Growing the Digital Repository of Mathematical Formulae with Generic LaTeX Sources

Howard Cohl*, Moritz Schubotz§, Marje McClain*, Bonita Saunders*, Alex Danoff†, Jimmy Li‡, Jake Migdall§§, Amber Liu§§, Cherry Zou§§, Azeem Mohammed§§, Shraeya Madhu§§

*Applied and Computational Mathematics Division, NIST, Gaithersburg, Maryland, U.S.A.
§Database Systems and Information Management Group, Technische Universität Berlin, Germany
†Thomas S. Wootton High School, Rockville, MD
‡ Richard Montgomery High School, Rockville, MD
§§Poolesville High School, Poolesville, MD

CICM 2015 Project

July 17, 2015

Table of contents

Collecting information on special functions

- 1964: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables
  - Milton Abramowitz and Irene Stegun
  - 1064 pages (book)
  - definitions, approximations, identities, plots and tables
- 2010: Digital Library of Mathematical Functions
  - NIST Handbook of Mathematical Functions as successor of A&S
  - F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors.
  - 968 pages (printed version), HTML version
  - Links, MathSearch, info boxes
- 2013: Digital Repository of Mathematical Formulae
  - Context-free full semantic information regarding individual formula
  - Additional Sources
  - Community interaction and collaboration

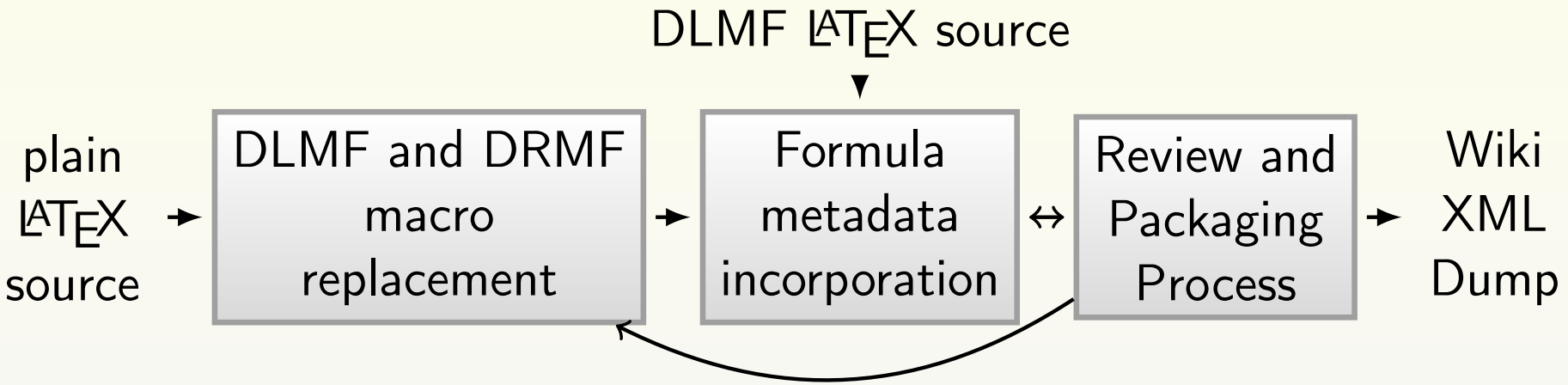
DRMF Goals

The **NIST Digital Repository of Mathematical Formulae** (DRMF) is designed for a mathematically literate audience and should:

1. facilitate interaction among a **community** of mathematicians and scientists interested in **compendia** formulae data for orthogonal polynomials and special functions;
2. be **expandable**, allowing the input of new formulae from the literature;
3. represent the context-free full **semantic** information concerning individual formulas;
4. have a user friendly, consistent, and hyperlinkable viewpoint and authoring **perspective**;
5. contain easily **searchable** mathematics; and
6. take advantage of modern **MathML** tools for easy to read, scalably rendered content driven mathematics.

Current DRMF Implementation

Seeding



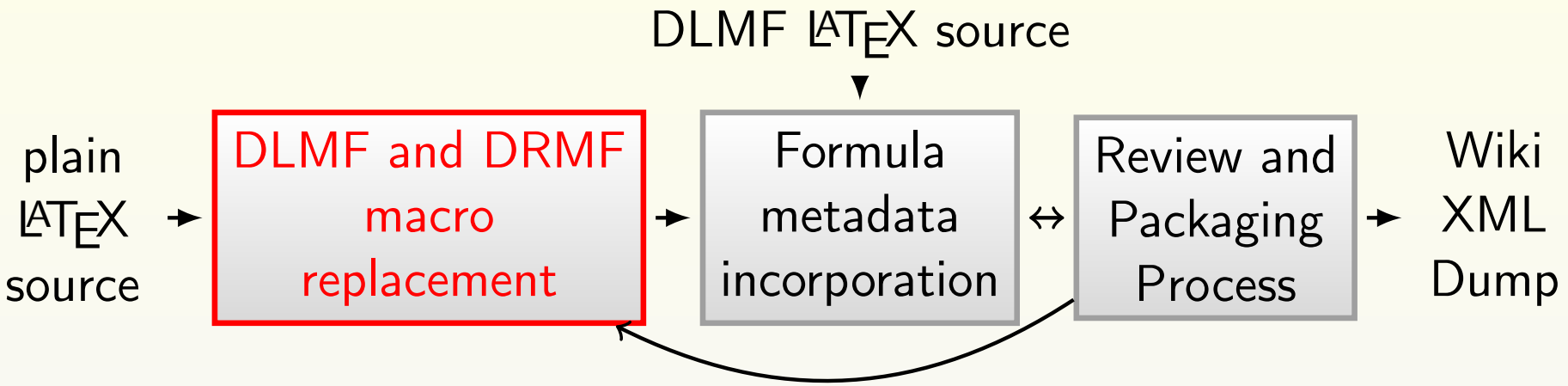
Display

- MediaWiki with Math and MathSearch extension
  - Table of contents pages
  - Lists of formulas pages
  - Formula home pages
  - Definition pages for DRMF macros

First three DRMF project stages

	Stage 1	Stage 2	Stage 3
Started in	2013	2014	2015
Dataset	DLMF, semantic LaTeX	KLS, plain LaTeX	eCF: Mathematica BMP: book images
Semantic enrichment	identify constraints, substitutions, notes, names, proofs, ...	add new semantic macros	image recognition macro suggestion
Technologies	manual review, rule-based approaches	improved rules	natural language processing and machine learning
Number of formula home pages	500	1500	5000
Human time per formula homepage	10 minutes	5 minutes	1 minute
Test corpora contribution	gold standard for constraint and proof detection	gold standard for macro replacement	evaluation metrics

Macro replacement



Semantic LaTeX macros – Some Examples

Some Examples

- Special Functions and Orthogonal polynomials:
  - Trigonometric sine function
  - Euler gamma function
  - Jacobi polynomial
  - little q-Laguerre/Wall polynomial
- Rendered as:
  - $\sin z$, $\Gamma(z)$, $P_n^{(\alpha,\beta)}(x)$, and $p_n(x;a|q)$.
- LaTeX presentations given by
  - `\sin z`, `\Gamma(z)`, `P_n^{(\alpha,\beta)}(x)`, `p_n(x;a|q)`.
- Semantic LaTeX representations
  - `\sin@@{z}`
    `\EulerGamma@{z}`
    `\Jacobi{\alpha}{\beta}{n}@{x}`
    `\littleqLaguerre{n}@{x}{a}{q}`

Semantic macro **breakdown**

685 semantic LaTeX macros

- 395 macros for real and complex valued functions
- 185 macros for polynomials (orthogonal and whatnot)
- 29 macros for integer valued functions
- 24 macros for various operators
- 16 macros for encapsulating semantic information
- 14 macros for quantifiers, set operators and symbols
- 9 macros for sets of numbers
- 5 macros for constants
- 5 macros for linear algebra
- 3 macros for distributions

Semantic LaTeX macro properties: lengths between 1 and 26 characters (median length is 8 characters), individual names capitalized, abbreviations utilized, macro names correspond with object names, Greek numerals

Semantic LaTeX Macro Glossary – csv format

For each macro we store:

- Example Macro calling sequence
- Name of object described by macro
- Object description
- Demonstration rendering of called semantic macros
- Description identifier
- Brief summary and description of calling options
- Link to url giving precise definition

`Glossary.csv` used in generation of symbols lists within Wikitext and for statistical purposes.

Macro Replacements for Generic LaTeX Source Datasets

For the 3 chapters of KLS as well as the KLSadd LaTeX source, **89** semantic macros were replaced a total of **3308** times, currently represented by **456** lines of regular expression code. Currently the six most common replacements are:

- q-Pochhammer symbol – replaced **659** times
- Euler gamma function – replaced **266** times
- q-hypergeometric function – replaced **237** times
- Pochhammer symbol – replaced **205** times
- Racah polynomial – replaced **117** times
- cosine function – replaced **82** times

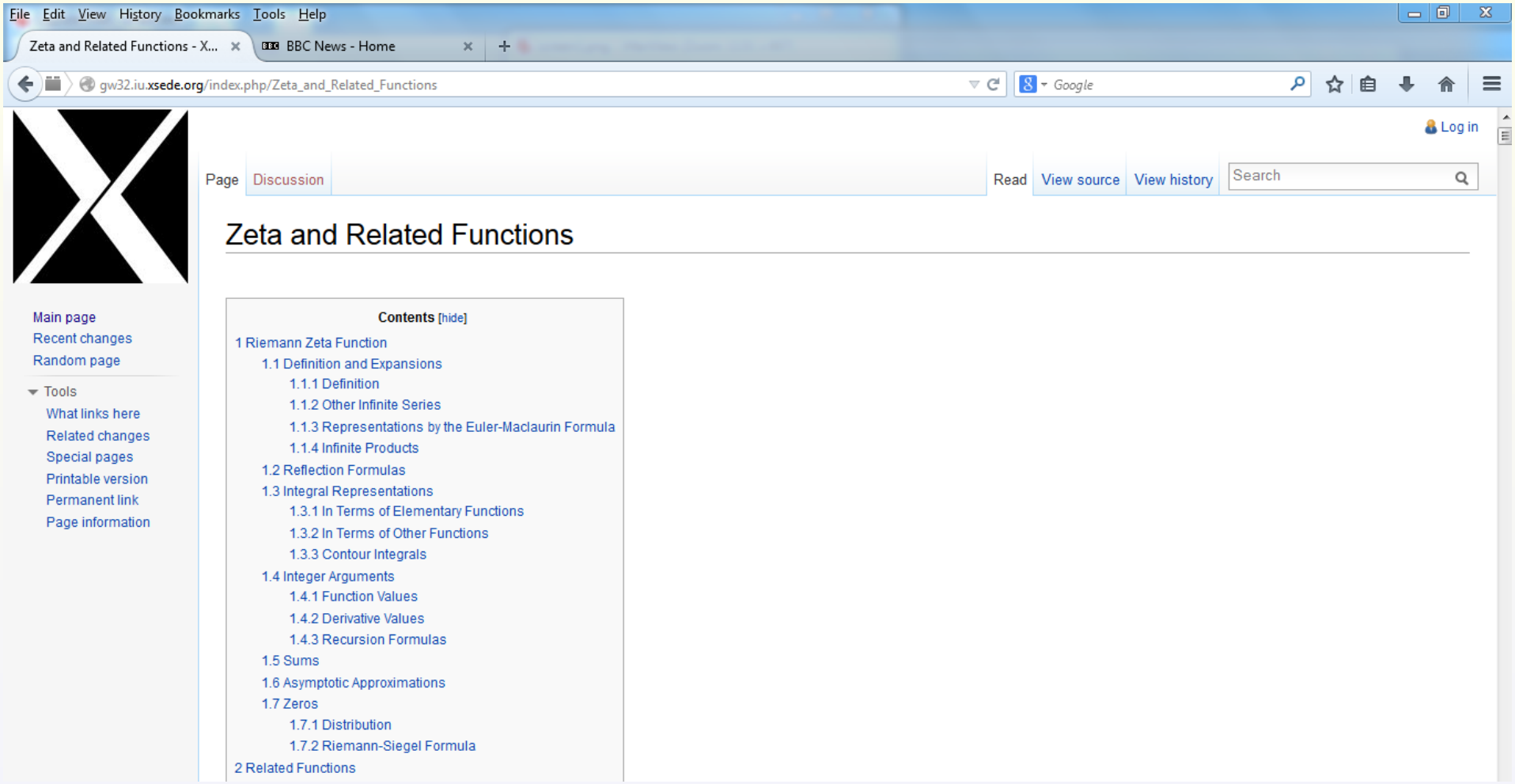
Example of complexity of the problem – KLSadd dataset

- After processing of the LaTeXinput, only formulas remain.
- The only metadata currently extracted (in an automated fashion) for this source are constraint and substitution annotations.
- Much more formula metadata should be extracted from the text (e.g., bibliographic metadata, references to KLS formulae, errata information, formula comments and notes, symmetries in parameters, sketches of proofs, etc.).
- Currently the macro replaced source for Koornwinder's KLS addendum is inserted at the end of each chapter (or section).
- Addendum formulae should actually be exhaustively inserted at their corresponding precise locations within the original KLS LaTeX source.
- There is a need to insert markers within the KLSadd source which identify and orient correct location insertions within the KLS dataset.

Semantic LaTeX Macro Functionality in the MediaWiki Math Extension through LaTeXml

- XSEDE LaTeXml server
- Provides automatic line breaking for long formulas
- Nicely displayed mathematics
- Presentation and Content MathML
- Processes a customized set of 685 semantic LaTeX macros
  - generates MathML – presentation and content
  - currently 685 DLMF and DRMF LaTeX macros
    - 529 DLMF macros
    - 156 DRMF macros

DRMF Zeta and Related Functions Page



DRMF Zeta and Related Functions Page (cont.)

$$\zeta(s) = \frac{(2\pi)^s e^{-s-(\gamma s/2)}}{2(s-1)\Gamma\left(\frac{1}{2}s+1\right)} \prod_{\rho} \left(1-\frac{s}{\rho}\right) e^{s/\rho}$$

Constraint(s): product over zeros ρ of ζ with $\Re\rho > 0$

Reflection Formulas

$$\zeta(1-s) = 2(2\pi)^{-s}\cos\left(\tfrac{1}{2}\pi s\right)\Gamma(s)\zeta(s)$$

Constraint(s): $s \neq 0, 1$

$$\zeta(s) = 2(2\pi)^{s-1}\sin\left(\tfrac{1}{2}\pi s\right)\Gamma(1-s)\zeta(1-s)$$

$$\xi(s) = \xi(1-s)$$

$$\xi(s) = \tfrac{1}{2}s(s-1)\Gamma\left(\tfrac{1}{2}s\right)\pi^{-s/2}\zeta(s)$$

$$(-1)^k\zeta^{(k)}(1-s) = \frac{2}{(2\pi)^s}\sum_{m=0}^{k}\sum_{r=0}^{m}\binom{k}{m}\binom{m}{r}\left(\Re\left(c^{k-m}\right)\cos\left(\tfrac{1}{2}\pi s\right) + \Im\left(c^{k-m}\right)\sin\left(\tfrac{1}{2}\pi s\right)\right)\Gamma^{(r)}(s)\zeta^{(m-r)}(s)$$

Substitution(s): $c = -\ln(2\pi) - \frac{1}{2}\pi i$

Constraint(s): $s \neq 0, 1$ & $k = 1, 2, 3, \ldots$

DLMF macros provide semantic content in formulas

- **DLMF OPSF Macros** via **LaTeXML-server**
  - 546 **semantic DLMF** LaTeX **OPSF** macros
  - **additional** 49 **semantic DRMF LaTeX** macros
- **Objects:** `\sum,\int,\deriv{f}{x},\qderiv[n]{q}@{z}`
- **Constants:** `\expe,\iunit,\cpi,\EulerConstant`
- **Special Functions and Orthogonal Polynomials**

$\Gamma(z)$	`\EulerGamma@{z}`	`http://dlmf.nist.gov/5.30#E1`
$J_\nu(z)$	`\BesselJ{\nu}@{z}`	`http://dlmf.nist.gov/10.2#E2`
$Q^\mu_\nu(z)$	`\LegendreQ[\mu]{\nu}@{z}`:	`http://dlmf.nist.gov/14.3#E7`
$P_n^{(\alpha,\beta)}(x)$	`\JacobiP{\alpha}{\beta}{n}@{x}`	`http://dlmf.nist.gov/18.3#T1.t1.r3`

Formula Home Pages

- Whereas **Wikipedia** and other web authoring tools **manifest notions or descriptions** as **first class objects**, the DRMF does that with **mathematical formulae**.
- **DRMF** provides for each formula, a **formula home page**:
  1. **Rendered description of the formula** (required);
  2. **Constraints** the formula must obey;
  3. **Substitutions** required to understand formula;
  4. **Bibliographic citation** (required);
  5. Open section for **proofs** (required) – *DLMF*;
  6. **List of symbols** and **links** to definitions (required) – *DLMF macros*;
  7. Open section for **notes** – *connections between formulas*; and
  8. Open section for **external links** – *computer generated proofs*;

Sample formula home page

HowardCohl Talk Preferences Watchlist Contributions Log out

Page Discussion Read Edit View history Search

Formula:DLMF:25.4:E5

<< Formula:DLMF:25.4:E4 formula in Zeta and Related Functions Formula:DLMF:25.5:E1 >>

$$(-1)^k\zeta^{(k)}(1-s)=\frac{2}{(2\pi)^s}\sum_{m=0}^{k}\sum_{r=0}^{m}\binom{k}{m}\binom{m}{r}\left(\Re\left(c^{k-m}\right)\cos\left(\tfrac{1}{2}\pi s\right)+\Im\left(c^{k-m}\right)\sin\left(\tfrac{1}{2}\pi s\right)\right)\Gamma^{(r)}(s)\zeta^{(m-r)}(s)$$

Contents [hide]

1 Substitution(s)
2 Constraint(s)
3 Proof
4 Symbols List
5 Bibliography
6 URL links

Substitution(s) [edit]

$c=-\ln(2\pi)-\frac{1}{2}\pi\,\mathbf{i}$

Constraint(s) [edit]

$s\neq 0,1$ & $k=1,2,3,\ldots$

Proof [edit]

We ask users to provide proof(s), reference(s) to proof(s), or further clarification on the proof(s) in this space.

Symbols List [edit]

(-1) : (-1) = \opminus : http://dlmf.nist.gov/5.7.E7
ζ : Riemann zeta function : http://dlmf.nist.gov/25.2#E1
π : the ratio of a circle's circumference to its diameter : http://dlmf.nist.gov/5.19.E4
$\binom{a}{b}$: binomial coefficient : http://dlmf.nist.gov/1.2#E1 http://dlmf.nist.gov/26.3#SS1.p1
$\Re\,a$: real part : http://dlmf.nist.gov/1.9#E2
$\cos$: cosine function : http://dlmf.nist.gov/4.14#E2
$\Im\,a$: imaginary part : http://dlmf.nist.gov/1.9#E2
$\sin$: sine function : http://dlmf.nist.gov/4.14#E1
Γ : Euler's gamma function : http://dlmf.nist.gov/5.2#E1
$\ln$: principal branch of logarithm function : http://dlmf.nist.gov/4.2#E2
$\mathbf{i}$: imaginary unit : http://dlmf.nist.gov/1.9.i

Bibliography [edit]

Equation (5), Section 25.4 of **DLMF**.

URL links [edit]

We ask users to provide relevant URL links in this space.

<< Formula:DLMF:25.4:E4 formula in Zeta and Related Functions Formula:DLMF:25.5:E1 >>

Further questions

- How does one **facilitate** effective **community interaction** & **contribution** with such a **resource**?
  - implement a high degree of **computer verification** of community input
  - ensure a degree of **moderation** in the Wiki
- Can one build a piece of **intelligent software** which is able to
  - scan in **books**;
  - produce **LaTeX source**;
  - replace commands for functions in the source with **semantic macros**;
  - **extract data** from the **text** (such as constraints)
  - **associate data** with relevant formulae and **removes** text;
  - **produce Wikitext**;
  - and **upload** Wikitext to a publicly accessible website?
- How does one **search** the resulting mathematical database?

Ongoing projects to investigate the above questions

- **Macro replacements** from well-constructed LaTeX source
- **Extraction** of **mathematical data** from **text** (keywords)
- **Wikitext generation**
- **Porting/building** a **mathematical search engine** in **MediaWiki**
- **Output of formula data** from **right-clickable menus** in a variety of **formats** so that formulas can be **used** and also **verified**
  - LaTeX expanded
  - LaTeX semantic
  - presentation MathML
  - content MathML
  - Mathematica
  - Maple
  - Sage

Past/Present/Related development team members

- **Moritz Schubotz** (TU-Berlin): **MediaWiki Math**
- Past/Present High School Students:
  - **Jake Migdall** : **MathJax menu**
  - **Alex Danoff** : **seeding/macro replacement**
  - **Amber Liu** : **MathJax menu customization**
  - **Cherry Zou** : **seeding/macro replacement**
  - **Jimmy Li** : **mathematical search**
  - (**Azeem Mohammed** : **LaTeX to Wikitext**)
  - (**Shraeya Madhu** : **Seeding Project**)
- **[Bruce Miller** (NIST) : (**DLMF macros/Search**)]
- **[Abdou Youssef** (NIST) : (**DLMF Math Search**)]

Ongoing project: Content MathML

- **Presentation MathML → Content MathML**
  - **LaTeXML** generates **presentation MathML** and **Content MathML (DLMF macros)** [symbol interaction]
- How can we **improve** the **Content MathML**?
- Resolve **ambiguities** associated with :
  - **Superscipts/subscripts**, e.g., x^0
  - **Sums/products/integrals/limits**, e.g., $\sum_{n=0}^{\infty} f(n)$
  - **Multiplication/function application**, e.g., $f(a+b)$
  - **Prime** notation (variable vs. derivative), e.g., $f'(a+b)$
- **Content Dictionaries** w/links to **macros** and **mathematical definitions** (e.g., **DLMF**)
- **Phrase Books** translate between different syntaxes
- Example: **LaTeX↔ Mathematica ↔ Wikitext** (while maintaining **Content MathML**)

... this can not be real ...

- There is a demo
- `http://drmf.wmflabs.org`