact the OpenMath
        society at http://www.openmath.org.
    </CDComment>

    <CDName>drmf-kls9</CDName>
    <CDBase>https://cd.formulasearchengine.com/</CDBase>
    <CDURL>https://cd.formulasearchengine.com/drmf-kls9.ocd</CDURL>
    <CDReviewDate>2018-07-10</CDReviewDate>
    <CDStatus>experimental</CDStatus>
    <CDDate>2018-07-10</CDDate>
    <CDVersion>0</CDVersion>
    <CDRevision>0</CDRevision>

    <Description>
        This CD defines symbols from KLS Chapter 9.
    </Description>

    <CDDefinition>
        <Name>normWilsonWtilde</Name>
        <Role>application</Role>
        <Description>
            normalized Wilson polynomial ${\tilde W}$
        </Description>
    </CDDefinition>
    <CDDefinition>
        <Name>monicWilson</Name>
        <Role>application</Role>
        <Description>
            monic Wilson polynomial
        </Description>
    </CDDefinition>
    <CDDefinition>
        <Name>monicRacah</Name>
        <Role>application</Role>
        <Description>
            monic Racah polynomial
        </Description>
    </CDDefinition>
    <CDDefinition>
        <Name>normctsHahnptilde</Name>
        <Role>application</Role>
        <Description>
            normalized continuous Hahn polynomial ${\tilde p}$
        </Description>
    </CDDefinition>
</CD>
```

Listing 1: The content dictionary