The Art of Mathematics Retrieval

Petr Sojka et al.

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<sojka@fi.muni.cz>

NLPlab seminar, Brno, Czech Republic
November 3rd, 2011

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Why Math Retrieval (TEX math search)?

Searching is crucial part of *accessibility* of the great stuff you all create, usually with the lot of *mathematics.*

How to pose questions about math? Math in TEX notation?

- compact and logical expression of formulas, quickest entering of it into a document or query
- picture is worth of thousands words, math formulae is worth of hundreds words (Ross Moore)

Why TEX math search is more relevant *now* than ever?

- Because of G? (G as in Google, Globalization,…).
- The *vast* treasure of mathematical papers; 140,000 new papers in Zentralblatt MATH expected this year, most of them authored in TEX math notation. All mathematics ever publisher is estimated at 100,000,000 pages (3,500,000 articles).
- Search – crucial part; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron.
- Text and keyword based search? No problem (Google, review databases); *success*.
- Mathematics formulae search? It *is* a problem (either in Google or in the review databases); more or less a *failure so far*.

Motivation for MSE (including formulae)

prof. James Davenport, CEIC member, MKM2011 PC chair, on panel at EuDML workshop in Bertinoro as a reply to the question “what functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?”:

“Math formulae search.”

Motivation for using a MSE (including formulae) – cont.

- Allowing formulas in queries helps to *disambiguate and narrow* search. Sometimes the only difference among set of notions/key words would be in a math formula.
- Example 1: knowing the solution of partial differential equation in $L^1(\mathbb{C}^3)$, is there one in $L^2(\mathbb{C}^5)$?
- Example 2: historians may want to follow the history of a (class of) formula(s) across languages and vocabularies (e.g. same objects studied/used by physicists and mathematicians under different names).
- Example 3: physicist looking for theorems about solitons, but mathematicians use these terms for something completely different from my perspective and I do not know how they call those I'm interested in. Putting the equation my solitons are solutions of might be the only way to locate relevant literature.
- Imagine your favourite ebook math textbook being TEX-search aware—e.g. your search app supports math formulae search.

Take-off message from the talk: *Yes, you can!*

Compare `google.com/search?q=Einstein` with math-aware search of `Einstein+$E=mc^2$` over arXiv.

The rest of the talk: how is it actually done (for EuDML).

Towards math search engine (MSE) – existing players

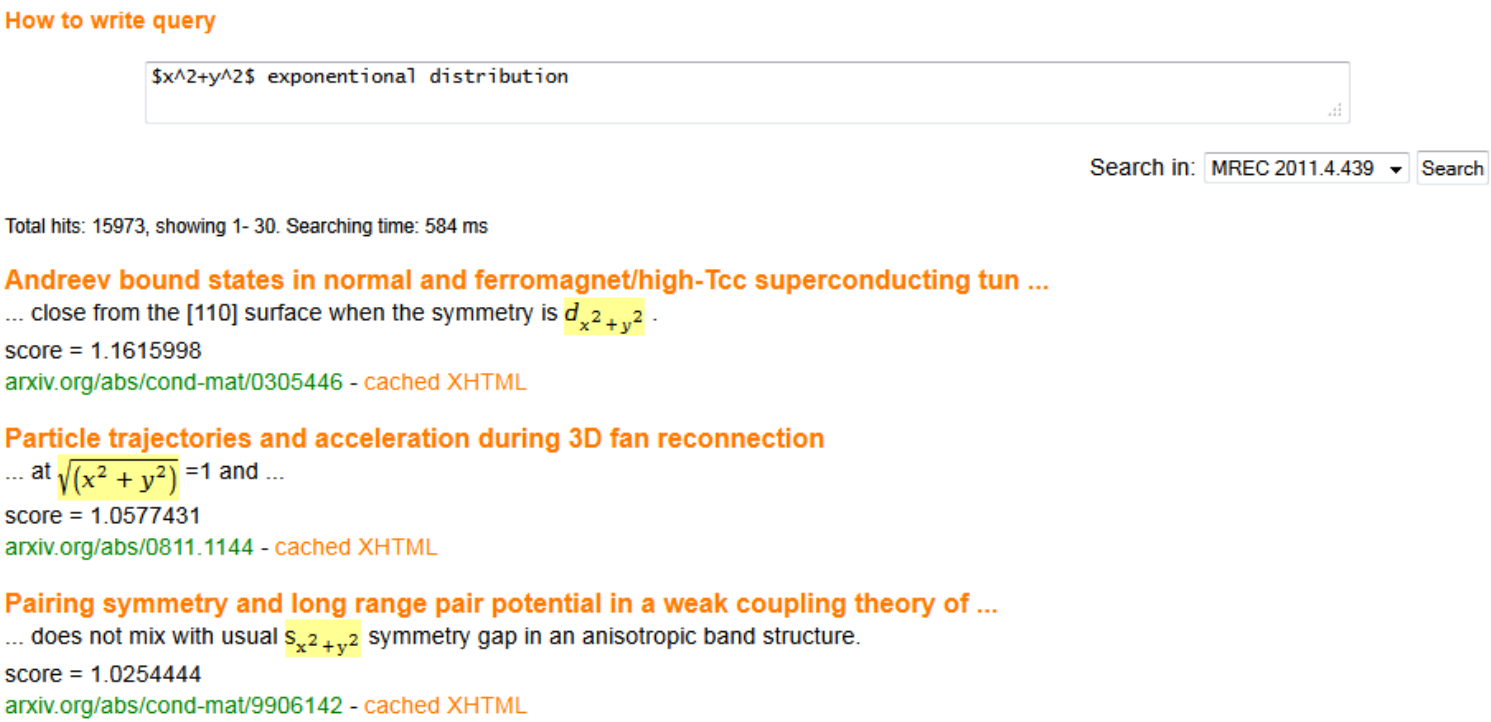
- Niche market for big players (as Google), attempts to solve by publishers (LaTeXSearch by Springer).
- Many challenges: heterogenity of math representation, notation, semantics handling, no established and accepted user interface and query language.
- Numerous attempts to solve the problem: MathDex, EgoMath, LaTeXSearch, LeActiveMath, DLMF equation search, MathWebSearch, but none accepted by the community as *the* MSE.

Existing systems – pros and cons

- **MathDex**: formerly MathFind * seven digit figure NSF grant by Design Science (Robert Miner) * Lucene based, indexing n-grams of presentation MathML * pioneering conversion effort
- **EgoMath and EgoMath2**: based on full text web search system Egothor * presentation MathML for indexing * idea of formulae augmentation, α-equivalence algorithms and relevance calculation
- **LaTeXSearch**: MSE offered by Springer * closed source * only for LaTeX math string approximate match based on strings * no formulae structure matching * small database: 3 million formulae from 'random' sources
- **LeActiveMath**: indexing string tokens from OMDoc with OpenMath semantic notation * *only* for documents authored for LeActiveMath learning environment
- **DLMF**: *only* for documents authored for DLMF in special markup * equation search
- **MathWeb Search**: semantic approach – uses substitution trees – not based on full text searching * supports Content MathML and OpenMath * problem with acquiring semantic data

MIaS — Math Indexer and Searcher

- math-aware, full-text based search engine
- joins textual and mathematical querying
- MathML *or* TEX input



Dual world of TEX and MathML

Math for people: TEX notation wins and is used by people (mostly AMSLATEX fits most needs).

Math for software applications: MathML wins and is used by most computer algebra systems, browsers, in workflow of DTP systems…

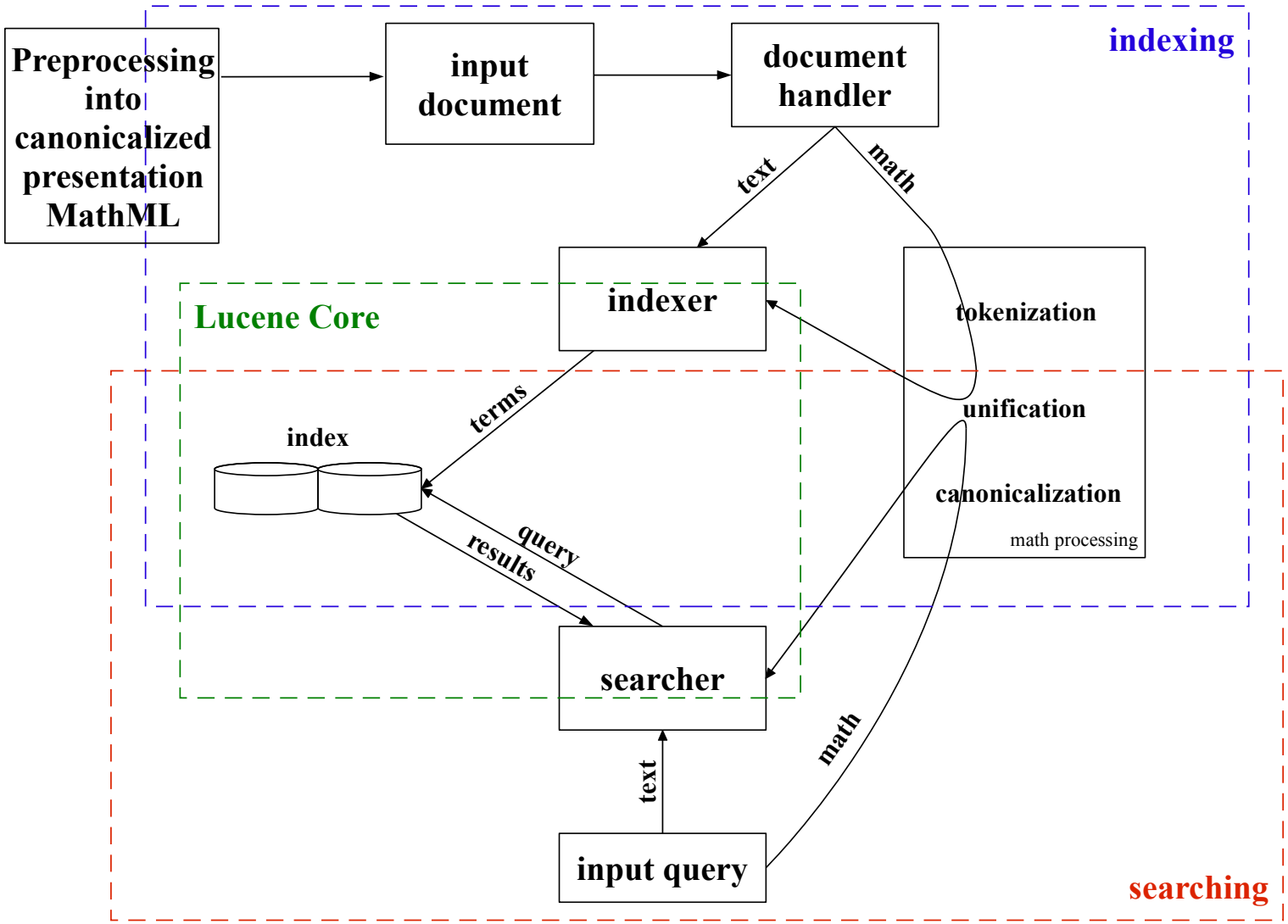
Dual world of query language and indexing language

In text retrieval: Indexing word stems only instead of word forms.

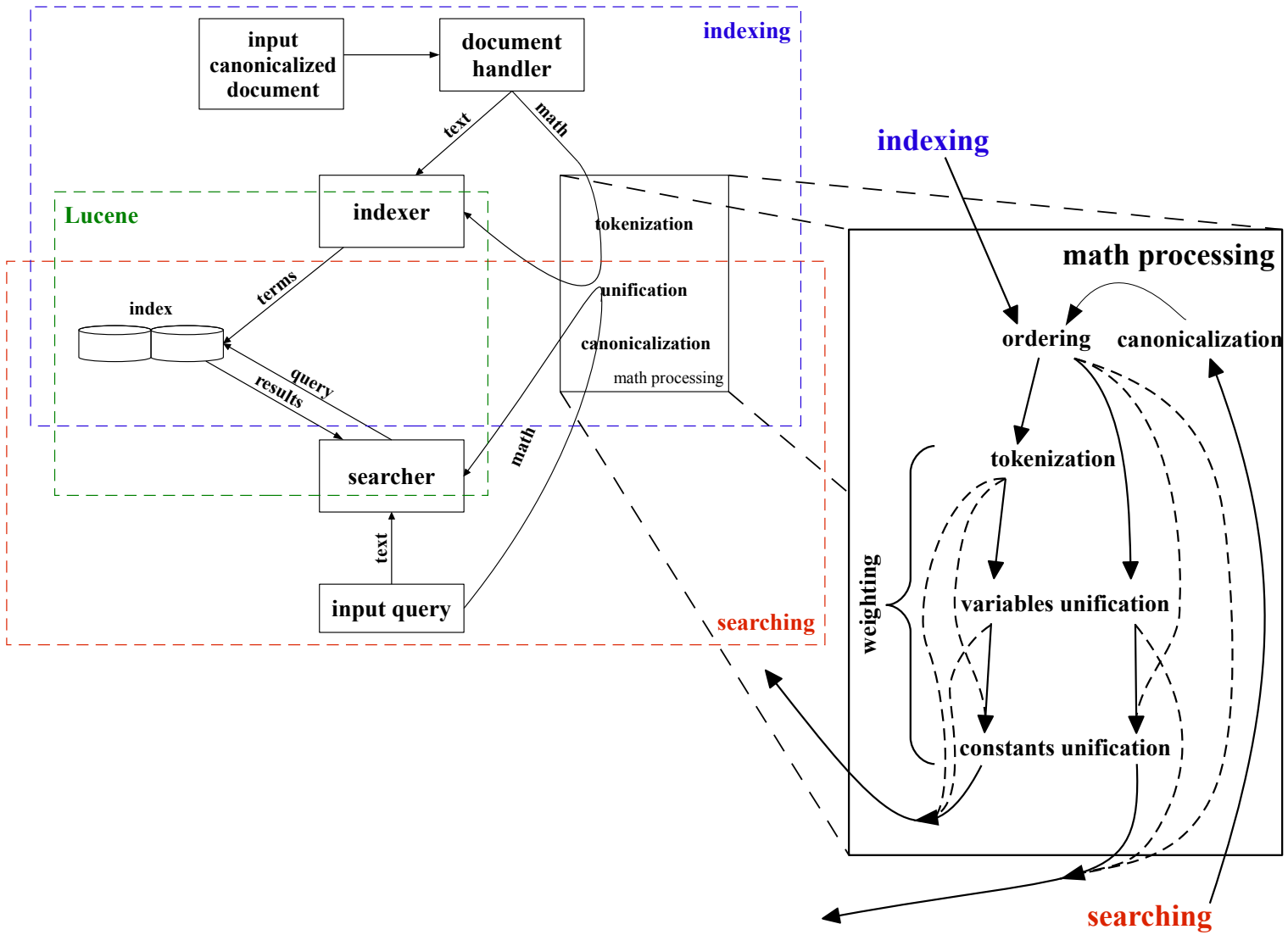
TEXbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

From text to math: the same idea explored for math (e.g. having dozen of representations of a formula in the index).

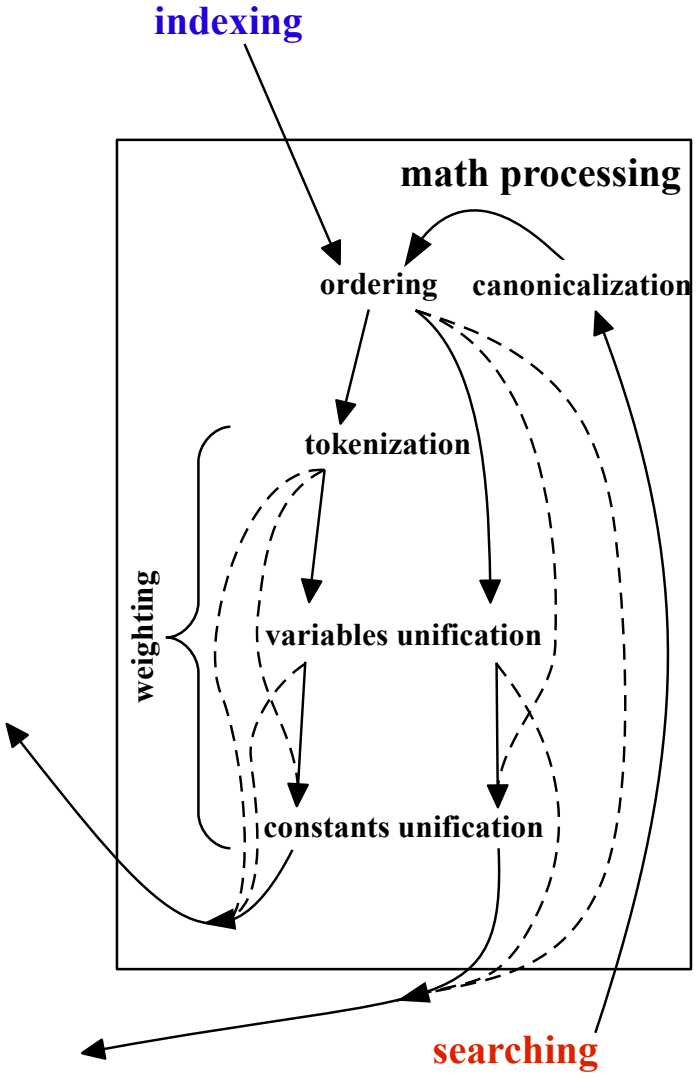
MSE overall design

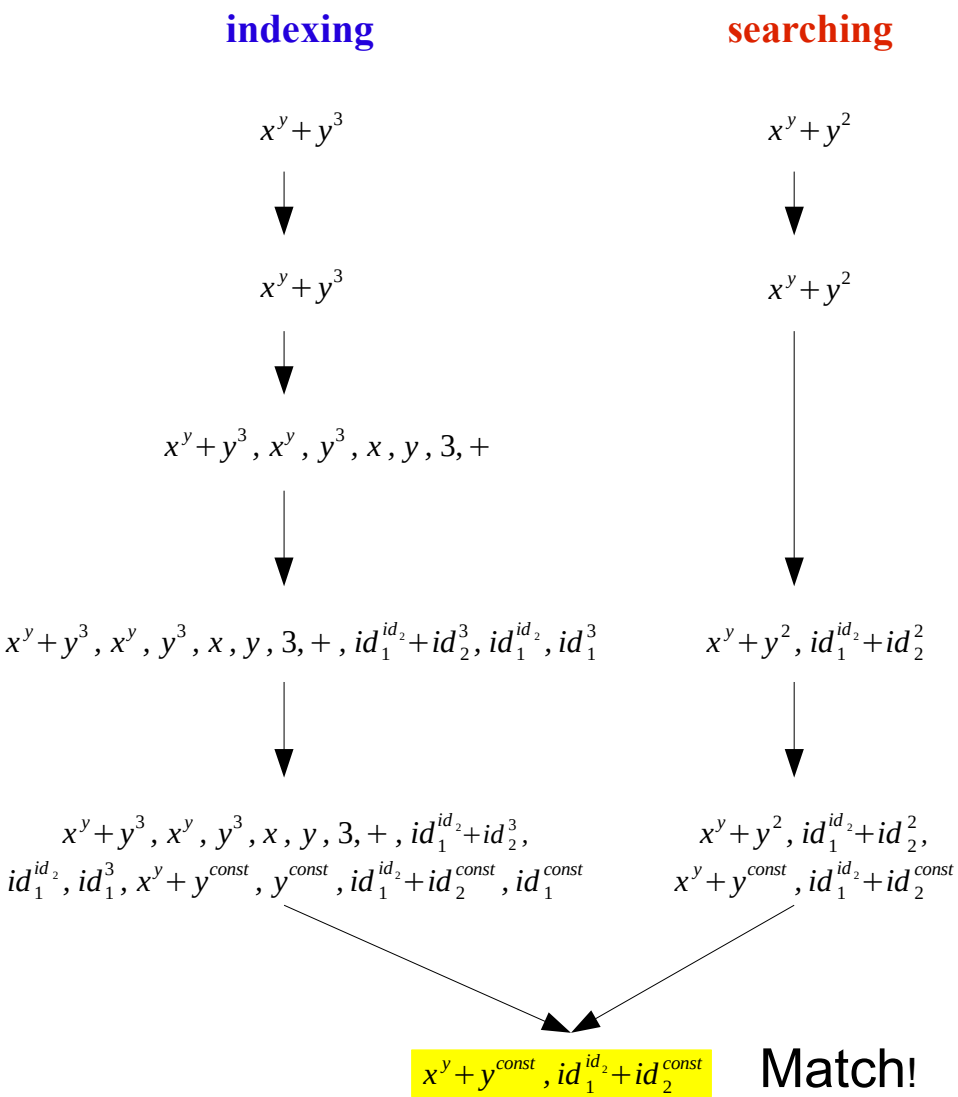


Math indexing design



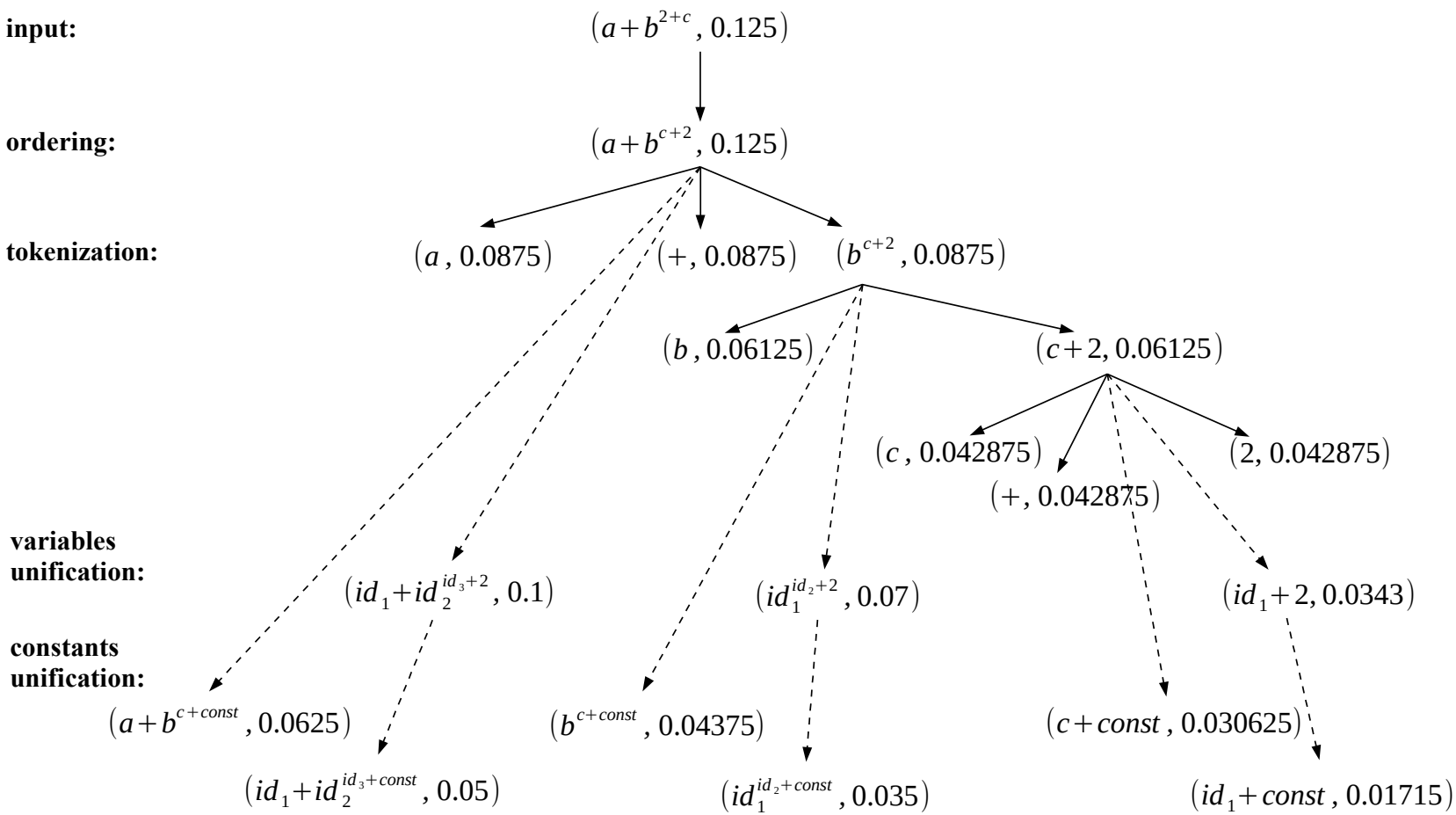
Example





Match!

Formula processing example – subformulae weighting



Implementation

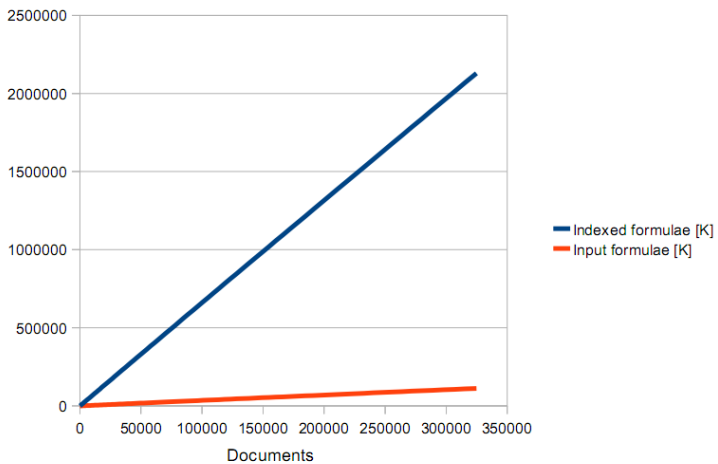
- Java
- Lucene 3.1.0
- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system
- MIaS4Solr plugin was created for the use in Solr in EuDML
- Textual content – processed by StandardAnalyzer

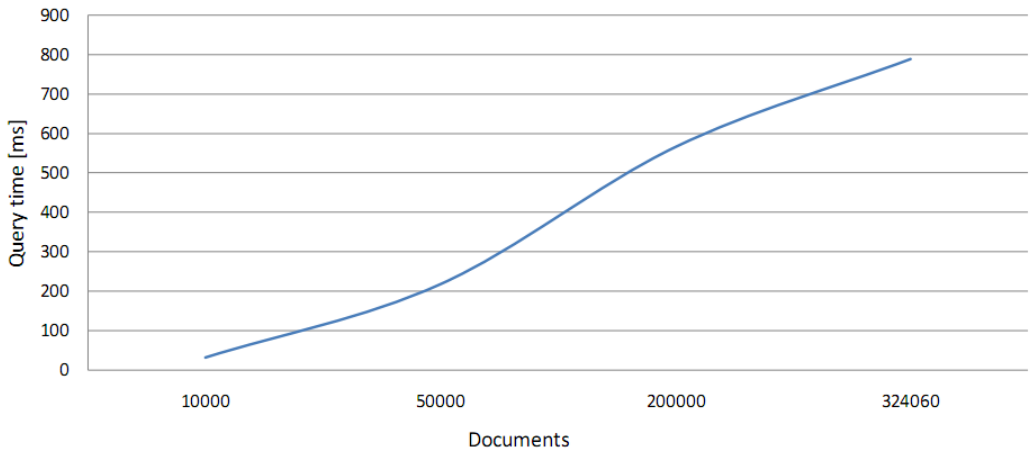
Data used for evaluation: MREC corpus

- Mathematics REtrieval Corpus (MREC, version 2011.4.439)
  - 439,423 documents (originated from arXMLiv [8], validated, enriched with metadata for snippet generation)
  - Uncompressed size 124 GB, compressed 15 GB
  - 158 million input formulae, 2.9 billion subexpressions indexed (Lucene index size 47 GB)
- For more information see paper (DML 2011, Bertinoro) [10] and home page of MREC subproject http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html

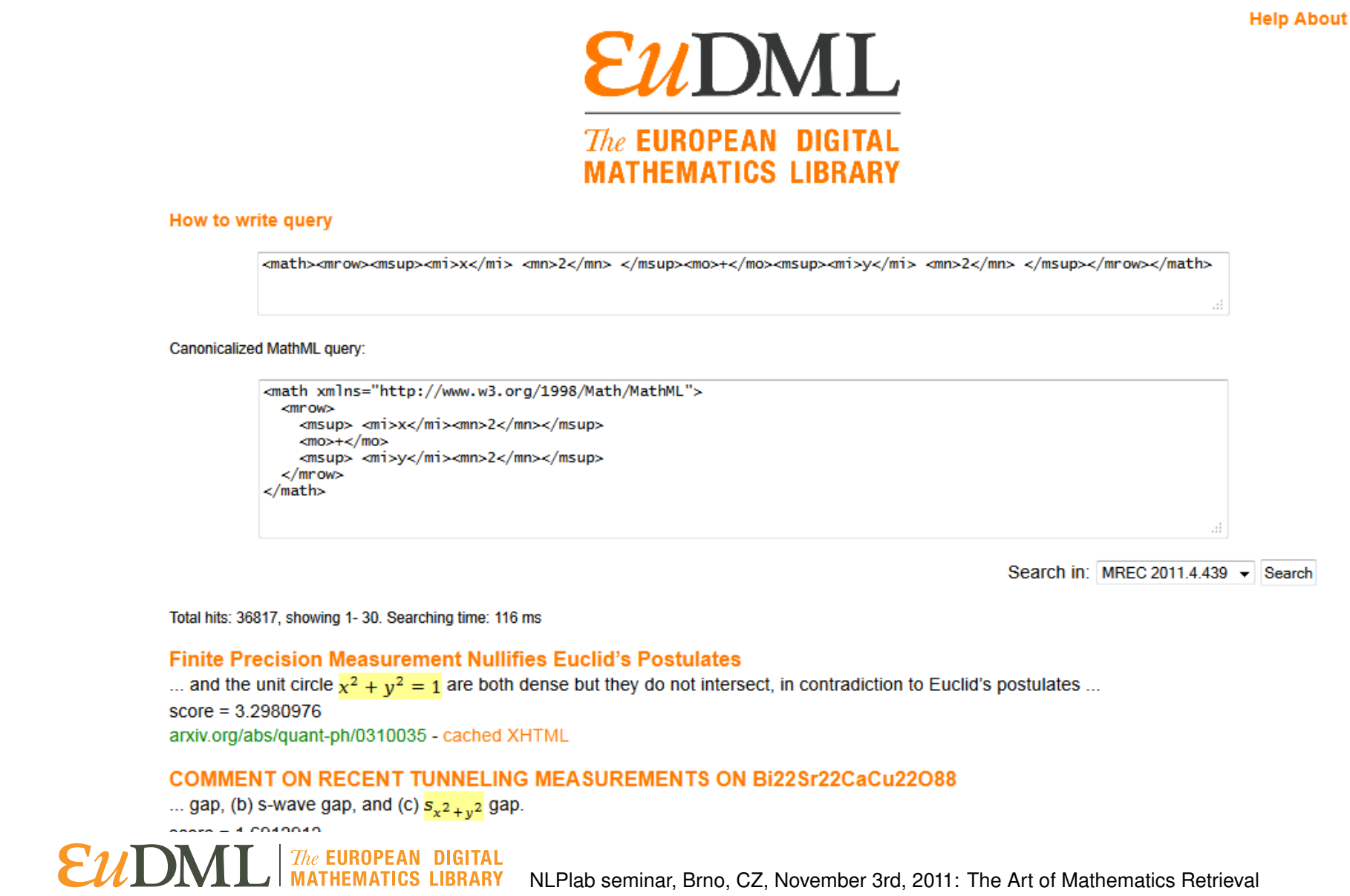
Scalability (tested on MREC 2011.4.439)

- Indexing time: 1,378.82 min (23 hours, down to 9 h with threads)
- Average query time: 469 ms
- Overall index size 47 GB (most of it math entries)
- Linear time scale – still seems feasible for a digital library





Search demonstration



Formulae search demonstration comments

Demo web interface: http://aura.fi.muni.cz:8085/webmias/

- MathML/TEX input (Tralics [2] for conversion to MathML [7])
- Canonicalization of the query – UMCL library [1]
- Matched document snippet generation
- MathJax for nicer math rendering and better portability

MIaS already integrated in the EuDML system.

Plea for further structure search developments in NLP

- shown approach allows structure search by using relatively small extension of current search technologies for text without support for structures.
- in tagged textual corpora there are *many* structures (grammatical or dependency trees) to be searched – Sketch computations are just some of applications
- method explorable in NLP – slight extension of Bonito/Manatee?

Conclusions

- Scalable solution for math formulae search researched, implemented, tested and integrated into current version of EuDML system!
- MIaS project pages – http://nlp.fi.muni.cz/projekty/eudml/mias

Future work

- Preprocessing from TEX, PDF,…
- `copypaste` package (storing TEX math code into PDF as second layer with `/ActualText` (for indexing purposes): typesetters may use in their workflows
- Improved MathML canonicalization and new preprocessing filters, test on new EuDML data
- Weighting optimization (by machine learning)
- Query relaxation (“Did you mean…”)
- Addition of Content MathML tree indexing?
- Mathematical equivalence computation via symbolic algebra system?

Summary

MIaS will hopefully become *the* MSE used by the community. Our hope is based on these features:

- *text+math IR compatible*, accepting both TEX and MathML formats (fits mathematician's needs)
- new math formulae similarity (weighting) approach compatible with *both presentation (structure) and content (semantic)* MathML
- *scalable* (index with almost 3 billion subformulae tested)
- *Lucene/Solr compatible* system employed and *used in EuDML will hit the masses* ;-).

For more information see papers in SpringerLink (MKM 2011, Bertinoro) [5] and ACM DL (DocEng 2011, Mountain View) [6].

Related work

Work motivated by projects of The European Digital Mathematics Library (EuDML) and The Digital Mathematics Library Czech Republic (DML-CZ).

Related topics researched at FI as part of projects above in LEMMA and NLP laboratories:

- gensim package (scalable document clustering) by Radim Řehůřek
- pdfRecompressor (JBIG2 enhancements by OCR,…) by Radim Hatlapatka
- TEX to MathML conversions, normalization and canonicalization by Michal Růžička
- MathML preprocessing (normalization and canonicalization) by Michal Růžička, Peter Mravec

Related work (cont.)

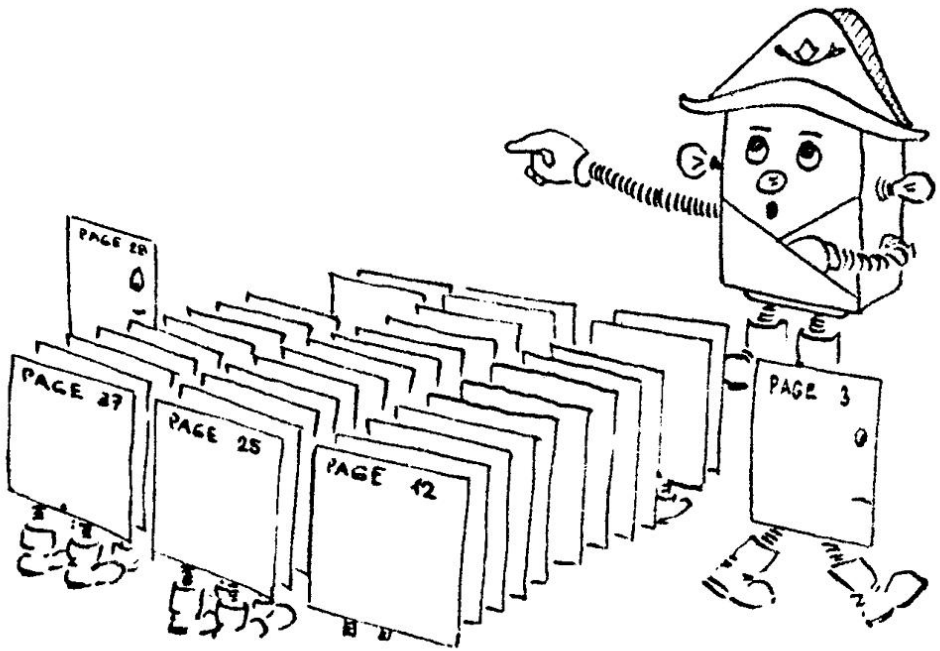
- Metadata Editor tool development, metadata enhancements by Petr Kovář, Mirek Bartošek, Vlastimil Krejčíř, Martin Šárfy
- (math) OCR by Masakazu Suzuki, Radim Hatlapatka, Radovan Panák, Tomáš Mudrák)
- (Meta)data vizualization (Visual Browser) by Zuzana Nevěřilová
- Czech Braille driver with math support by Martin Jarmar
- and a lot more

Acknowledgments

- EuDML and DML-CZ project funding
- Martin Líška (search implementation)
- Michal Růžička, Radim Hatlapatka, Zuzana Nevěřilová
- Martin Jarmar, Petr Mravec, Radovan Panák, Tomáš Mudrák, Vítězslav Dostál, Martin Kacvinský
- Mirek Bartošek, Petr Kovář, Vlastimil Krejčíř, Martin Šárfy
- Infty group (led by Suzuki)
- numerous authors and contributors of numerous tools used
- numerous people discussing and supporting our work

Questions?

Thank you for your attention.




Archambault, D., Moço, V.: Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) Computers Helping People with Special Needs, Lecture Notes in Computer Science, vol. 4061, pp. 1191–1198. Springer Berlin / Heidelberg (2006), <http://dx.doi.org/10.1007/11788713_172>


Grimm, J.: Producing MathML with Tralics. In: Sojka [4], pp. 105–117, <http://dml.cz/dmlcz/702579>


MREC – Mathematical REtrieval Collection, <http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html>


Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <http://www.fi.muni.cz/ sojka/dml-2010-program.html>


Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Urban, J., Rabe, F., (eds.) Proceedings of CICM Conference 2011 (Calculemus/MKM). Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 228–243. Springer-Verlag, Berlin, Germany (Jul 2011), <http://dx.doi.org/10.1007/978-3-642-22673-1_16>


Sojka, P., Líška, M.: The Art of Mathematics Retrieval. In: Tompa, F., Hardy, M. (eds.) Proceedings of DocEng 2011 Conference. pp. 57–60. ACM. Mountain View, September 2011.


Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: MathML-aware Article Conversion from LATEX. In: Sojka, P. (ed.) Proceedings of DML 2009. pp. 109–120. Masaryk University, Grand Bend, Ontario, CA (Jul 2009), <http://dml.cz/dmlcz/702561>


Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science 3, 299–307 (2010), <http://dx.doi.org/10.1007/s11786-010-0024-7>


Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [4], pp. 11–24, <http://dml.cz/dmlcz/702569>


Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec.
Web Interface and Collection for Mathematical Retrieval.
In Petr Sojka and Thierry Bouche, editors, *Proceedings of DML 2011*, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University. <http://dml.cz/dmlcz/702604>.