Towards Machine-Actionable Modules of a Digital Mathematics Library

Michal Růžička, Petr Sojka, Vlastimil Krejčíř

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

DML/CICM 2013, Bath
July 11th, 2013, 3:00 PM

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Outline

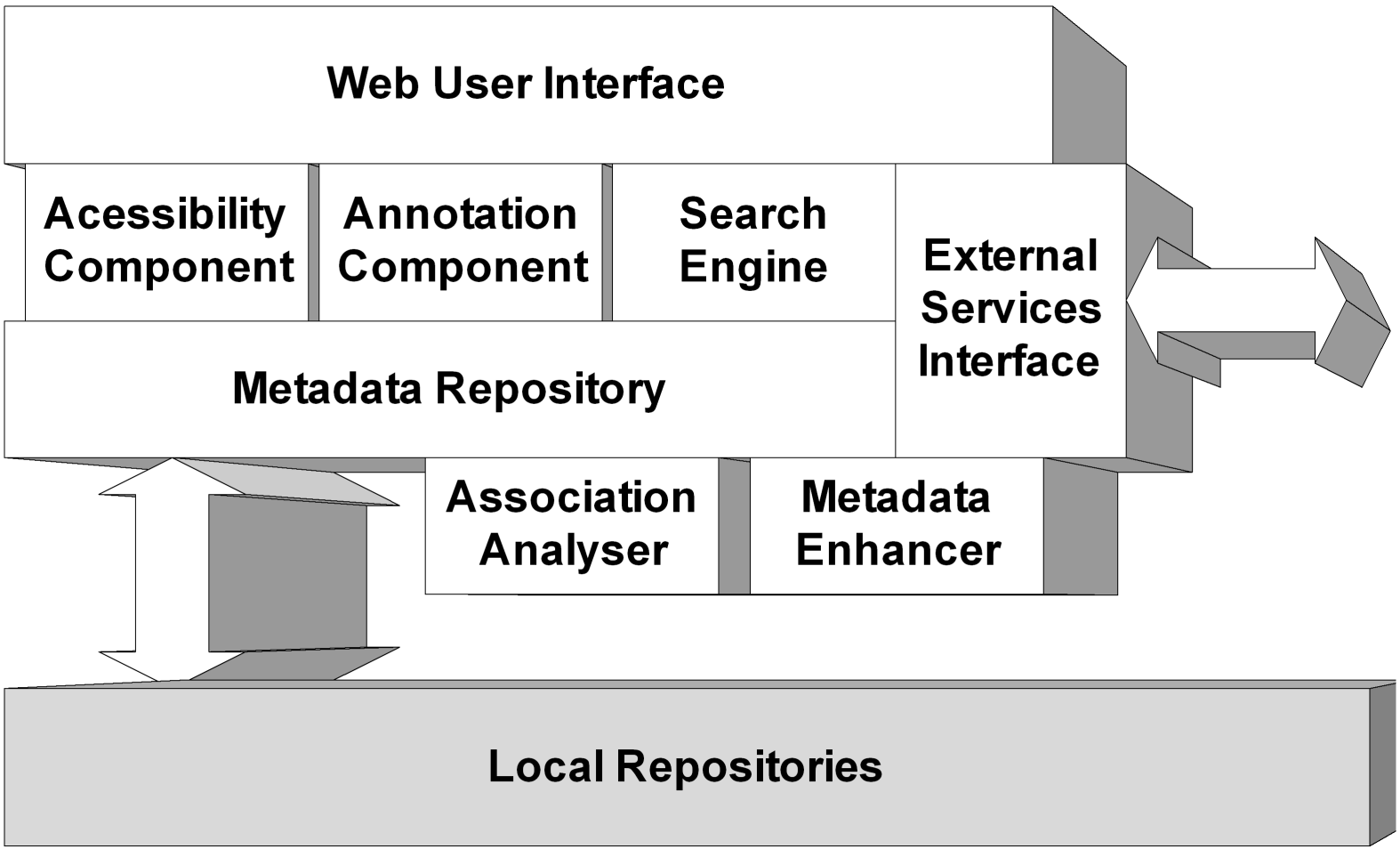
Vision of European Digital Mathematics Library

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 publishers') *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 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,…

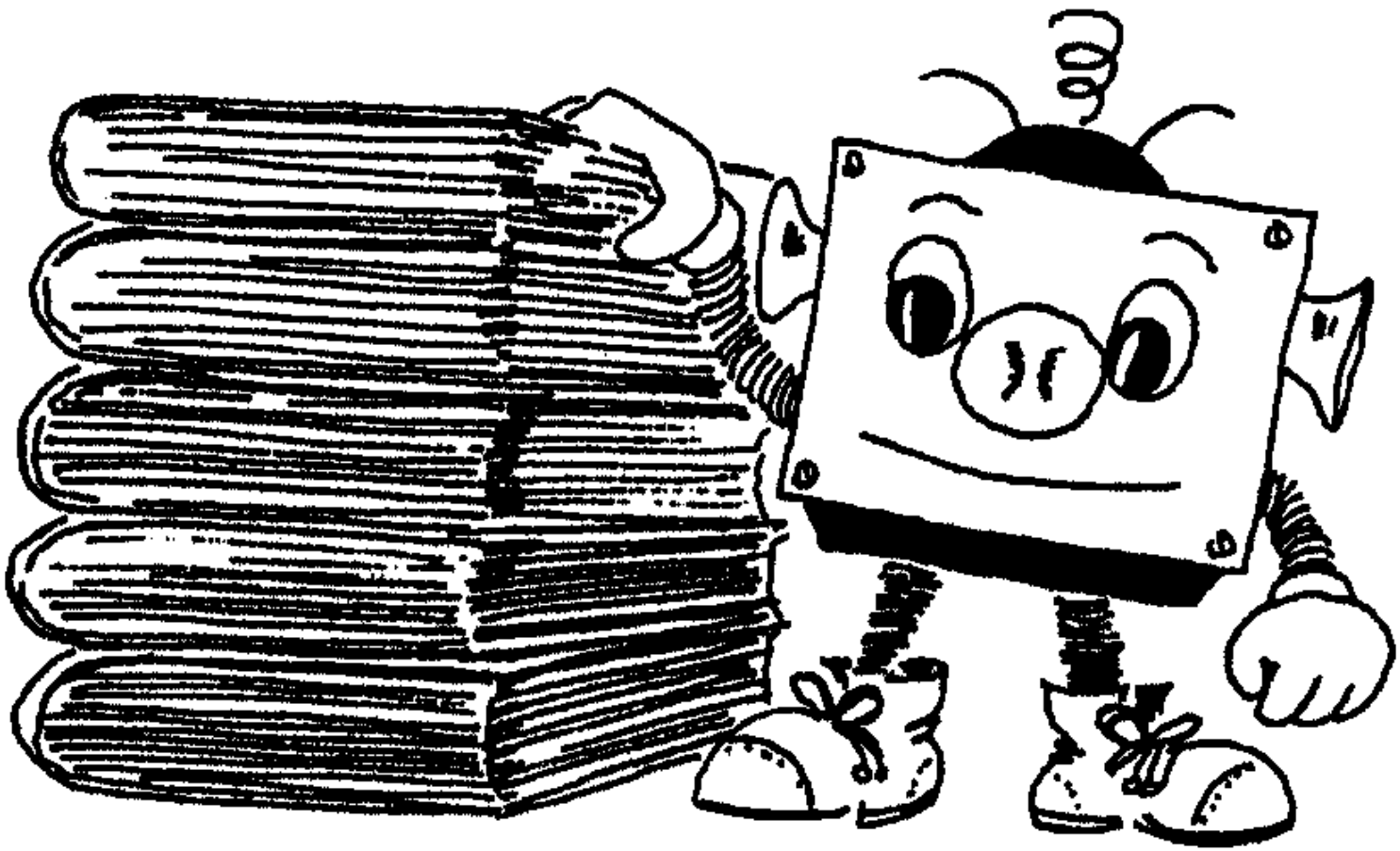
DML content providers serve mostly publisher's or regional more or less established DML repositories: The Czech Digital Mathematics Library DML-CZ, NUMDAM, DML-PL, DML-PT, DML-GR, DML-BG, DML-ES,…

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) and consequently extened. It allows also math-awareness (e.g. math stored both in TEX and MathML), and fully fledged reference lists. NLM generation is supported by developed tools funded by NiH.

Innovation, rather than research.

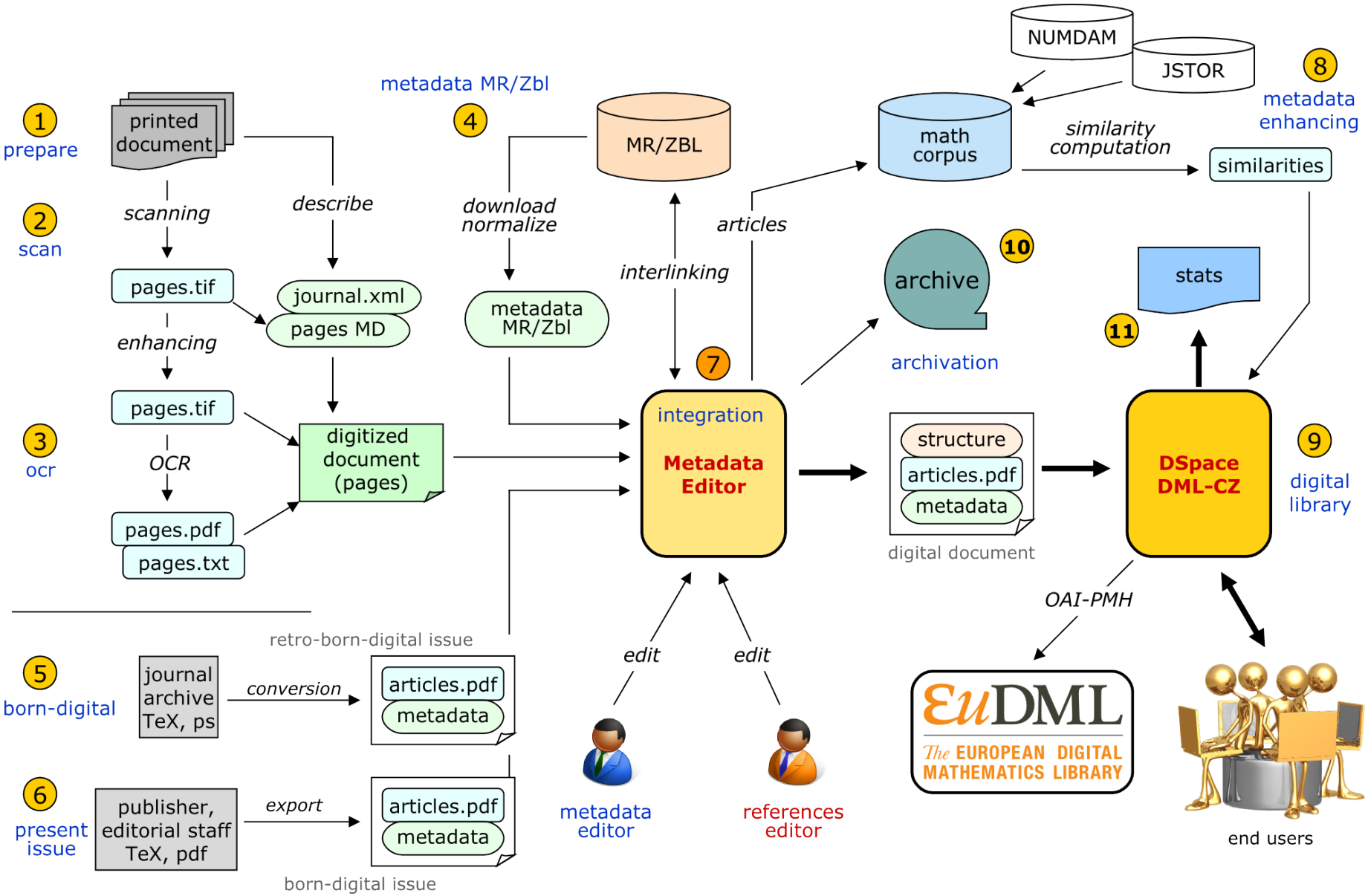
From paper to digital processing, from local to the global DML



Example of DML-CZ

- <http://dml.cz> with 33,000+ papers (330,000+ pages)
- agreements with the primary content providers (mostly *small* publishers)
- metadata validation and editing, often also digitization
- who
- what
- browse, browse similar
- how to search
- export (to EuDML, Google Scholar), metadata conversions

Do as much as possible you can do locally, providing as rich and validated content as possible.


1 prepare
printed document
2 scan
scanning
pages.tif
enhancing
pages.tif
3 ocr
OCR
pages.pdf
pages.txt
describe
journal.xml
pages MD
digitized document (pages)
retro-born-digital issue
5 born-digital
journal archive TeX, ps
conversion
articles.pdf
metadata
6 present issue
publisher, editorial staff TeX, pdf
export
articles.pdf
metadata
born-digital issue
metadata MR/Zbl
4
MR/ZBL
download normalize
metadata MR/Zbl
interlinking
articles
7
integration
Metadata Editor
edit
edit
metadata editor
references editor
NUMDAM
JSTOR
math corpus
similarity computation
8 metadata enhancing
similarities
archive
10
archivation
stats
11
structure
articles.pdf
metadata
digital document
DSpace DML-CZ
9 digital library
OAI-PMH
EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY
end users

The approach used in DML-CZ

A successfully built repository (e.g. set of *workflows*) needs a *coordinated* effort of *librarians*, *IT specialists* and representatives of users—*content specialists*: (D+M+L)=success 'equation'.

Design, technical and *political decisions* behind building the *Czech Digital Mathematics Library DML-CZ* (<http://dml.cz>) in the context of other thematical community projects (PubMed Central, ADS, INSPIRE, SCOAP3 and EuDML) have been solved. *No wheel reinvention.*

Our framework integrates workflow for the articles scanned from a paper (*math OCR*), for documents from retro-born digital period (data available in some type of electronic form) and for born-digital ones. To sustain, minimize manual labor, automate as much as possible.

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 $K \in \mathfrak{A}$ 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

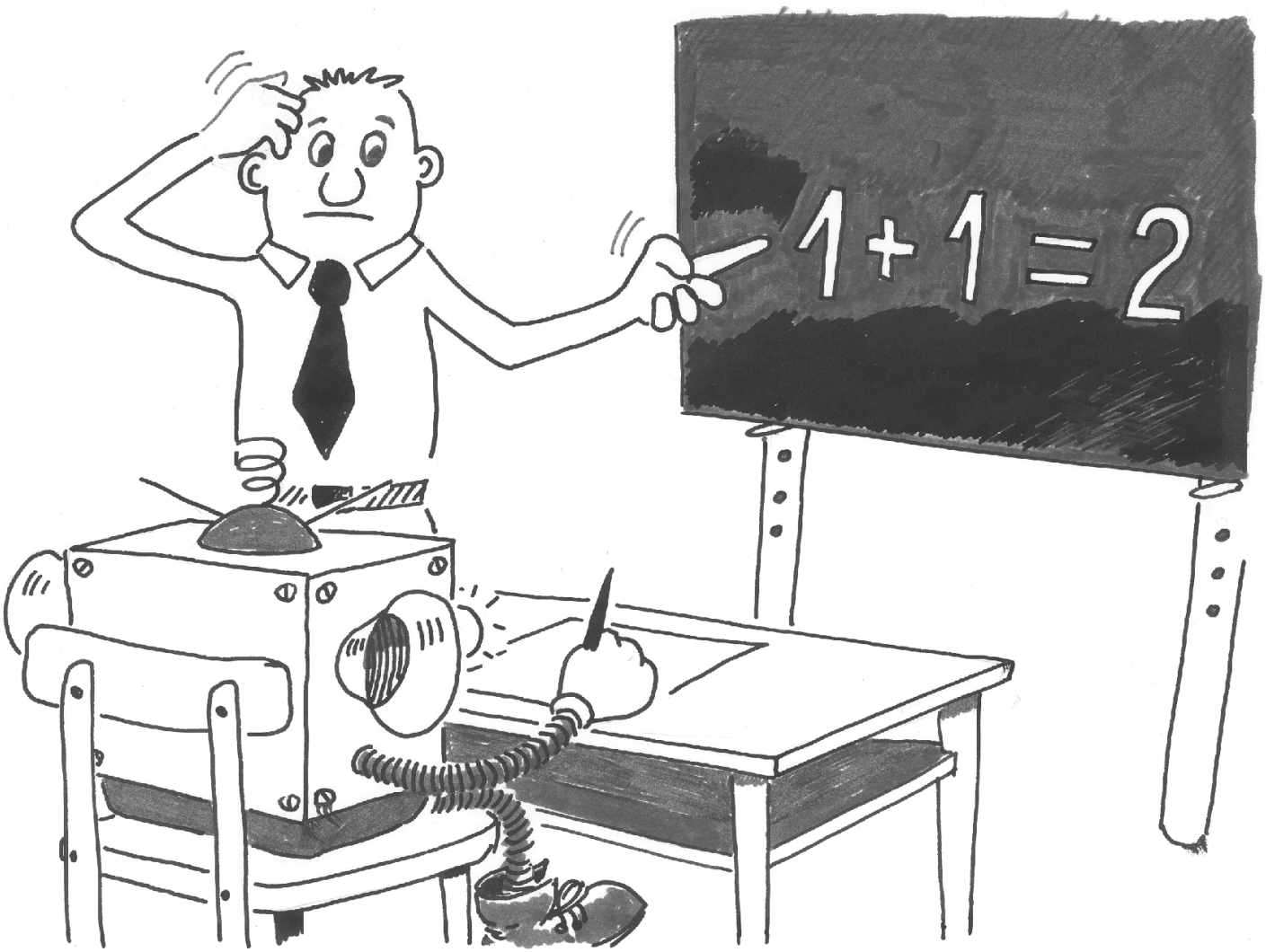
ЯН МАРЖИК (Jan Mařík), Прага.

(Поступило в редакцию 10/X 1955 г.)

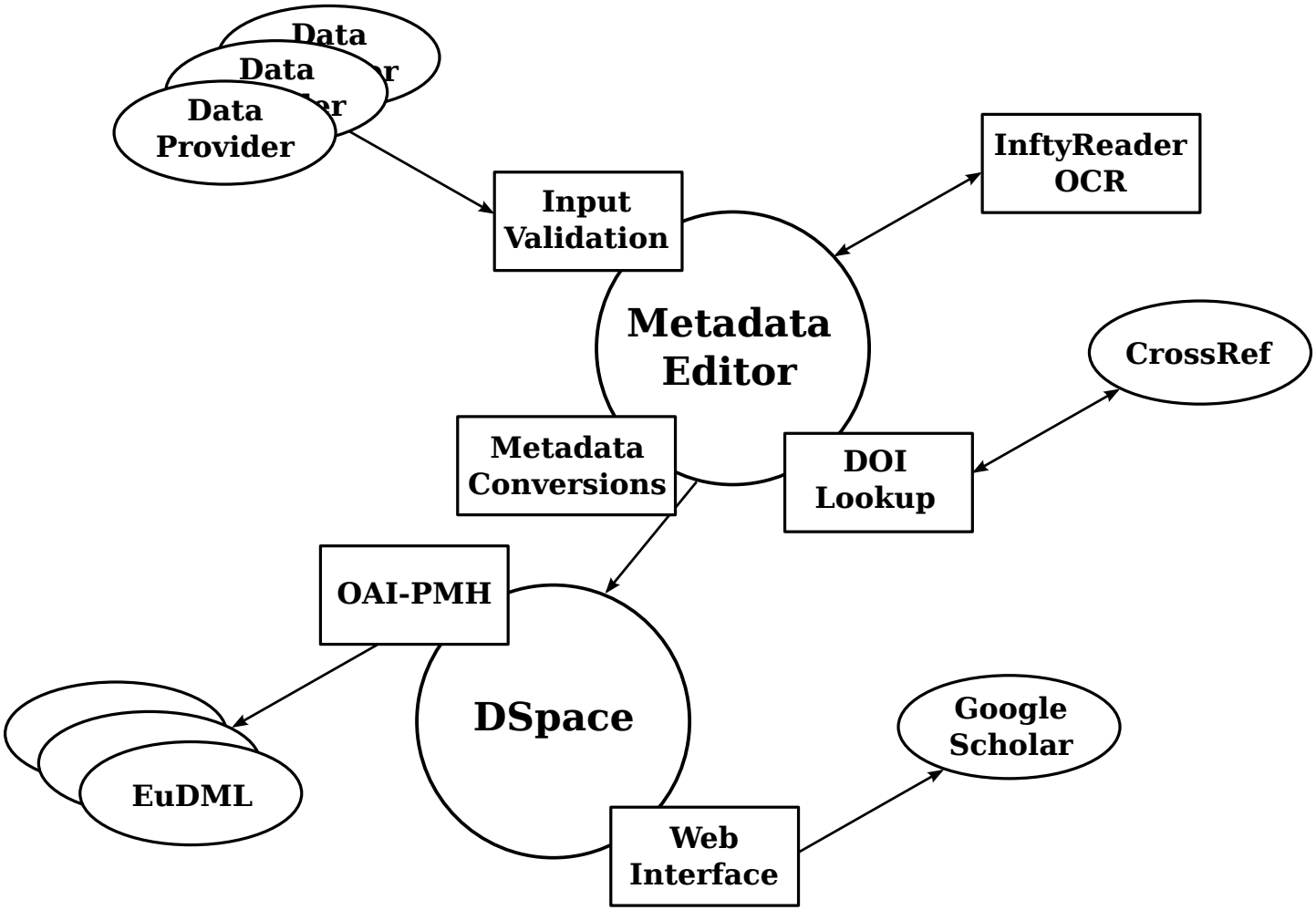
Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^{m} v_i^2(x) \leqq 1$ 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

Challenges of Math handling: OCR, indexing, search…



Challenges of automation, validation, metadata mapping in DML-CZ



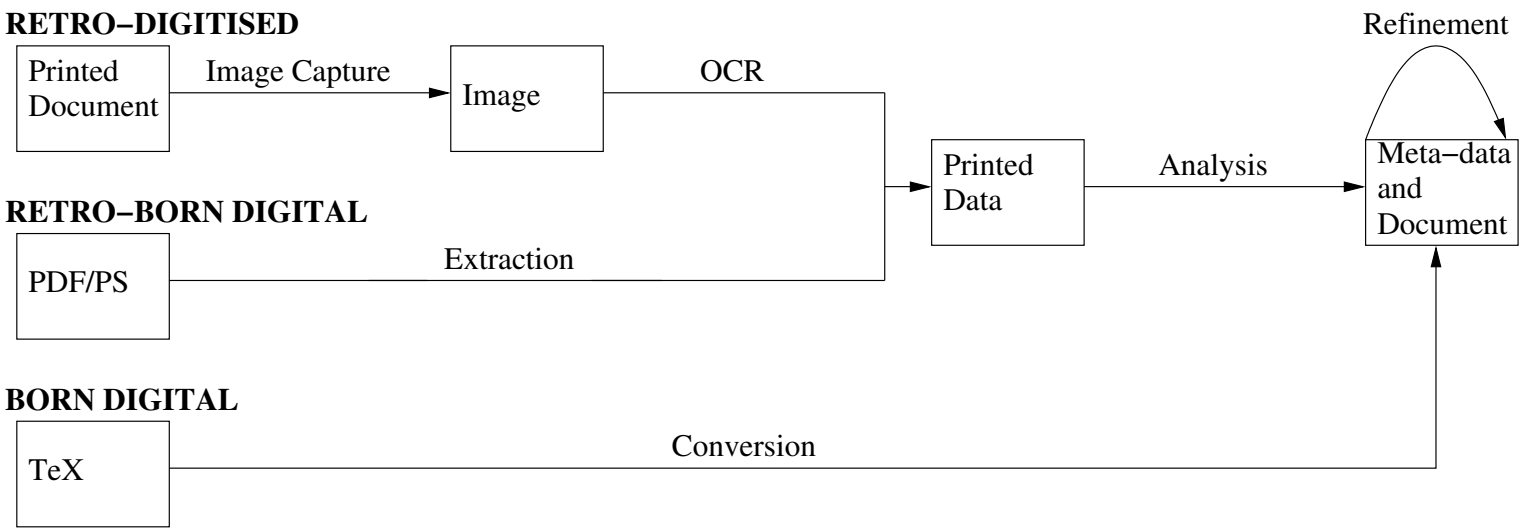
Document accessibility 4 DML processing challenges

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 fulltexts 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 to get fulltexts with math to enhance DML services

In EuDML: Run in parallel in Brno (DML-CZ), Grenoble (NUMDAM) and Lisbon (other) 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,…). Re-OCRing DML-CZ takes two weeks.

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)
```

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

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

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

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

```
<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>
```

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]).

Enhanced PDF serving issues (rights from data providers, errors).

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: $\rightarrow$ high *automation* to save costs: converting to MathML (via Tralics) to allow discoverability and indexing (formulae similarity search). 130+K fulltexts with MathML, and growing….

Reference Parsing for linking and validation

- References needed to segment and parse in fulltexts: manual editing (in Metadata editor) costly
- Born-digital needed to check against CrossRef
- ParsCit <http://wing.comp.nus.edu.sg/parsCit/>
- DOI lookup

Parsing with ParsCit

From OCR we get:

```
[5] Lambe, L., Stasheff, J.: Applications
of perturbation theory to iterated fibrations.
Manuscripta Math. 58 (1987), 363-376.
```

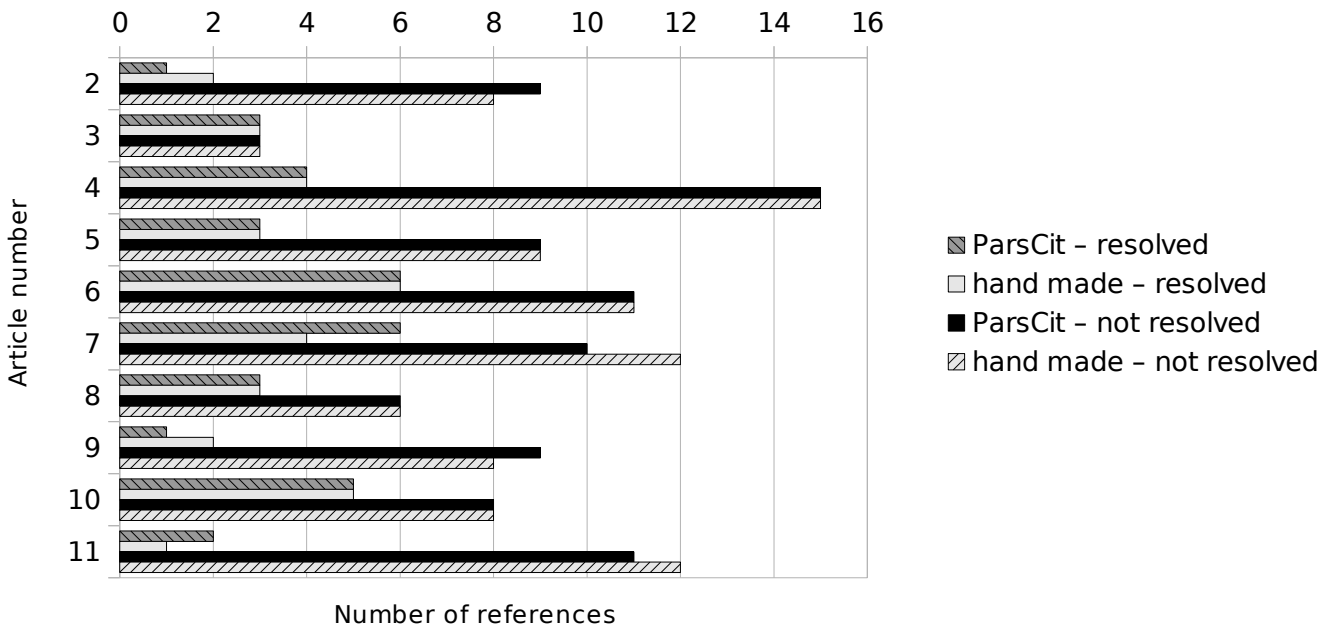
Parsing citations with ParsCit

```
<algorithms version="110505">
  <algorithm name="ParsCit" version="110505">
    <citationList>
      <citation valid="true">
        <authors>
          <author>L Lambe</author>
          <author>J Stasheff</author>
        </authors>
        <title>Applications of perturbation theory to iterated
               fibrations.</title>
        <date>1987</date>
        <journal>Manuscripta Math.</journal>
        <volume>58</volume>
        <pages>363--376</pages>
        <marker>[5]</marker>
        <rawString>Lambe, L., Stasheff, J.: Applications of
                   perturbation theory to iterated fibrations.
                Manuscripta Math. 58 (1987), 363-376.</rawString>
      </citation>
    </citationList>
  </algorithm>
</algorithms>
```

DOI lookup success rates

ParsCit as a preprocessor for CrossRef DOI look up by HTTP XML Query

Articles of volume 48, issue 5 of the *Archivum Mathematicum* journal (<http://dml.cz/handle/10338.dmlcz/143106>)



DOI lookup success rates

	number of refs.	ParsCit resolved	ParsCit not resolved	hand made resolved	hand made not resolved
article #2	10	1 (10.00%)	9 (90.00%)	2 (20.00%)	8 (80.00%)
article #3	6	3 (50.00%)	3 (50.00%)	3 (50.00%)	3 (50.00%)
article #4	19	4 (21.05%)	15 (78.95%)	4 (21.05%)	15 (78.95%)
article #5	12	3 (25.00%)	9 (75.00%)	3 (25.00%)	9 (75.00%)
article #6	17	6 (35.29%)	11 (64.71%)	6 (35.29%)	11 (64.71%)
article #7	16	6 (37.50%)	10 (62.50%)	4 (25.00%)	12 (75.00%)
article #8	9	3 (33.33%)	6 (66.67%)	3 (33.33%)	6 (66.67%)
article #9	10	1 (10.00%)	9 (90.00%)	2 (20.00%)	8 (80.00%)
article #10	13	5 (38.46%)	8 (61.54%)	5 (38.46%)	8 (61.54%)
article #11	13	2 (15.38%)	11 (84.62%)	1 (7.69%)	12 (92.31%)

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*
- DSpace exports via OAI-PMH directly in EuDML NLM
- negotiations with Google Scholar towards better visibility
- fulltext export including math (for indexing/search and similarity)
- alternative visualization

DML-CZ visibility

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

Searching (semantically) similar papers

Exploration of a DML: browsing (semantically) similar papers

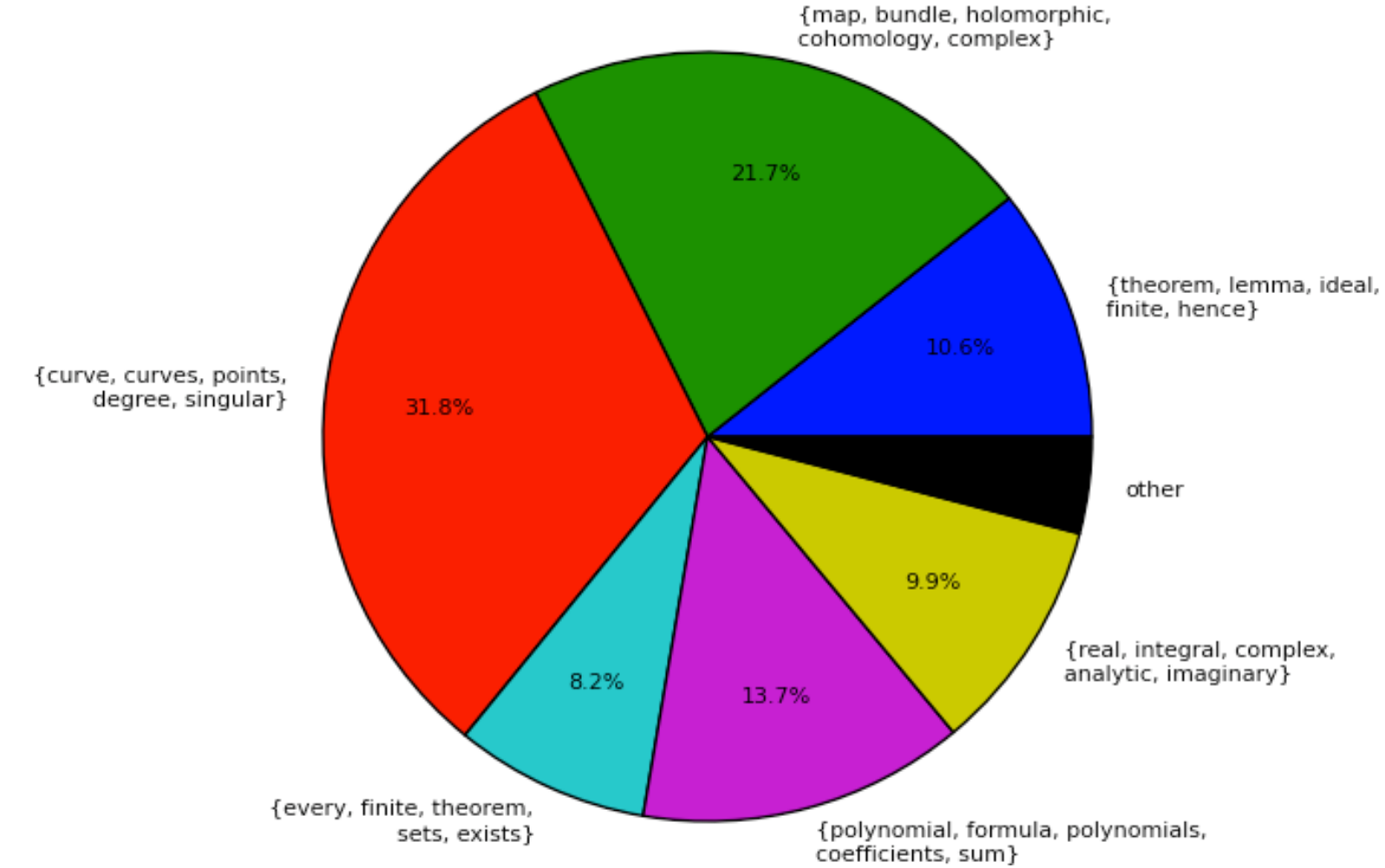
Semantic search via topic modeling: Latent Semantic Indexing, Latent Dirichlet Allocation

Leading Edge Example: Automated Meaning Picking from Texts

LDA Topics Pie Chart for math.0406240:

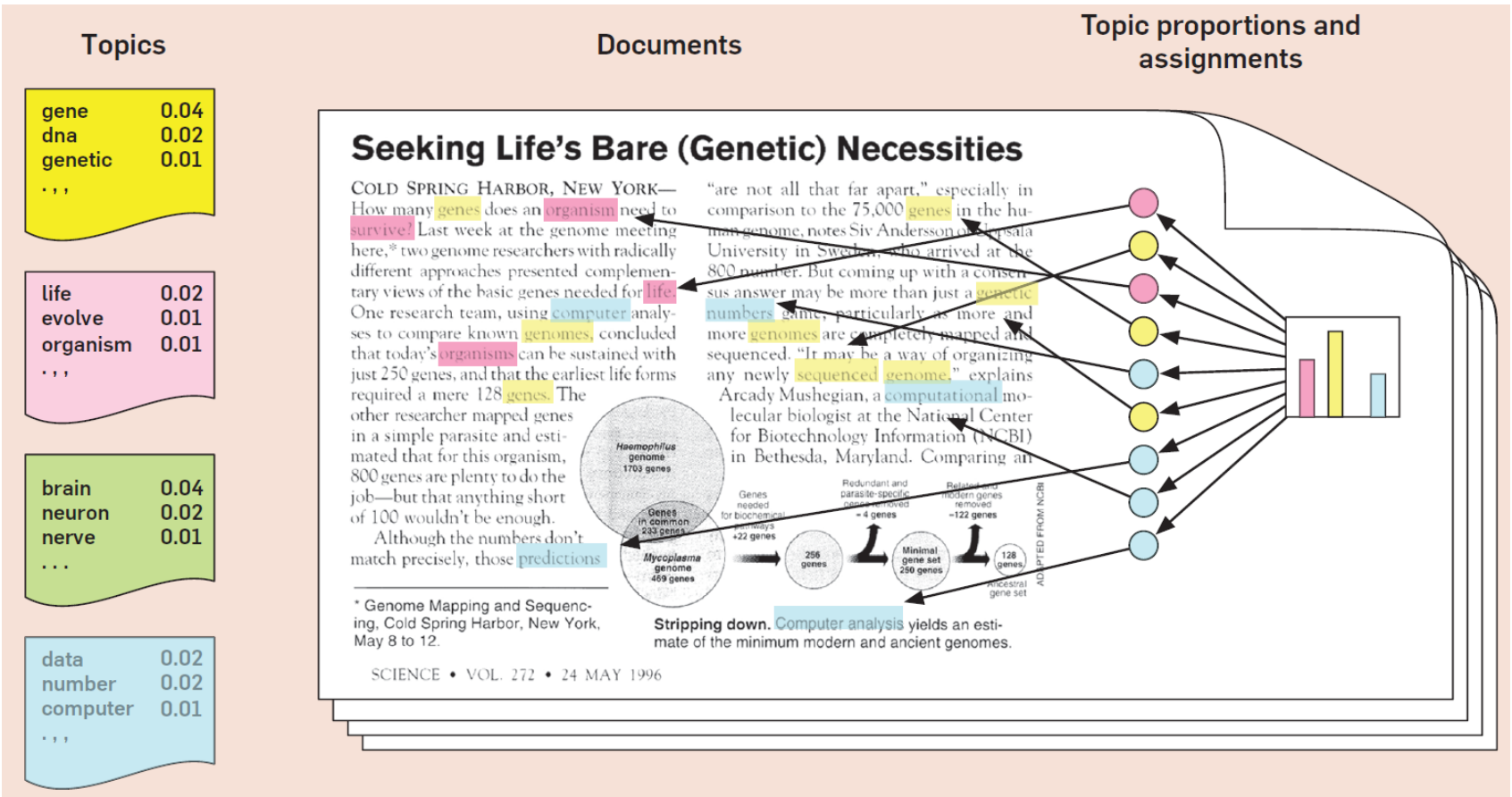
Each slice represents a different topic. The size of the slice corresponds to "how much is the article about this topic?". Topics which contribute <6% to the above document are aggregated under "other".

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



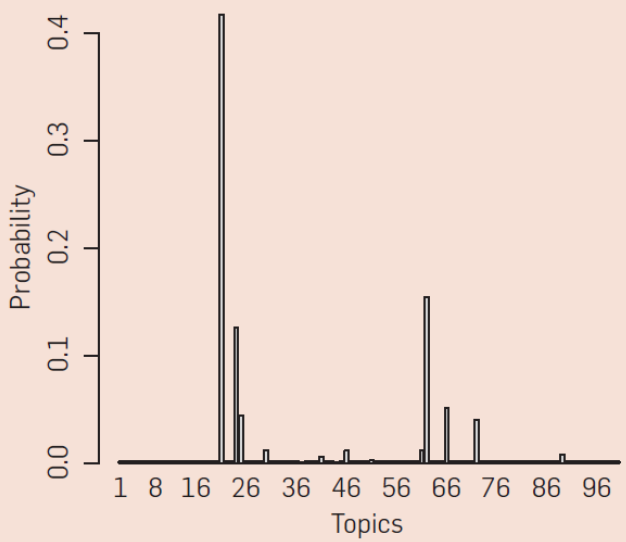
Probabilistic Topical Modeling: Latent Dirichlet Allocation

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



Topical Modeling: Latent Dirichlet Allocation II

- all topics computed automatically from document corpora



"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.

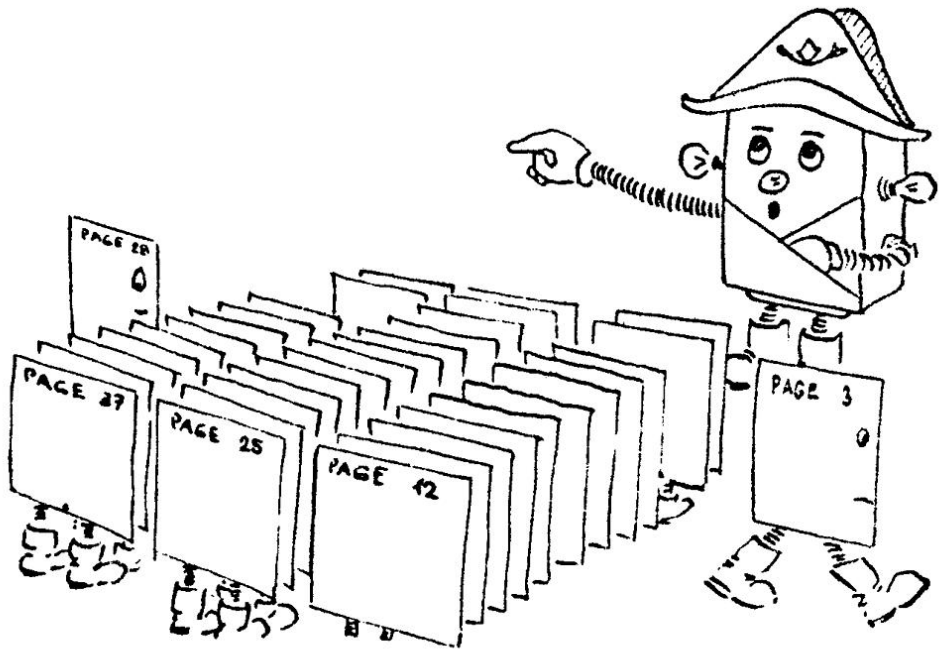
Summary

- DML-CZ and EuDML is up and running, with several novel math-aware approaches developed and *in use*, with high degree of automation
- verified complex workflow and proven technologies and tools for [local] DMLs
- scalable solution for fulltext processing including math formulae, math OCR, search researched, implemented, tested and integrated into current version of EuDML system!
- content aggregation and remapping via REPOX and DSpace
- math-aware methods for document similarity (MathML2text, gensim)
- a lot more on <http:/project.eudml.org> and <http://project.dml.cz>

Future work

- Improving further automation and OCR.
- Improving math-aware search: MathML canonicalization and preprocessing filters, evaluation with the use of EuDML math query log (database of intentions).
- Math mining

Acknowledgments and questions?



Acknowledgements: EuDML project (funding), DML-CZ project, EuDML and DML-CZ 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.


Credit for DML-CZ workflow picture goes to Mirek Bartošek.