The European Digital Mathematical Library: An Overview of Math Specific Technologies

Petr Sojka

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

National Institute of Informatics, Tokyo
June 24th, 2013, 1:30PM

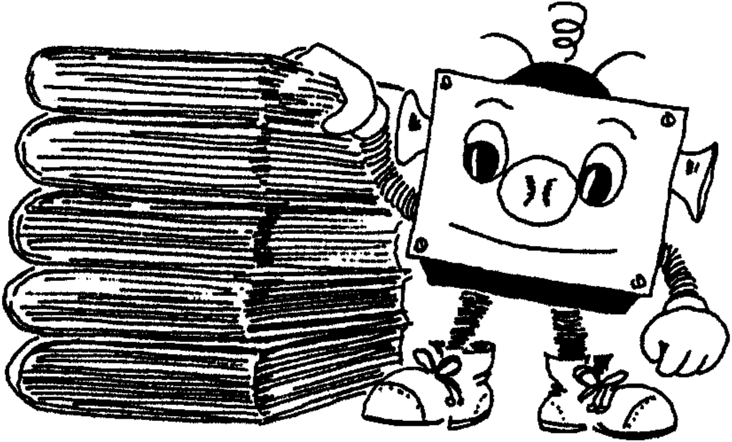
EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Outline and take-home message

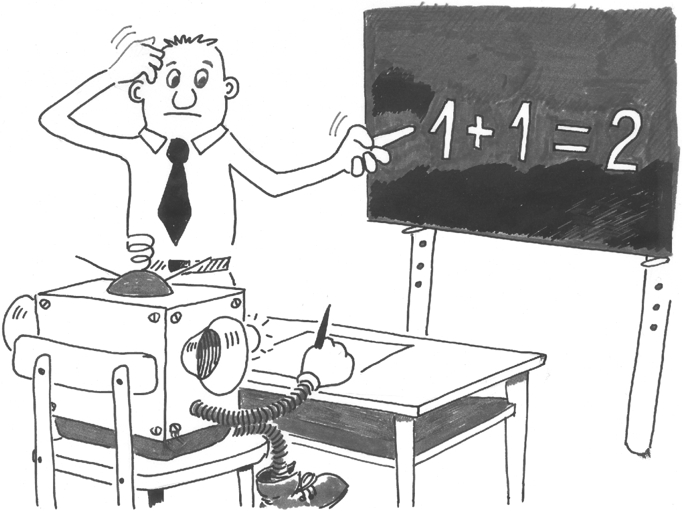
1. Pictorial overview
2. Motivation, vision of WDML, PubMed Central for Mathematics
3. Data aggregation from local DMLs
4. Conversions
5. Search
6. Similarity
7. Conclusions

Towards the dream of *math-aware* WDML: EuDML



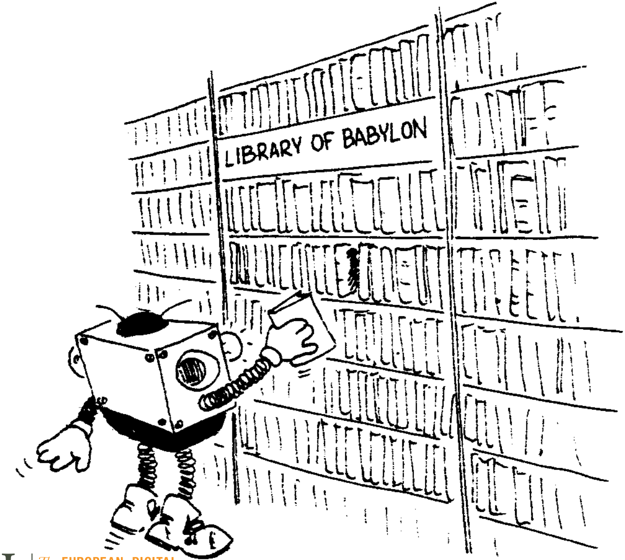
Information overload in globalized *scientific* world

Mathematics should follow other sciences (HEP, PMC,...)

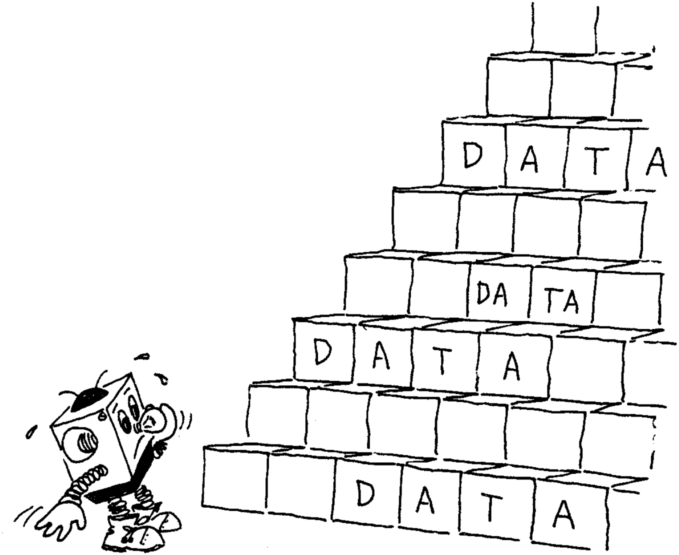


The European Digital Mathematics Library: *EuDML*

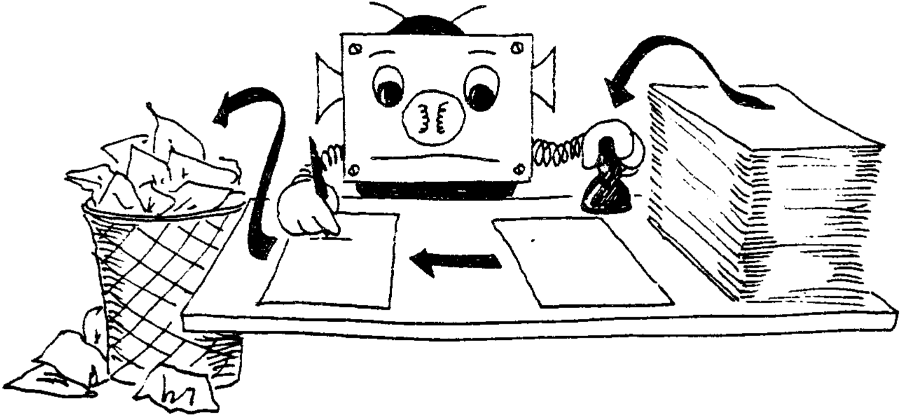
‘Bottom up’ deployment towards EU or *worldwide scale*



EuDML: from local data collections to the virtual DL

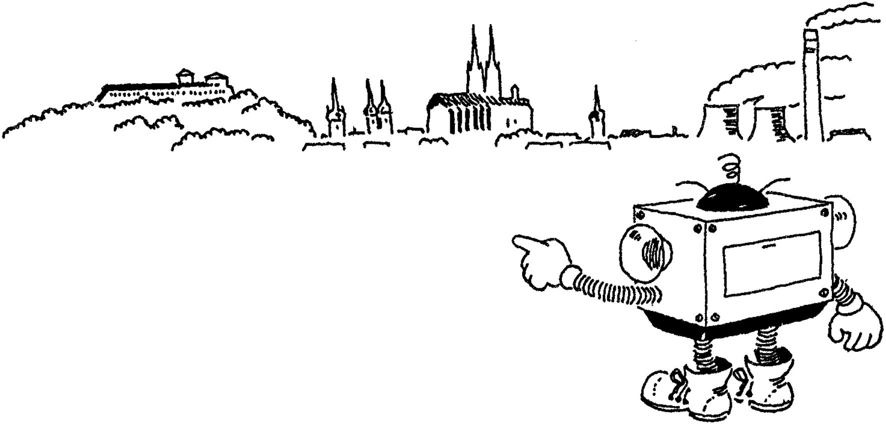


From paper to digital *workflow*

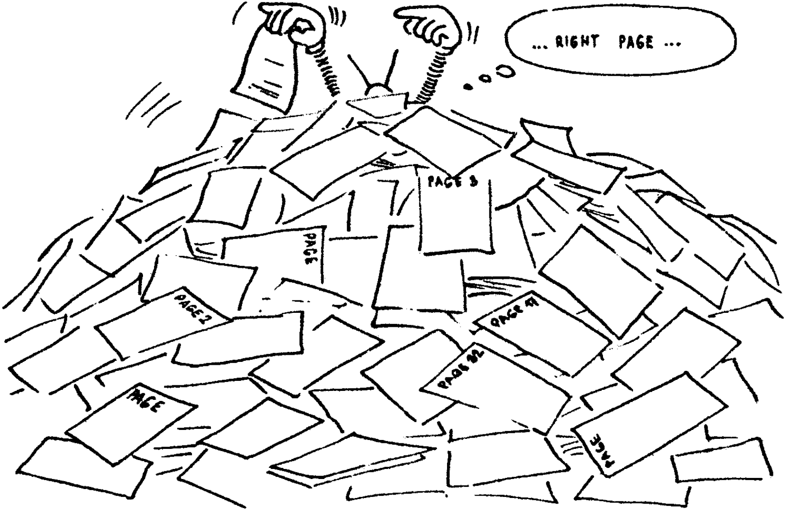


Retro-digitization, *accessible* digital library development

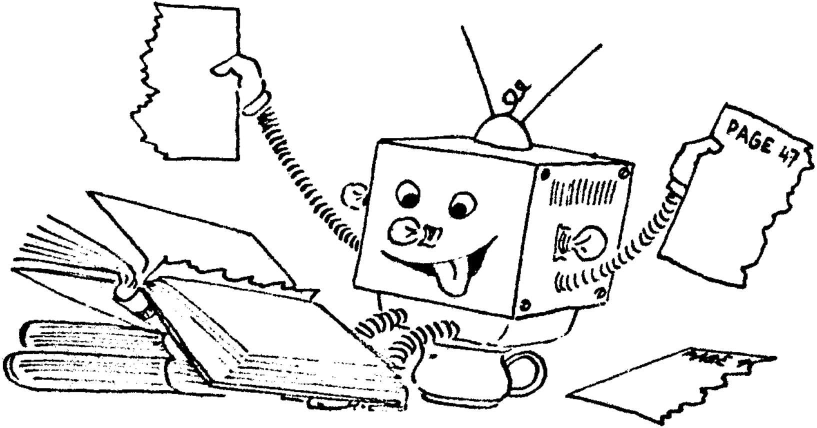
Experiences from project *DML-CZ* for EuDML (Brno, CZ)



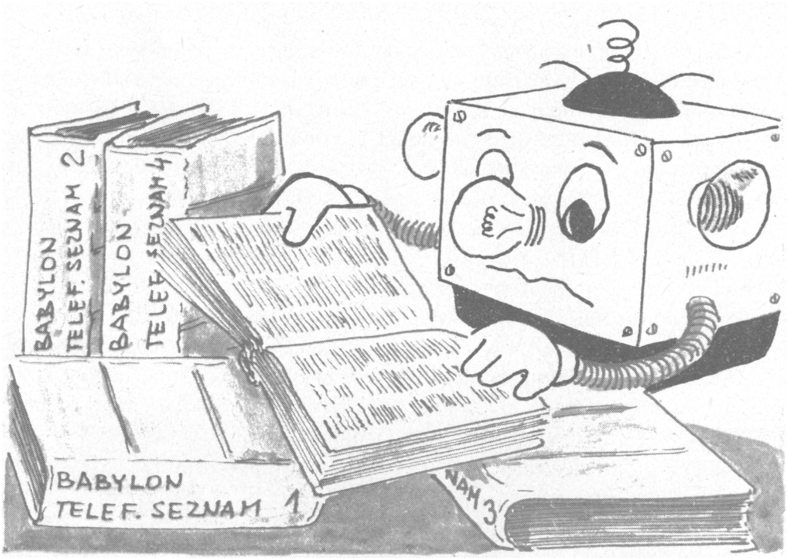
EuDML: new approaches to *math document retrieval*



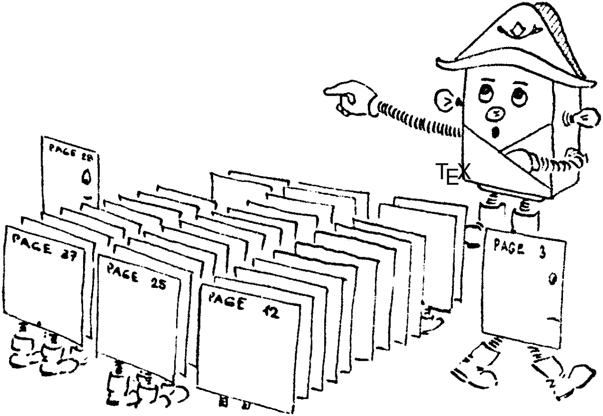
New approaches to *math-aware similarity, clustering and accessibility*



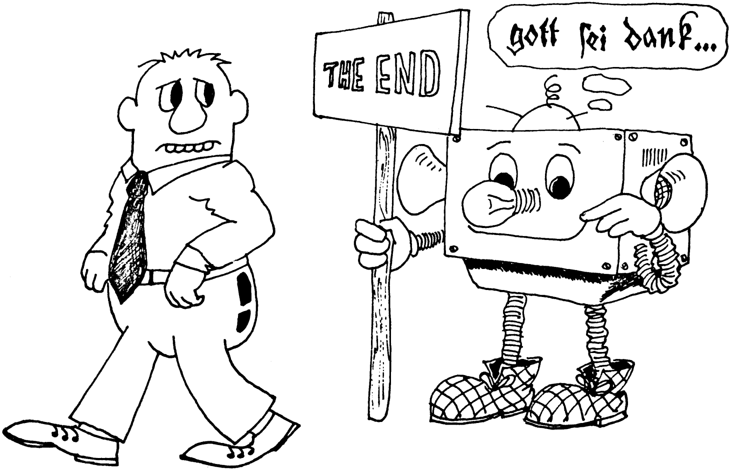
Tools for *automated math extraction* from PDF



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



End of talk overview



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.

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.

Even other attempts on the European level (FP5, FP6) were not successful.

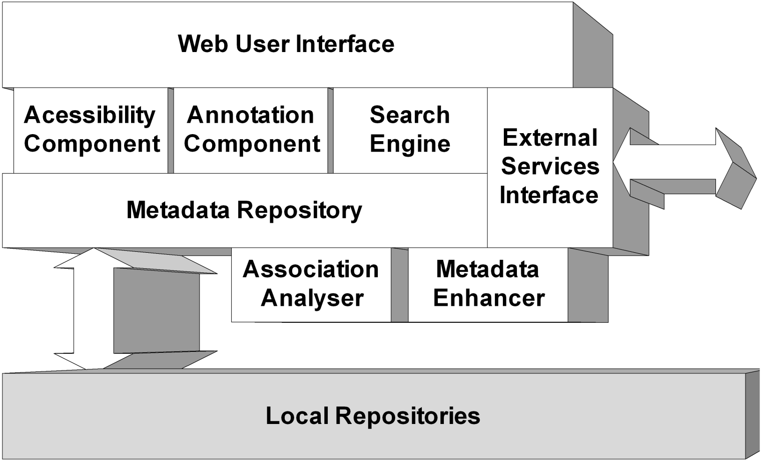
Vision of European Digital Mathematics Library

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, DLs and publishers:



One portal: European Digital Mathematics Library

Aggregation of data from building bricks of regional repositories

14 data and technology providers plus associated partners as ZMath, Göttingen library,…

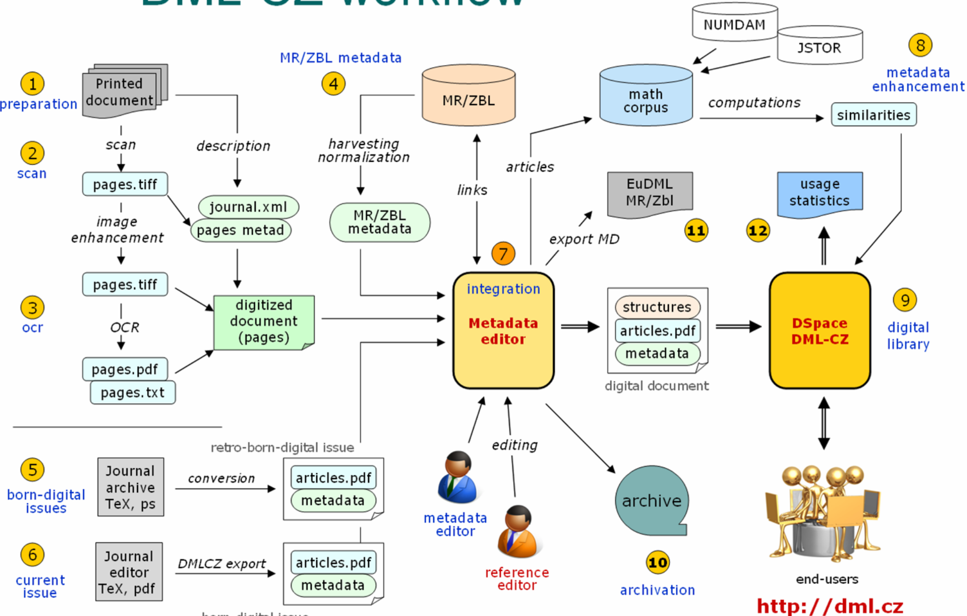
DML content providers serve mostly publisher's or regional more or less established DML repositories: The Czech Digital Mathematics Library DML-CZ, NUMDAM, DML-PL, DML-PT, DML-GR, DML-BG, DML-ES,…

Aggregation via standard OAI-PMH protocol (OAI servers run by data providers).

EuDML metadata schema(s) was borrowed from NLM (heavily funded by US NiH), as it allows also math-awareness (e.g. math stored both in TEX and MathML), and fully fledged reference lists.

Inovation, rather than research. Example of DML-CZ: <http://dml.cz> with 30,000+ papers (300,000+ pages). For more, see (who, what, browse, browse similar, how to search).

DML-CZ workflow



1 preparation
Printed document
scan
description
pages.tiff
journal.xml
pages metad

2 scan
image enhancement
pages.tiff

3 ocr
OCR
pages.pdf
pages.txt
digitized document (pages)

4 MR/ZBL metadata
MR/ZBL
harvesting normalization
MR/ZBL metadata
links

5 born-digital issues
Journal archive TeX, ps
conversion
retro-born-digital issue
articles.pdf
metadata

6 current issue
Journal editor TeX, pdf
DMLCZ export
articles.pdf
metadata
born-digital issue

7
integration
Metadata editor
editing
metadata editor
reference editor

articles
NUMDAM
JSTOR
math corpus
computations
similarities

8 metadata enhancement

export MD
EuDML MR/Zbl
11

structures
articles.pdf
metadata
digital document

DSpace DML-CZ
9 digital library

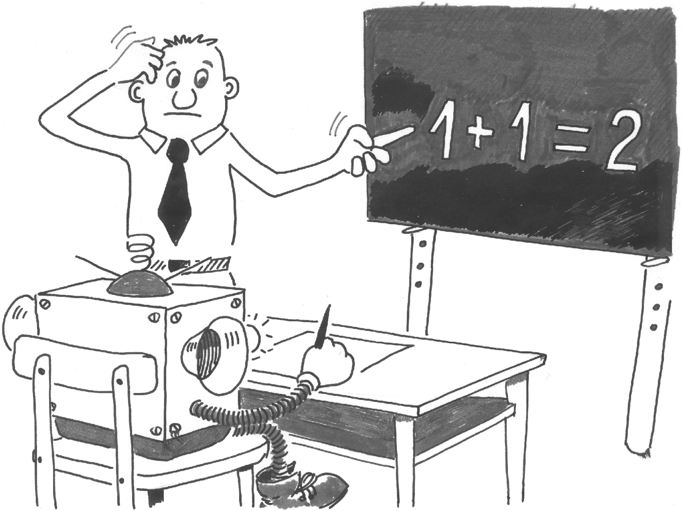
usage statistics
12

archive
10 archivation

end-users
http://dml.cz

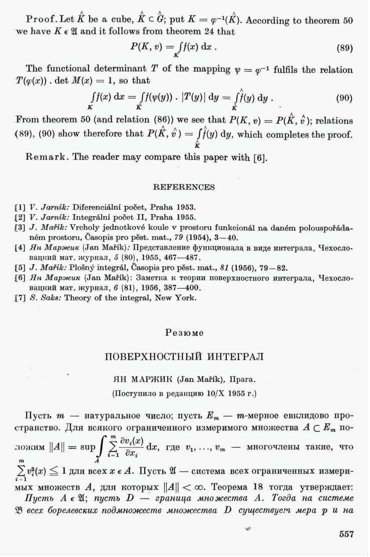
© DML-CZ, 2009

Challenges of Math handling: OCR, indexing, search…



Take care! “God is in the details.” (Mies van der Rohe)

Data heterogenity, specificity: no free lunch to unify



Proof. Let $\hat{K}$ be a cube, $\hat{K} \subset \hat{G}$; put $K = \varphi^{-1}(\hat{K})$. According to theorem 50 we have $K \epsilon \mathfrak{A}$ and it follows from theorem 24 that

$$P(K, v) = \int_K f(x)\,\mathrm{d}x\,. \tag{89}$$

The functional determinant T of the mapping $\psi = \varphi^{-1}$ fulfils the relation $T(\varphi(x))\,.\,\det M(x) = 1$, so that

$$\int_K f(x)\,\mathrm{d}x = \int_{\hat{K}} f(\psi(y))\,.\,|T(y)|\,\mathrm{d}y = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y\,. \tag{90}$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y)\,\mathrm{d}y$, which completes the proof.

Remark. The reader may compare this paper with [6].

REFERENCES

[1] *V. Jarník:* Diferenciální počet, Praha 1953.

[2] *V. Jarník:* Integrální počet II, Praha 1955.

[3] *J. Mařík:* Vrcholy jednotkové koule v prostoru funkcionál na daném polouspořádaném prostoru, Časopis pro pěst. mat., *79* (1954), 3—40.

[4] *Ян Маржик* (Jan Mařík): Представление функционала в виде интеграла, Чехословацкий мат. журнал, *5* (80), 1955, 467—487.

[5] *J. Mařík:* Plošný integrál, Časopis pro pěst. mat., *81* (1956), 79—82.

[6] *Ян Маржик* (Jan Mařík): Заметка к теории поверхностного интеграла, Чехословацкий мат. журнал, *6* (81), 1956, 387—400.

[7] *S. Saks:* Theory of the integral, New York.

Резюме

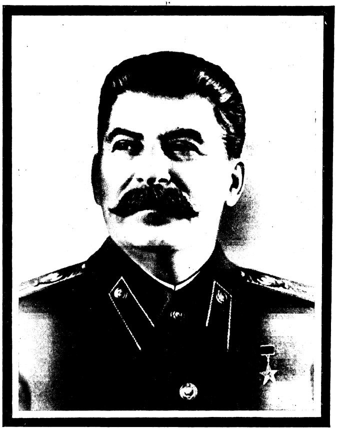
ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЖИК (Jan Mařík), Прага.

(Поступило в редакцию 10/X 1955 г.)

Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^m \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^m v_i^2(x) \leqq 1$ для всех $x \epsilon A$. Пусть $\mathfrak{A}$ — система всех ограниченных измеримых множеств A, для которых $\|A\| < \infty$. Теорема 18 тогда утверждает: *Пусть $A \epsilon \mathfrak{A}$; пусть D — граница множества A. Тогда на системе $\mathfrak{B}$ всех борелевских подмножеств множества D существует мера p и на*

557



ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

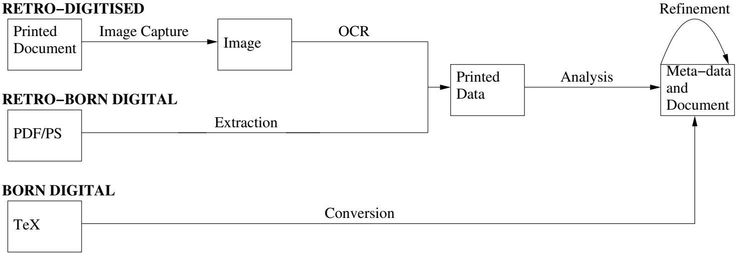
Document accessibility 4 DML processing challenges

Conversions (inversion of authoring+typesetting) needed from:

born-digital period: typesetting by TEX with export of [meta]data into digital library: maxTract

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), *two-layer PDF*

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution: finally Tesseract



From PDF to MathML (via LaTeX)

Most fulltext available as PDF only, often as low quality scanned volume pages. Aggregation via IP protected OAI-PMH, including the PDFs behing moving wall.

Workflow based in the case of:

born-digital PDFs: on maxTract, otherwise on PDFBox (plain text);

bitmap PDFs: on Infty, otherwise on Tesseract (no math).

Infty from Fukuoka

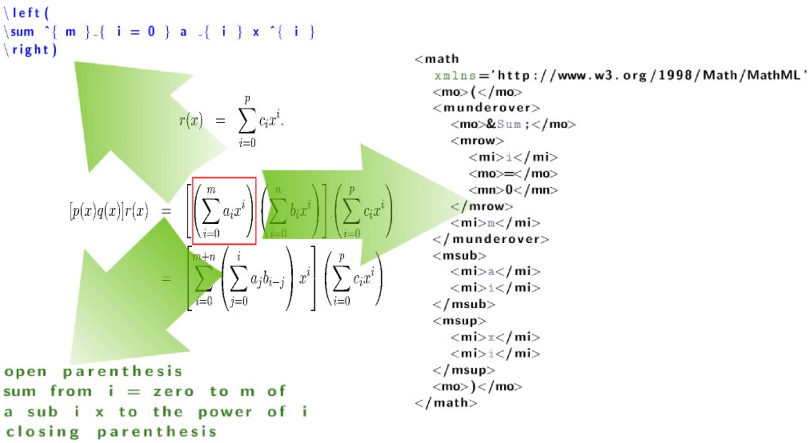
Run in parallel in Brno, Grenoble and Lisbon to speed up. Almost 200K papers (more than 1M pages and still running).

Working with prof. Suzuki to improve further (automation, support for Russian, LaTeX driver,...).

Automated only, no time (and money) to fix OCR errors.

MathML output used for [internal] indexing and similarity computations only, not for metadata or export.

maxTract from Birmingham



```
\left(
\sum ^{ m }_{ i = 0 } a _{ i } x ^{ i }
\right)
```

$$r(x) = \sum_{i=0}^{p} c_i x^i.$$

$$[p(x)q(x)]r(x) = \left[\boxed{\left(\sum_{i=0}^{m} a_i x^i\right)}\left(\sum_{i=0}^{n} b_i x^i\right)\right]\left(\sum_{i=0}^{p} c_i x^i\right)$$

$$= \left[\sum_{i=0}^{m+n}\left(\sum_{j=0}^{i} a_j b_{i-j}\right) x^i\right]\left(\sum_{i=0}^{p} c_i x^i\right)$$

```
<math
  xmlns='http://www.w3.org/1998/Math/MathML'
  <mo>(</mo>
  <munderover>
    <mo>&Sum;</mo>
    <mrow>
      <mi>i</mi>
      <mo>=</mo>
      <mn>0</mn>
    </mrow>
    <mi>m</mi>
  </munderover>
  <msub>
    <mi>a</mi>
    <mi>i</mi>
  </msub>
  <msup>
    <mi>x</mi>
    <mi>i</mi>
  </msup>
  <mo>)</mo>
</math>
```

open parenthesis
sum from i = zero to m of
a sub i x to the power of i
closing parenthesis

maxTract from Birmingham II: adding accessibility

Adding accessibility to mathematical documents on multiple levels:

- access to content for print impaired users, such as those with visual impairments, dyslexia or dyspraxia
- output compatible with web browsers, screen readers and tools such as copy and paste, which is achieved by enriching the regular text with mathematical markup. The output can also be used directly, within the limits of the presentation MathML produced, as machine readable mathematical input to software systems such as Mathematica or Maple.

On EuDML 10k+ fulltexts are served, mostly for reading in Chrome (HTML5 output) and/or Adobe Acrobat Reader (as multiple-layer PDFs, [no tagged PDFs yet]).

Metadata and conversions: 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*

REPOX engine (by IST Lisbon) to remap different metadata formats to unique representation.

Metadata on the web—W3C standards: MathML, WAI-ARIA (Web Accessibility Initiative—Acessible Rich Internet Applications), WCAG (Web Content Accessibility Guidelines) 2.0.

Big volumes: → high *automation* to save costs: converting to MathML (via Tralics) to allow discoverability and indexing (formulae similarity search). 130+K fulltexts with MathML, Infty still running….

Why Search?

Vast amounts of [moving] contents in digital libraries: from browsing to *search*; from static links to indirect search links, or even semantic search.

Searching is crucial part of *accessibility* and *exploration* of the great ideas around, carved into 0s and 1s.

Pragmatic decisions on math indexing level: *presentation* vs. *content* vs. *semantic*. In EuDML first step: scalable presentation (structural), with methods (tree indexing and weighting) extendable for content or semantic.

Why Math Search (MIR)?

A picture is worth thousands words.

“A math formulae is worth of hundreds of words.“ (Ross Moore)

There are papers with more formulae than plain text.

Precision vs. Recall optimisation: optimizing recall is better for exploratory searching (we have not opted for precision as holy grail at the moment).

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.

We did not start from scratch

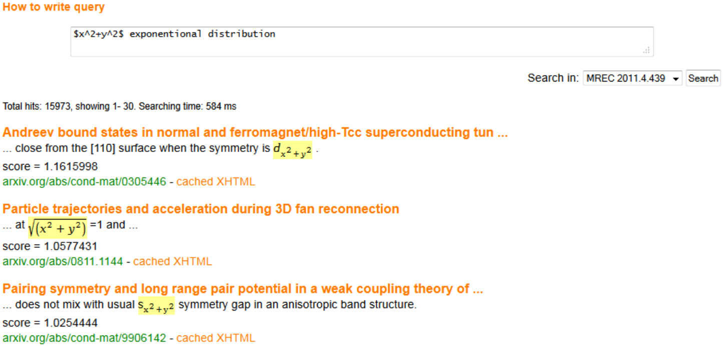
Compare `google.com/search?q=Einstein` with math-aware search of `Einstein+$E=mc^2$` over arXiv.

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
```

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

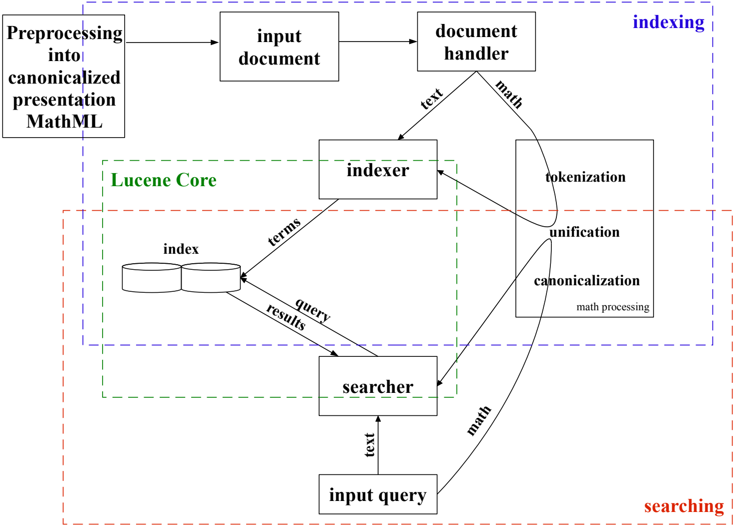
Pairing symmetry and long range pair potential in a weak coupling theory of ...

... does not mix with usual $s_{x^2+y^2}$ symmetry gap in an anisotropic band structure.

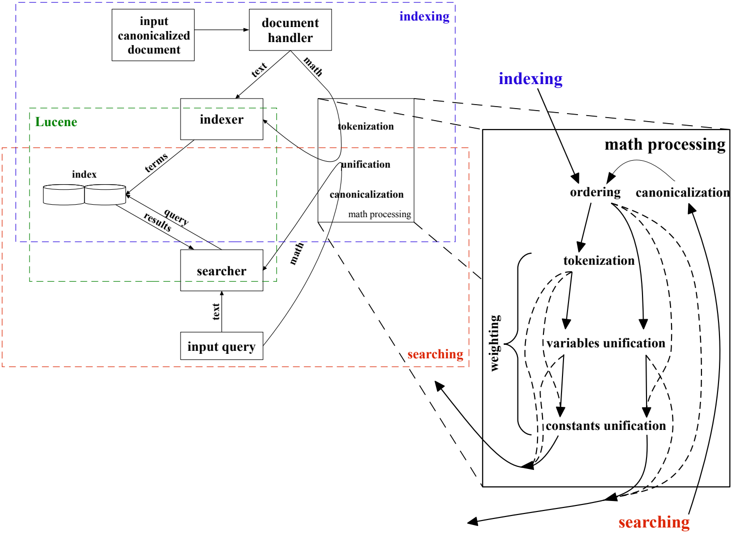
score = 1.0254444

arxiv.org/abs/cond-mat/9906142 - cached XHTML

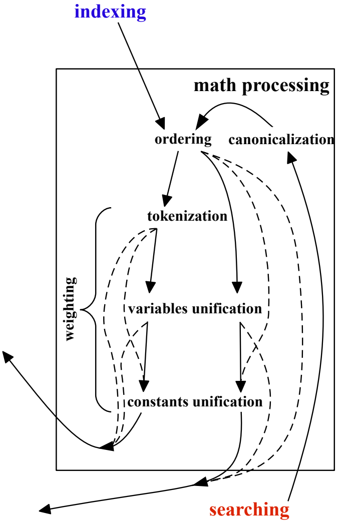
MSE overall design

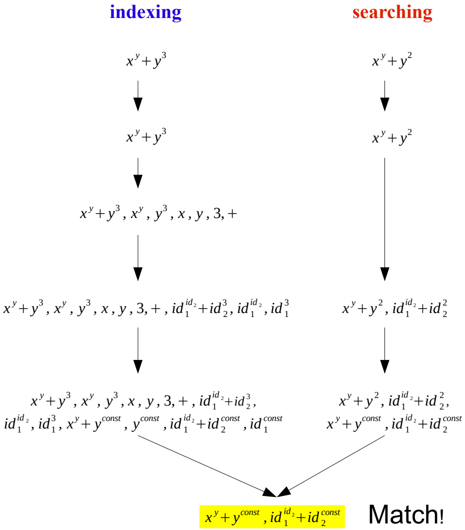


Math indexing design



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

$$x^y+y^2$$

$$x^y+y^2$$

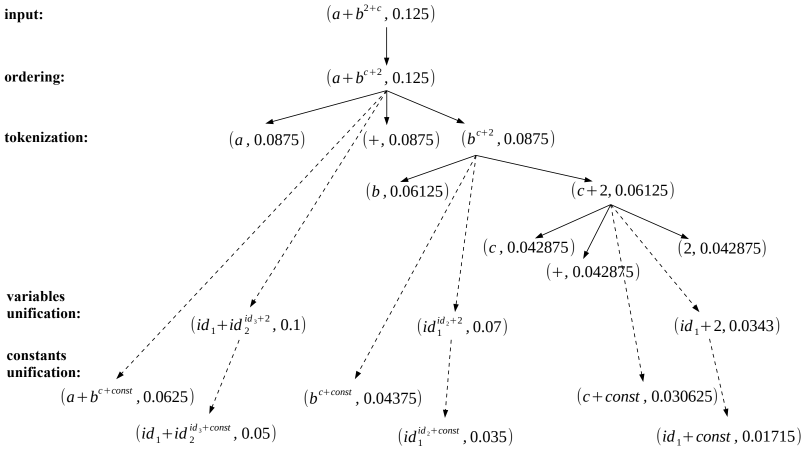
$$x^y+y^2, id_1^{id_2}+id_2^2$$

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

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

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

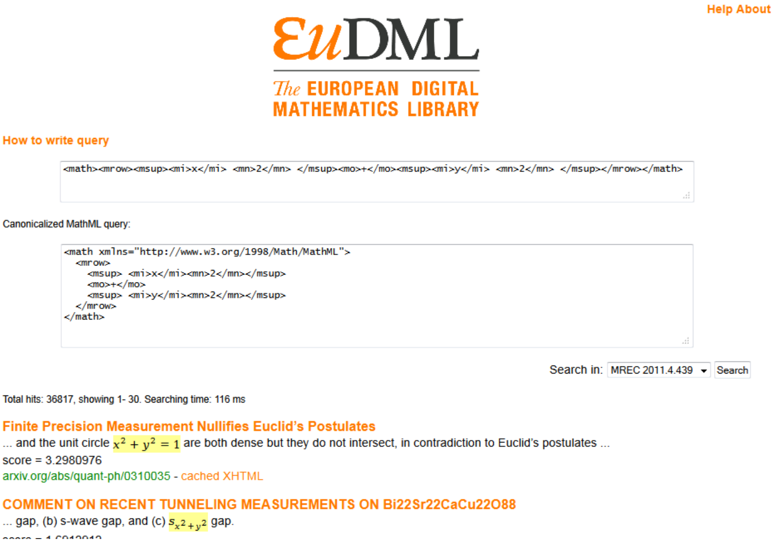
constants unification: $(a+b^{c+const}, 0.0625)$ $(b^{c+const}, 0.04375)$ $(c+const, 0.030625)$

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

Implementation

- Java
- Lucene 3.1.0, now switching to Lucene/Solr 4
- Mathematical part implements Lucene'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



How to write query

```
<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 comments

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

- MathML/TEX input (Tralics [2] for conversion to MathML [**?**])
- Canonicalization of the query – problems with UMCL library [1]
- Matched document snippet generation
- MathJax for nicer math rendering and better portability
- Snuggle TeX for on-the-fly as-you-type rendering

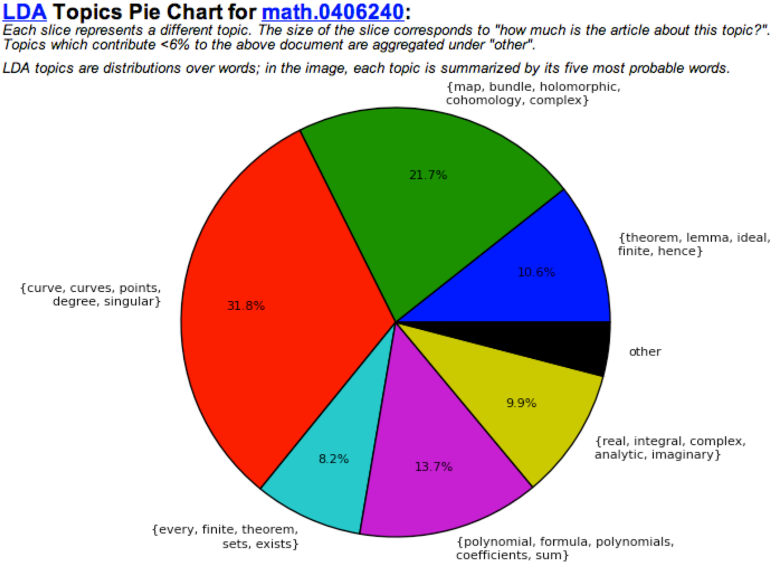
All up and ready on the EuDML system.

Searching (semantically) similar papers

Exploration of a DML: browsing (semantically) similar papers

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

Leading Edge Example: Automated Meaning Picking from Texts



LDA Topics Pie Chart for math.0406240:

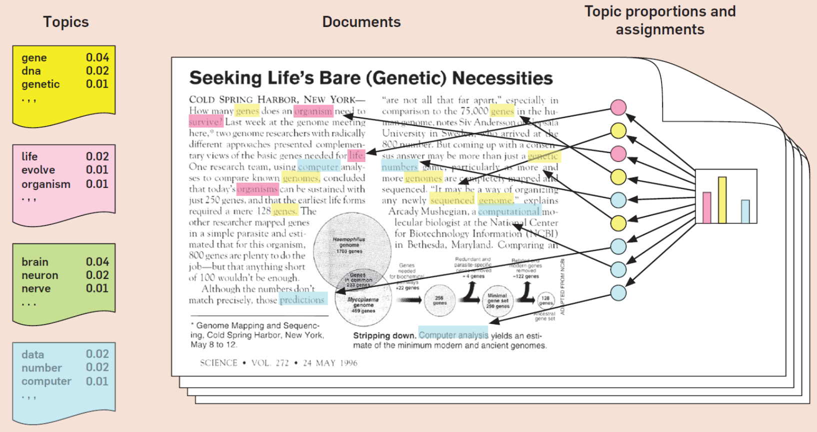
Each slice represents a different topic. The size of the slice corresponds to "how much is the article about this topic?". Topics which contribute <6% to the above document are aggregated under "other".

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

{map, bundle, holomorphic, cohomology, complex}
21.7%
{theorem, lemma, ideal, finite, hence}
10.6%
{curve, curves, points, degree, singular}
31.8%
other
9.9%
{real, integral, complex, analytic, imaginary}
8.2%
13.7%
{every, finite, theorem, sets, exists}
{polynomial, formula, polynomials, coefficients, sum}

Probabilistic Topical Modeling: Latent Dirichlet Allocation

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


Topics
Documents
Topic proportions and assignments
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
. , ,
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

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

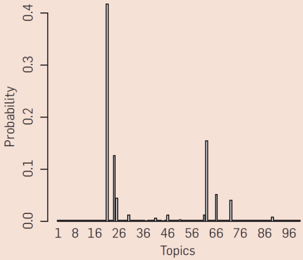
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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

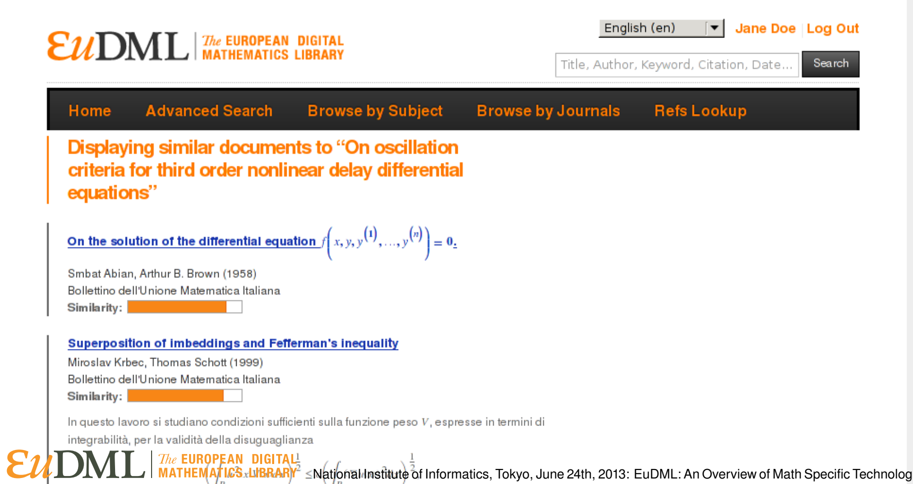
- all topics computed automatically from document corpora



"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

[illegible]

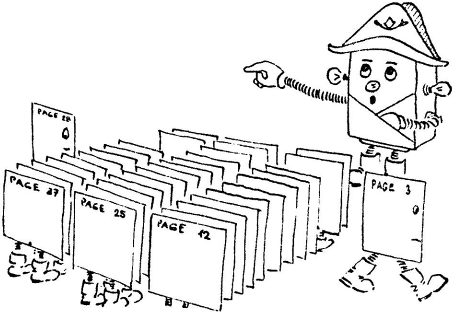
Summary

- EuDML is up and running, with several novel math-aware approaches developed and *in use*
- 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!
- MIR/MIaS project pages – https://mir.fi.muni.cz/
- math-aware methods for document similarity (MathML2text, gensim)
- a lot more on <http:/project.eudml.org> (e.g. PDF size reduction of 62% of original already CCITT-G4 compressed PDFs, etc.)

Future work

- DML workshop series, join us at DML 2013 c/o CICM Bath in UK in July 2013
- Activities towards WDML (Sloan funding,…)
- EuDML initiative consortium, further sustainability solutions (grant proposal writing).
- Improved MathML canonicalization and new preprocessing filters, search developed and evaluated with the use of EuDML math query database of intentions.
- Addition of Content MathML tree indexing.
- NCTIR 11!

Acknowledgments and questions?



Acknowledgements: EuDML project (funding), ELIAS (trip here), EuDML colleagues, and authors and contributors of tools used.

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>

Líška, 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. Formánek, M. Líška, 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 Líška. 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>

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.