Towards Digital Mathematical Libraries: Collection and Search

Petr Sojka et al.

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

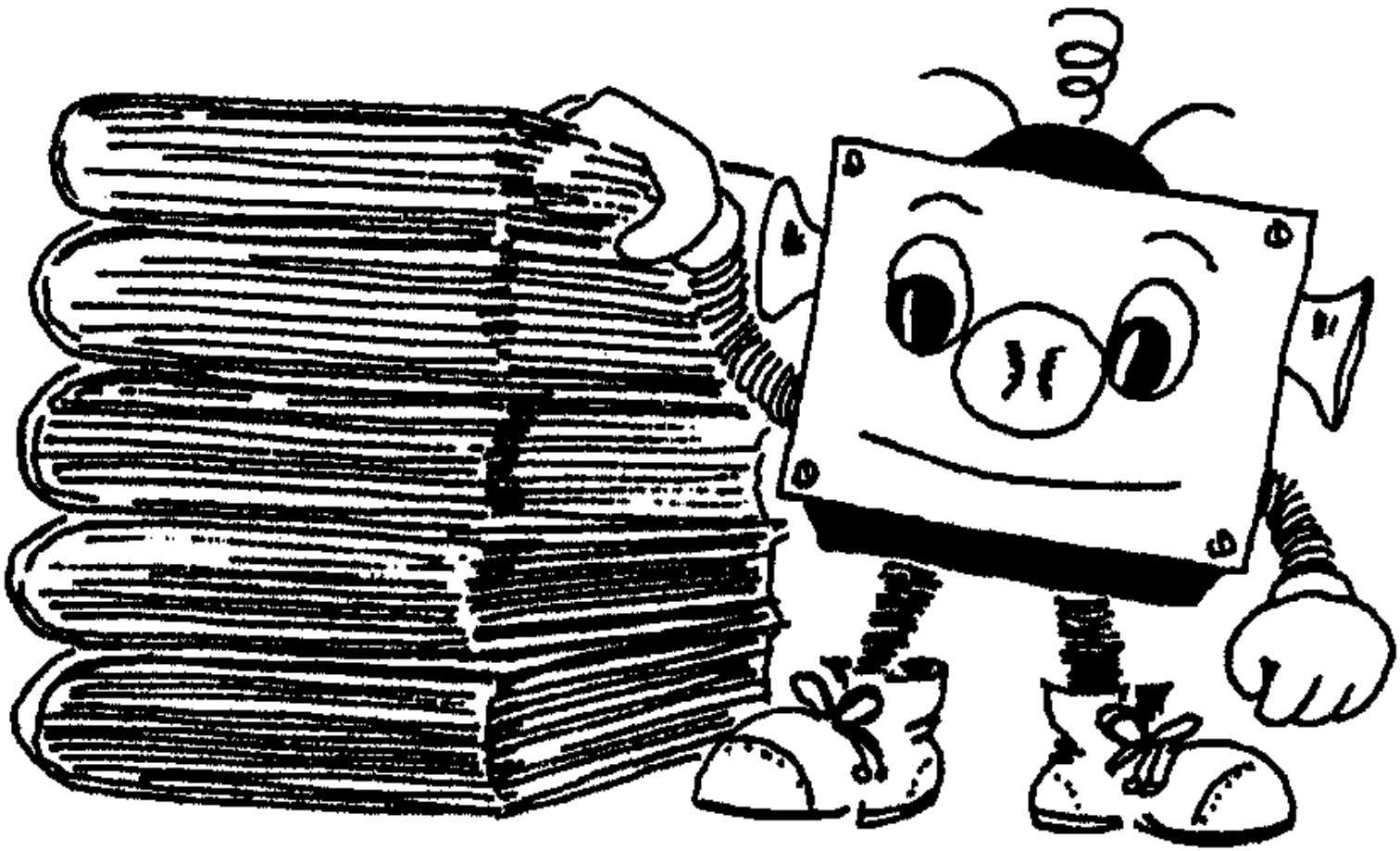
University of Coimbra, Math Department
September 7th, 2012, 11AM

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Outline and two take-home messages

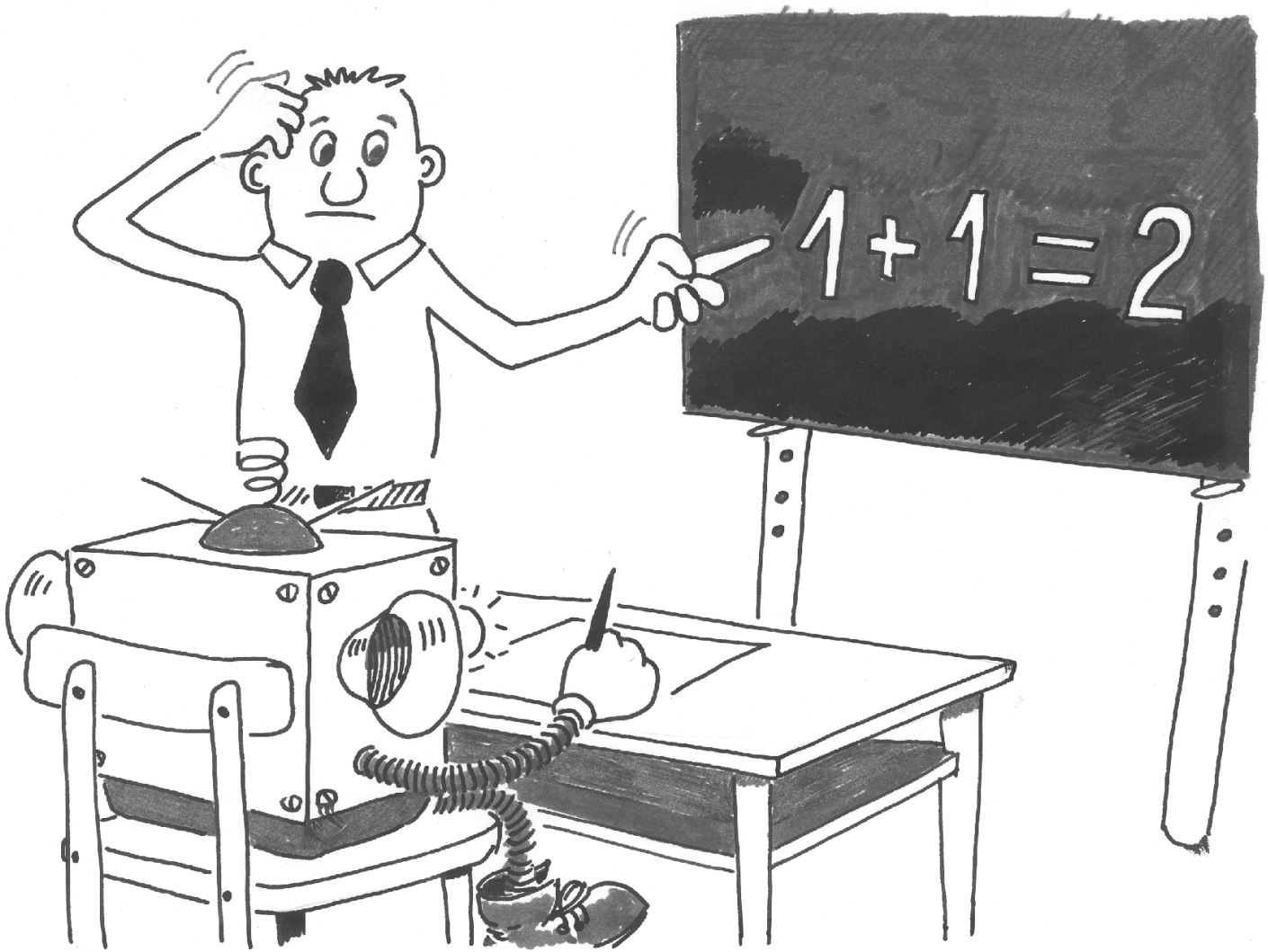
1. Pictorial overview
2. Motivation, vision of WDML, PubMed Central for Mathematics
3. Complexity of digitization workflow of The Czech Digital Mathematics Library DML
4. Search
5. Math Indexer and Searcher (MIaS)
6. Search evaluation
7. Conclusions

Towards *Digital* Mathematical Libraries: DML, EuDML, WDML

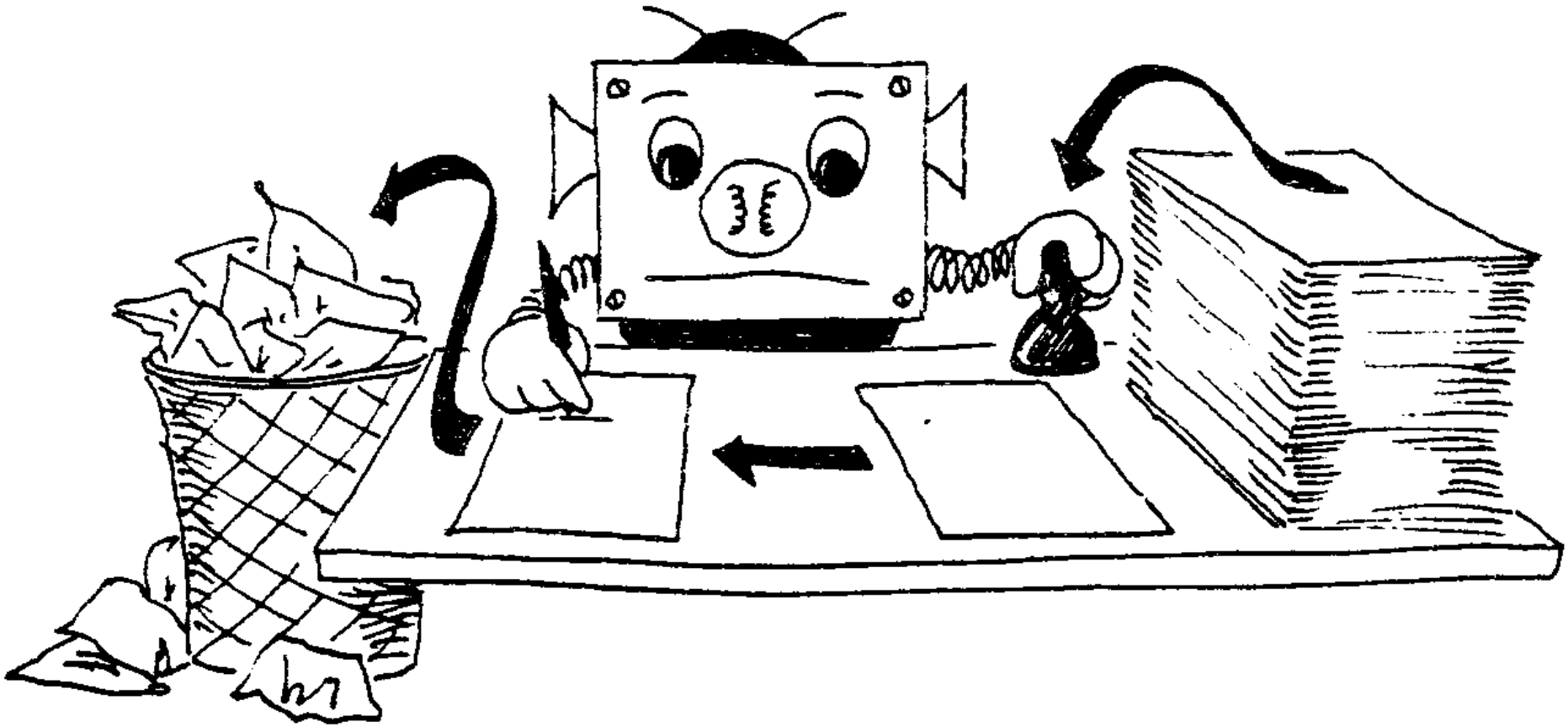


Information overload in globalized *scientific* world

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

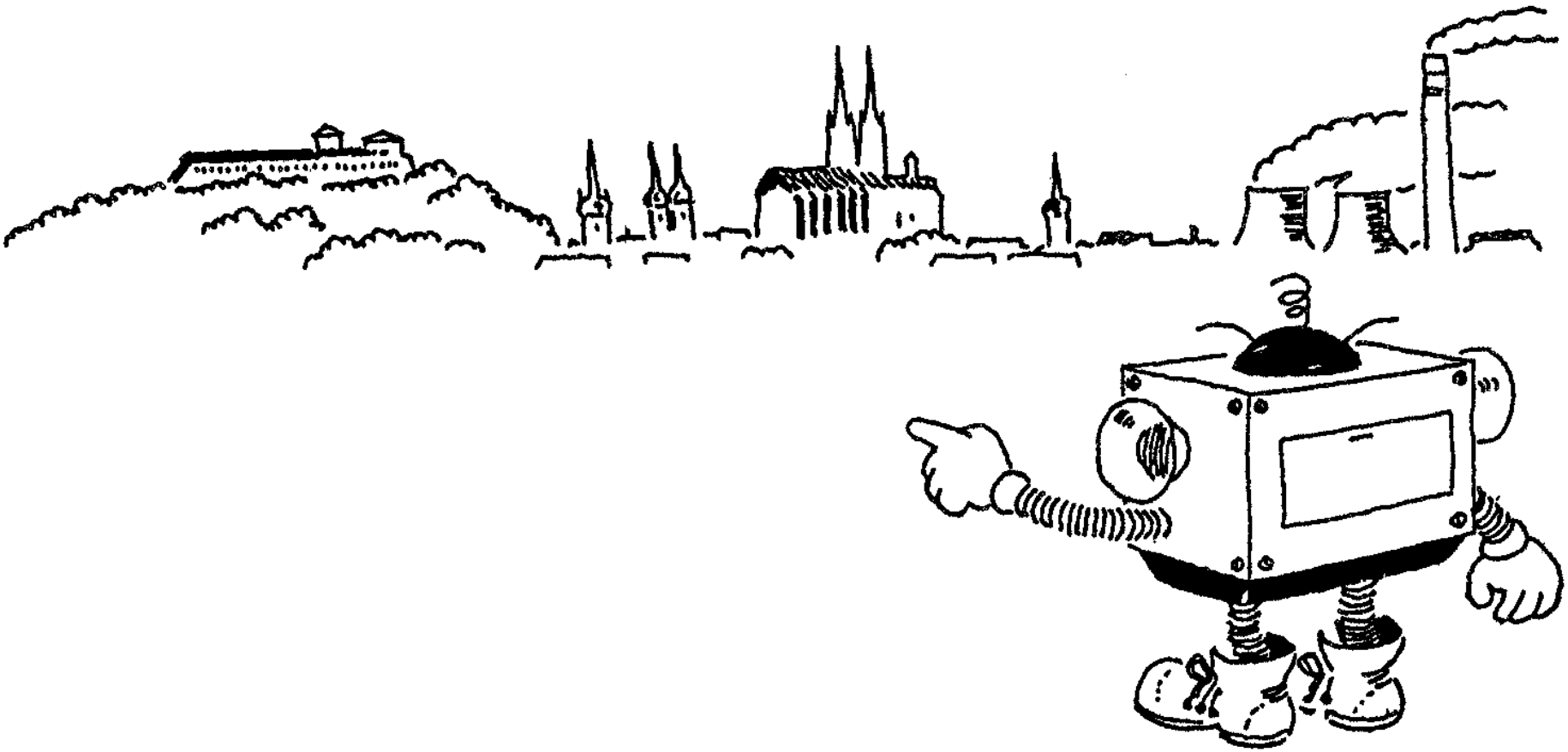


From paper to digital *workflow*

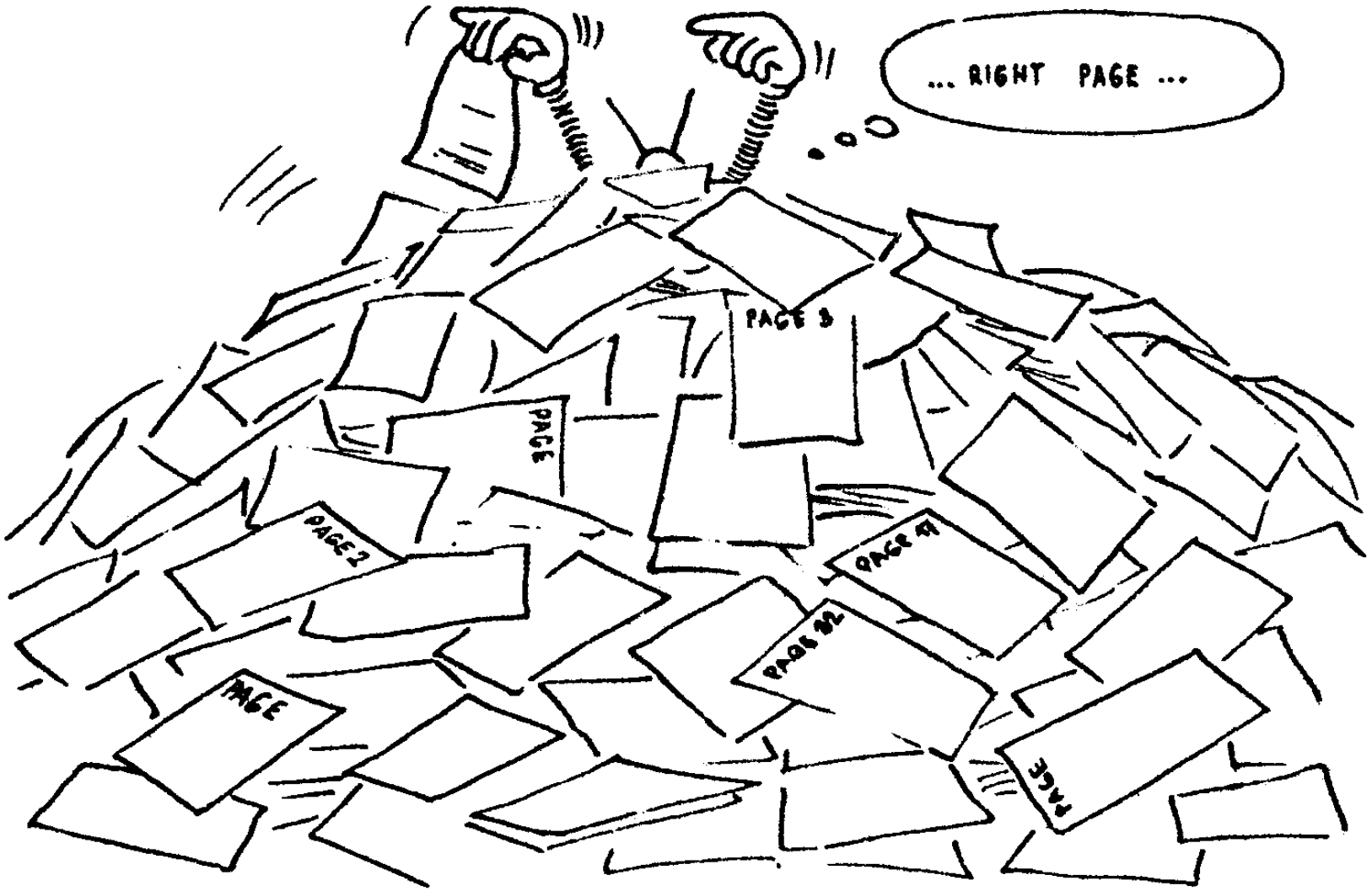


Retro-digitization, *accessible* digital library development

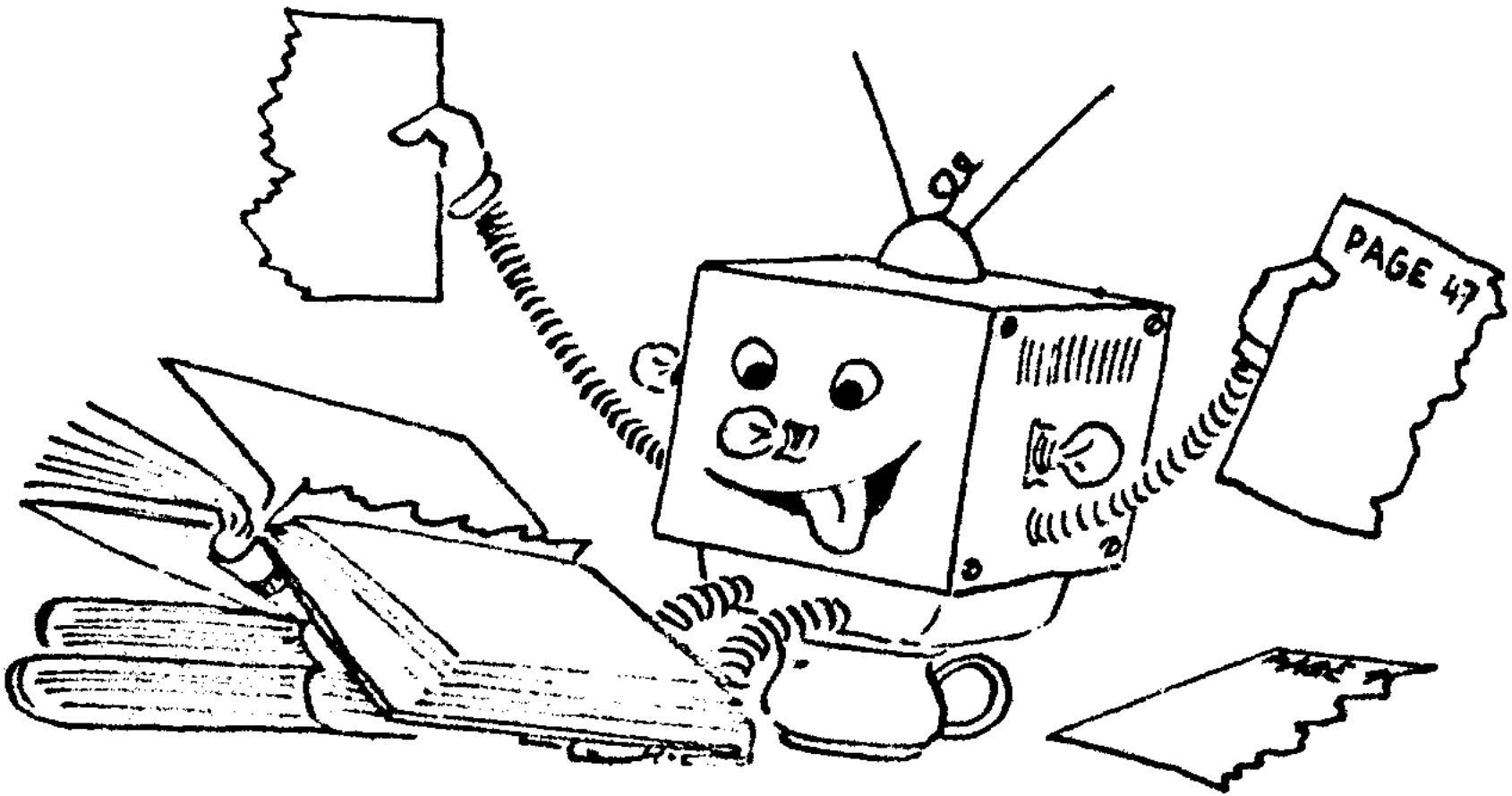
Experiences from project *DML-CZ* (Brno, CZ)



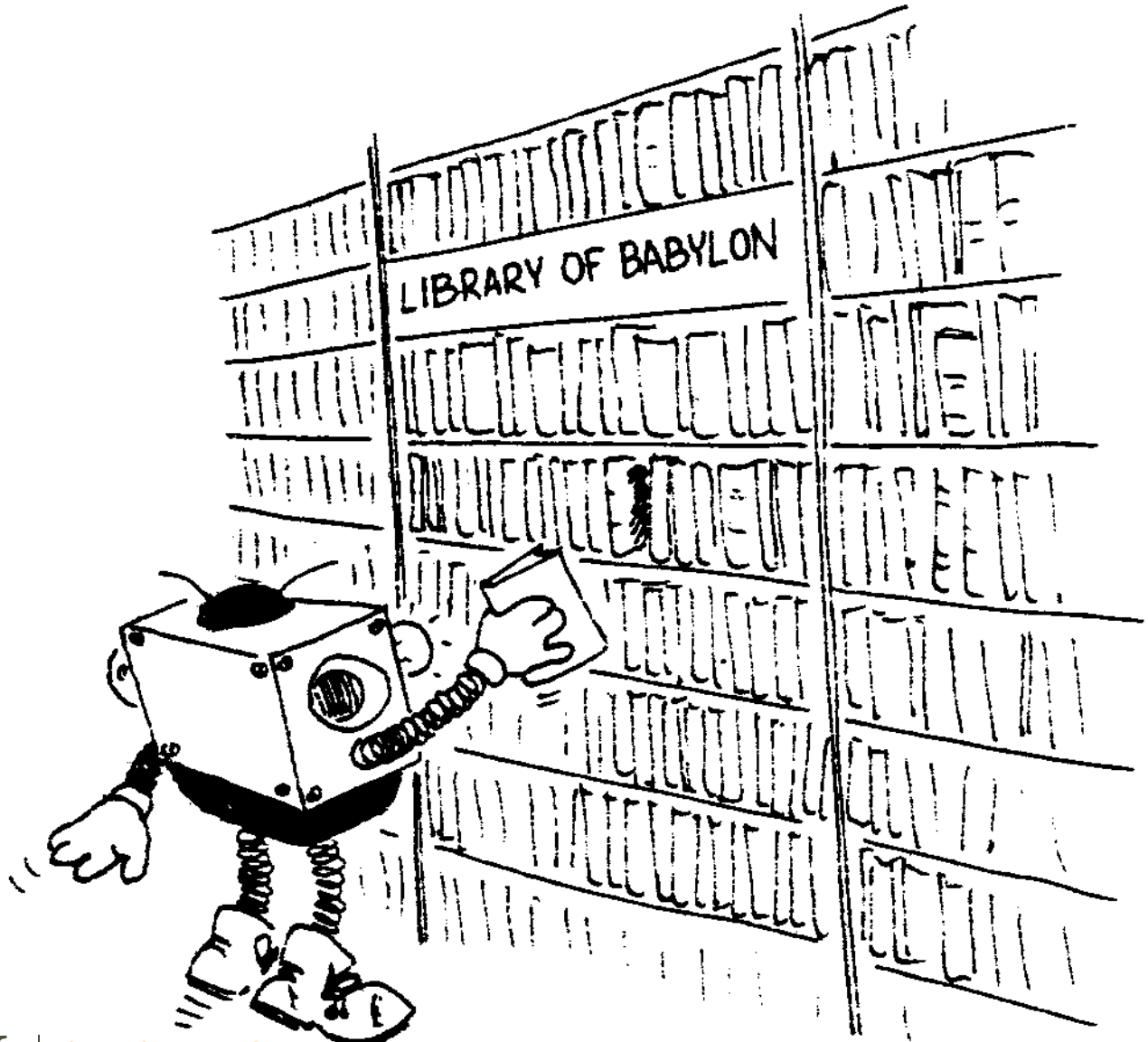
DML-CZ: new *workflows* and math data indexing



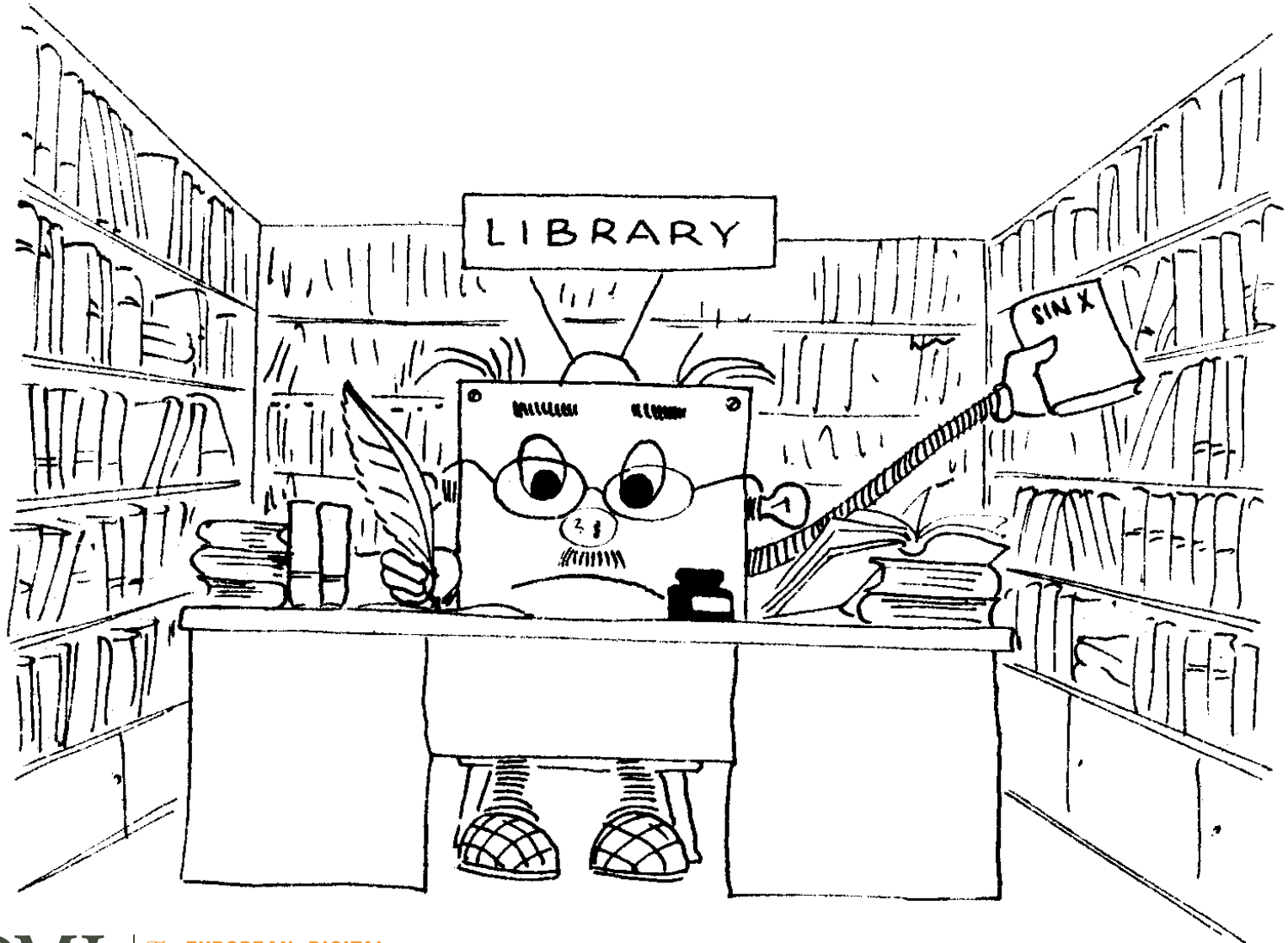
New approaches to *math document retrieval*



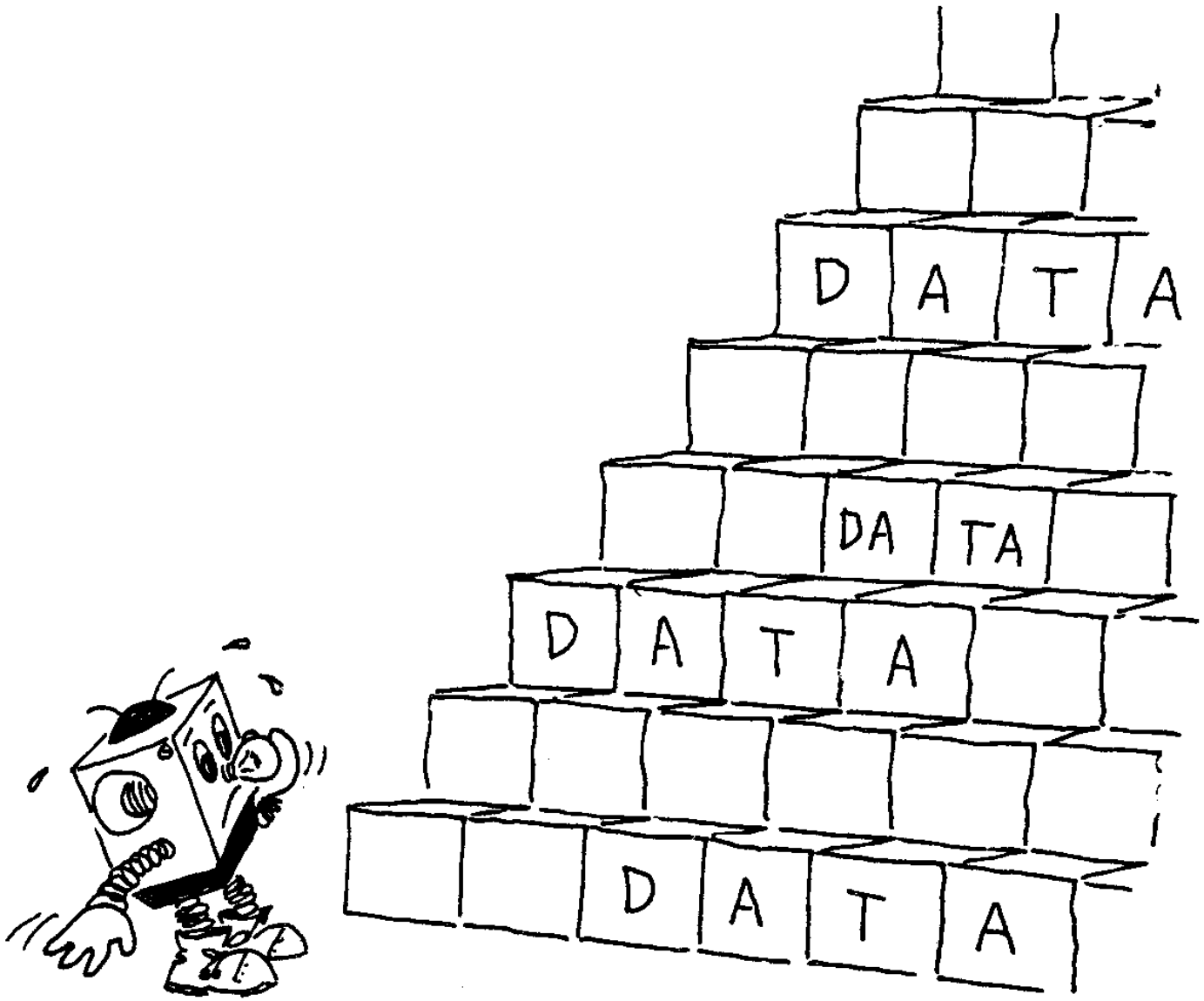
‘Bottom up’ deployment towards EU or *worldwide scale*



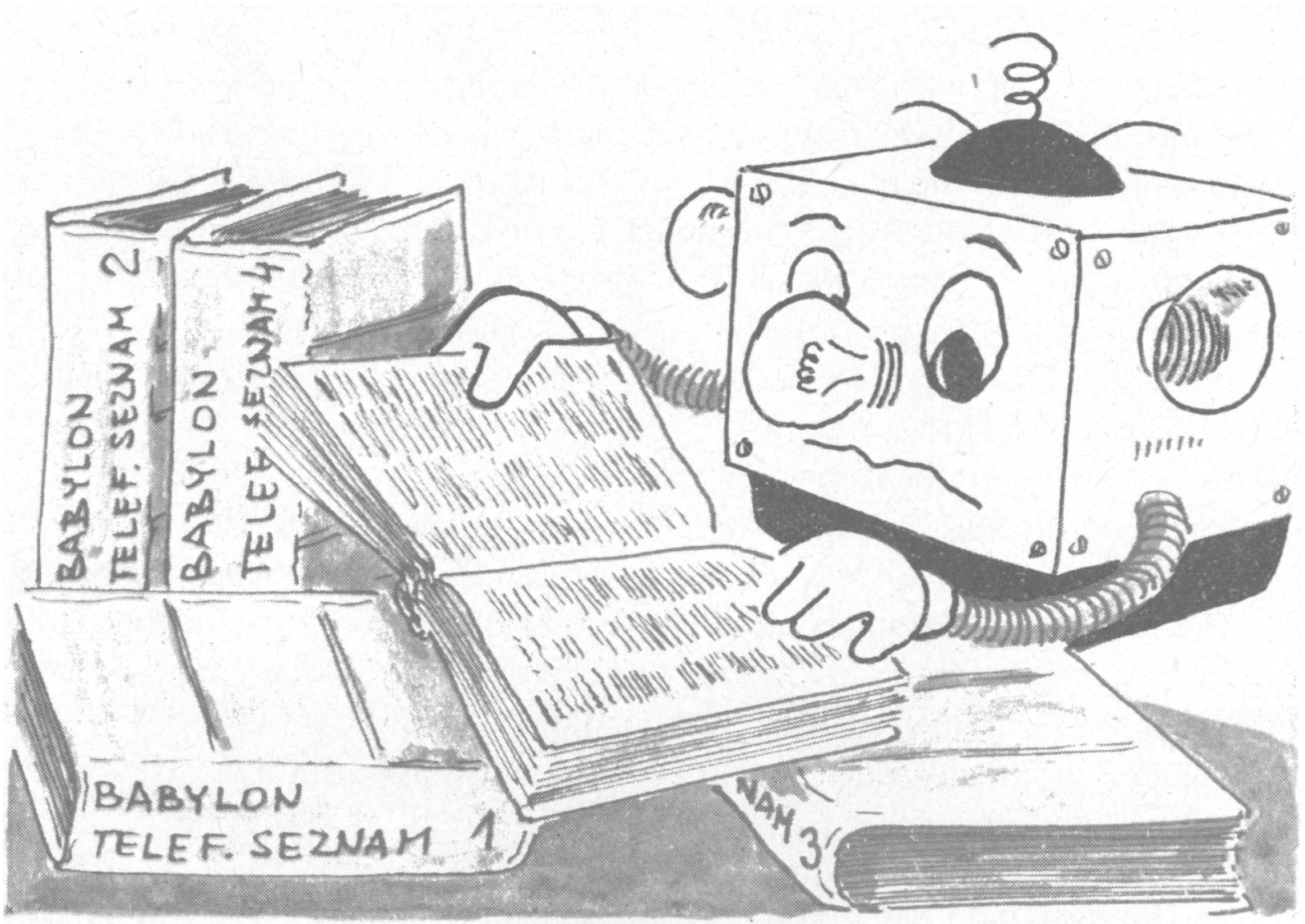
The European Digital Mathematics Library: *EuDML*



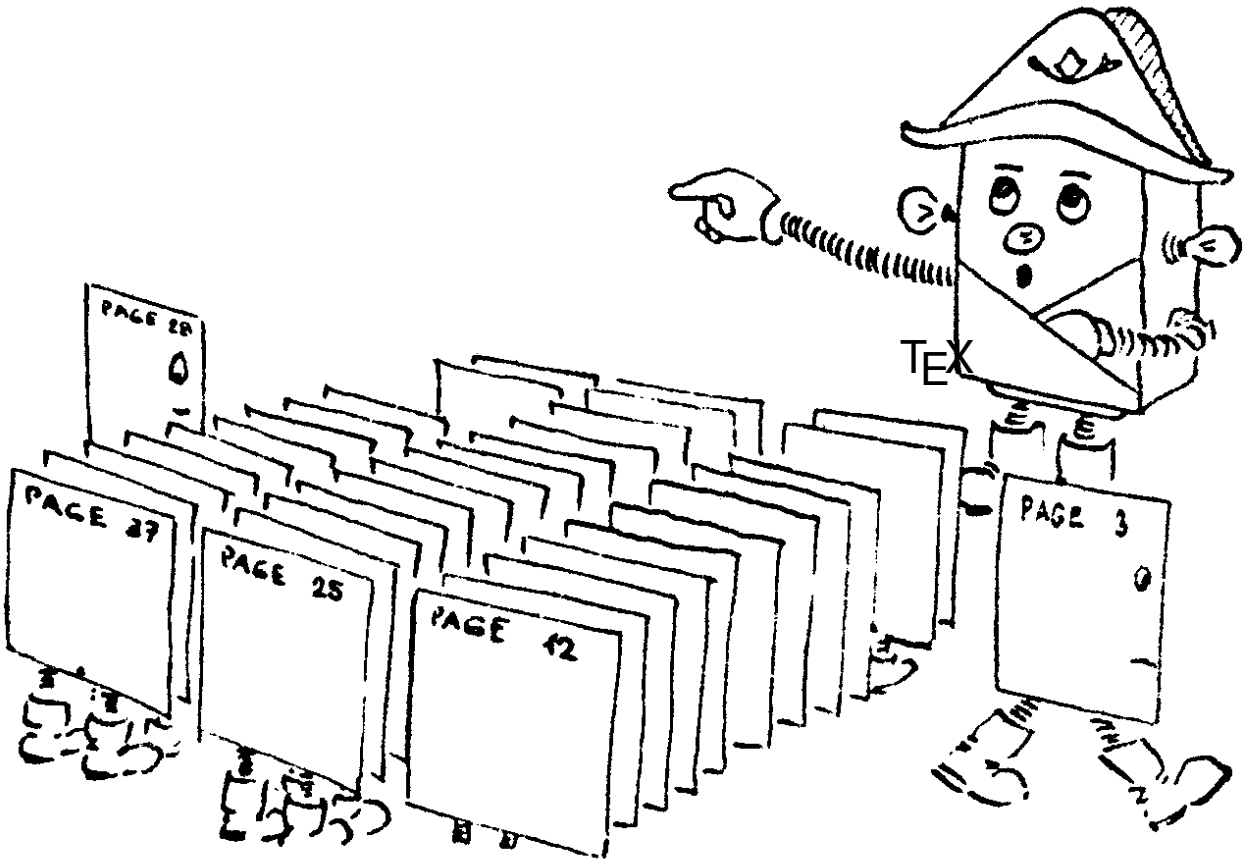
EuDML: from local data collections to the virtual DL



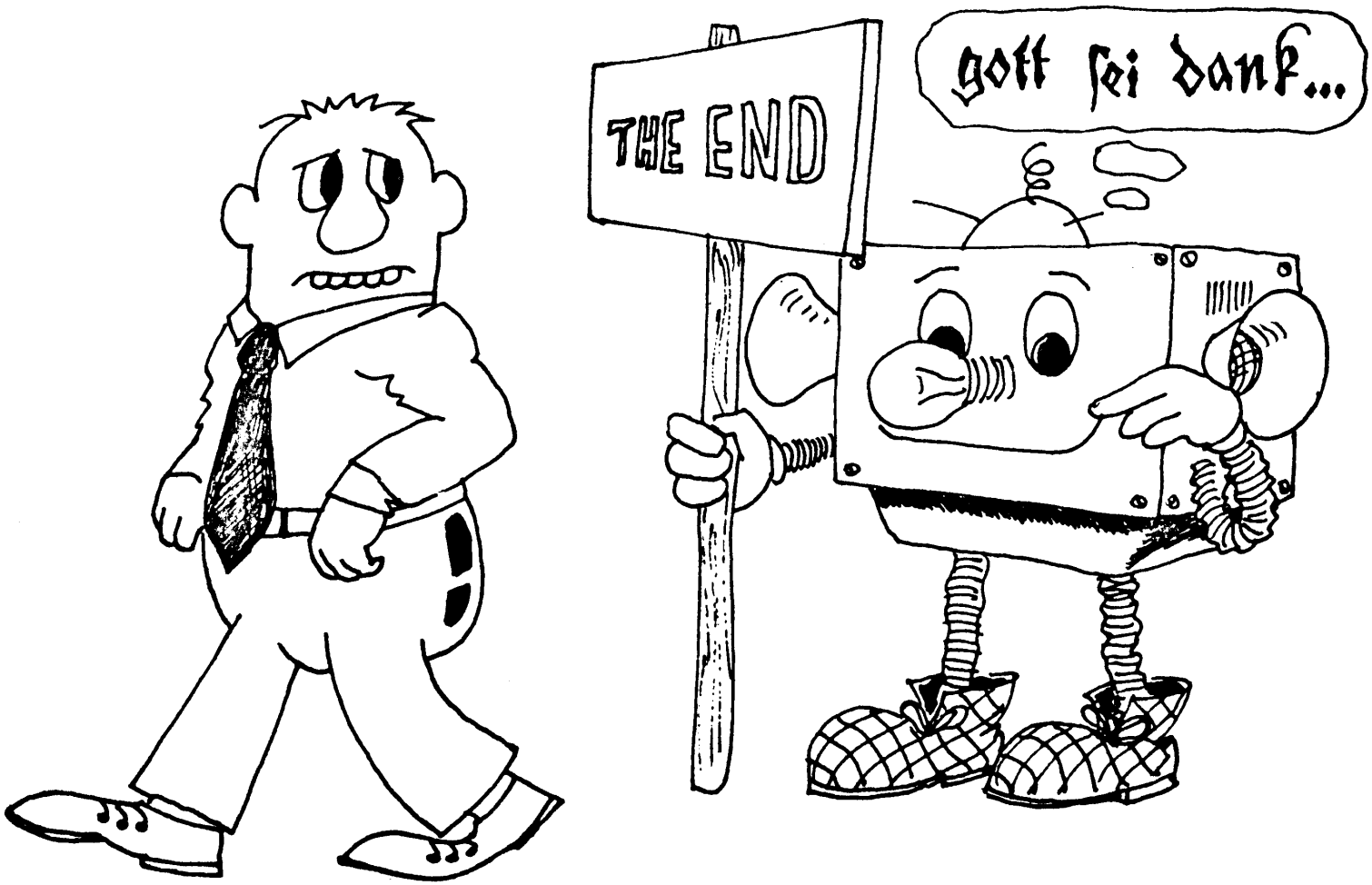
Tools for *automated math extraction* from PDF



Yes, you can! You can have highly automated workflow to get accessible math, search, visibility, scalability,…



End of talk overview



Decade of the 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.

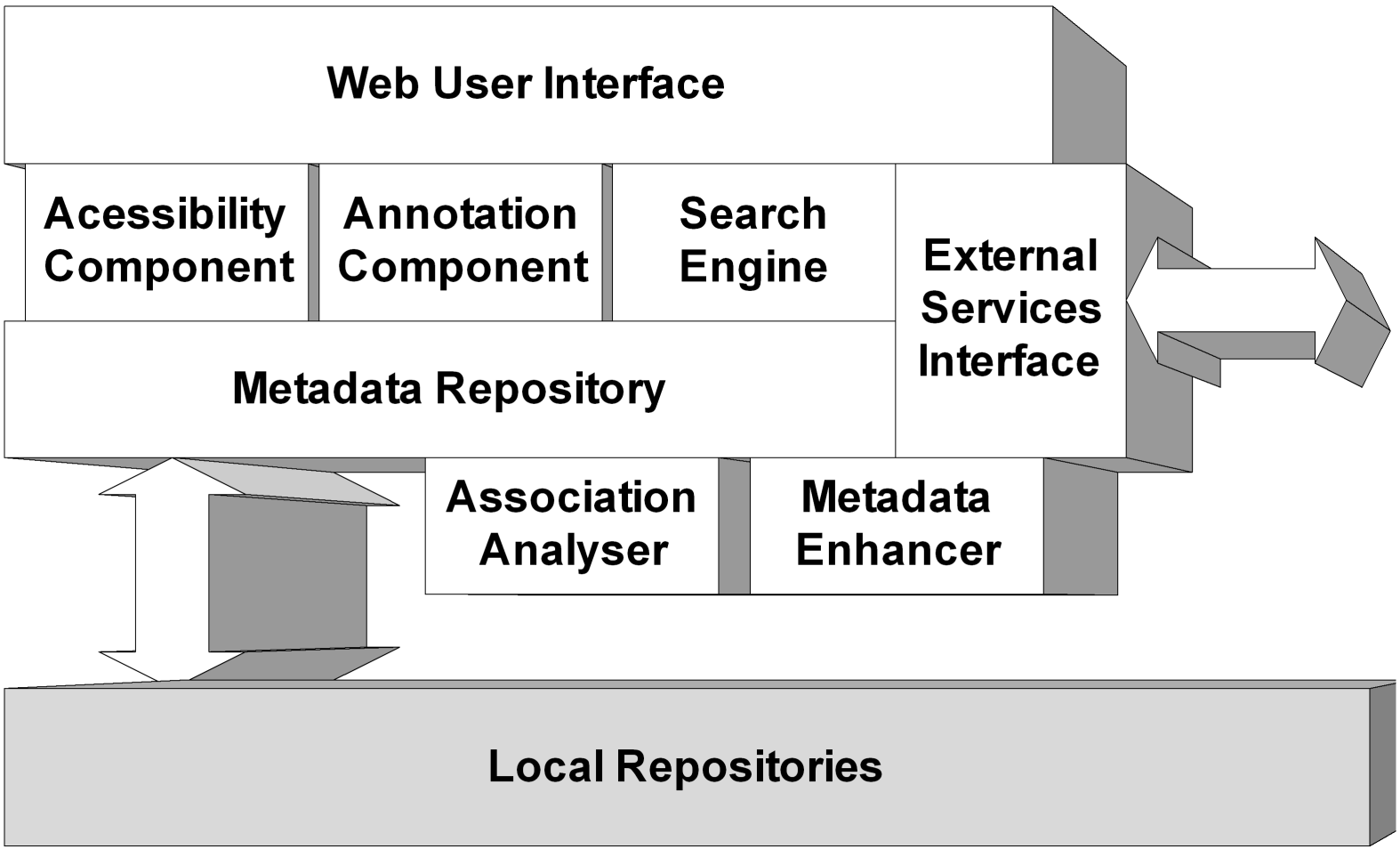
Vision of European Digital Mathematics Library

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)* started from February 2010. The strategy of EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY is:

- 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 $\rightarrow$ 'a must/me too' effect then as with PubMed Central.

EuDML as a virtual library portal

EuDML will be a *virtual* library based on data from smaller data providers, DLs and publishers:



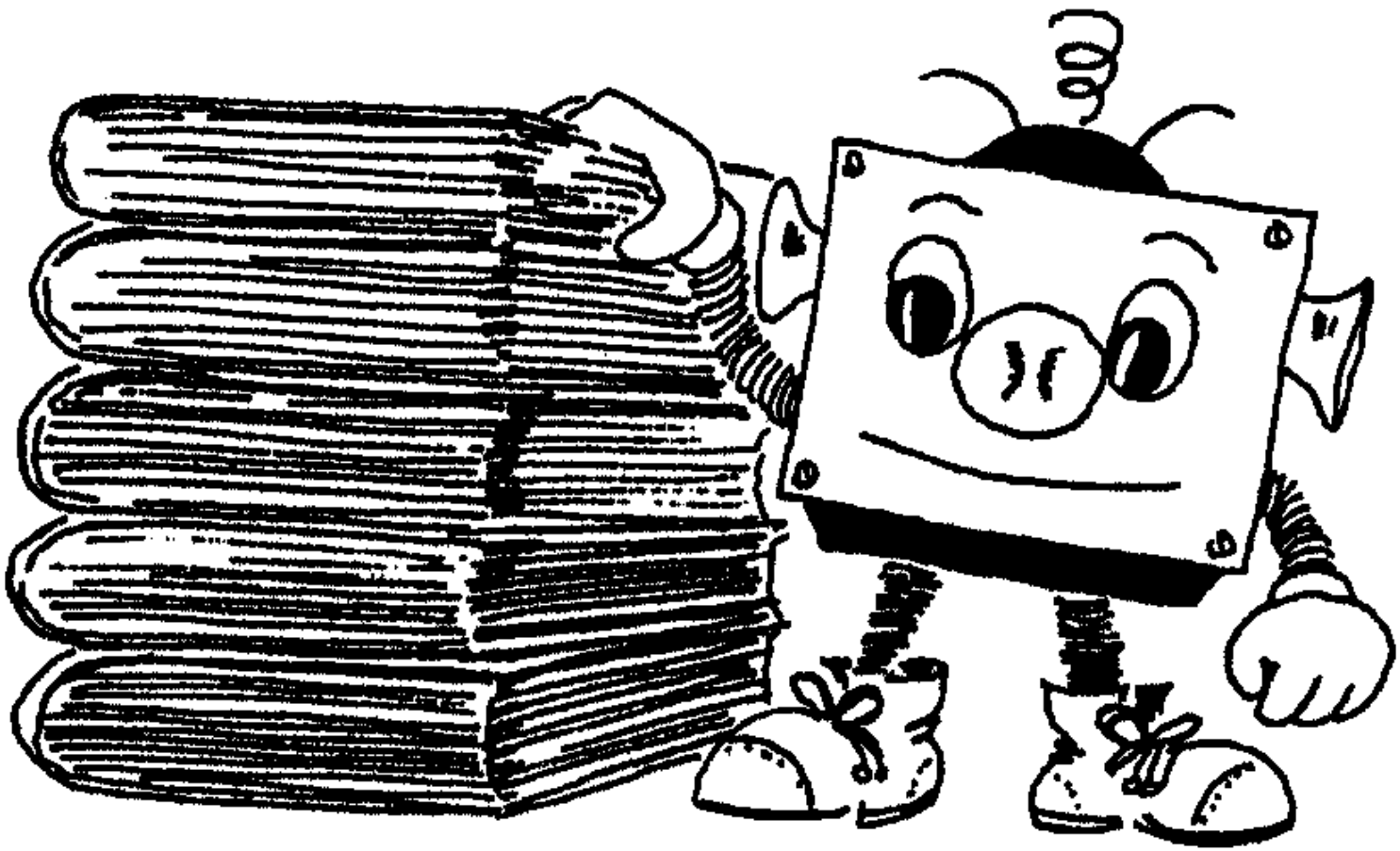
European Digital Mathematics Library

Bottom up—from building bricks of regional repositories

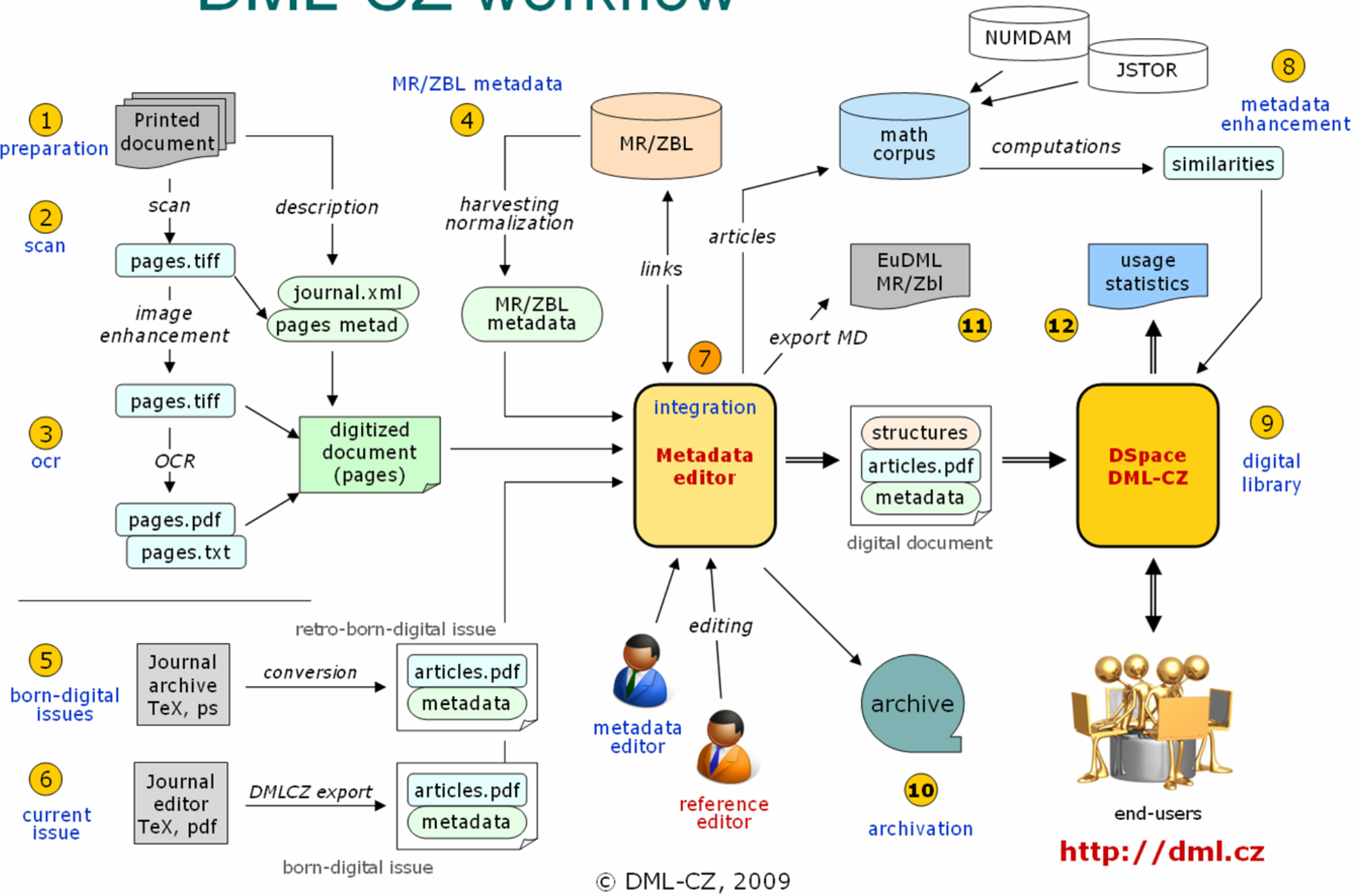
As DML content providers serve mostly publisher's or regional DML repositories as The Czech Digital Mathematics Library DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,…: aggregating content from local repositories to build the bigger (global?) DML.

Example of DML-CZ: up and running digital mathematics library <http://dml.cz> with 30,000+ papers (300,000+ pages). For more, see (who, what, browse, browse similar, how to search).

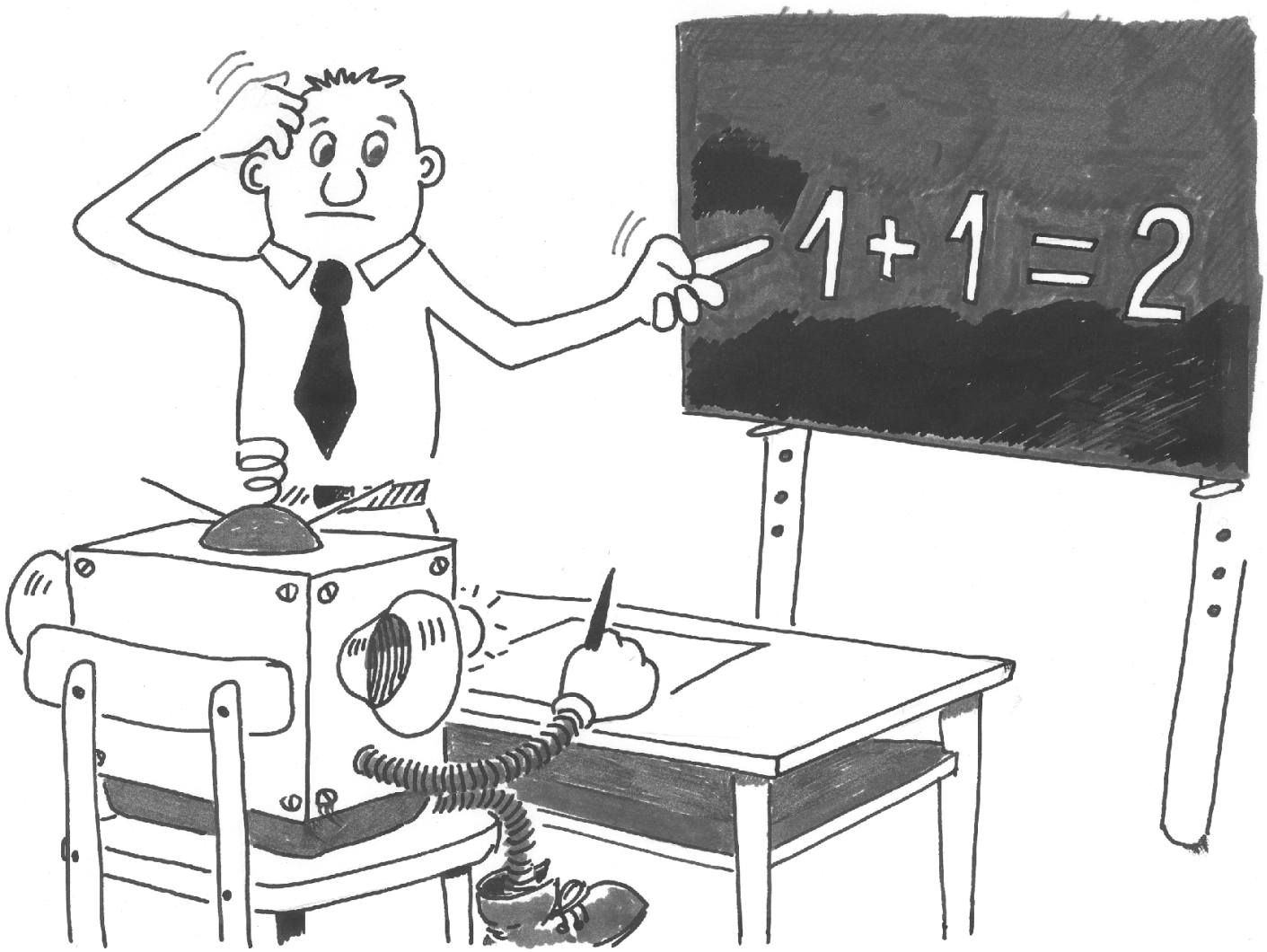
From paper to digital processing, from local to the global DML



DML-CZ workflow



Challenges of Math handling: OCR, indexing, search…



DML-CZ—data: scientific math published in CZ/SK

Proof. Let $\hat{K}$ be a cube, $\hat{K} \subset \hat{G}$; put $K = \varphi^{-1}(\hat{K})$. According to theorem 50 we have $K \in \mathfrak{A}$ 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 $x \in A$. Пусть $\mathfrak{A}$ — система всех ограниченных измеримых множеств A, для которых $\|A\| < \infty$. Теорема 18 тогда утверждает: *Пусть $A \in \mathfrak{A}$; пусть D — граница множества A. Тогда на системе $\mathfrak{B}$ всех борелевских подмножеств множества D существует мера p и на*

ИОСИФ ВИССАРИОНОВИЧ СТАЛИН
1879—1953

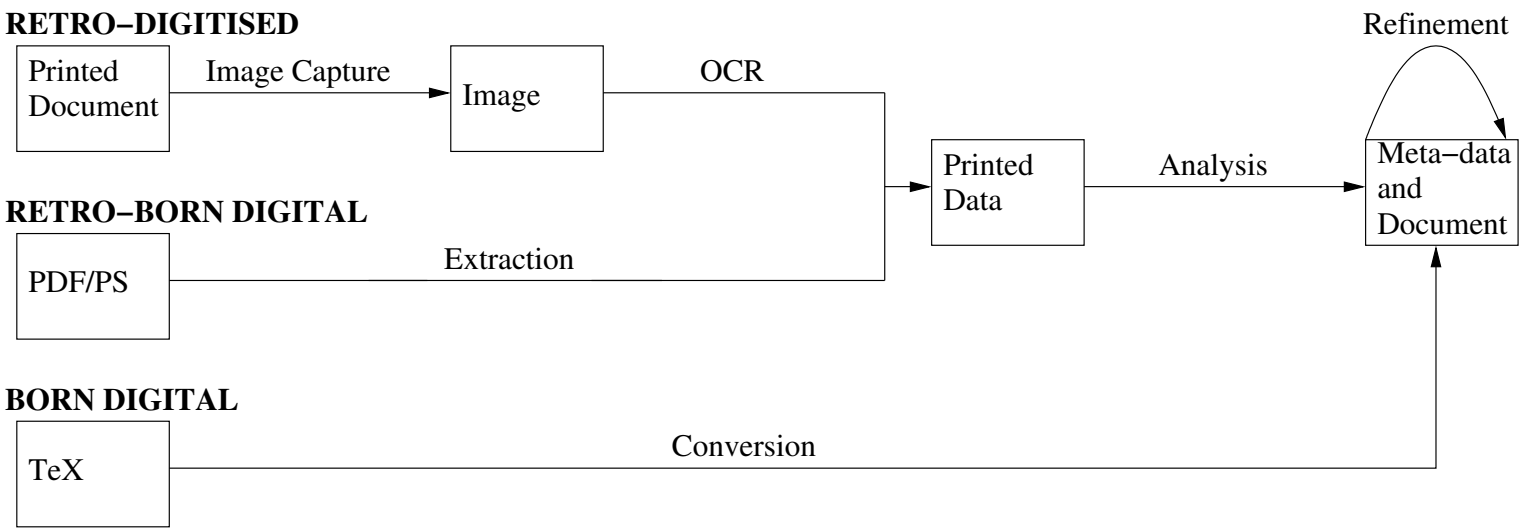
Document accessibility 4 DML processing challenges

Data heterogenity, plethora of formats, validation and conversions:

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: PDF2Math (PDF2NLM?)

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



MathML or LaTeX? MathML and LaTeX!

Data heterogenity, plethora of formats, validation and conversions:

world of authors: LaTeX, TeX notation of mathematics

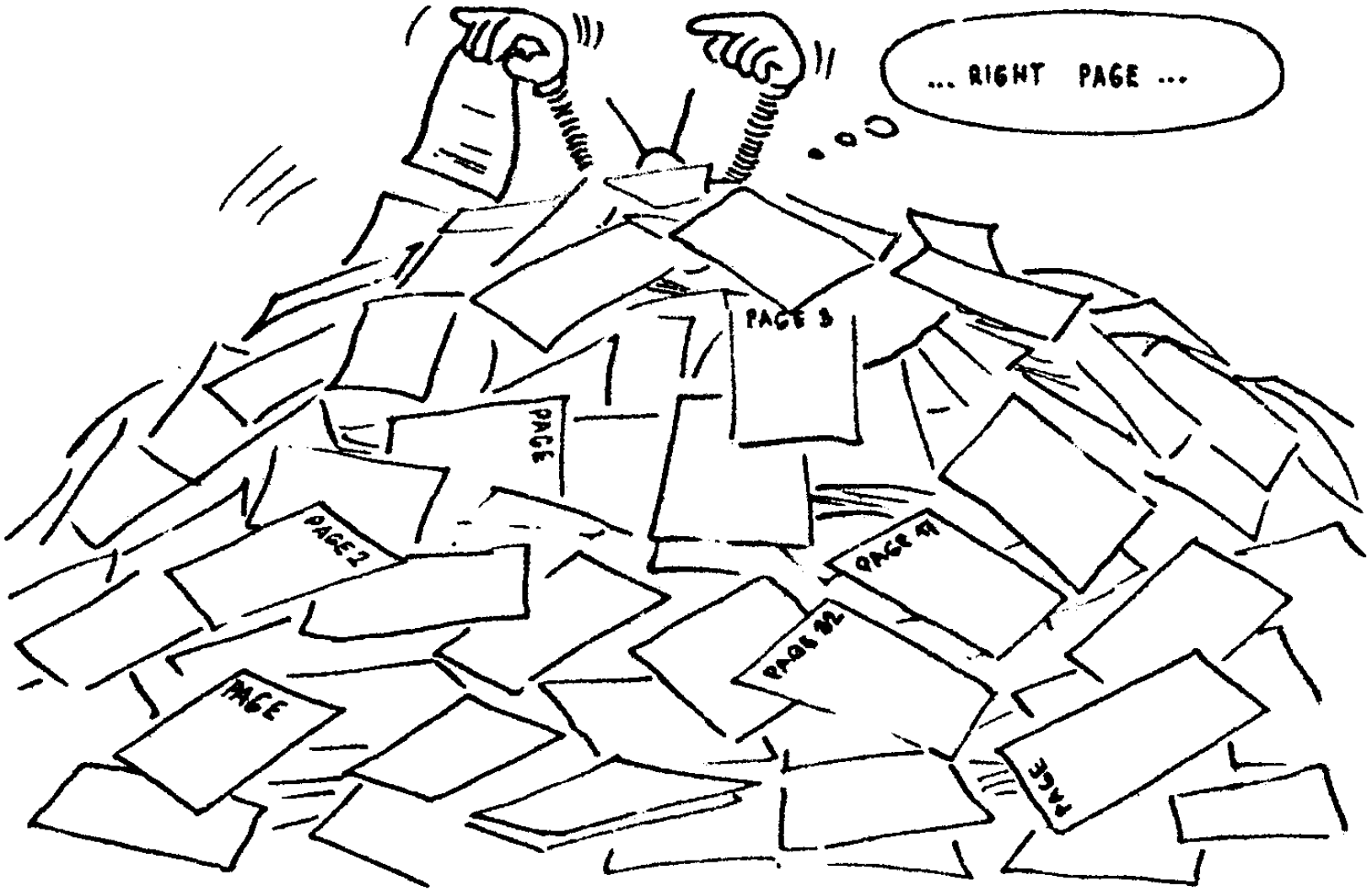
world of applications/data exchange: XML, *MathML*

Big volumes: $\rightarrow$ high *automation* to save costs

Exchange on the web—W3C standards: MathML, WAI-ARIA (Web Accessibility Initiative—Acessible Rich Internet Applications), WCAG (Web Content Accessibility Guidelines) 2.0 or Dogma W4.

By converting to MathML to allow discoverability and indexing (formulae fuzzy search).

DML-CZ document engineering—data processing



DML-CZ challenges and lessons learned

DML-CZ, the Czech Digital Mathematics Library, now serves amore than *300,000 pages of more than 30,000 math papers*. Challenges were

- *migration of existing workflows (retro-digital, retro-digital and born-digital) into the repository*
- negotiations with Google Scholar towards better visibility
- math indexing and search
- alternative visualization
- space and processing demands
- ....

DML-CZ is according to The Ranking Web of World Repositories the best repository in CZ, 91. in EU and 203. in the world.

Why Search?

I have a dream. [M.L. King, Jr., 1963]

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

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

Why Math Search (MIR)?

A picture is worth thousands words.

The speeding up world is moving towards graphics (user interfaces and visualization).

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

There are papers with more formulae than plain text.

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."

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 do not start from scratch

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

Take-off message from this part of talk: *Yes, you can!*

I have a dream. [M.L. King, Jr.]

I hope

Yes, we can!

Towards math search engine (MSE) – existing players

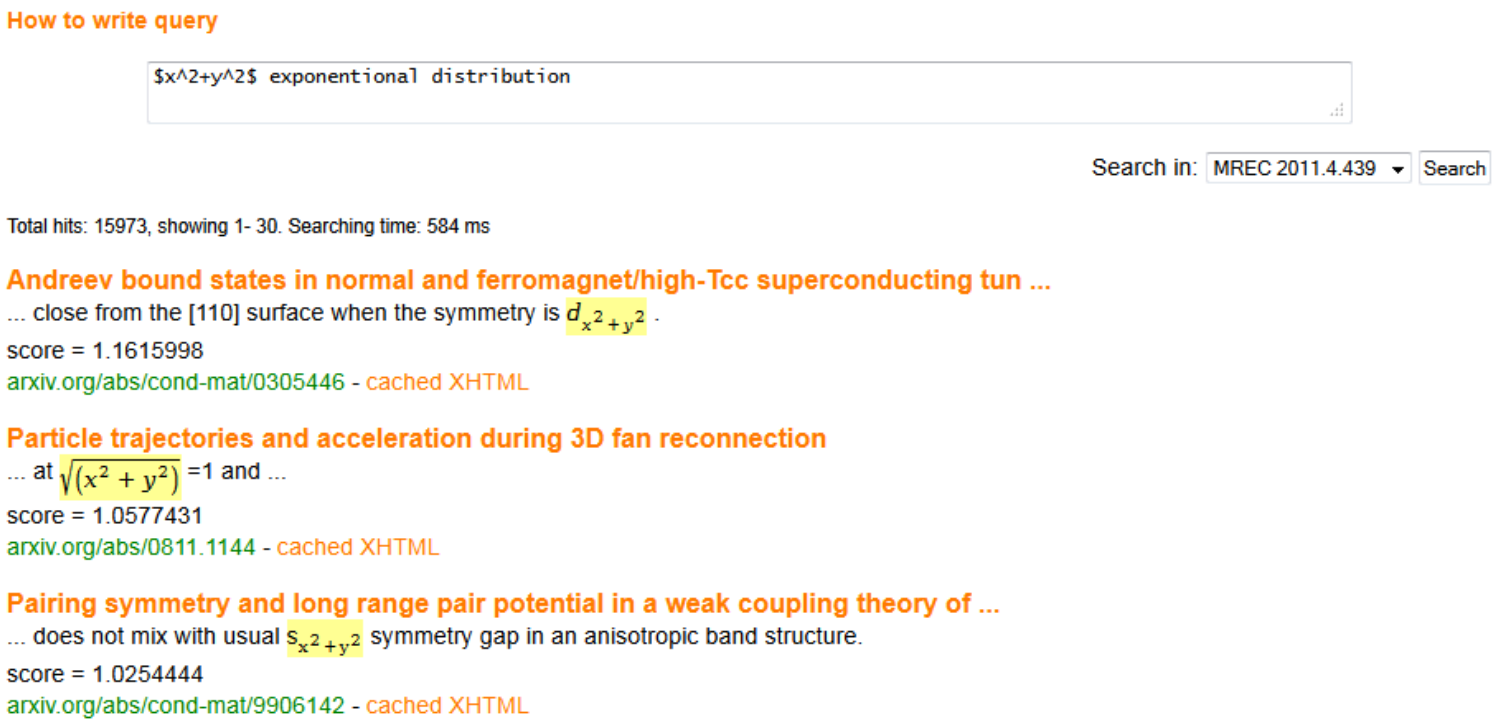
- Niche market for big players (as Google), attempts to solve by publishers (LaTeXSearch by Springer).
- Many challenges: heterogenity of math representation, notation, semantics handling, no established and accepted user interface and query language.
- Numerous attempts to solve the problem: MathDex, EgoMath, LaTeXSearch, LeActiveMath, DLMF equation search, MathWebSearch, but none accepted by the community as *the* MSE.

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



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…

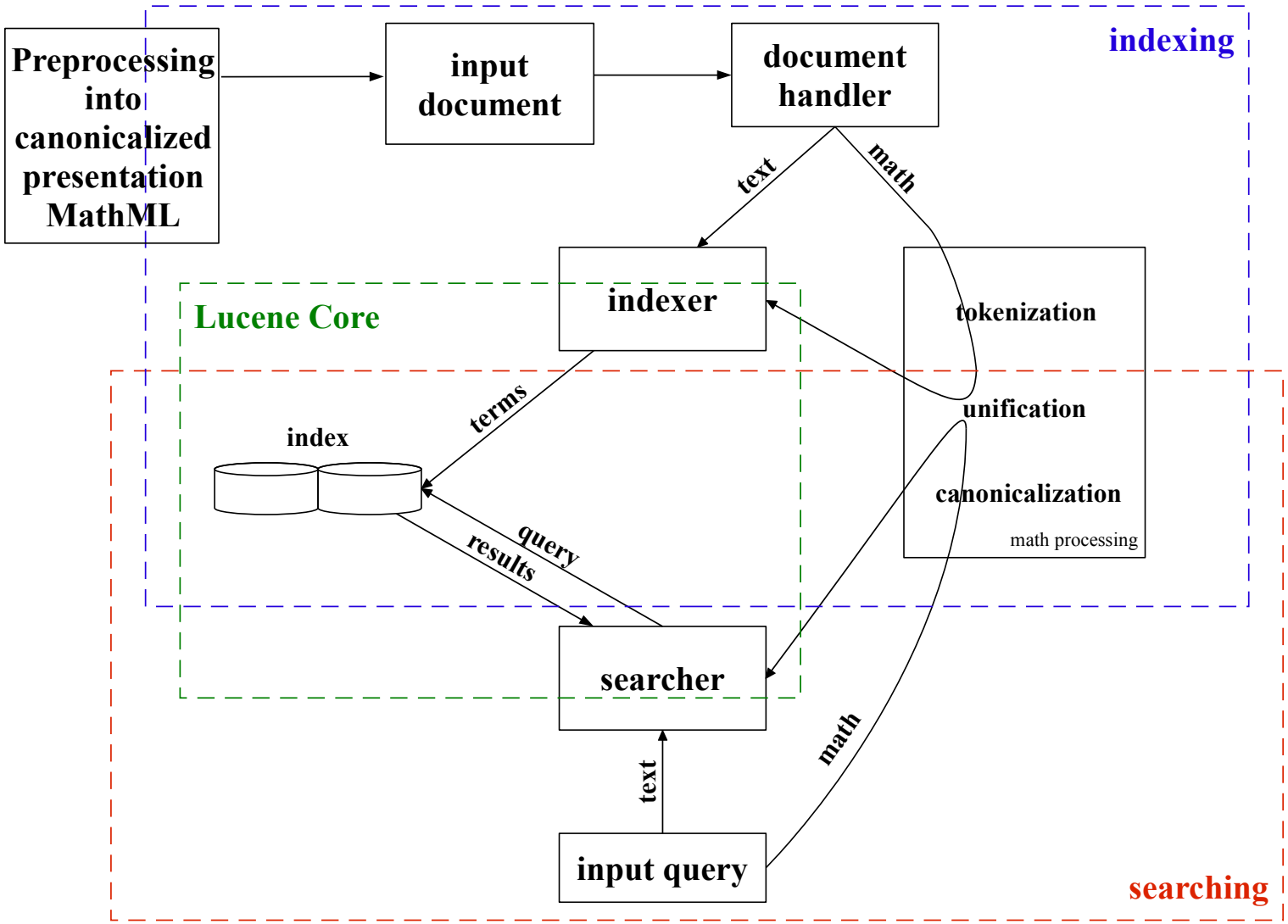
Dual world of query language and indexing language

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

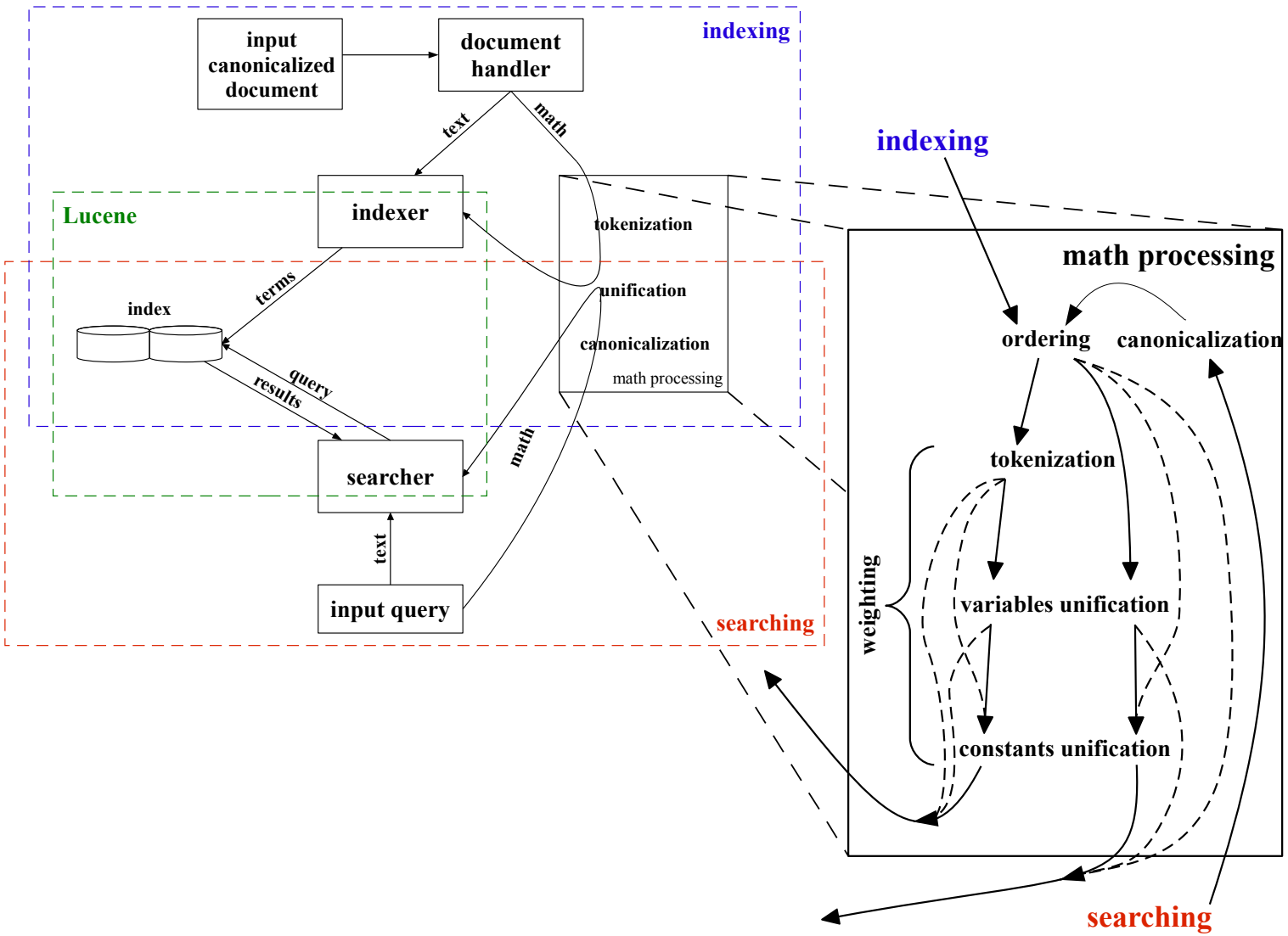
TEXbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

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

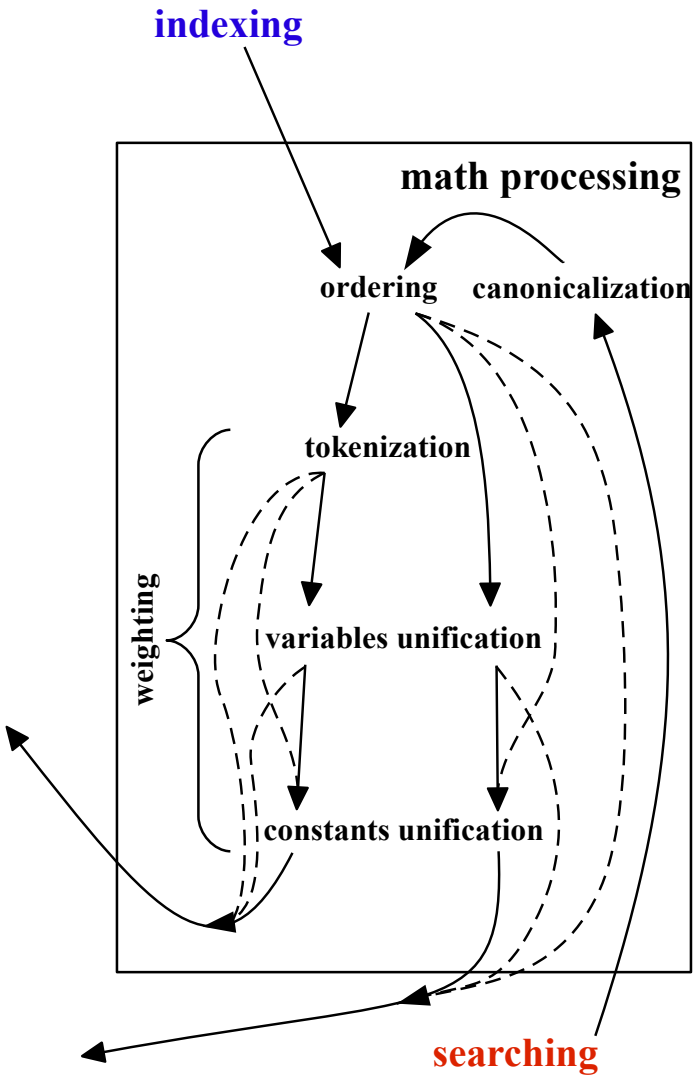
MSE overall design

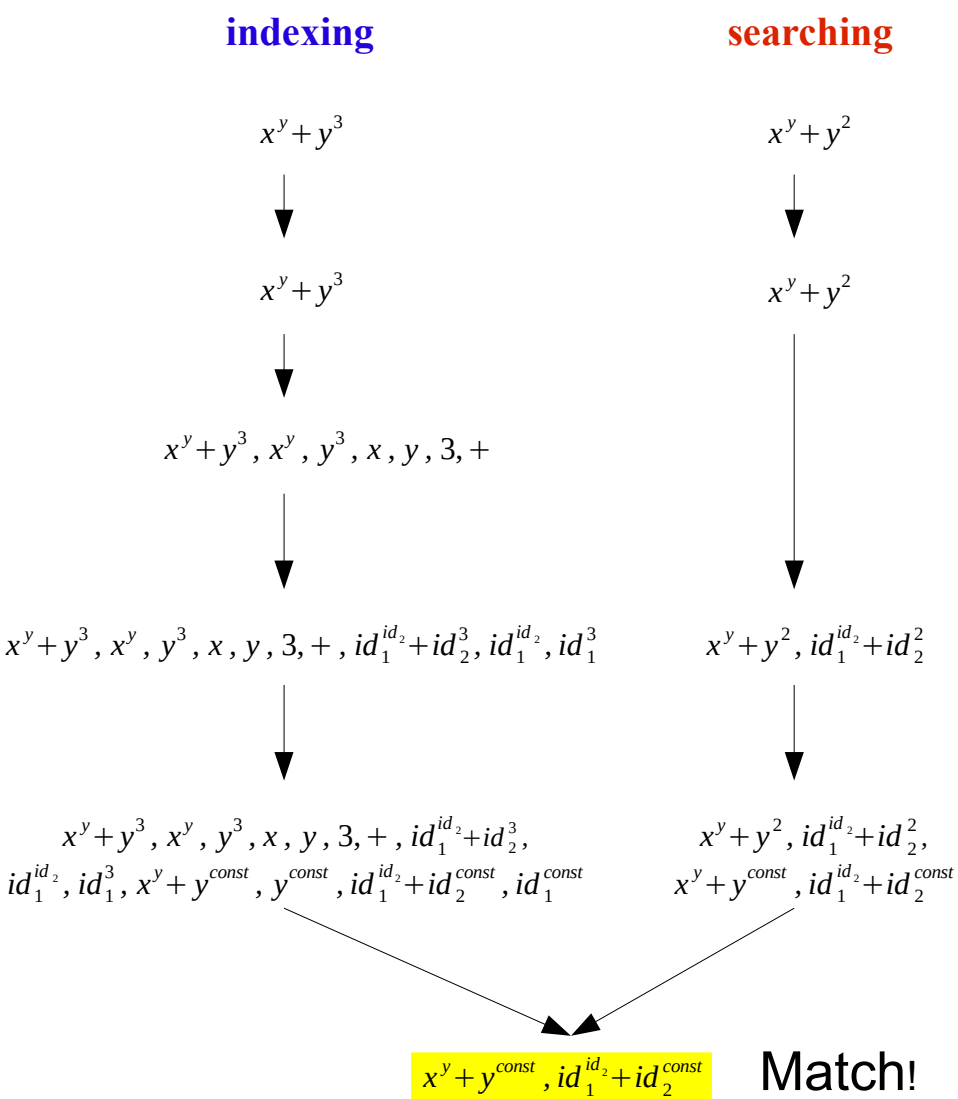


Math indexing design

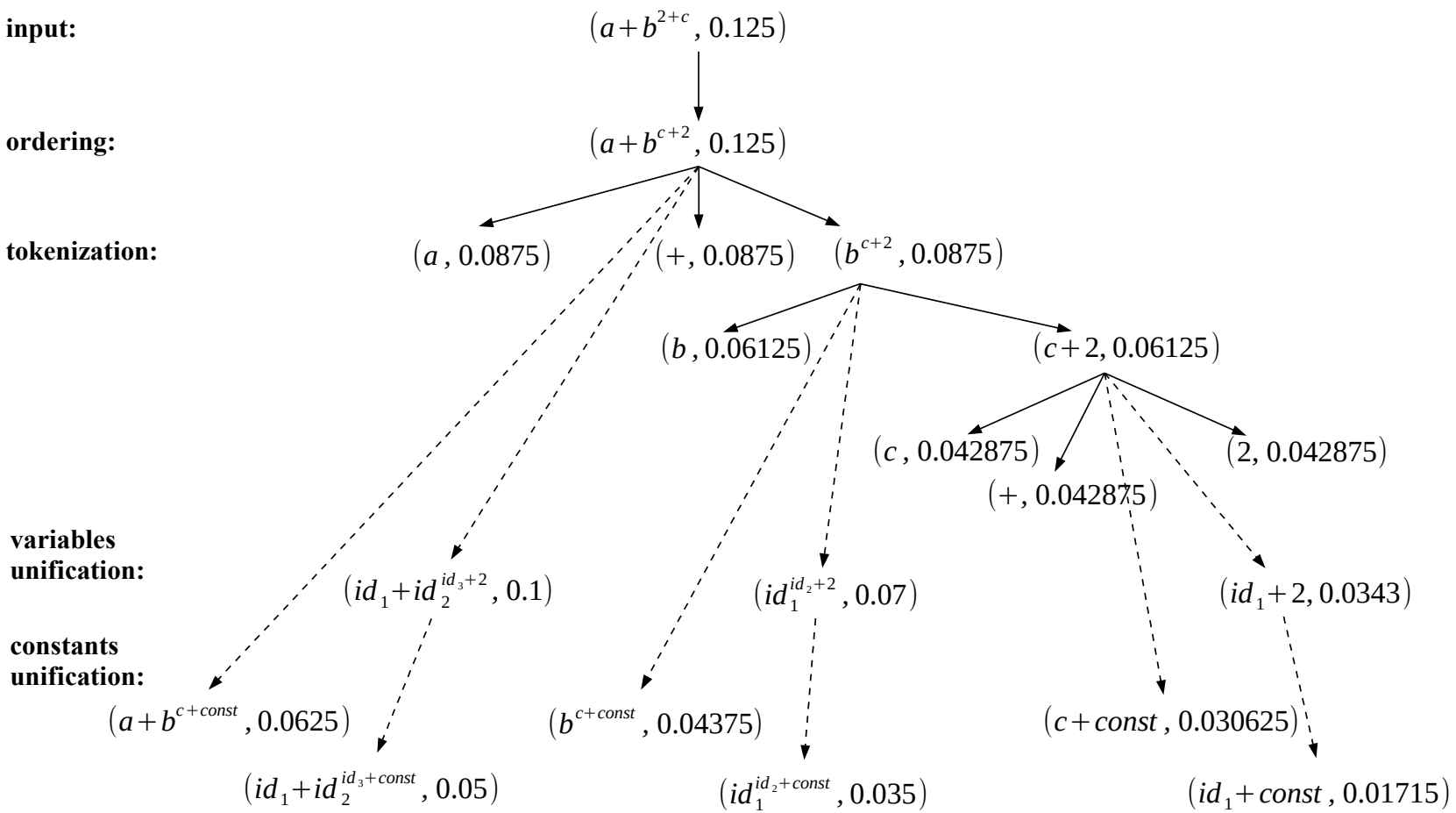


Example


indexing
math processing
ordering
canonicalization
tokenization
weighting
variables unification
constants unification
searching


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



Implementation

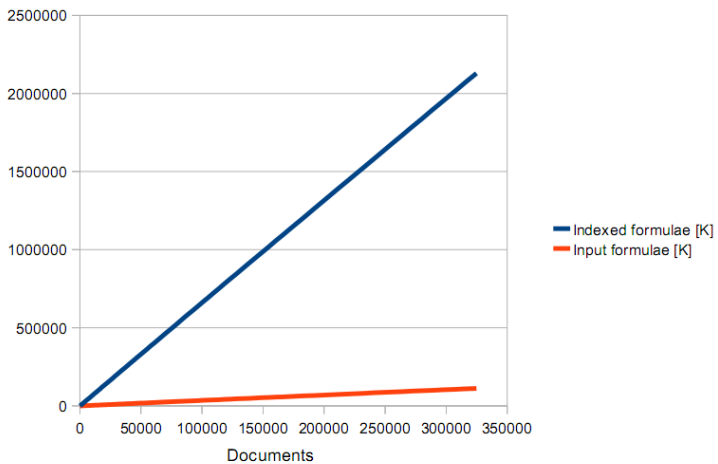
- Java
- Lucene 3.1.0
- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system
- MIaS4Solr plugin was created for the use in Solr in EuDML
- Textual content – processed by StandardAnalyzer

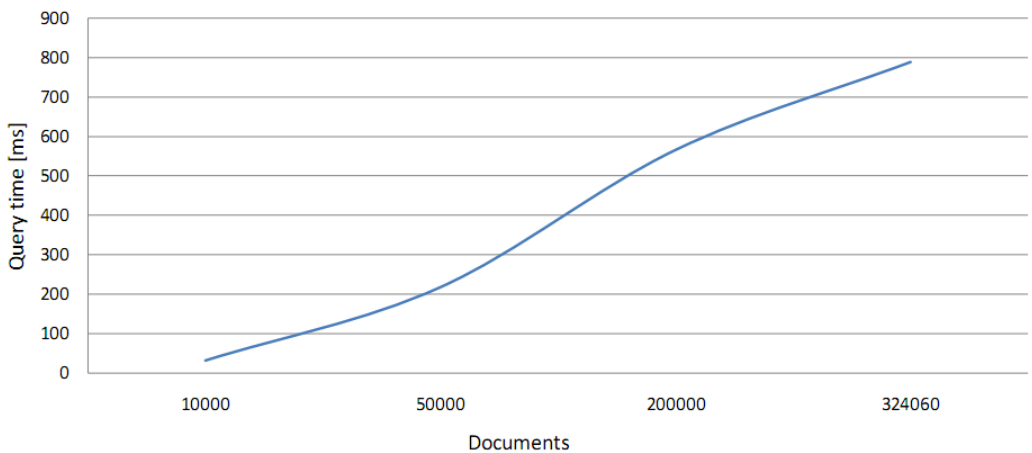
Data used for evaluation: MREC corpus

- Mathematics REtrieval Corpus (MREC, version 2011.4.439)
  - 439,423 documents (originated from arXMLiv [8], validated, enriched with metadata for snippet generation)
  - Uncompressed size 124 GB, compressed 15 GB
  - 158 million input formulae, 2.9 billion subexpressions indexed (Lucene index size 47 GB)
- For more information see paper (DML 2011, Bertinoro) [10] and home page of MREC subproject http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html

Scalability (tested on MREC 2011.4.439)

- Indexing time: 1,378.82 min (23 hours, down to 9 h with threads)
- Average query time: 469 ms
- Overall index size 47 GB (most of it math entries)
- Linear time scale – still seems feasible for a digital library





Search demonstration

Formulae search demonstration comments

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

- MathML/TEX input (Tralics [2] for conversion to MathML [7])
- Canonicalization of the query – problems with UMCL library [1]
- Matched document snippet generation
- MathJax for nicer math rendering and better portability

MIaS already integrated in the the beta version of EuDML system.

Summary – part I – Collection

- Verified complex DML-CZ digitization workflow and proven technologies and tools for math DL
- EuDML: Towards *accessible* wordwide digital mathematical library, based on DML-CZ know-how and tools
- DML workshop series, join us at DML 2013 c/o CICM Bath in UK in July 2013
- Activities towards WDML (Washington meeting, Sloan funding,...)

Summary - part 2 – math search

- Scalable solution for math formulae search researched, implemented, tested and integrated into current version of EuDML system!
- MIaS project pages – http://nlp.fi.muni.cz/projekty/eudml/mias

Future work

- Preprocessing from TEX, PDF,...
- `copypaste` package (storing TEX math code into PDF as second layer with `/ActualText` (for indexing purposes): typesetters may use in their workflows.
- Improved MathML canonicalization and new preprocessing filters, test on more EuDML data.
- Weighting optimization (by machine learning).
- Query relaxation ("Did you mean..").
- Addition of Content MathML tree indexing
- Mathematical equivalence computation via symbolic algebra system?
- NCTIR challenge

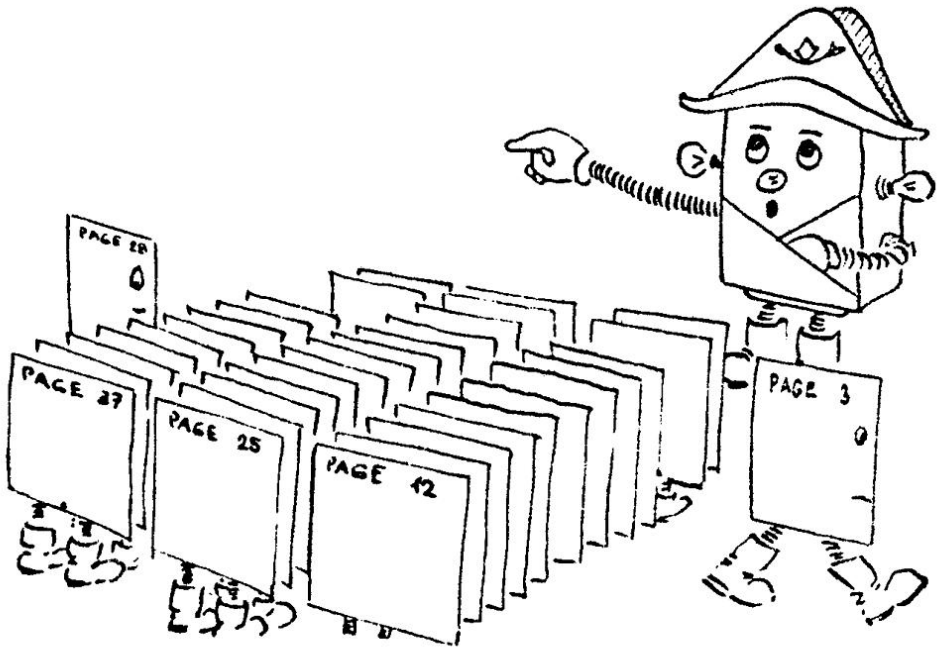
Summary

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) [6].

Acknowledgments and questions?



Acknowledgements: EuDML project (funding), EuDML colleagues, Martin Líška, Michal Růžička, 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>


Sojka, P., Líška, M.: The Art of Mathematics Retrieval. In: Tompa, F., Hardy, M. (eds.) Proceedings of DocEng 2011 Conference. pp. 57–60. ACM. Mountain View, September 2011.


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>.