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
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.
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 was:

- to master the technology, develop tools and offer them;
- concept of *moving wall* to motivate and engage commercial publishers without Open Access business 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:

![Diagram of EuDML architecture]

- **Web User Interface**
- **Accessibility Component**
- **Annotation Component**
- **Search Engine**
- **External Services Interface**
- **Metadata Repository**
- **Association Analyser**
- **Metadata Enhancer**
- **Local Repositories**
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
       - pages.tiff
     - Description
       - journal.xml
         - pages metadata
       - MR/ZBL metadata
     - Harvesting normalization
       - MR/ZBL
       - Links
         - articles
           - EuDML MR/ZBL

2. Scan
   - pages.tiff
   - Image enhancement
     - pages.pdf
     - pages.txt

3. OCR
   - Digitized document (pages)

4. MR/ZBL
   - MR/ZBL metadata
     - Articles
       - Links
         - Export MD
           - DSpace
             - DML-CZ
               - Metadata editor
                 - Structures
                   - Articles.pdf
                     - Metadata
               - Digital document
                 - Editing
                   - Retro-born-digital issue
                     - Conversion
                       - TeX, ps
                         - Journal archive
                       - DMLCZ export
                         - TeX, pdf
                           - Current issue
                             - Metadata editor
                               - Reference editor
                                 - Archivation
                                   - End-users
                                     - http://dml.cz

© DML-CZ, 2009

National Institute of Informatics, Tokyo, June 24th, 2013: EuDML: An Overview of Math Specific Technologies
Challenges of Math handling: OCR, indexing, search…
Take care! “God is in the details.” (Mies van der Rohe)
Data heterogeneity, specificity: no free lunch to unify

Proof. Let \( \hat{K} \) be a cube, \( \hat{K} \subset \hat{O} \); put \( K = q^{-1}(\hat{K}) \). According to theorem 50 we have \( K \in \mathcal{F} \) and it follows from theorem 24 that

\[
P(K, v) = \int_{K} f(x) \, dx.
\]

(89)

The functional determinant \( T \) of the mapping \( \psi = q^{-1} \) fulfills the relation \( T(q(x)) \cdot \det M(x) = 1 \), so that

\[
\int_{K} f(x) \, dx = \int_{K} f(q(y)) \cdot |T(y)| \, dy = \int_{K} \psi' \, dy.
\]

(90)

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

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

References


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 \( A \subset E_m \) по- лезным \( |A| = \sup \sum \overline{d}_{i} \cdot \overline{d}_{i} \) \( \sum \overline{d}_{i} \cdot \overline{d}_{i} \), где \( v_1, \ldots, v_m \) — многочлены такие, что

\[
\sum \overline{d}_{i} \cdot \overline{d}_{i} \leq 1 \text{ для всех } x \in A.
\]

Пусть \( \mathcal{F} \) — система всех ограниченных измеримых множеств \( A \), для которых \( |A| < \infty \). Теорема 18 тогда утверждает:

Пусть \( A \in \mathcal{F} \); пусть \( D \) — граница множества \( A \). Тогда на системе \( \mathcal{F} \) всех борелевских подмножеств множества \( D \) существует мера \( \mu \) и на

557
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
Most fulltext available as PDF only, often as low quality scanned volume pages. Aggregation via IP protected OAI-PMH, including the PDFs behind 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

\[
\left( \sum_{i=0}^{m} a_i x^i \right)^p
\]

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

\[
[p(x)q(x)]r(x) = \left[ \sum_{i=0}^{m} a_i x^i \right] \left[ \sum_{i=1}^{n} b_i x^i \right] \left[ \sum_{i=0}^{p} c_i x^i \right]
\]

\[
= \sum_{i=0}^{m+n} \left( \sum_{j=0}^{i} c_j b_{i-j} \right) x^i \left( \sum_{i=0}^{p} c_i x^i \right)
\]
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 heterogeneity, 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.


Big volumes: \(\rightarrow\) high \textit{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.”
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, $\alpha$-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

$e^{x^2+y^2}$ exponential distribution

Search in: MREC 2011.4.39  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
MSE overall design

Preprocessing into canonicalized presentation MathML -> input document -> document handler

input query -> searcher

indexer

index

terms

query

results

indexing

Lucene Core

text

math

text

math

text

searching

query

results

math processing

canonicalization

unification

tokenization

math

Indexer

Input query

Math processing

Canonicalization

Unification

Tokenization

searching

indexing

Lucene Core

searcher

input query

input document

document handler

Preprocessing into canonicalized presentation MathML

Overview
Motivation, EuDML
Aggregation
Conversions
Search
Similarity
Conclusions

Overview
Motivation, EuDML
Aggregation
Conversions
Search
Similarity
Conclusions

EuDML: The European Digital Mathematics Library

National Institute of Informatics, Tokyo, June 24th, 2013: EuDML: An Overview of Math Specific Technologies
Overview

Motivation, EuDML

Aggregation

Conversions

Search

Similarity

Conclusions

Math indexing design

Overview

Motivation, EuDML

Aggregation

Conversions

Search

Similarity

Conclusions

Math indexing design
Example

math processing

ordering
canonicalization

tokenization

variables unification

constants unification

indexing

searching

Match!

Overview
Motivation, EuDML
Aggregation
Conversions
Search
Similarity
Conclusions

National Institute of Informatics, Tokyo, June 24th, 2013: EuDML: An Overview of Math Specific Technologies
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)

variables

unification:

(id_1 + id_2^{id_1+2}, 0.1)

(constants

unification:

(a + b^{c+const}, 0.0625)

(b^{c+const}, 0.04375)

(c + const, 0.030625)

(id_1 + id_2^{id_5+const}, 0.05)

(id_1^{id_5+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

Canonicalized MathML query:

```xml
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup><mi>x</mi><mn>2</mn></msup> + <msup><mi>y</mi><mn>2</mn></msup> = 1
  </mrow>
</math>
```

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


**COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88**
... gap, (b) s-wave gap, and (c) $s_{\pm}^2$ gap.

[arxiv.org/abs/12120310](http://arxiv.org/abs/12120310)
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.

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

- topic: weighted list of words
- document: weighted list of topics
Topical Modeling: Latent Dirichlet Allocation II

- all topics computed automatically from document corpora

![Graph showing topics and their probabilities]

- "Genetics": human, genome, genetic, genes, sequence, gene, molecular, sequencing, map, information, genetics, mapping, project, sequences
- "Evolution": evolution, evolutionary, species, organisms, life, origin, biology, groups, phylogenetic, living, diversity, group, new, two, common
- "Disease": disease, host, bacteria, diseases, resistance, bacterial, new, strains, control, infectious, malaria, parasite, parasites, united, tuberculosis
- "Computers": computer, models, information, data, computers, system, network, systems, model, parallel, methods, networks, software, new, 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 $f(x, y, y^{(1)}, ..., y^{(n)}) = 0$.

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

• 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!
Acknowledgements and questions?

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


Credits for LDA pictures goes to David M. Blei.

Credits for illustrations goes to Jiří Franek.