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

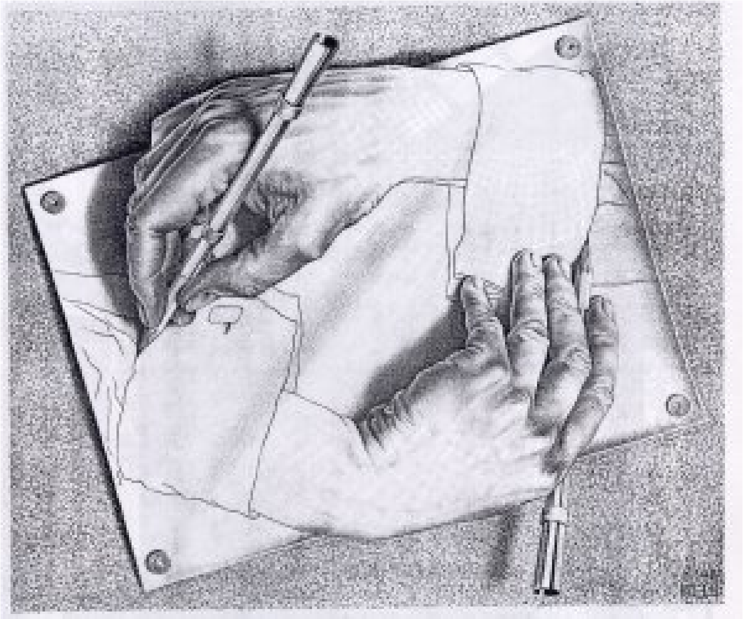
EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

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.

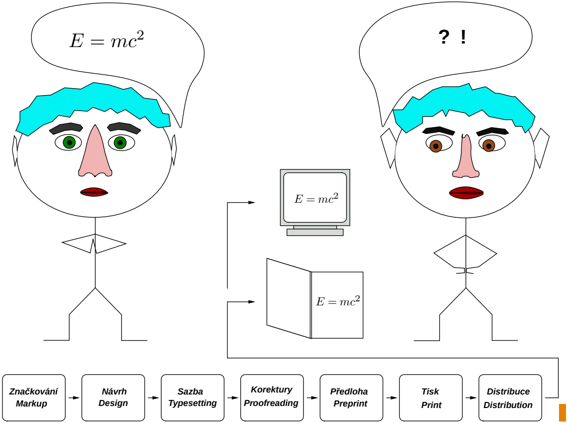


Math-aware Akiko writes and publishes

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



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

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.

Notes worth mentioning: on the receiver side

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 :-)).

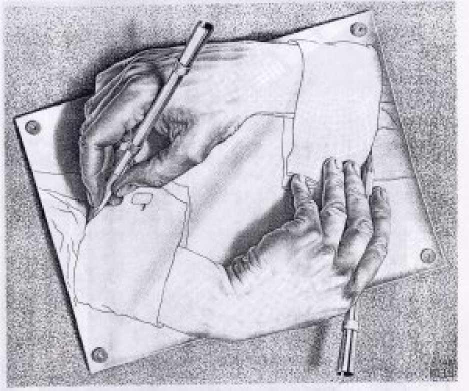
Bruce may have used *navigational search* rather than *research search* during search process. (Guha et al. distinguish two major forms of search: Navigational and Research.)

Navigational vs. Research search

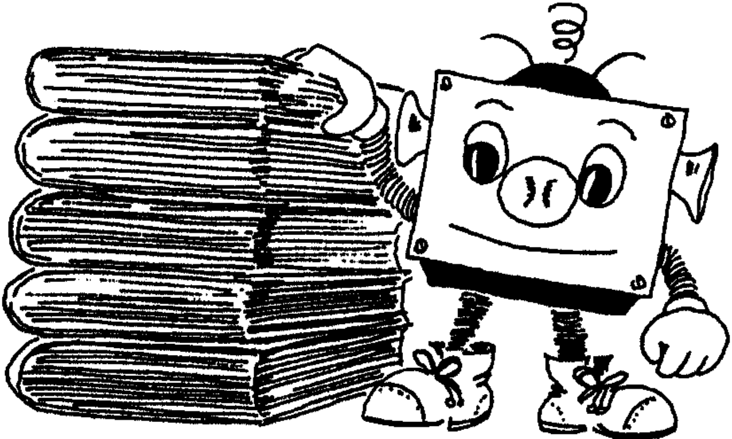
“In navigational search, the user is using the search engine as a navigation tool to navigate to a particular intended document. Semantic Search is not applicable to navigational searches. In Research Search, the user provides the search engine with a phrase which is intended to denote an object about 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

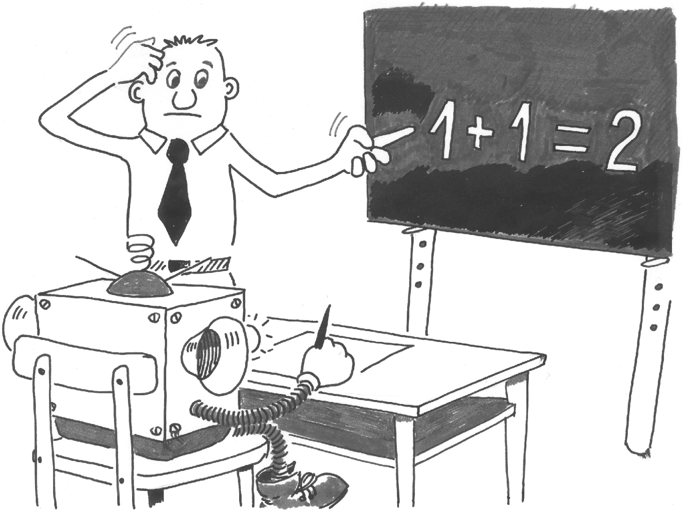


From paper to *digital* library: instead of going tothe classical library going to the web: *D*ML-CZ since 2004

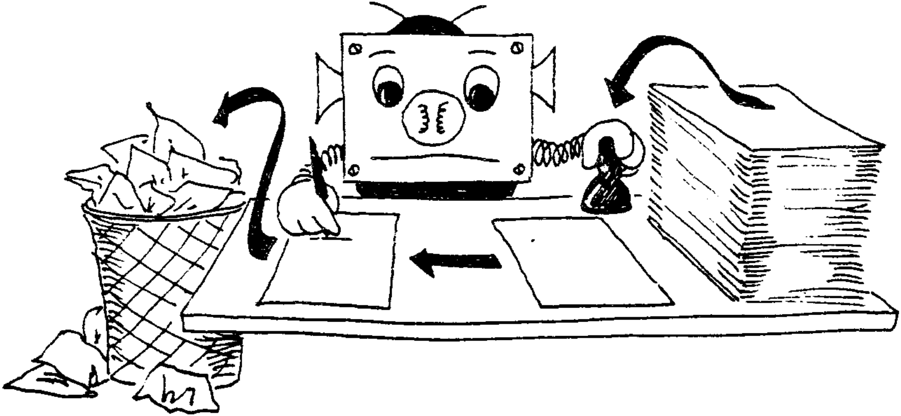


Information overload in globalized scientific world: navigational search needed in DML-CZ

Information overload also in specific domains (mathematics): research search needed only ocassionaly

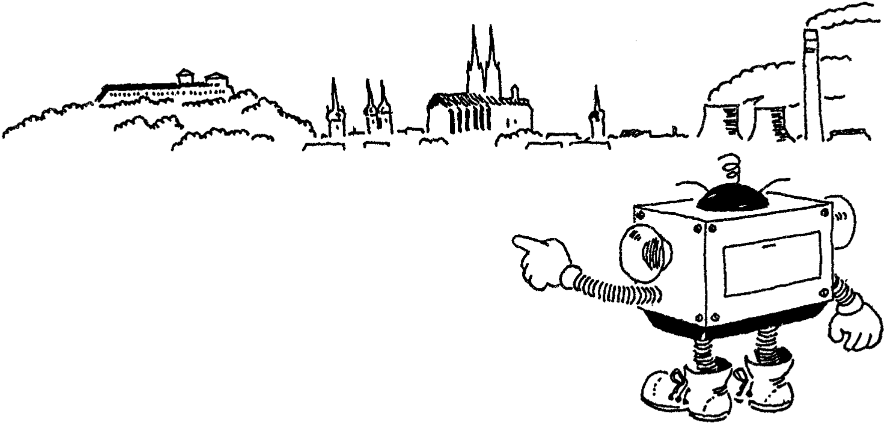


From paper to digital workflow: radical changes and consequences for MIR



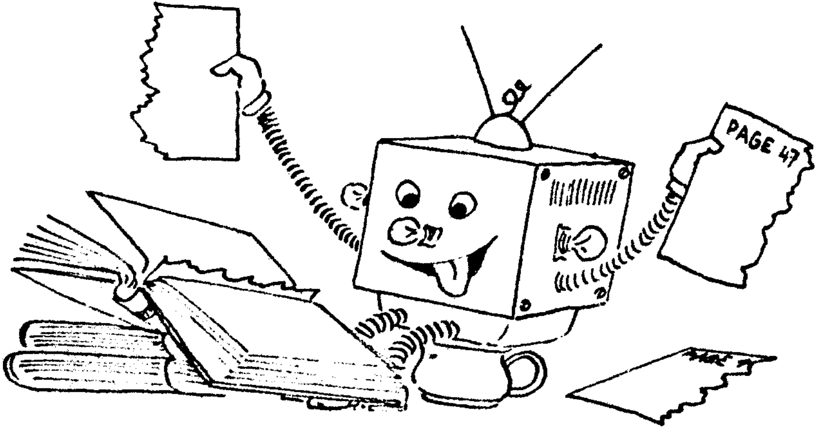
Retro-digitization, digital library developments

Specific needs in MIR: new tools developed in Brno, CZ for MIR purposes, DML-CZ is running there

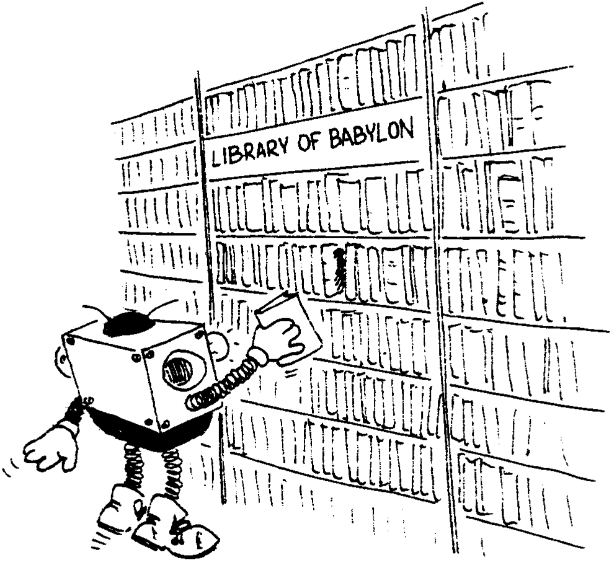


New workflows and data processing: DSpace and Lucene for fulltext search not sufficient, MIaS needed → thesis series on MIR

New tools: [math-aware] semantic similarity engine gensim (Řehůřek)

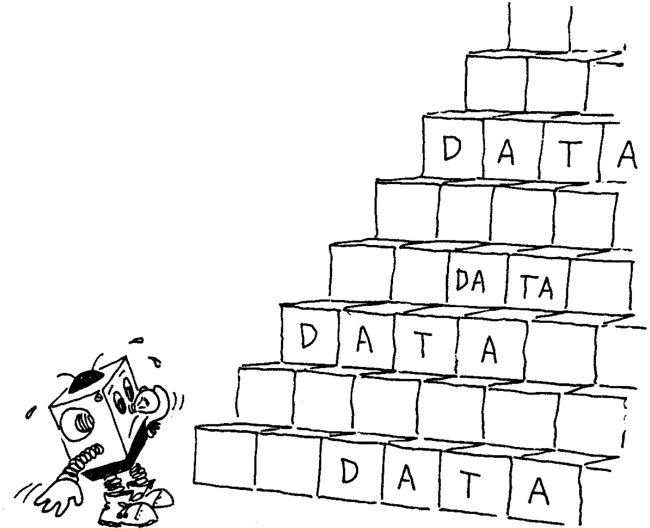


From local DMLs like DML-CZ to bigger ones: EuDML since 2010

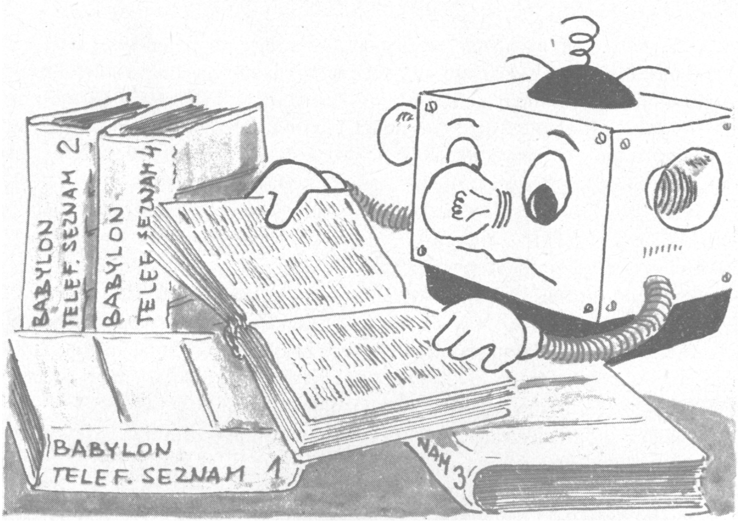


The European Digital Mathematics Library: EuDML

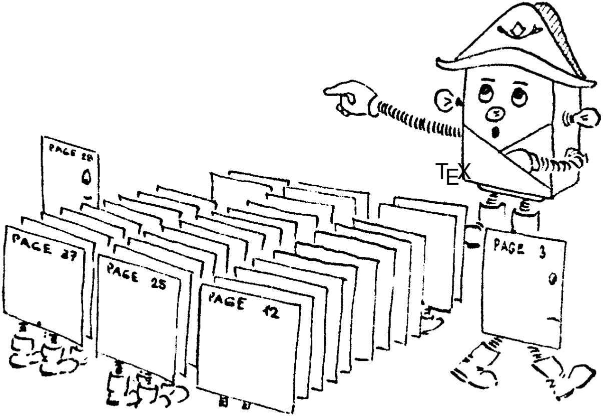
EuDML lesson: heterogenity in data: in markup, formats, collections, working attitude,…



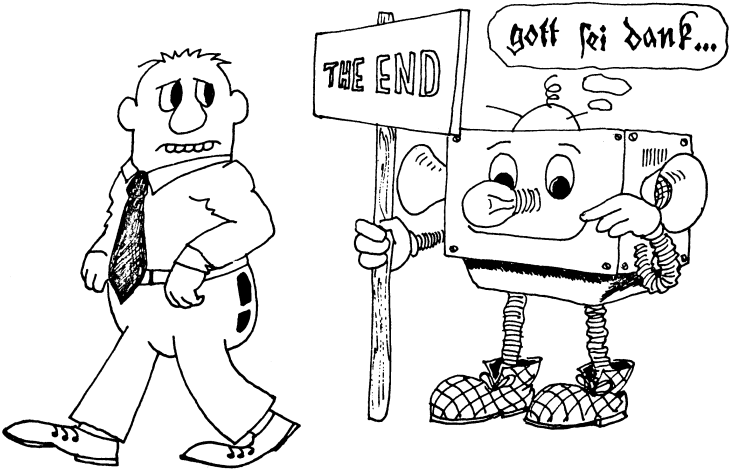
MIR for EuDML: new scalable tools development (reported CICM 2011)



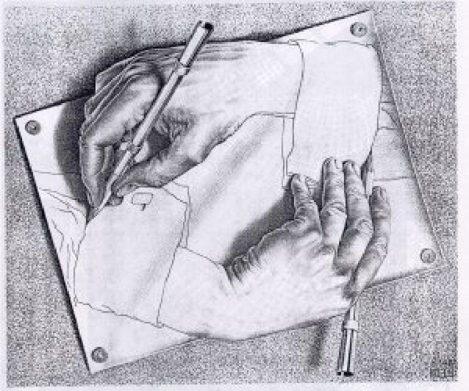
Yes, you can! Navigational MIR NOW, once you collect and normalize data,…



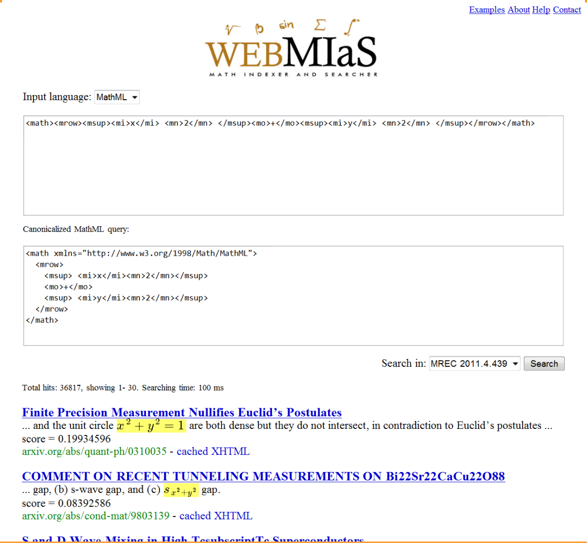
End of historical overview



The Present



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



Input language: MathML

```
<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

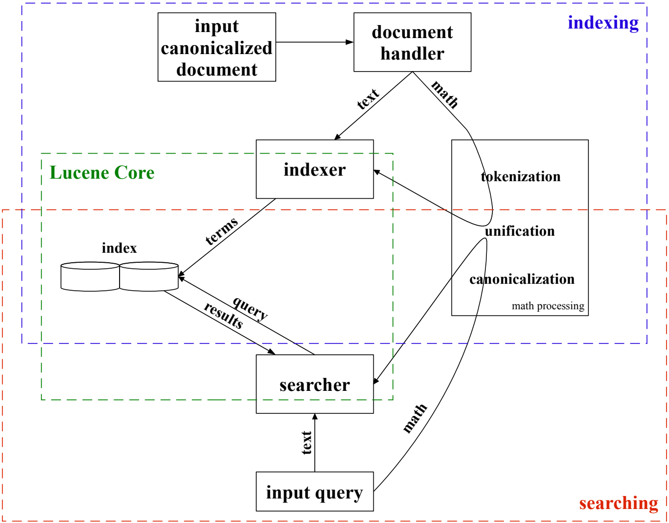
Total hits: 36817, showing 1- 30. Searching time: 100 ms

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) $s_{x^2+y^2}$ gap.
score = 0.08392586
arxiv.org/abs/cond-mat/9803139 - cached XHTML

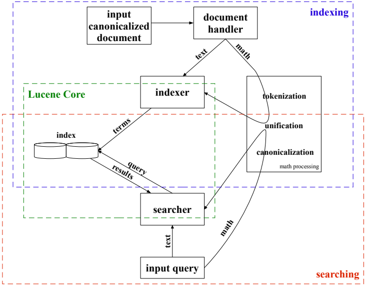
S and D Wave Mixing in High TcsubscriptTc Superconductors

WebMIaS Workflow based on the state-of-the-art tools (Lucene)



MIaS/WebMIaS gory details

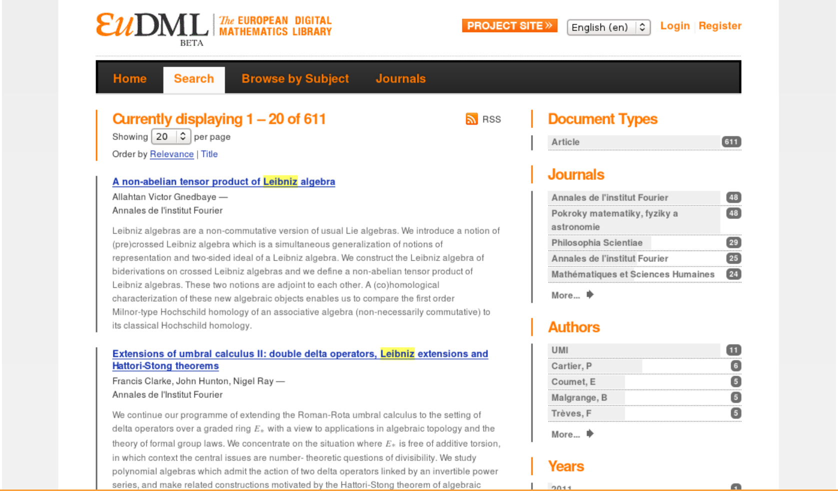
Martin Líška will disclose the MIaS details in the follow-up talk.



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>



EuDML BETA | The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Currently displaying 1 – 20 of 611

Showing 20 per page

Order by Relevance | Title

A non-abelian tensor product of Leibniz algebra

Allahtan Victor Gnedbaye —
Annales de l'institut Fourier

Leibniz algebras are a non-commutative version of usual Lie algebras. We introduce a notion of (pre)crossed Leibniz algebra which is a simultaneous generalization of notions of representation and two-sided ideal of a Leibniz algebra. We construct the Leibniz algebra of biderivations on crossed Leibniz algebras and we define a non-abelian tensor product of Leibniz algebras. These two notions are adjoint to each other. A (co)homological characterization of these new algebraic objects enables us to compare the first order Milnor-type Hochschild homology of an associative algebra (non-necessarily commutative) to its classical Hochschild homology.

Extensions of umbral calculus II: double delta operators, Leibniz extensions and Hattori-Stong theorems

Francis Clarke, John Hunton, Nigel Ray —
Annales de l'Institut Fourier

We continue our programme of extending the Roman-Rota umbral calculus to the setting of delta operators over a graded ring E_* with a view to applications in algebraic topology and the theory of formal group laws. We concentrate on the situation where E_* is free of additive torsion, in which context the central issues are number- theoretic questions of divisibility. We study polynomial algebras which admit the action of two delta operators linked by an invertible power series, and make related constructions motivated by the Hattori-Stong theorem of algebraic

Document Types

Article	611

Journals

Annales de l'institut Fourier	48
Pokroky matematiky, fyziky a astronomie	48
Philosophia Scientiae	29
Annales de l'Institut Fourier	25
Mathématiques et Sciences Humaines	24

More...

Authors

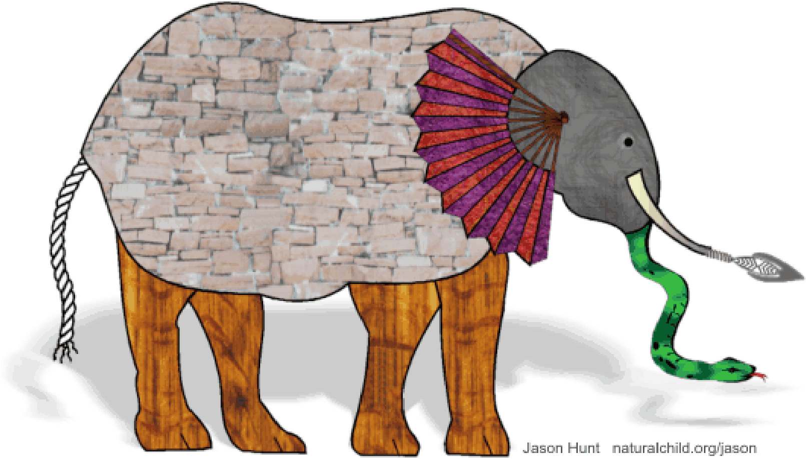
UMI	11
Cartier, P	6
Coumet, E	5
Malgrange, B	5
Trèves, F	5

More...

Years

2011	1

衆瞽
探象之図



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).

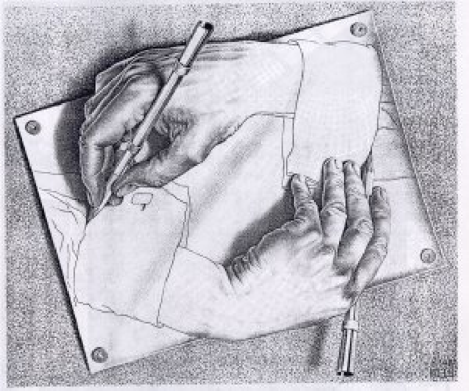
Problem of message (content+form) representation (of math when transporting the message over the web).

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



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.

Formulae images indexing. (Dostál, based on <http://mufin.fi.muni.cz>).

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).

Math NLP: Math Sketch engine <http://ske.fi.muni.cz>

Sumary

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

Let it happen.

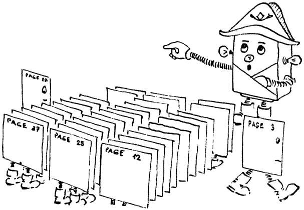
It has happened in the Past.

It is happening in the Present, now.

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

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).

Archambault, D., Berger, F., Moço, V.: Overview of the “Universal Maths Conversion Library”. In: Pruski, A., Knops, H. (eds.) Assistive Technology: From Virtuality to Reality: Proceedings of 8th European Conference for the Advancement of Assistive Technology in Europe AAATE 2005, Lille, France. pp. 256–260. IOS Press, Amsterdam, The Netherlands (Sep 2005)

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>

Baker, J.B., Sexton, A.P., Sorge, V.: A linear grammar approach to mathematical formula recognition from PDF. In: Proceedings of the Conferences in Intelligent Computer Mathematics, CICM 2009. LNAI, vol. 5625, pp. 201–216. Springer (2009)

Baker, J.B., Sexton, A.P., Sorge, V.: Towards reverse engineering of PDF documents. In: Sojka, P., Bouche, T. (eds.) Towards a Digital Mathematics Library, DML 2011. pp. 65–75. Masaryk University Press, Bertinoro, Italy (July 2011), <http://hdl.handle.net/10338.dmlcz/702603>

Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: Project EuDML—A First Year Demonstration. In: Davenport, J.H., Farmer, W.M., Urban, J., Rabe, F. (eds.) Intelligent Computer Mathematics. Proceedings of 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011. Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 281–284. Springer-Verlag, Berlin, Germany (Jul 2011), <http://dx.doi.org/10.1007/978-3-642-22673-1_21>

Grimm, J.: Producing MathML with Tralics. In: Sojka [13], pp. 105–117, <http://dml.cz/dmlcz/702579>

Jarmar, M.: Conversion of Mathematical Documents into Braille. Master’s thesis, Faculty of Informatics (Jan 2012), <https://is.muni.cz/th/172981/fi_m/?lang=en>

Líška, M., Sojka, P., Růžička, M., Mravec, P.: Web Interface and Collection for Mathematical Retrieval. In: Sojka, P., Bouche, T. (eds.) Proceedings of DML 2011. pp. 77–84. Masaryk University, Bertinoro, Italy (Jul 2011), <http://www.fi.muni.cz/ sojka/dml-2011-program.html>

Maplesoft, a division of Waterloo Maple Inc.: MathML – Maple Help (Apr 2012), <http://www.maplesoft.com/support/help/Maple/view.aspx?path=MathML>

- Munavalli, R., Miner, R.: MathFind: A Math-Aware Search Engine. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 735–735. SIGIR '06, ACM, New York, NY, USA (2006), <http://doi.acm.org/10.1145/1148170.1148348>
- Řehůřek, R., Sojka, P.: Automated Classification and Categorization of Mathematical Knowledge. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) Intelligent Computer Mathematics—Proceedings of 7th International Conference on Mathematical Knowledge Management MKM 2008. Lecture Notes in Computer Science LNCS/LNAI, vol. 5144, pp. 543–557. Springer-Verlag, Berlin, Heidelberg (Jul 2008)
- Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), <http://is.muni.cz/publication/884893/en>, software available at <http://nlp.fi.muni.cz/projekty/gensim>
- 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 (Mar 2011), submitted to MKM 2011
- Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W.M., Urban, J., Rabe, F. (eds.) Intelligent Computer Mathematics. Proceedings of 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011. 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>
- 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>
- Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: INFTY — An integrated OCR system for mathematical documents. In: Vanoirbeek, C., Roisin, C., Munson, E. (eds.) Proceedings of ACM Symposium on Document Engineering 2003. pp. 95–104. ACM, Grenoble, France (2003)
- Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [13], pp. 11–24, <http://dml.cz/dmlcz/702569>
- The LaTeXML project: The LaTeXML Developer Portal (Apr 2012), <https://trac.mathweb.org/LaTeXML/>

The MathWorks, Inc.: MuPAD – Matlab (May 2012), <http://www.mathworks.com/discovery/mupad.html>

Watt, S.M.: Mathematical Document Classification via Symbol Frequency Analysis. In: Sojka, P. (ed.) Towards Digital Mathematics Library—Proceedings of DML 2008. pp. 29–40. Masaryk University, Birmingham, UK (Jul 2008), <http://www.fi.muni.cz/ sojka/dml-2008-program.xhtml>

Wolfram: Mathematica Import/Export Format : MathML (Apr 2012), <http://reference.wolfram.com/mathematica/ref/format/MathML.html>

Wolfram Alpha LLC: Wolfram Alpha (Apr 2012), <http://www.wolframalpha.com/>