MIR.fi.muni.cz: Past, Present and Future

Petr Sojka

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

July 8th, 2012

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY
Math-aware Akiko thinks and wants to convey a message

Let A, B are mathematicians. Let us name them Akiko and Bruce.
Let Akiko has a thought, ideas in math, she wants to convey.
Let she linearize it, mark it up, and express it disambiguated in markup language.
<table>
<thead>
<tr>
<th>Motivation</th>
<th>The Past</th>
<th>The Present</th>
<th>The Future</th>
<th>Conclusions and Future Work</th>
</tr>
</thead>
</table>

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
Math-aware Akiko writes and publishes

Let she runs a typesetting program on marked up electronic data and let she creates a *paper* in digital format.

Let she post it on her webpage in PDF.
Math-aware Akiko communicates to Bruce

$E = mc^2$

?!
Math-aware Bruce does research

Bob is a working mathematician and let he does research.

Doing re-search is search.

He is redoing search, crucial operation for him.

Let he searches for information in WDML (3,000,000+ reviewed papers in total, 140,000+ new reviewed papers a year in Zbl).
Math-aware Bruce does navigational search and reads

He puts words [and formulae] into search box of his web browser and finds Akiko’s paper.

His brain does paper layout analysis and recognizes pictures, letters, words, formulae and other objects on the Akiko paper page.

He interprets recognized language, using his common sense, by processing language syntax, semantics and pragmatics.

He has got the Akiko’s message finally!
Levels of [math] retrieval

Images?
Strings?
Words?
Collocations, phrases, formulae (syntax)?
Collocations, phrases, formulae meaning (semantics)?
Information, ideas in context (pragmatics)?
By other means (telepathy)?
On all layers available, plus processing using MKM techniques!
Notes worth mentioning: on the sender side

Akiko provided PDF as mean of delivery (either as scanned bitmap or born-digital), e.g. *no explicitly marked/disambiguated/rich embedded semantics*.

No flexiforms or the like were used in the process. The message has to be “decrypted” on-the-fly during the process or after retrieval.

Web technologies and indexing from PDF was used.
User interface, query language and query debugging important, so is speed of search.

Bruce used his pragmatic competence to get the message, even though his semantic understanding of some words and collocations were different.

Even word meaning is subjective and moving target, nothing to be carved in stone (ontology :-)).

Bruce may have used navigational search rather than research search during search process. (Guha et al. distinguish two major forms of search: Navigational and Research.)
Navigational vs. Research search

“In navigational search, the user is using the search engine as a navigation tool to navigate to a particular intended document. Semantic Search is not applicable to navigational searches. In Research Search, the user provides the search engine with a phrase which is intended to denote an object about which the user is trying to gather/research information. There is no particular document which the user knows about that s/he is trying to get to. Rather, the user is trying to locate a number of documents which together will give him/her the information s/he is trying to find.”

(Wikipedia under ‘semantic search’)
The Past
From paper to *digital* library: instead of going to the classical library going to the web: *DML-CZ* since 2004
Information overload in globalized scientific world: navigational search needed in DML-CZ
Information overload also in specific domains (mathematics): research search needed only occasionally
From paper to digital workflow: radical changes and consequences for MIR

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
Retro-digitization, digital library developments
Specific needs in MIR: new tools developed in Brno, CZ for MIR purposes, DML-CZ is running there
New workflows and data processing: DSpace and Lucene for fulltext search not sufficient, MIaS needed → thesis series on MIR
New tools: [math-aware] semantic similarity engine gensim (Řehůřek)
From local DMLs like DML-CZ to bigger ones: EuDML since 2010
The European Digital Mathematics Library: EuDML
EuDML lesson: heterogeneity in data: in markup, formats, collections, working attitude,...
MIR for EuDML: new scalable tools development (reported CICM 2011)
Yes, you can! Navigational MIR NOW, once you collect and normalize data,…
End of historical overview
The Present
Motivation

The Past

The Present

The Future

Conclusions and Future Work

Math Indexer and Searcher – MIaS: WebMIaS with MREC (arXMLiv)

Input language: MathML

Canonicalized MathML query:

Search in: MREC 2011.4.439 

Total hits: 36817, showing 1-30. Searching time: 100 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 = 0.19934596
arxiv.org/abs/quant-ph/0310035 - cached XHTML

COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88
... gap, (b) s-wave gap, and (c) \( s_\pm + g \) gap.
score = 0.08392586
arxiv.org/abs/cond-mat/9803139 - cached XHTML

S and D-Wave Mixing in High-Tc Superconductors

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
WebMIaS Workflow based on the state-of-the-art tools (Lucene)
Martin Líška will disclose the MIaS details in the follow-up talk.
EuDML use case

80% scanned/bitmaps, only 20% born-digital, no fully marked NLM source.

Math context only now starts to appear on <http://eudml.org>

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
Q: Is elephant a wall (belly), hand fan (ear), solid pipe (tusk), pillar (leg), rope (tail) or tree branch (trunk)?
Levels of text/math understanding/processing

1.0 lexical – words, *strings* of characters/TeX’s $$.  
2.0 syntactical – phrases, *parsed* formulas (trees/MathML).  
3.0 semantical – *meaning* of parsed phrases (cloud tags/ontologies/OpenMath).

Problem of message (content+form) representation (of math when transporting the message over the web).  
Google around 1.5 now (no semantics, but for the purpose are people happy).
Many valid but different purposes for processing math

- Format choice *depends* on application’s *purpose*.
- Most applications have its own internal format anyway.
- For *exchange* it seems that *XML/MathML* (but which one?) currently wins (cut&paste in Windows 7, CAS).
- For authoring it seems that (La)TeX is preferred.
- Quite different requirements have theorem proving systems and computer algebra systems.
The Future

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
MIR.fi.muni.cz development plans and extensions based on MIaS

New canonicalization (DML talk on Monday) influence MIR experience considerably.

Wider unification and Content MathML indexing needed when moving towards research search. Is it really needed? if yes, then big research area of Math-aware NLP.


Semantic profiles (based on awesome Řehůřek’s gensim).

Ranking based on semantic profiling (e.g. MSC based).

Interactive User Experience [have you tried SearchPoint demo?] (Kacvinsky).

Sumary

Let A, B are mathematicians. Let they communicate via MIR systems like MIR.fi.muni.cz.

Let it happen.

It has happened in the Past.

It is happening in the Present, now.

It will happen in the Future. More to happen (e.g. at MIR happenning :-).
Questions

Questions?

Acknowledgements: This work was partially supported by the European Union through its Competitiveness and Innovation Programme (Information and Communication Technologies Policy Support Programme, ‘Open access to scientific information’, Grant Agreement No. 250503, a project of the European Digital Mathematics Library, EuDML).

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

MIR c/o CICM 2012, Jacobs University, Bremen, Germany
Motivation

The Past


The Present


The Future

Conclusions and Future Work

P. Sojka: MIR.fi.muni.cz: Past, Present and Future

<table>
<thead>
<tr>
<th>Motivation</th>
<th>The Past</th>
<th>The Present</th>
<th>The Future</th>
<th>Conclusions and Future Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>The LaTeXML project: The LaTeXML Developer Portal (Apr 2012), <a href="https://trac.mathweb.org/LaTeXML/">https://trac.mathweb.org/LaTeXML/</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>The Past</td>
<td>The Present</td>
<td>The Future</td>
<td>Conclusions and Future Work</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>-----------------------------</td>
</tr>
</tbody>
</table>