Dual World of TEX Math and MathML

Petr Sojka

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

TEXperience 2013, Brejlov, Czech Republic
September 27th, 2013

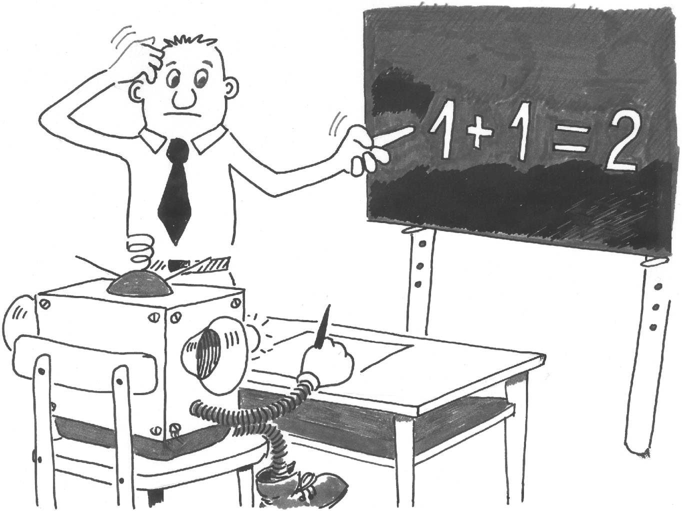
EuDML

The EUROPEAN DIGITAL
MATHEMATICS LIBRARY

Outline and take-home message

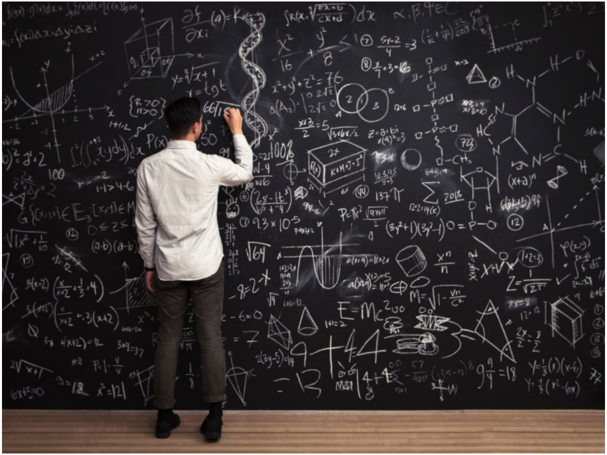
1. Specifics of Mathematics
2. TEX Math
3. Math on the Web
4. Math Search
5. Conversions
6. Search in Digital Mathematics Libraries
7. Math Indexer and Searcher (MIaS)
8. Conversions
9. Similarity
10. Conclusions

Mathematics is *specific and challenging* in many aspects



DEK was first to allow switching [math] typesetting from metallurgy to the *digital* world on an author's desktop computer

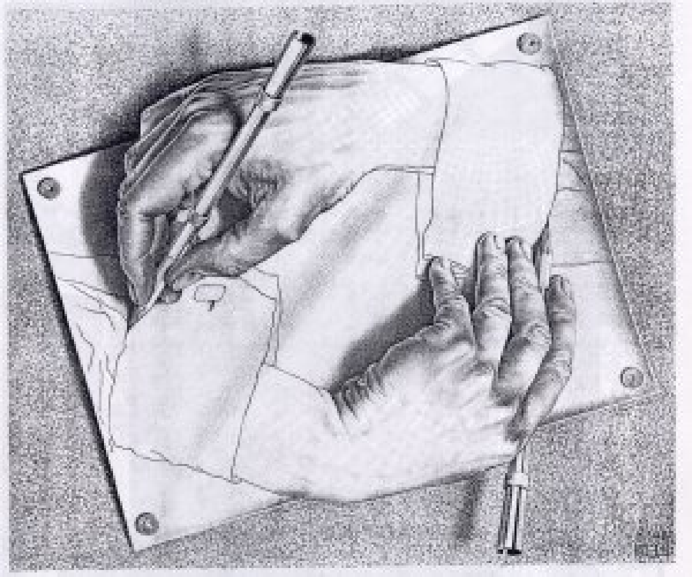
Complexity of Math (compared to text handling)



Complexity of Math compared to text

- typesetting of math has always been for masters
- structured objects within [linear] text
- different notations, fonts, growing symbols, spacing
- many levels of abstractions

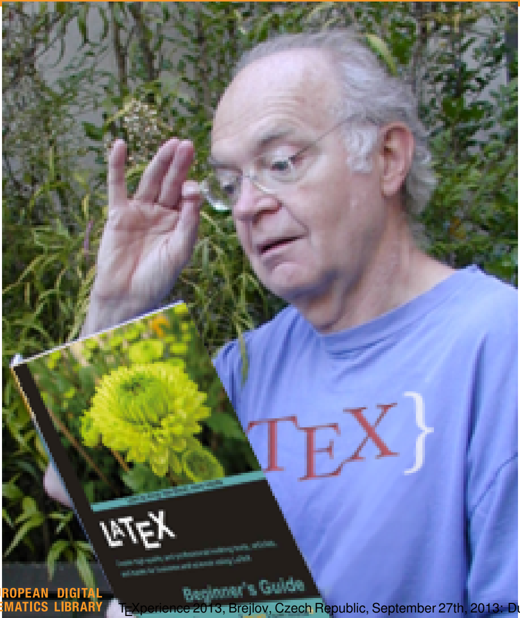
Many levels of abstractions



Math in TEX

- math-related stuff in the TEXbook sums up to one third of the book
- two math *modes* in TEX: inline math and display math
- TEX's design *just* to mimic the old way (no dissemination in digital form meant): `$x\over y$`
- presentation (visual, structural) aspect predominant, CONtinuous and disCRETE, concrete mathematics

From \over to \frac



Math in LaTeX

Q: *Three LaTeX mistakes that people should stop making?*

A: (by Lamport:)

1. Worrying too much about formatting and not enough about content.
2. Worrying too much about formatting and not enough about content.
3. Worrying too much about formatting and not enough about content.

Math in LATEX (cont.)

Q: *What's your view on mathematical typesetting in the future? Quantum leaps ahead?*

A: (by Lamport in 2000):

Standards are being driven by the marketplace, which cares only about the masses. So, mathematicians have no place in the brave new world of computing.

plainTEX, AMSTEX, LAMSTEX, AMSLATEX, LATEX+AMS

- packages by AMS became de facto standard in math publishing industry; Context goes by My Way
- logical structure uniformly marked, variants (Nath et al.)
- widespread adoption and support where *validation* needed

Automated Math, Math Exchange

- for automated processing validation needed
- non-extensible markup preferred
- exchange and rendering on the Web
- XML: MathML by W3C

Presentation MathML

```
<math xmlns=”http://www.w3.org/1998/Math/MathML”
      display=”inline”>
  <mrow>
    <mrow>
      <mi>x</mi>
      <mo>+</mo>
      <mi>y</mi>
    </mrow>
    <mo>=</mo>xml
    <mn>2</mn>
  </mrow>
```

Presentation MathML (cont.)

```
</math>
```

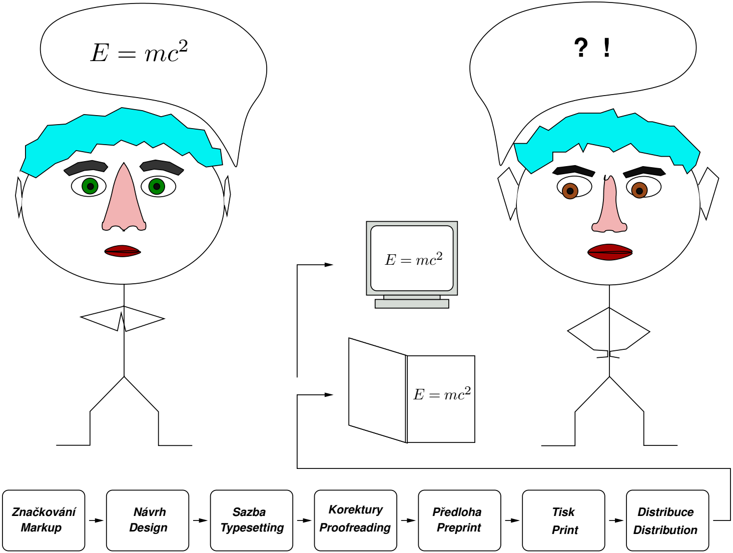
Presentation MathML (cont.)

```
<mstyle mathbackground=”yellow” mathcolor=”navy”
        mathsize=”16pt” mathvariant=”bold”>
  <mrow>
    <mi>x</mi>
    <mo>+</mo>
    <mi>y</mi>
  </mrow>
  <mo>=</mo>
  <mn mathcolor=”red”>2</mn>
</mstyle>
```

Content MathML

```
<m:math>
  <m:apply>
  <m:eq/>
    <m:apply>
      <m:plus/>
         <m:cn>2</m:cn>
         <m:cn>2</m:cn>
    </m:apply>
    <m:cn>4</m:cn>
  </m:apply>
</m:math>
```

Math exchange from the author’s brain to the readers’ one



Značkování* *Markup → ***Návrh* *Design*** → ***Sazba* *Typesetting*** → ***Korektury* *Proofreading*** → ***Předloha* *Preprint*** → ***Tisk* *Print*** → ***Distribuce* *Distribution***

Math Search

- MathJax
- MathML 3.0, WAI-ARIA (Web Accessibility Initiative—Acessible Rich Internet Applications), WCAG (Web Content Accessibility Guidelines) 2.0.
- direct MathML support in [tagged] PDF by Adobe, but nobody able to take advantage of it but Ross Moore
- ChromeVox
- Wolfram Alpha
- Systems for symbolic computation (Mathematica, Maple,…)
- everywhere, where *software* deals with math

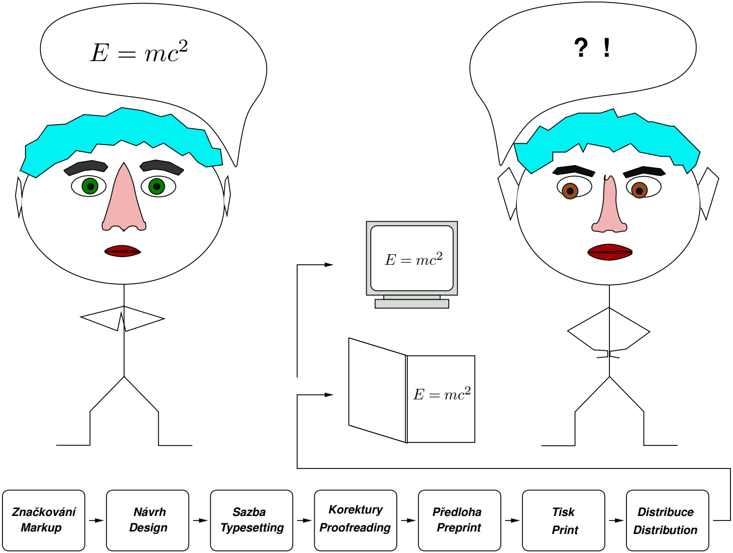
Dual worlds: MathML and LaTeX!

Data heterogenity, plethora of formats, validation and conversions:

world of authors: LaTeX, TeX notation of mathematics

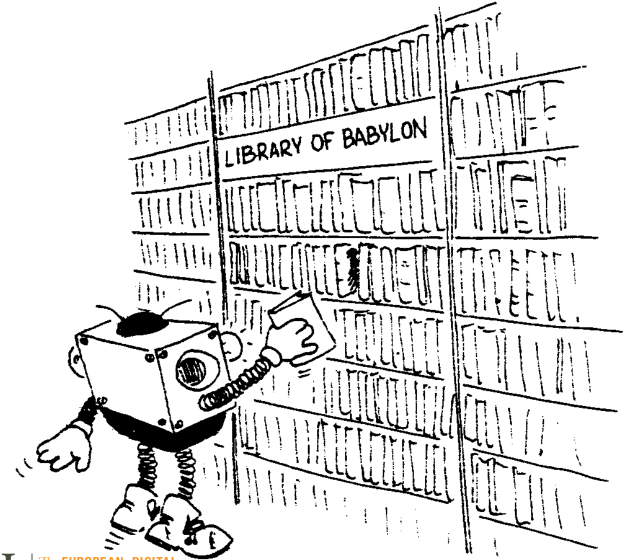
world of applications/data exchange: XML, *MathML*

Math exchange from the author's brain to the readers' one



Značkování ***Markup*** → ***Návrh*** ***Design*** → ***Sazba*** ***Typesetting*** → ***Korektury*** ***Proofreading*** → ***Předloha*** ***Preprint*** → ***Tisk*** ***Print*** → ***Distribuce*** ***Distribution***

Most scientific valuable content end up in the *Digital Library*



Math content is not an exception: the dream of the World Digital Mathematical Library

History of the dream: vision of WDML as PubMed 4 Math

In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

Progress of IT, connectivity, cheap storage, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation ($100,000,000 requested). Application was *not* successful.

Publishers started massive digitization themselves.

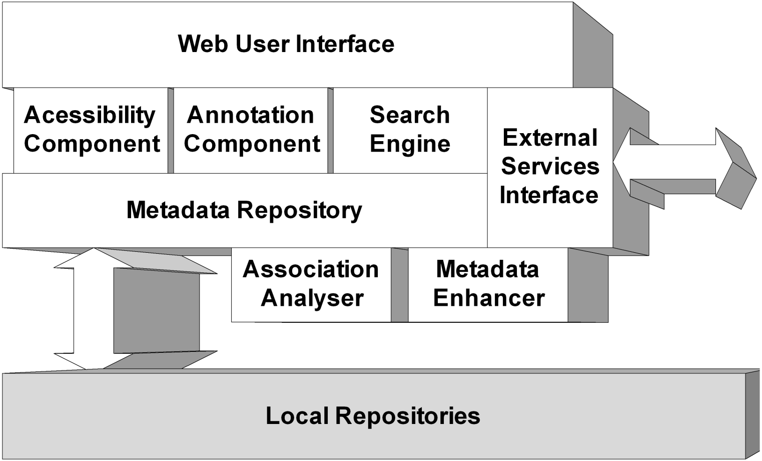
Vision of European Digital Mathematics Library

Finally three year project or *European Digital Mathematics Library, EuDML* (programme EU CIP-ICT-PSP, type Pilot B, EU contribution *(1.6 MEur, 50% of total budget only)* February 2010–January 2013. The strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY was:

- to master the technology, develop tools and offer them;
- concept of *moving wall* to motivate and engage commercial publishers without Open Access bussiness model;
- to collect data (from existing local or publisher's) *digital libraries* into 'one-stop shop' and achieve critical mass in the domain → 'a must/me too' effect then as with PubMed Central.

EuDML as a virtual library portal

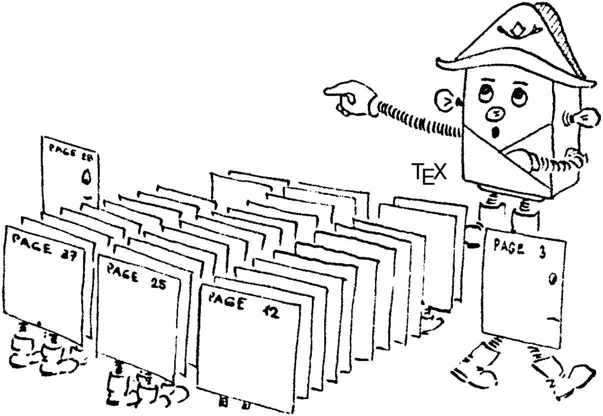
EuDML provides a *virtual* library based on data from smaller data providers (as DML-CZ, <http://dml.cz>), DLs and publishers:



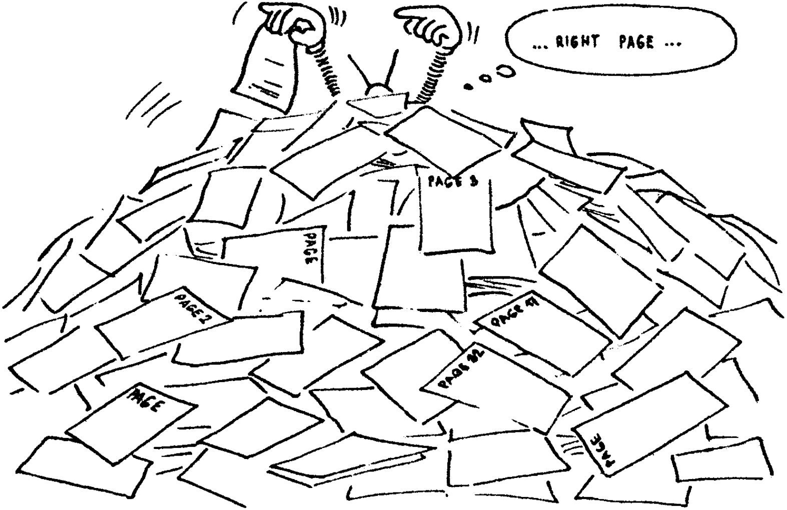
One portal: European Digital Mathematics Library

Digital Library without [Math] Search is an oxymoron

Yes, you can! <http://eudml.org>: accessible math, search, visibility, scalability,. . .



How to query and index math?



Going from author's heads to reader's via Math [formulae]?

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

Why math search is more relevant *now* than ever? (cont.)

- Because of G? (G as in Google, Globalization,…).
- The *vast* treasure of mathematical papers; 140,000 new papers in Zentralblatt MATH expected this year. 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? Even picture search? No problem (Google, review databases); *success*.
- Mathematics formulae (structure) 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) – cont.

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.”

Topics/questions:
1) What functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?
2) What are the most decisive factors on this crossroad?

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

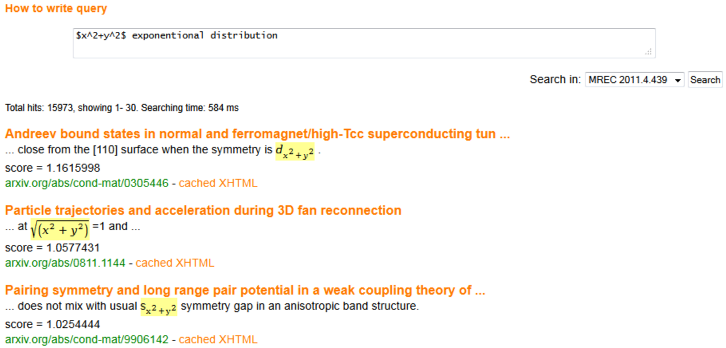
- Allowing formulas in queries helps to *disambiguate and narrow* search. Sometimes the only difference among set of notions/key words would be in a math formula.
- Example 1: knowing the solution of partial differential equation in $L^1(\mathbb{C}^3)$, is there one in $L^2(\mathbb{C}^5)$?
- Example 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).
- Imagine your favourite ebook math textbook being [TEX]-search aware—e.g. your search app supports math formulae search.

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



How to write query

```
$x^2+y^2$ exponentional distribution
```

Search in: MREC 2011.4.439 ▾ Search

Total hits: 15973, showing 1- 30. Searching time: 584 ms

Andreev bound states in normal and ferromagnet/high-Tcc superconducting tun ...
... close from the [110] surface when the symmetry is $d_{x^2+y^2}$.
score = 1.1615998
arxiv.org/abs/cond-mat/0305446 - cached XHTML

Particle trajectories and acceleration during 3D fan reconnection
... at $\sqrt{(x^2+y^2)}$ =1 and ...
score = 1.0577431
arxiv.org/abs/0811.1144 - cached XHTML

Pairing symmetry and long range pair potential in a weak coupling theory of ...
... does not mix with usual $s_{x^2+y^2}$ symmetry gap in an anisotropic band structure.
score = 1.0254444
arxiv.org/abs/cond-mat/9906142 - cached XHTML

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…

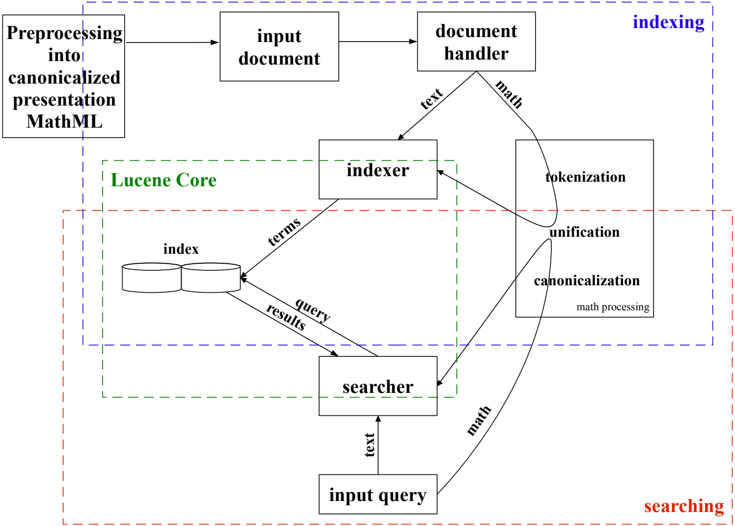
Dual world of query language and indexing language

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

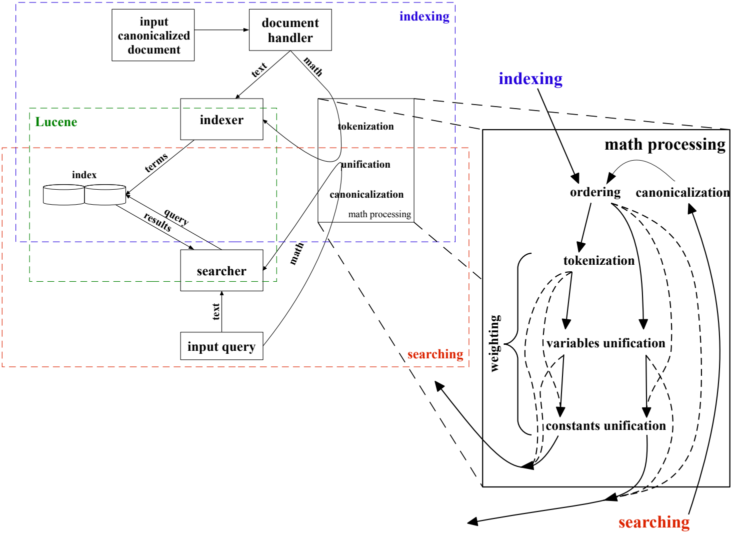
TEXbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

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

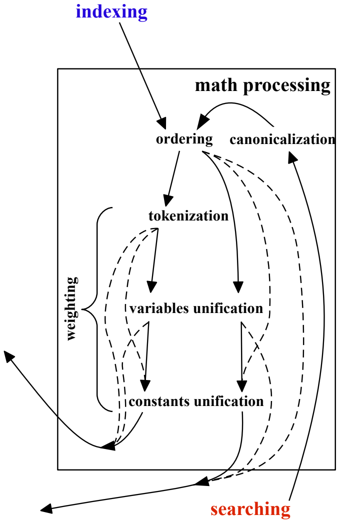
MSE overall design

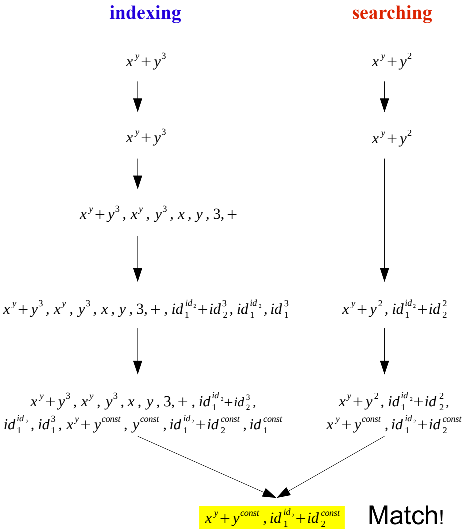


Math indexing design



Example





indexing

x^y+y^3

x^y+y^3

$x^y+y^3, x^y, y^3, x, y, 3, +$

$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3$

$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3, x^y+y^{const}, y^{const}, id_1^{id_2}+id_2^{const}, id_1^{const}$

searching

x^y+y^2

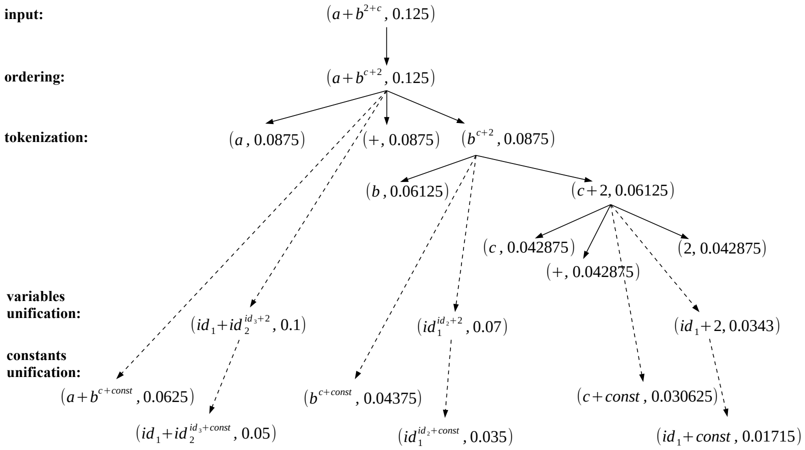
x^y+y^2

$x^y+y^2, id_1^{id_2}+id_2^2$

$x^y+y^2, id_1^{id_2}+id_2^2, x^y+y^{const}, id_1^{id_2}+id_2^{const}$

$x^y+y^{const}, id_1^{id_2}+id_2^{const}$ Match!

Formula processing example – subformulae weighting



input: $(a+b^{2+c}, 0.125)$

ordering: $(a+b^{c+2}, 0.125)$

tokenization: $(a, 0.0875)$ $(+, 0.0875)$ $(b^{c+2}, 0.0875)$

$(b, 0.06125)$ $(c+2, 0.06125)$

$(c, 0.042875)$ $(+, 0.042875)$ $(2, 0.042875)$

variables unification: $(id_1+id_2^{id_3+2}, 0.1)$ $(id_1^{id_2+2}, 0.07)$ $(id_1+2, 0.0343)$

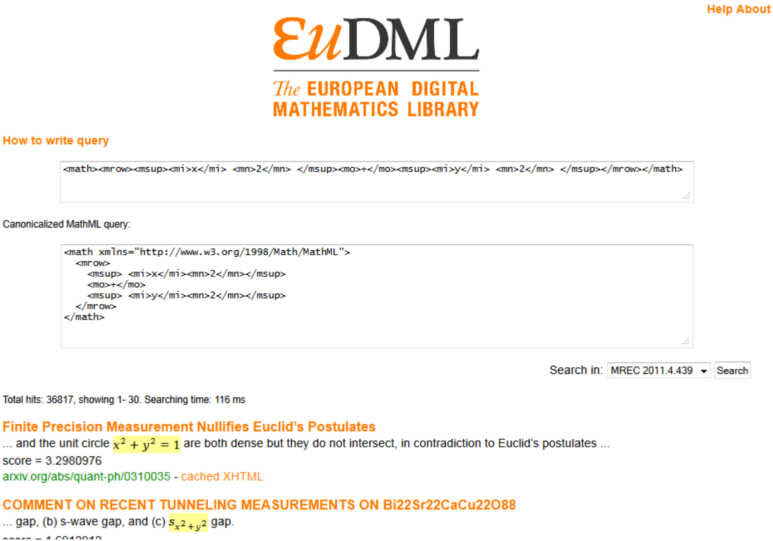
constants unification: $(a+b^{c+const}, 0.0625)$ $(b^{c+const}, 0.04375)$ $(c+const, 0.030625)$

$(id_1+id_2^{id_3+const}, 0.05)$ $(id_1^{id_2+const}, 0.035)$ $(id_1+const, 0.01715)$

Implementation

- Java
- Lucene 3.1.0, now switching to Lucene/Solr 4
- Mathematical part implements Luceneâ s interface Tokenizer – able to integrate to any Lucene based system
- MIaS4Solr plugin was created for the use in Solr
- Textual content – processed by StandardAnalyzer
- easily deployable in Java/Lucene based system or as a web service

Search demonstration



```
<math><mrow><msup><mi>x</mi> <mn>2</mn> </msup><mo>+</mo><msup><mi>y</mi> <mn>2</mn> </msup></mrow></math>
```

Canonicalized MathML query:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup> <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup> <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

Search in: MREC 2011.4.439 Search

Total hits: 36817, showing 1- 30. Searching time: 116 ms

Finite Precision Measurement Nullifies Euclid's Postulates

... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...

score = 3.2980976

arxiv.org/abs/quant-ph/0310035 - cached XHTML

COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88

... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.

score = 1.6912912

Formulae search demonstration

EuDML interface: http://eudml.org/search

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

- Snuggle TeX for on-the-fly as-you-type rendering
- Matched document snippet generation
- MathJax for nicer math rendering and better portability
- Canonicalization of the query – problems with UMCL library [1], now our own canonicalizer

All up and ready on the EuDML system.

Switching between the worlds: conversions

- in EuDML: MathML/TEX input (Tralics [2] for conversion to MathML [9])
- internal representation in Lucene index
- experiments with LaTeXML (both PML and ambiguous CML output)
- translation from Presentation to Content needed
- the need for normalization

Direct typesetting of MathML in Context

```
\usemodule[mathml]
\starttext
\startXMLdata
<math>
 <mrow>
  <msup> <mi>x</mi><mn>2</mn> </msup>
  <mo>+</mo>
  <mrow>
    <mn>4</mn><mo>InvisibleTimes;</mo><mi>x</mi>
  </mrow>
  <mo>+</mo>
  <mn>4</mn>
 </mrow>
</math>
\stopXMLdata
\stoptext
```

Searching (semantically) similar papers

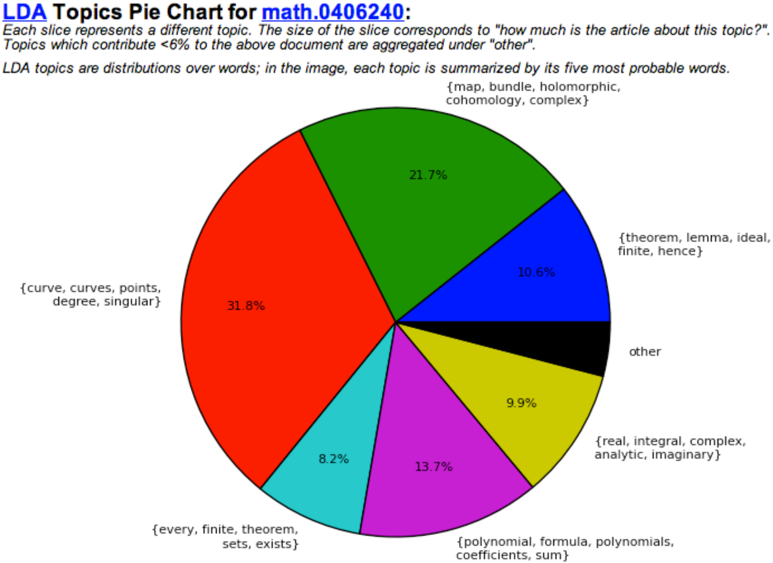
Exploration of a DML: browsing (semantically) similar papers

Semantic search via topic modeling: Latent Semantic Indexing, Latent Dirichlet Allocation

Methods developed for text, how to incorporate Math?

Which formulae are semantically similar (do have same/similar meaning)?

Leading Edge Example: Automated Meaning Picking from Texts



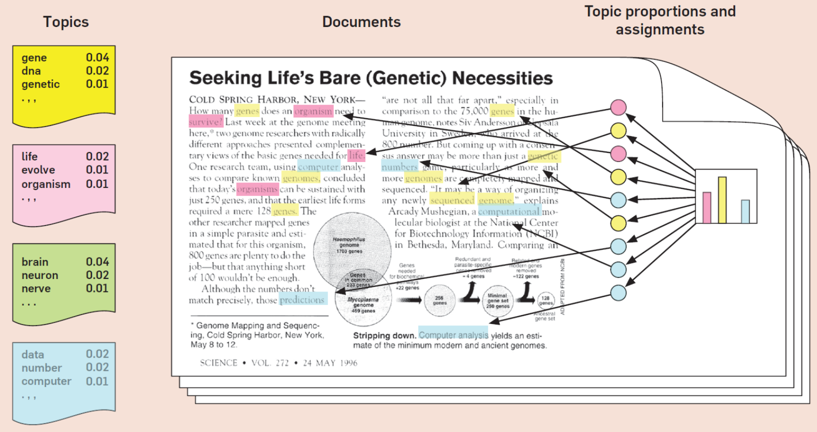
LDA Topics Pie Chart for math.0406240:

Each slice represents a different topic. The size of the slice corresponds to "how much is the article about this topic?". Topics which contribute <6% to the above document are aggregated under "other".

LDA topics are distributions over words; in the image, each topic is summarized by its five most probable words.

Probabilistic Topical Modeling: Latent Dirichlet Allocation

- topic: weighted list of words
- document: weighted list of topics



Topics

gene	0.04
dna	0.02
genetic	0.01
., ,	

life	0.02
evolve	0.01
organism	0.01
., ,	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
., ,	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

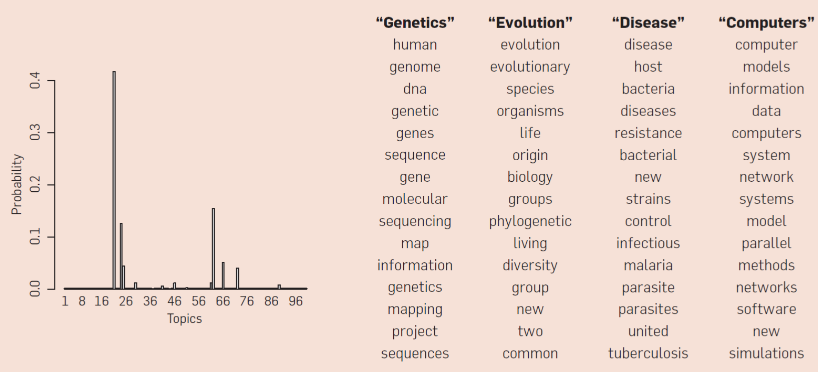
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topical Modeling: Latent Dirichlet Allocation II

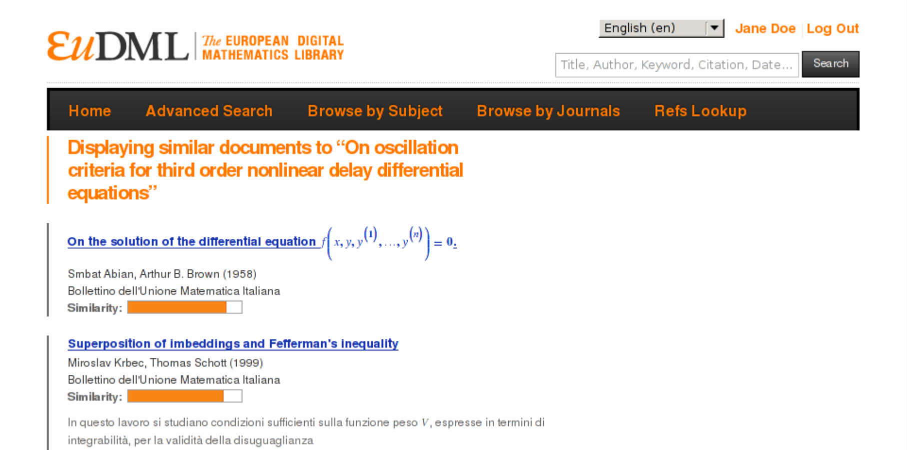
- all topics computed automatically from document corpora



“Genetics”	“Evolution”	“Disease”	“Computers”
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Content Similarity Results in <http://eudml.org>

We have developed and delivered technology for *similarity* (gensim), document *conversions* (to Braille or to text: Mathml2text) and math content *normalization*. Different formulae representations for similarity computation.



Displaying similar documents to “On oscillation criteria for third order nonlinear delay differential equations”

On the solution of the differential equation $f\left(x,y,y^{(1)},\ldots,y^{(n)}\right)=0$.

Smbat Abian, Arthur B. Brown (1958)
Bollettino dell'Unione Matematica Italiana
Similarity:

Superposition of imbeddings and Fefferman's inequality

Miroslav Krbec, Thomas Schott (1999)
Bollettino dell'Unione Matematica Italiana
Similarity:

In questo lavoro si studiano condizioni sufficienti sulla funzione peso V, espresse in termini di integrabilità, per la validità della disuguaglianza

Summary

- duality of math representation discussed
- verified complex workflow and proven technologies and tools for DML
- scalable solution for math formulae search researched, implemented, tested and integrated into current version of EuDML system!
- novel scalable Math search in EuDML is up and running, with several novel math-aware approaches developed and *in production use*
- MIR/MIaS project pages – https://mir.fi.muni.cz/
- math-aware methods for document similarity (MathML2text, gensim)

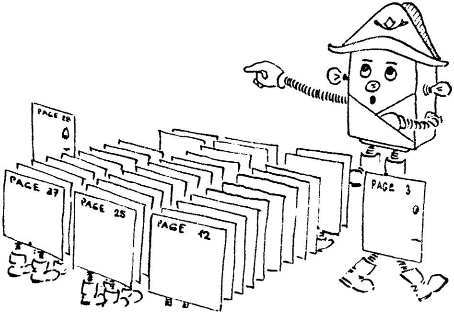
Summary and future work

Accepting duality of MIR interface 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* ;-).

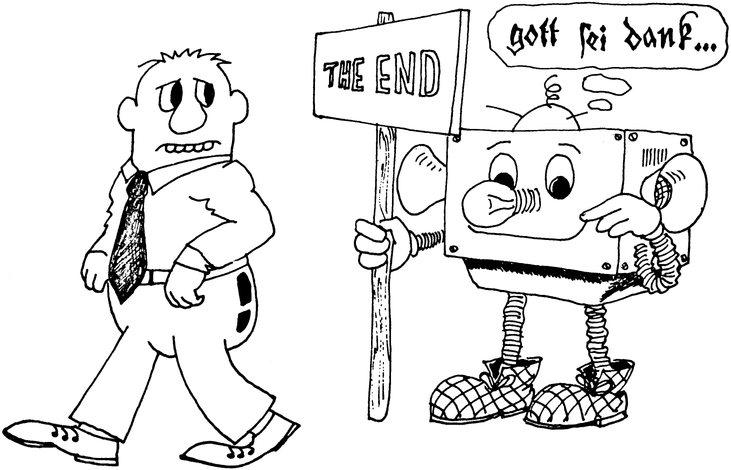
For more information see papers in SpringerLink (MKM 2011, Bertinoro) [5] and ACM DL (DocEng 2011, Mountain View) [8].

Acknowledgments and questions?



Acknowledgements: EuDML project (funding), EuDML colleagues, Martin LĂška, Michal Růžička, David Formánek and authors and contributors of other tools used or mentioned.

End of talk



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>

LĂĹĄka, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math T ask. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task. pp. 686-691. NII, Tokyo, 2013. PDF

D. FormĂĄnek, M. LĂĹĄka, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pp. 91–103, Aachen, 2012.

Sojka, Petr and Martin LĂĹĄka. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>

- 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>.
- Credits for LDA pictures goes to David M. Blei.
- Credits for illustrations goes to Jiří Franek.