Math Information Retrieval in the Past, Present and Future

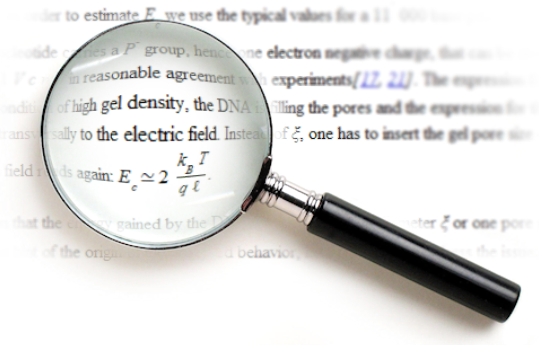
Petr Sojka

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

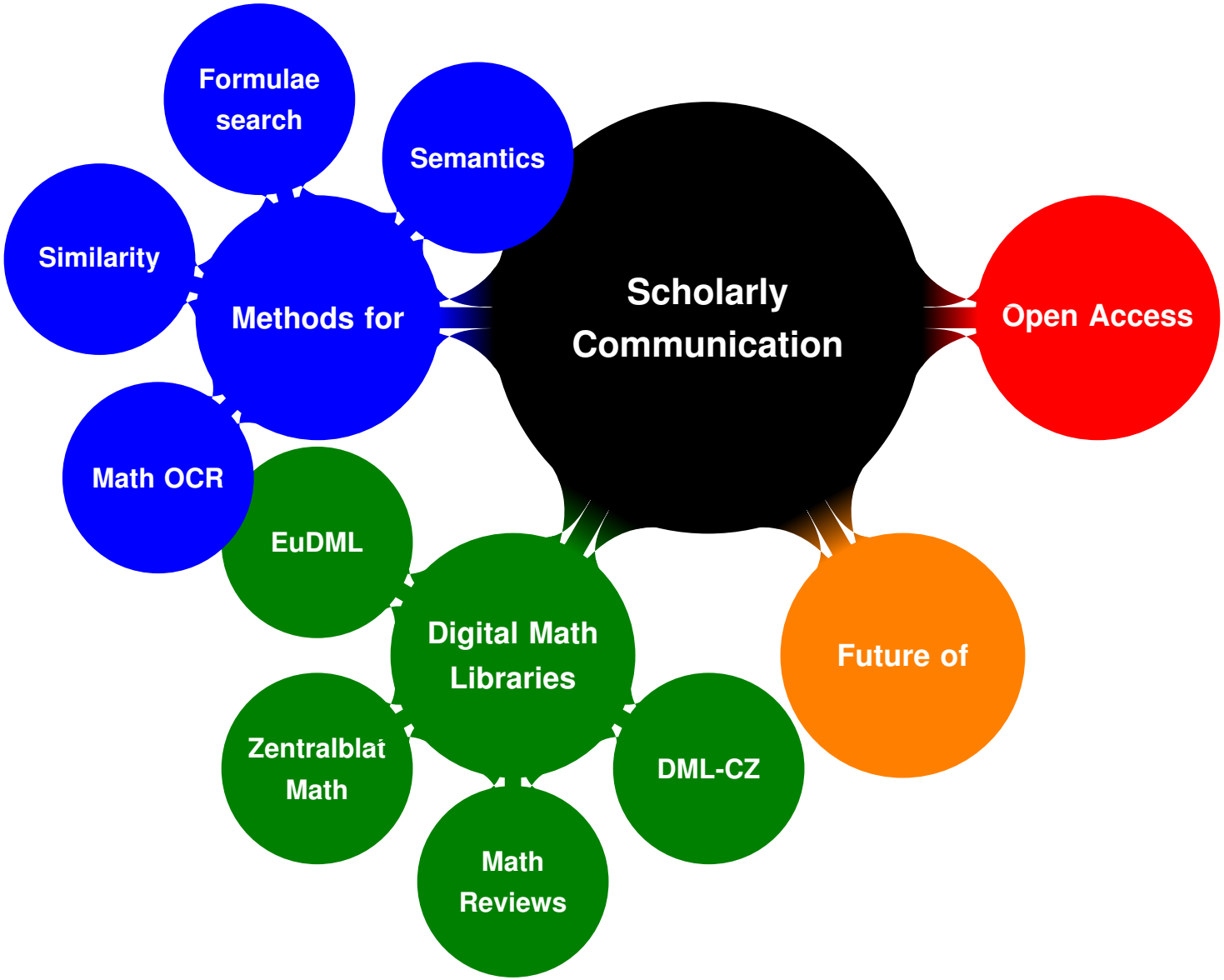
NLP Seminar, Faculty of Informatics, Brno, Czech Republic
October 7th, 2014

EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

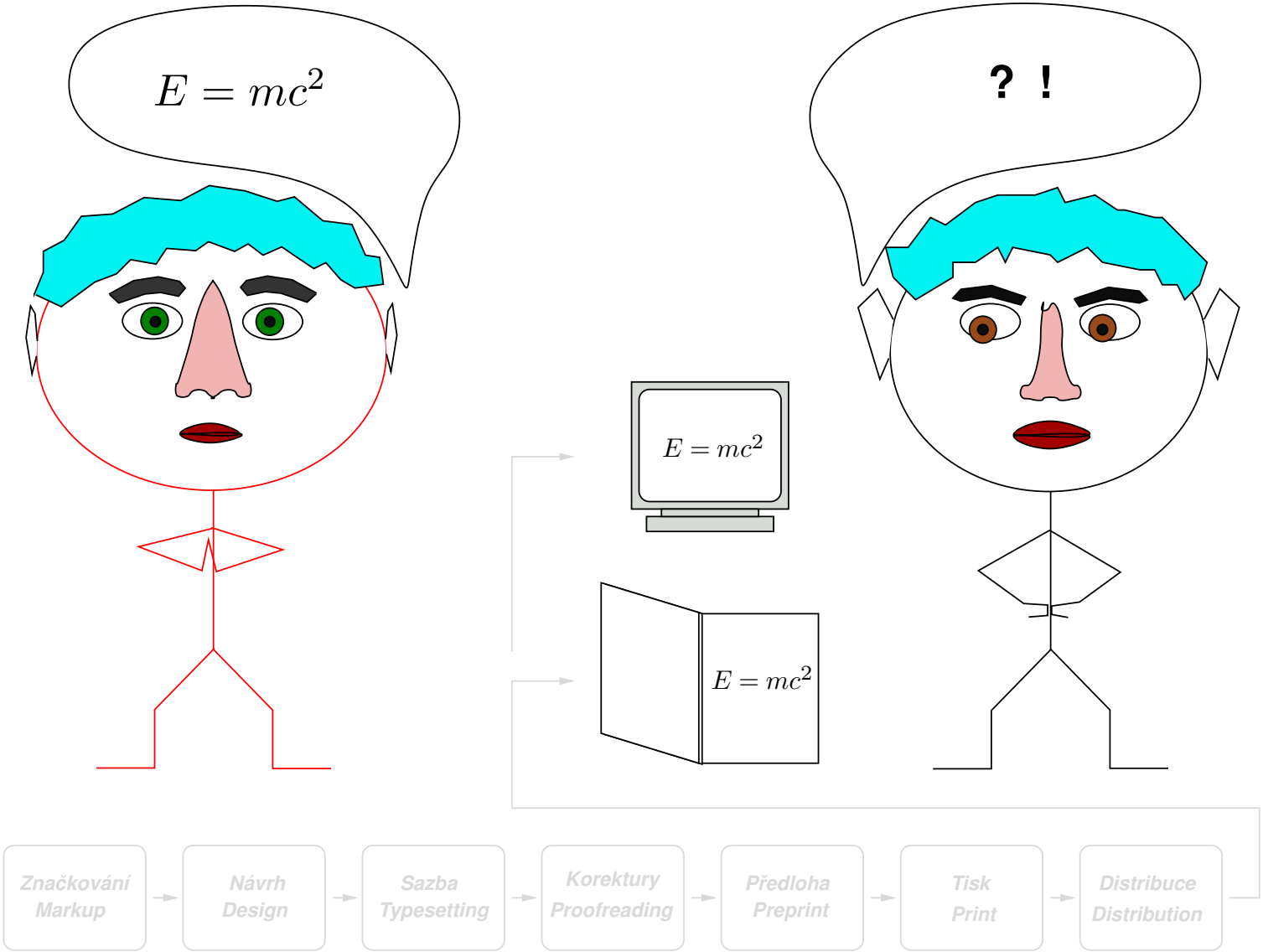
FACULTAS ARTIS INFORMATICAE • UNIVERSITAS MASARYKIANA •



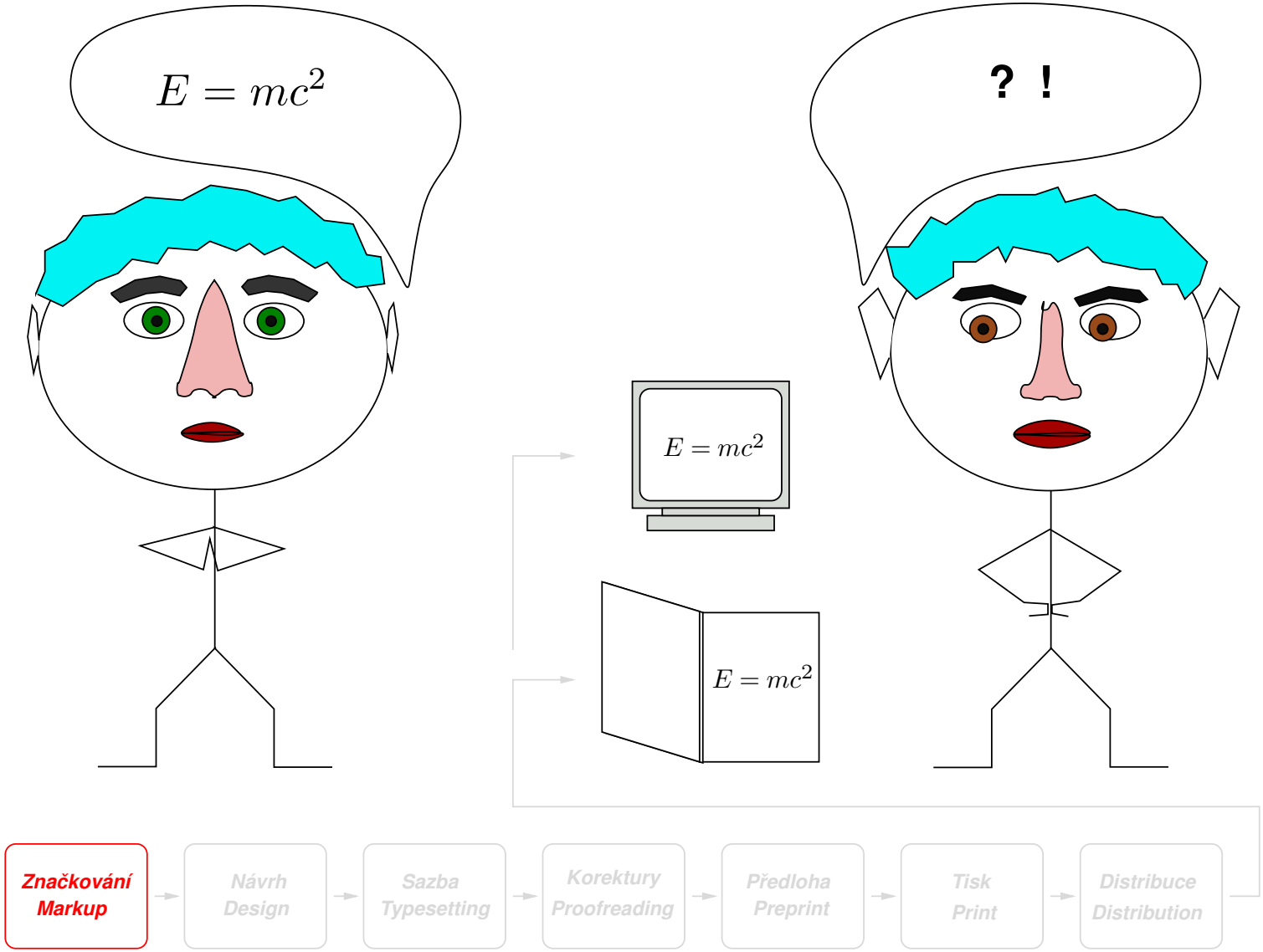
Talk topics and take-home message



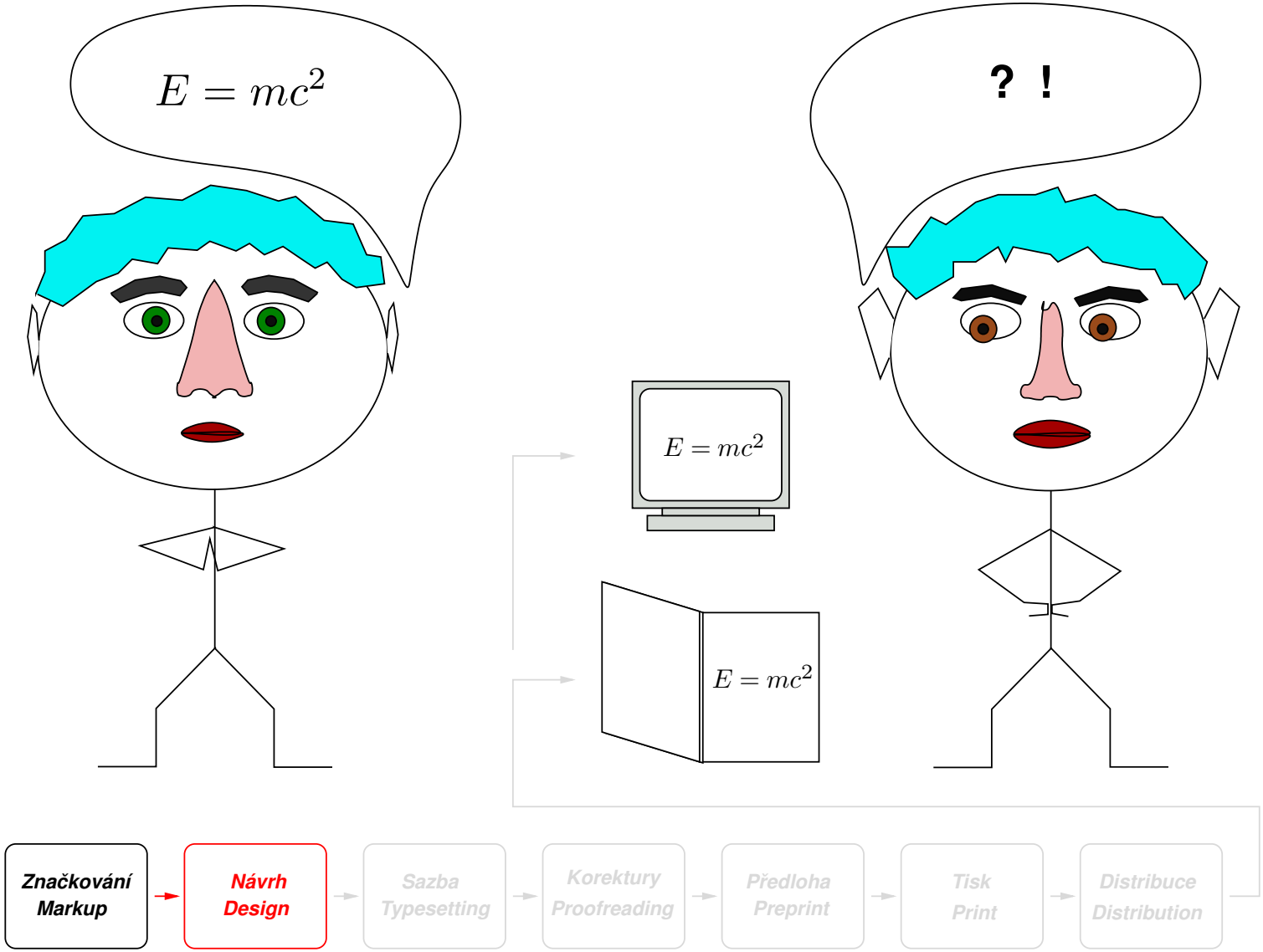
The Past: Scholarly Communication *going digital* from paper



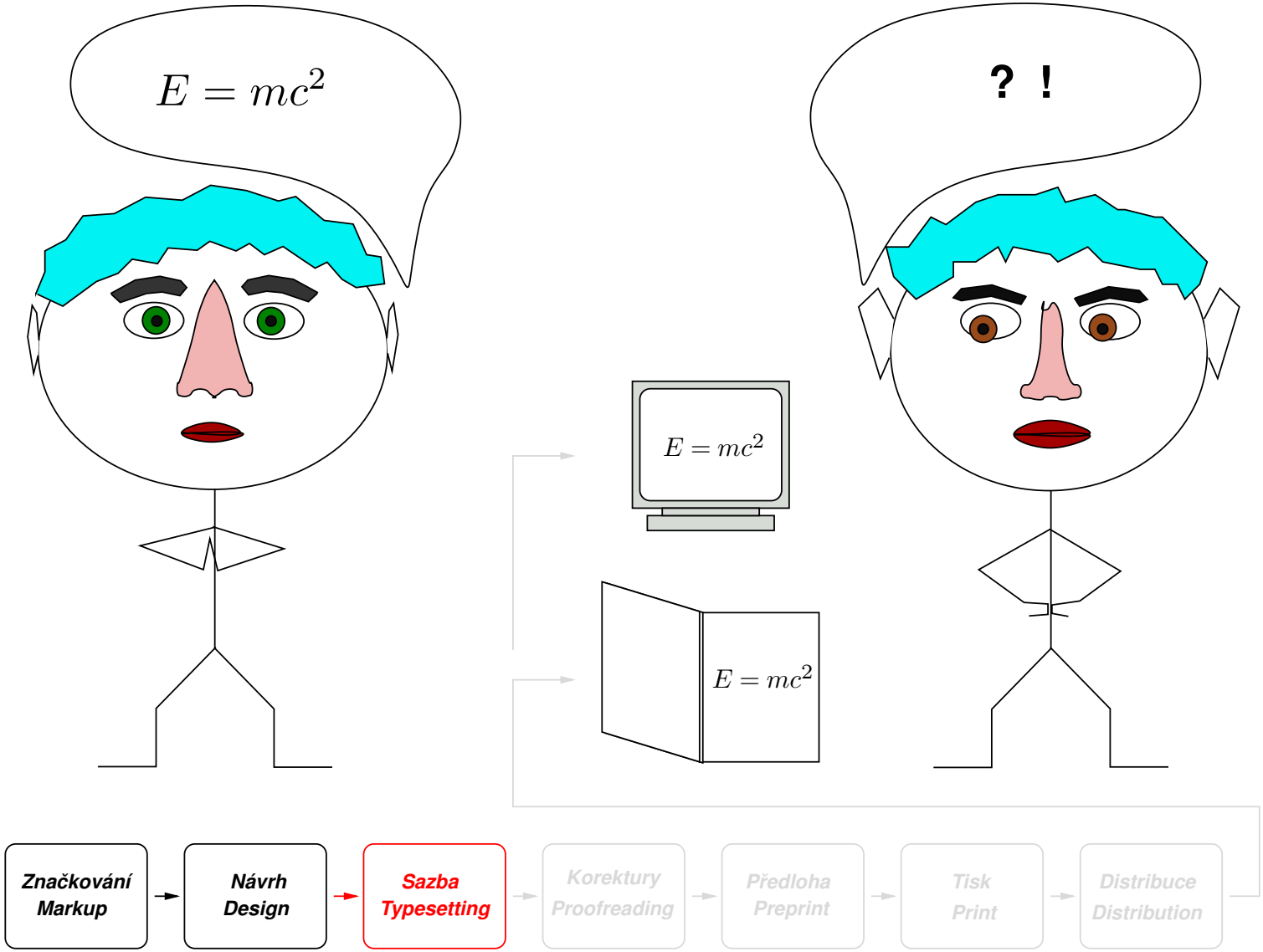
The Past: Scholarly Communication *going digital* from paper



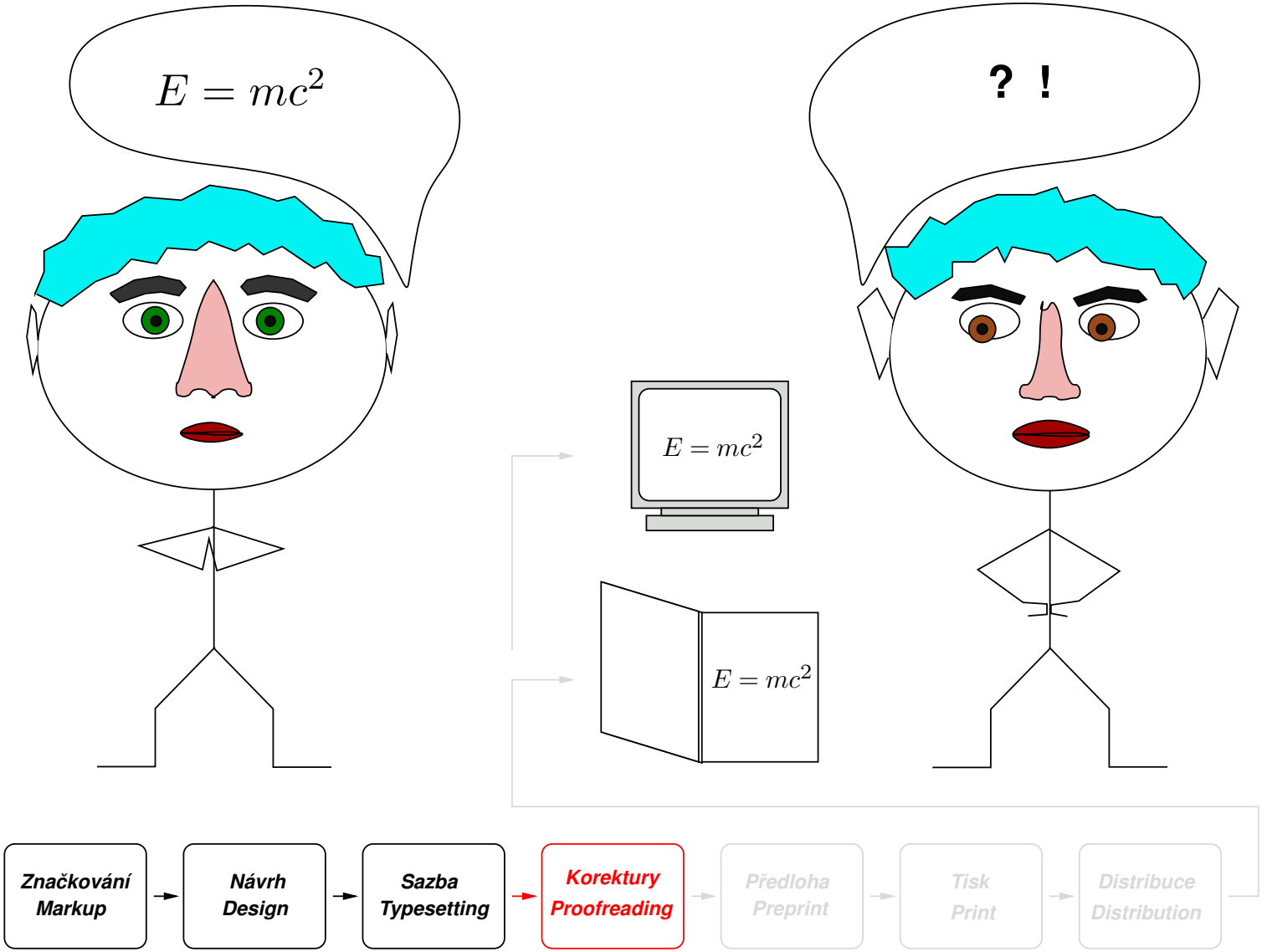
The Past: Scholarly Communication *going digital* from paper



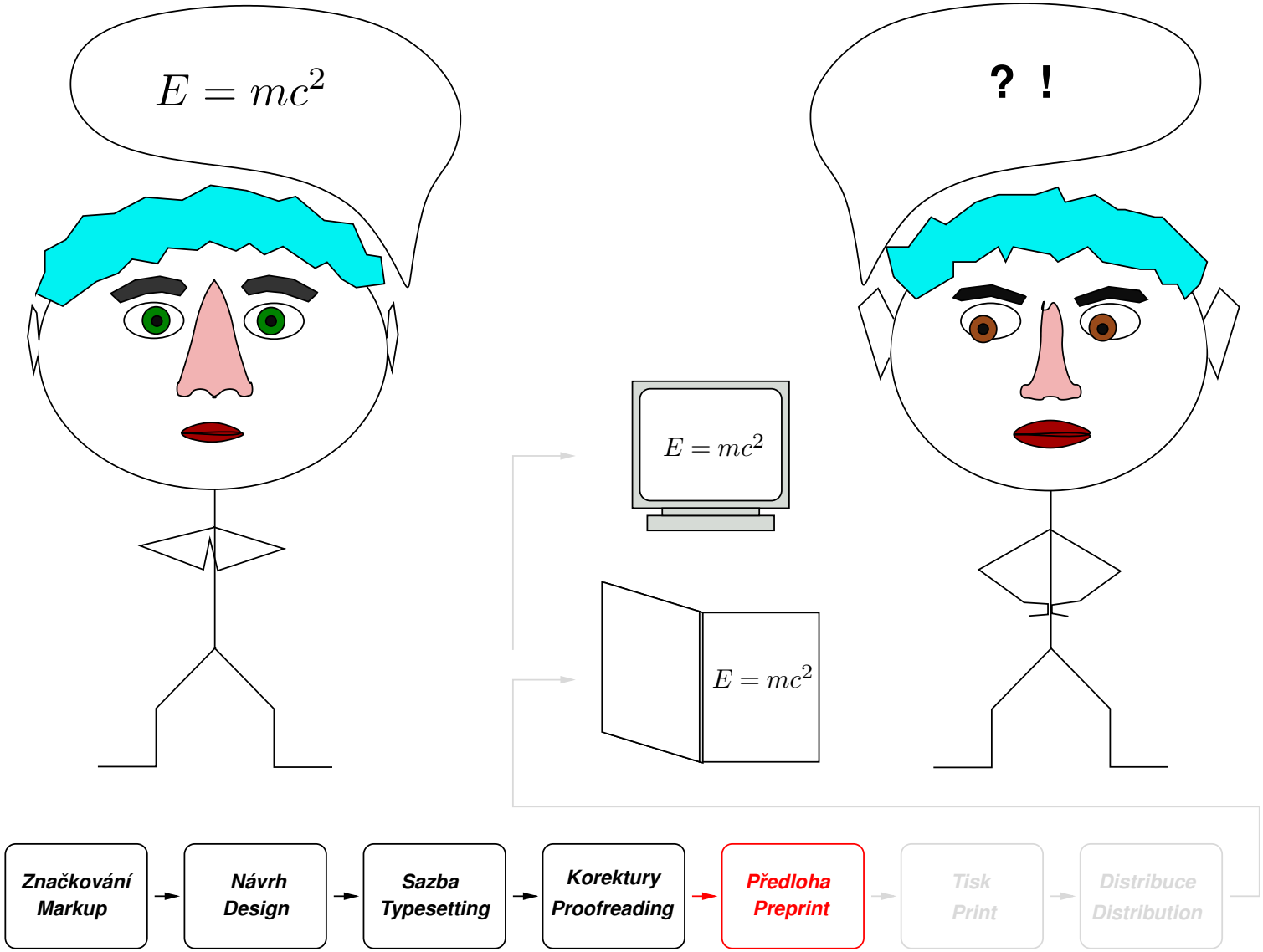
The Past: Scholarly Communication *going digital* from paper



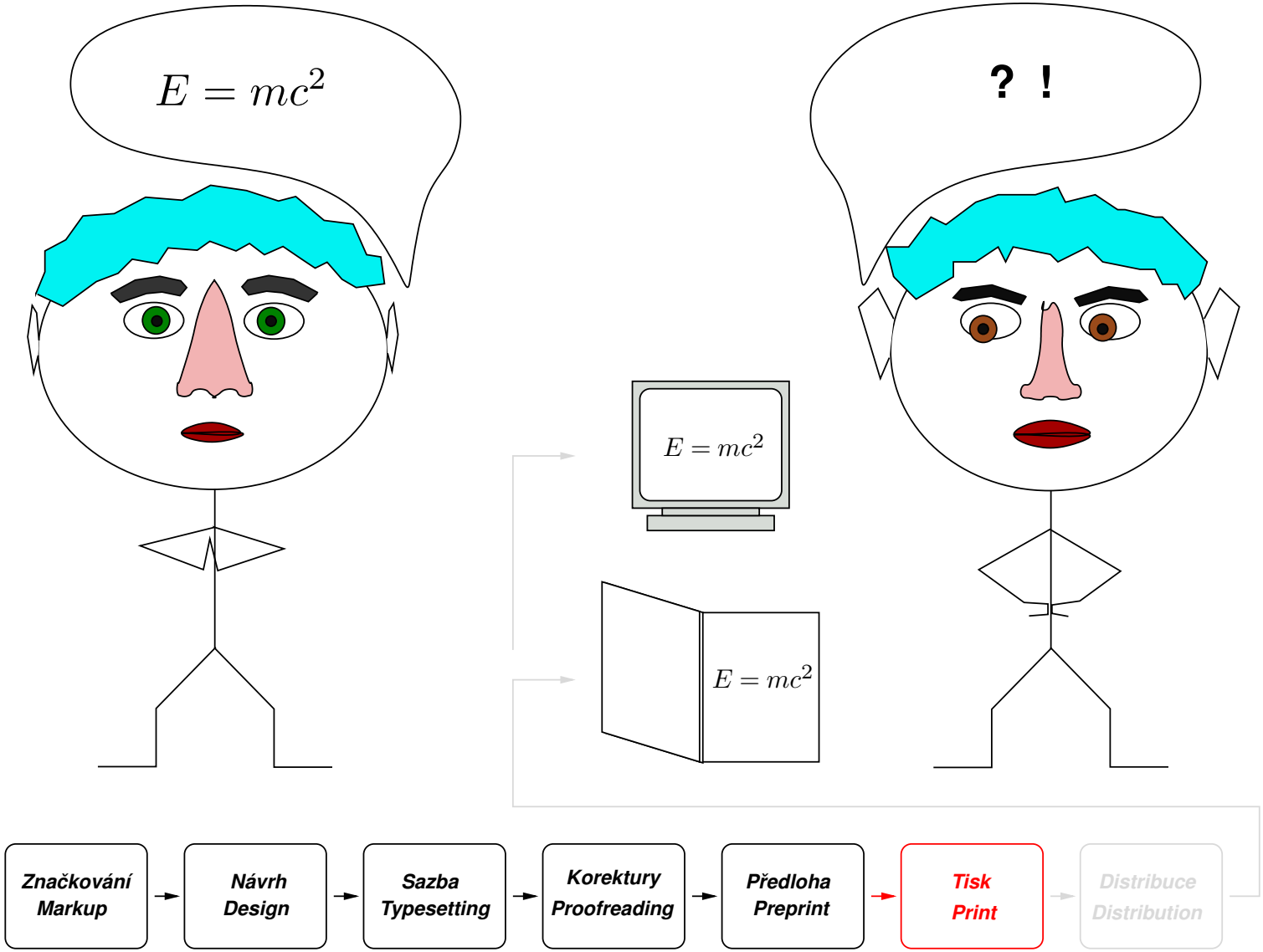
The Past: Scholarly Communication *going digital* from paper



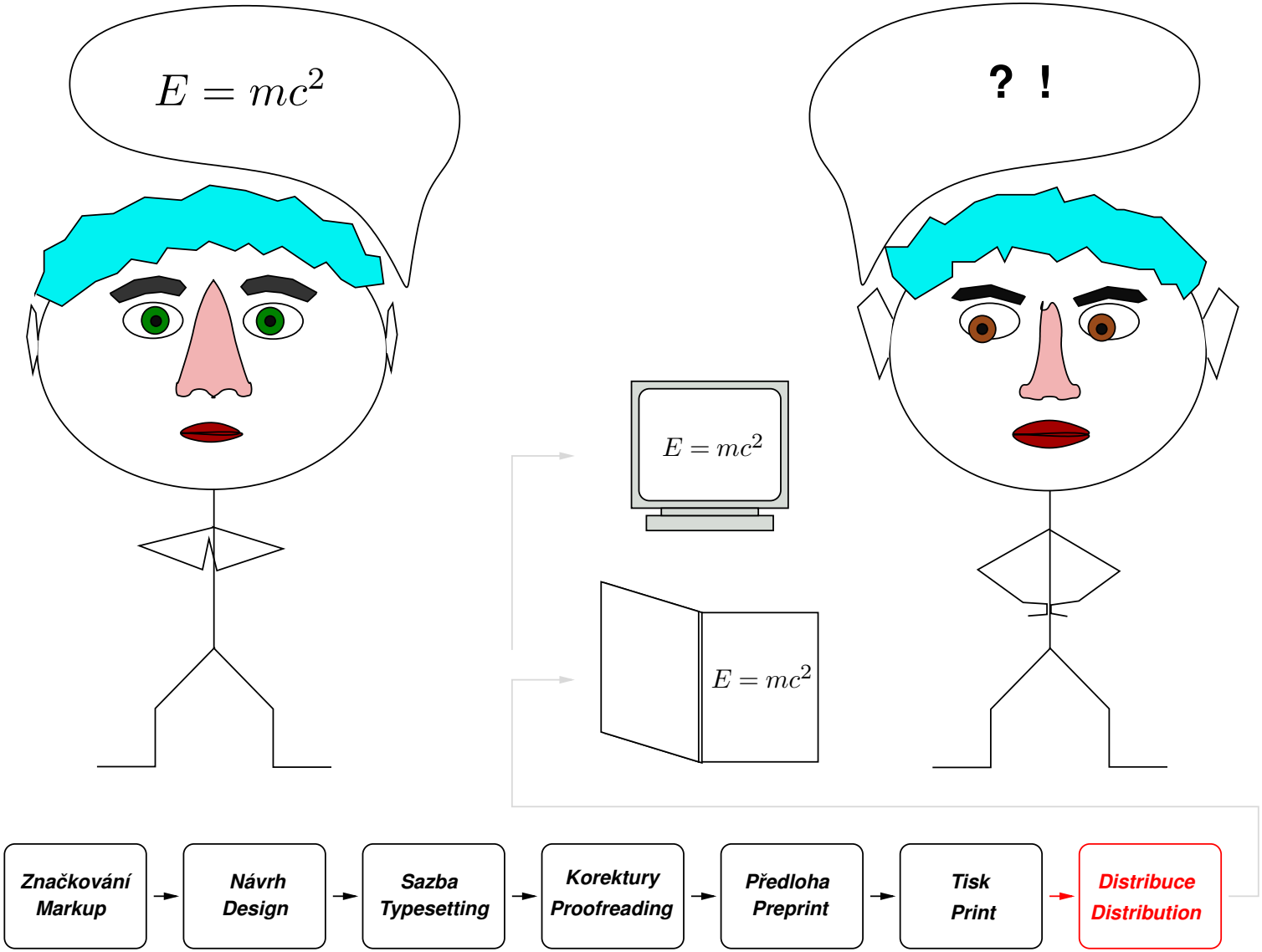
The Past: Scholarly Communication *going digital* from paper



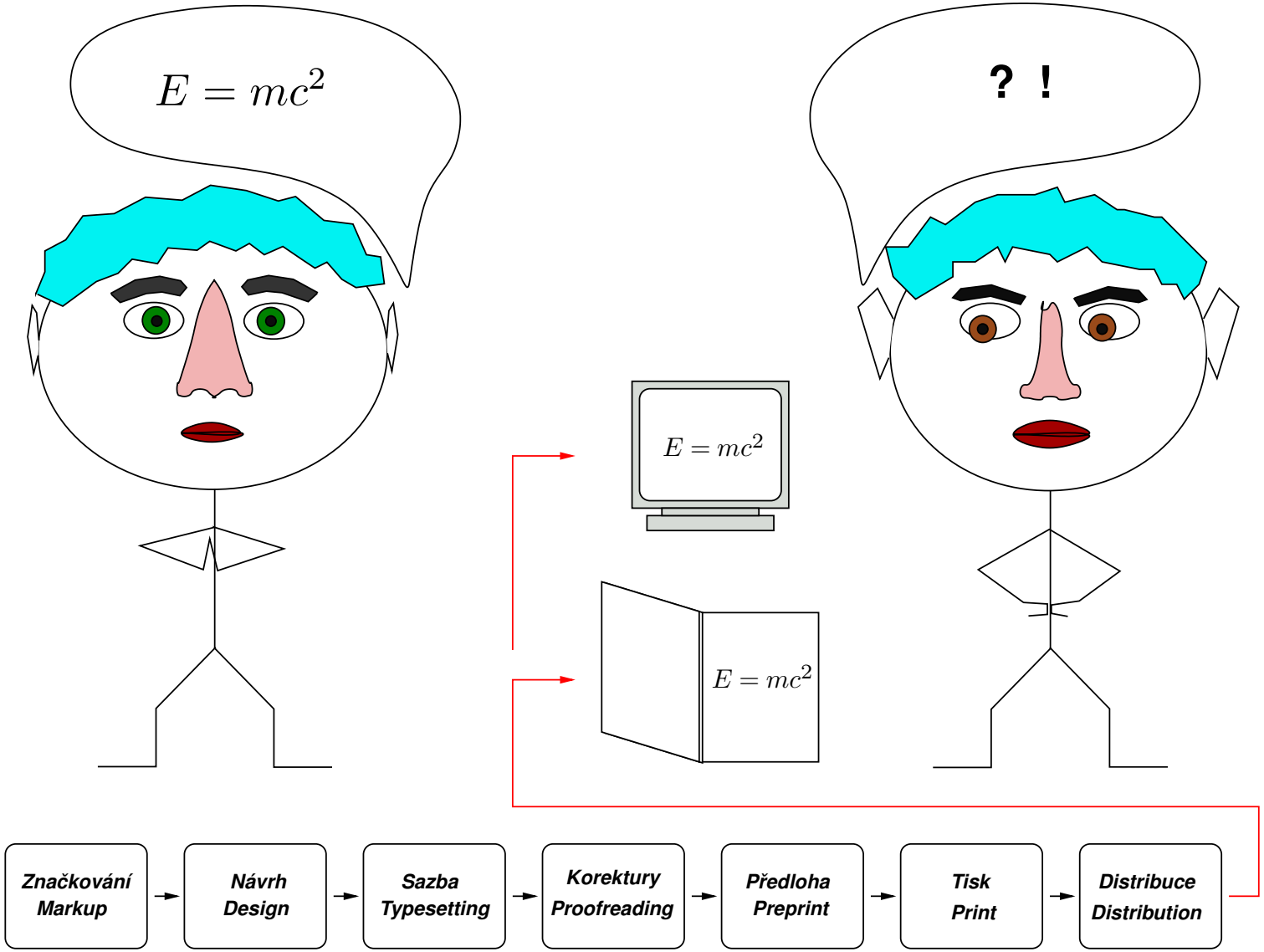
The Past: Scholarly Communication *going digital* from paper



The Past: Scholarly Communication *going digital* from paper



The Past: Scholarly Communication *going digital* from paper



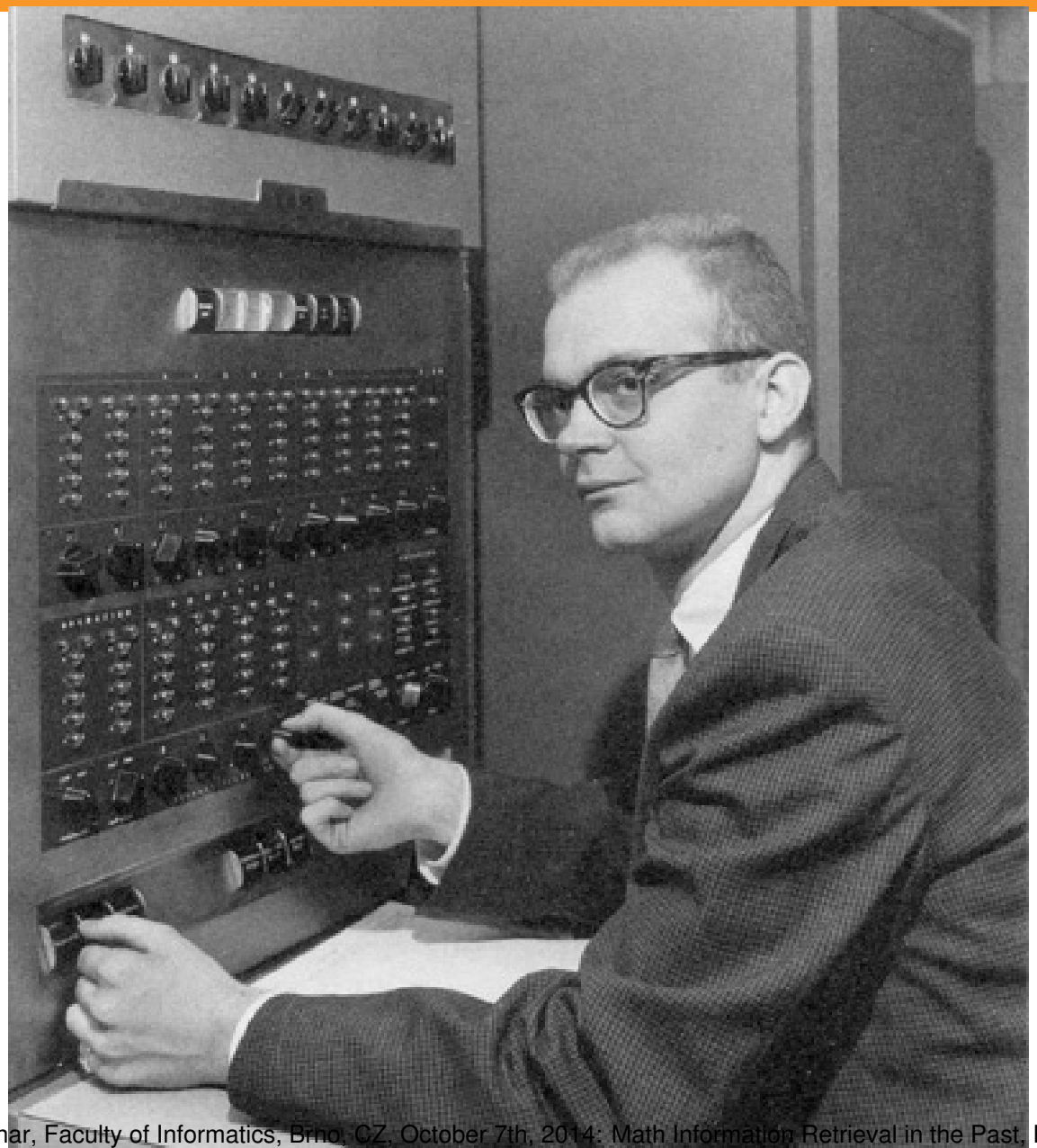
1931, Zentralblatt für Mathematik und ihre Grenzgebiete

- longest running *abstracting and reviewing service* in pure and applied mathematics
- still running, having almost 3,000,000 records of math literature published since 1868, most of them with independent *peer review*
- 7,000 reviewers, 120,000+ new records per year
- commercial service (paid access now)

1940, January: volume 1, number 1 of Mathematical Reviews

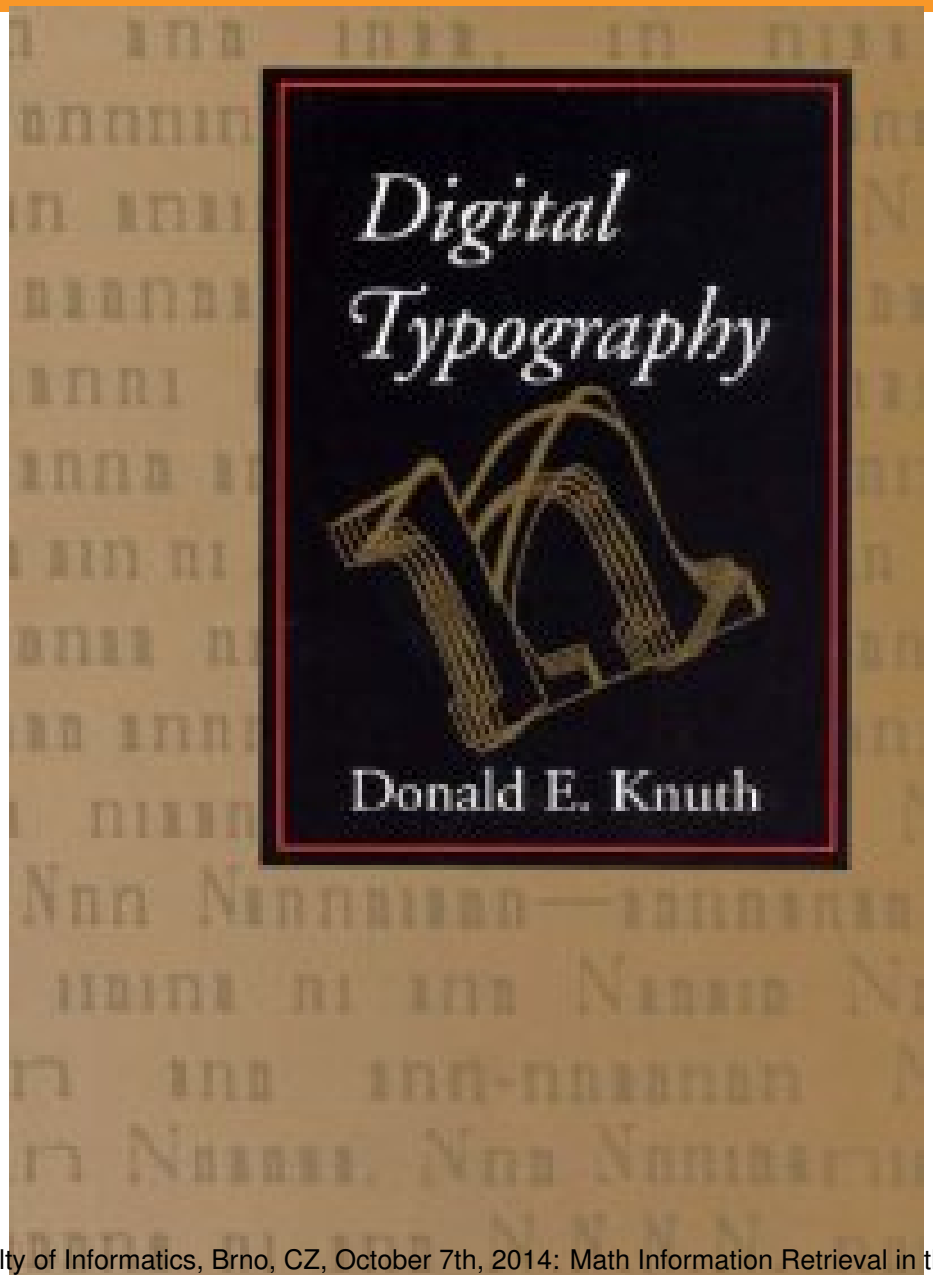
- US competitor to Zentralblatt, also commercial, paid access
- monthly issues of peer reviews of the world's current mathematical literature
- 25fold increase since 1940: 400 $\longrightarrow$ 10,000 reviews monthly
- MSC – Mathematics Subject Classification developed and shared with ZMath: most papers MSC-classified as part of review (or since 1990 by author)

1958: Stanford Computer Science pioneer uses IBM 650

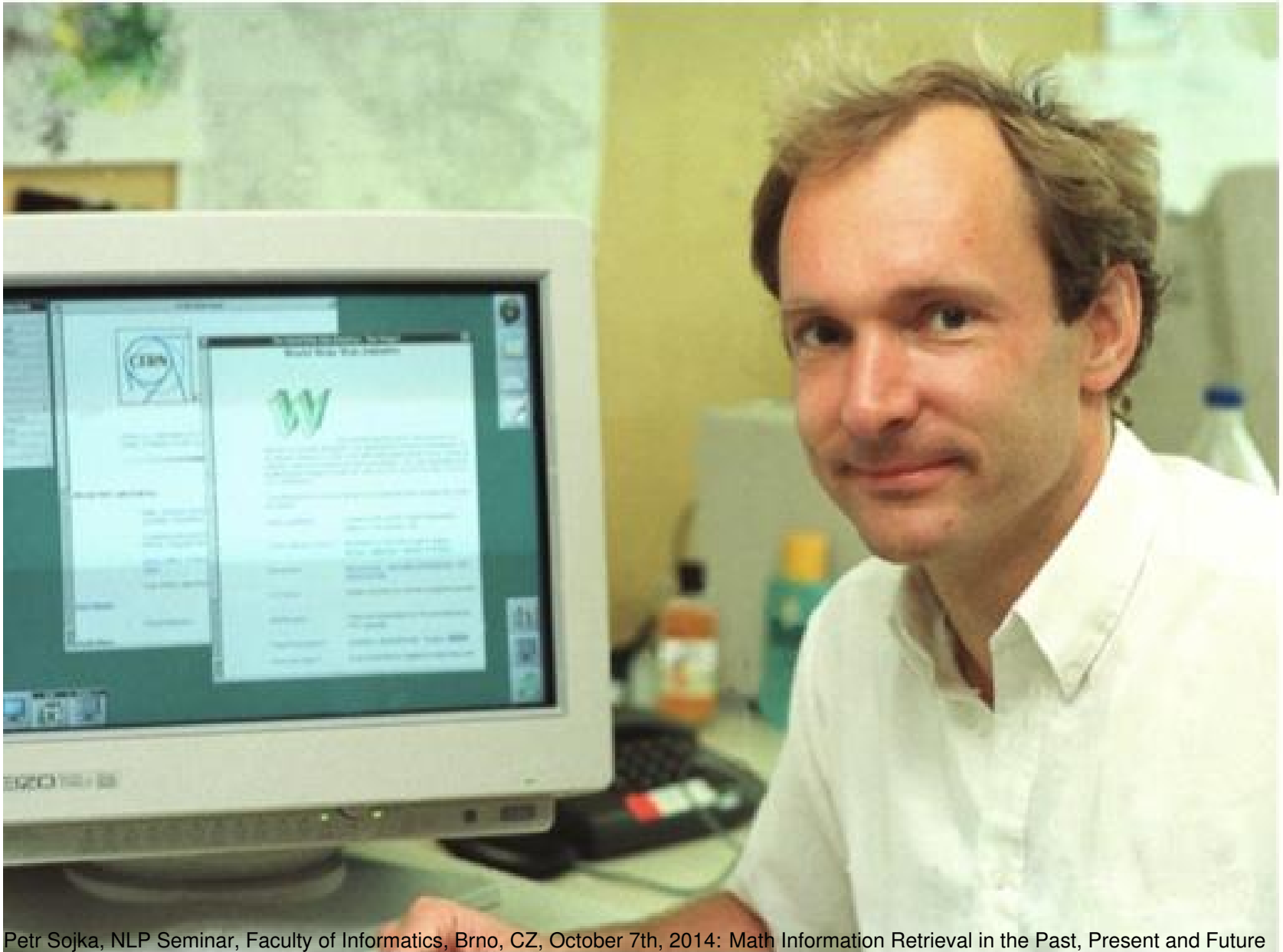


1967–69, Stanford Public Information REtrieval System: SPIRES

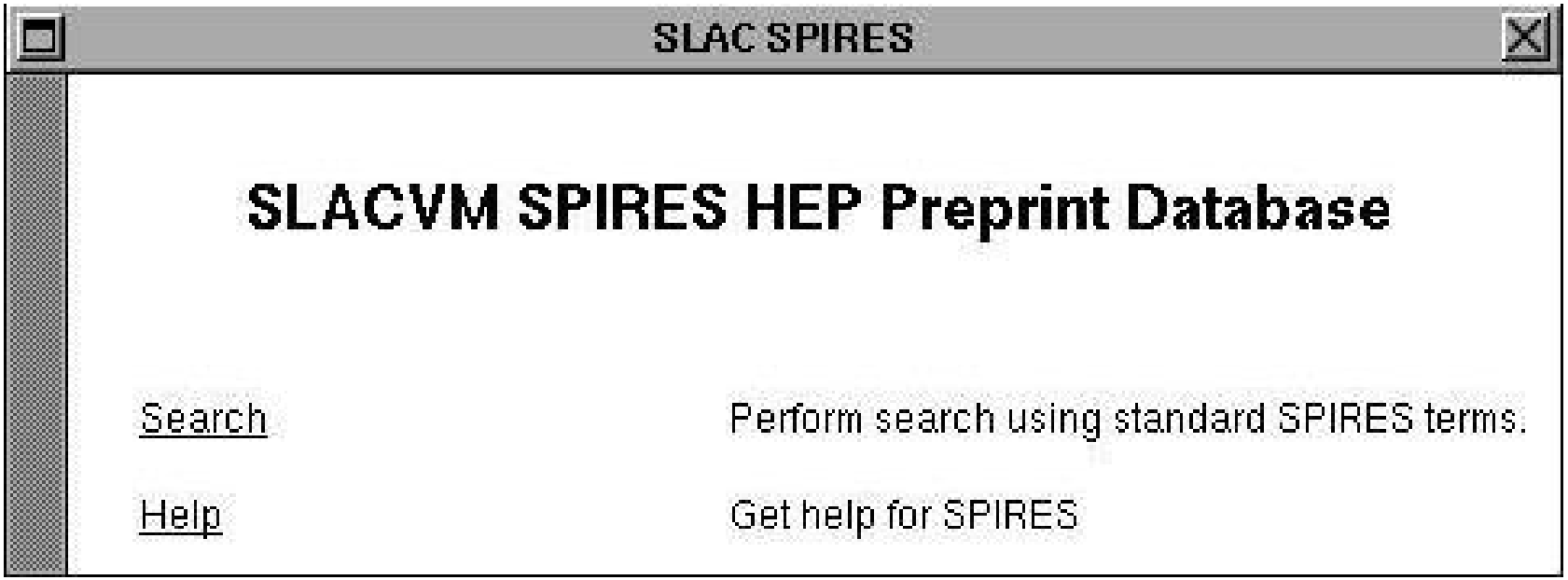
1977, Stanford, typesetting going digital, including math (TEX)



1989 @ CERN: hypertext to link the documents



1991, Stanford, SLAC: first web interface to DL catalogue



- first web site outside Europe
- first database on the web, “first web killer app” (T. Berners Lee)

1991, Los Alamos, preprint archive arXiv

An archive for electronic preprints of scientific papers in the fields of mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance which can be accessed online.

- Paul Ginsparg: speed up of scholarly communication
- submissions in TEX source: smooth switch to hypertext and PDF later
- now (2013) 7,000 submissions per month, almost 1,000,000 fulltexts in total
- hosted by Cornell library, costs ($400,000 per year) shared by the community of users (institutions of users)

1997, Stanford, stanford.google.com, going *global*

1997, “bringing order to the web”: Larry Page’s ranking

- *global* citation analysis: “pages that are well cited from many places around the web are worth looking at”
- random walker/surfer metaphor ($d = .85$):

$$PR(A) = (1 - d) + d \times \left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)} \right)$$

1998, CiteSeer: an automatic citation indexing system (for CS)

```
@inproceedings{Giles:1998:CAC:276675.276685,
 author = {Giles, C. Lee and Bollacker, Kurt D. and Lawrence
 title = {CiteSeer: an automatic citation indexing system},
 booktitle = {Proceedings of the third ACM
         conference on Digital libraries},
 series = {DL '98},
 year = {1998},
 isbn = {0-89791-965-3},
 location = {Pittsburgh, Pennsylvania, USA},
 pages = {89--98},
 url = {http://doi.acm.org/10.1145/276675.276685},
 doi = {10.1145/276675.276685},
 acmid = {276685},
 publisher = {ACM},
 address = {New York, NY, USA}}
```

Automated parsing of paper full texts (from PostScript) using regular expressions; now: DBLP, CiteseerX, ACM-DL,...

2000: History of the dream: vision of WDML

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.

Even other attempts on the European level (FP5, FP6) were not successful.

Publishers and local bodies started massive digitization *themselves*.

2000: History of the dream: vision of WDML

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.

Even other attempts on the European level (FP5, FP6) were not successful.

Publishers and local bodies started massive digitization *themselves*.

2000: History of the dream: vision of WDML

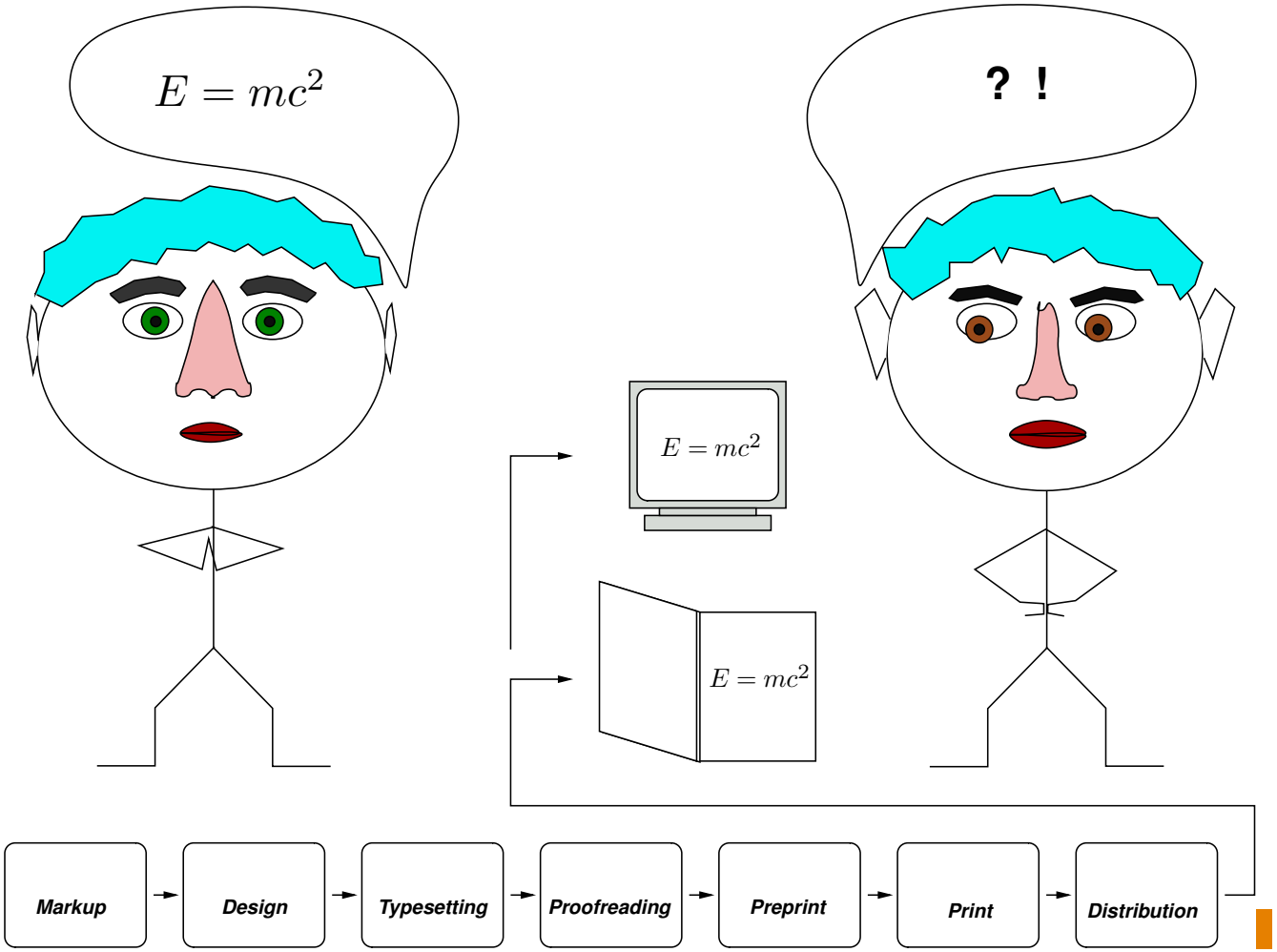
In the beginning was vision of all mathematical knowledge, *peer reviewed, verified* (100,000,000 pages) and engineered into one-stop e-shop/DL.

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.

Even other attempts on the European level (FP5, FP6) were not successful.

Publishers and local bodies started massive digitization *themselves*.

The Present: Scholarly Communication *via Digital Libraries*



Stakeholders in scholarly communication

- authors: provably benefit from open access – distribution optimization
- content mediators: publishers, digital libraries owners – [non-]profit optimization
- readers: benefit from accessibility – discoverability, search and presentation/understanding optimization

All benefit from *digitization*.

2005: The Czech Digital Mathematics Library

Project 1ET200190513 –- funded by the Academy of Sciences of the Czech Republic. Programme “Information Society” (National Research Programme, 2005–-2009), *full* (retro)digitization of 50,000 pages of mathematical literature per year, 8M CZK in total (≈ 1\$ per digitized page).

The goal: to investigate, develop and apply techniques, methods and tools that would allow the creation of a suitable infrastructure and conditions for establishing the Czech Digital Mathematics Library (DML-CZ). The library content: scholarly mathematical literature which has been published throughout history in the Czech lands.

<http://dml.cz>

2005: The Czech Digital Mathematics Library

Project 1ET200190513 –- funded by the Academy of Sciences of the Czech Republic. Programme “Information Society” (National Research Programme, 2005–-2009), *full* (retro)digitization of 50,000 pages of mathematical literature per year, 8M CZK in total (≈ 1\$ per digitized page).

The goal: to investigate, develop and apply techniques, methods and tools that would allow the creation of a suitable infrastructure and conditions for establishing the Czech Digital Mathematics Library (DML-CZ). The library content: scholarly mathematical literature which has been published throughout history in the Czech lands.

<http://dml.cz>

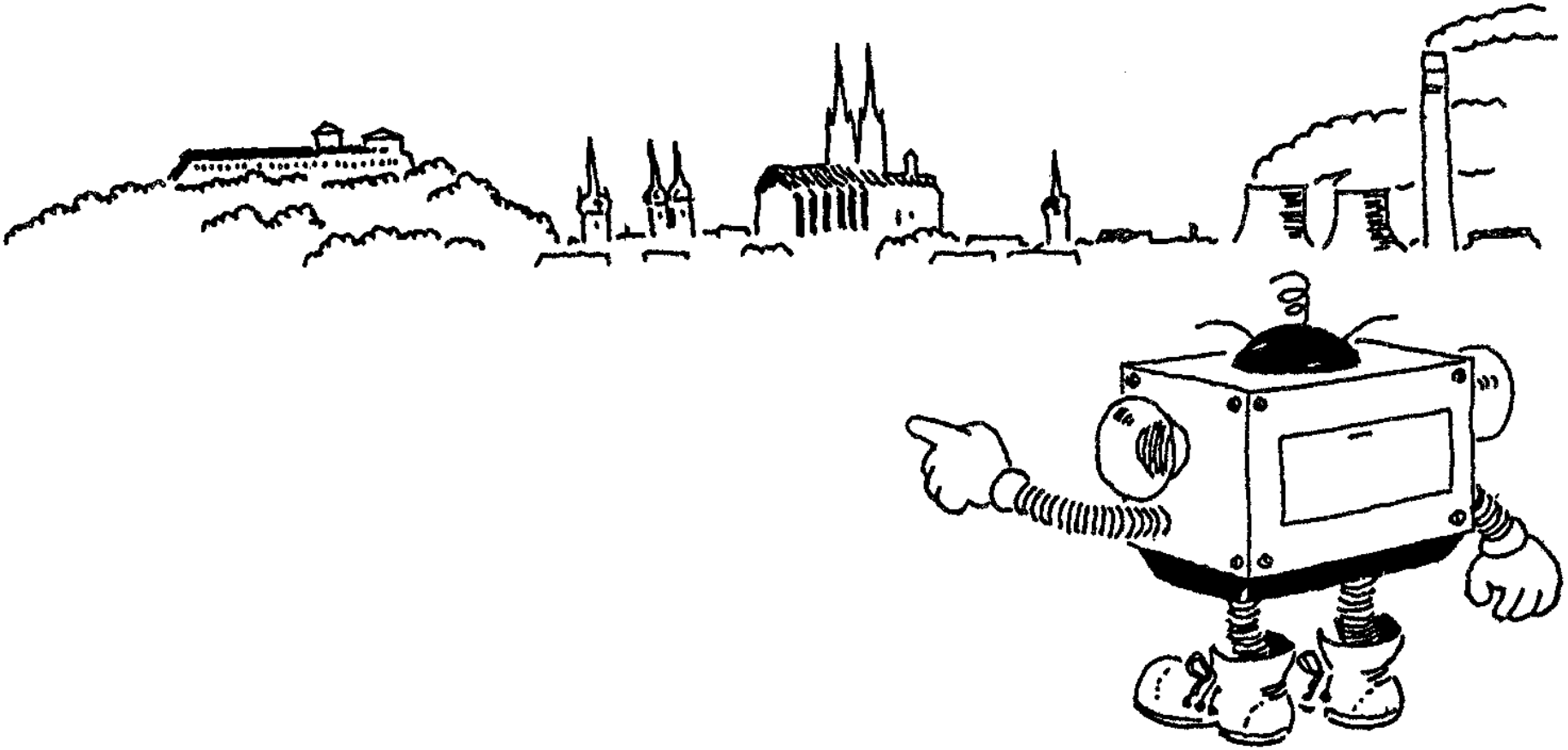
DML-CZ challenges: from pixels to minds

- Research part: **1)** gradual enhancement of the digital material by 'knowledge enhancing' filters on markup-rich XML data. **2)** New methods for (semantic) text processing tested on the available data. **3)** The design of the work-flow aiming at mathematical knowledge stored in digital library.
- IPR part: sharing/delivery (economic models for knowledge sharing due to interests of content owners/publishers).

DML-CZ challenges: from pixels to minds

- Research part: **1)** gradual enhancement of the digital material by ‘knowledge enhancing’ filters on markup-rich XML data. **2)** New methods for (semantic) text processing tested on the available data. **3)** The design of the work-flow aiming at mathematical knowledge stored in digital library.
- IPR part: sharing/delivery (economic models for knowledge sharing due to interests of content owners/publishers).

Bottom up processing—local (Brno, CZ) document engineering



DML-CZ: who?

Four contractors (all from Czech Republic):

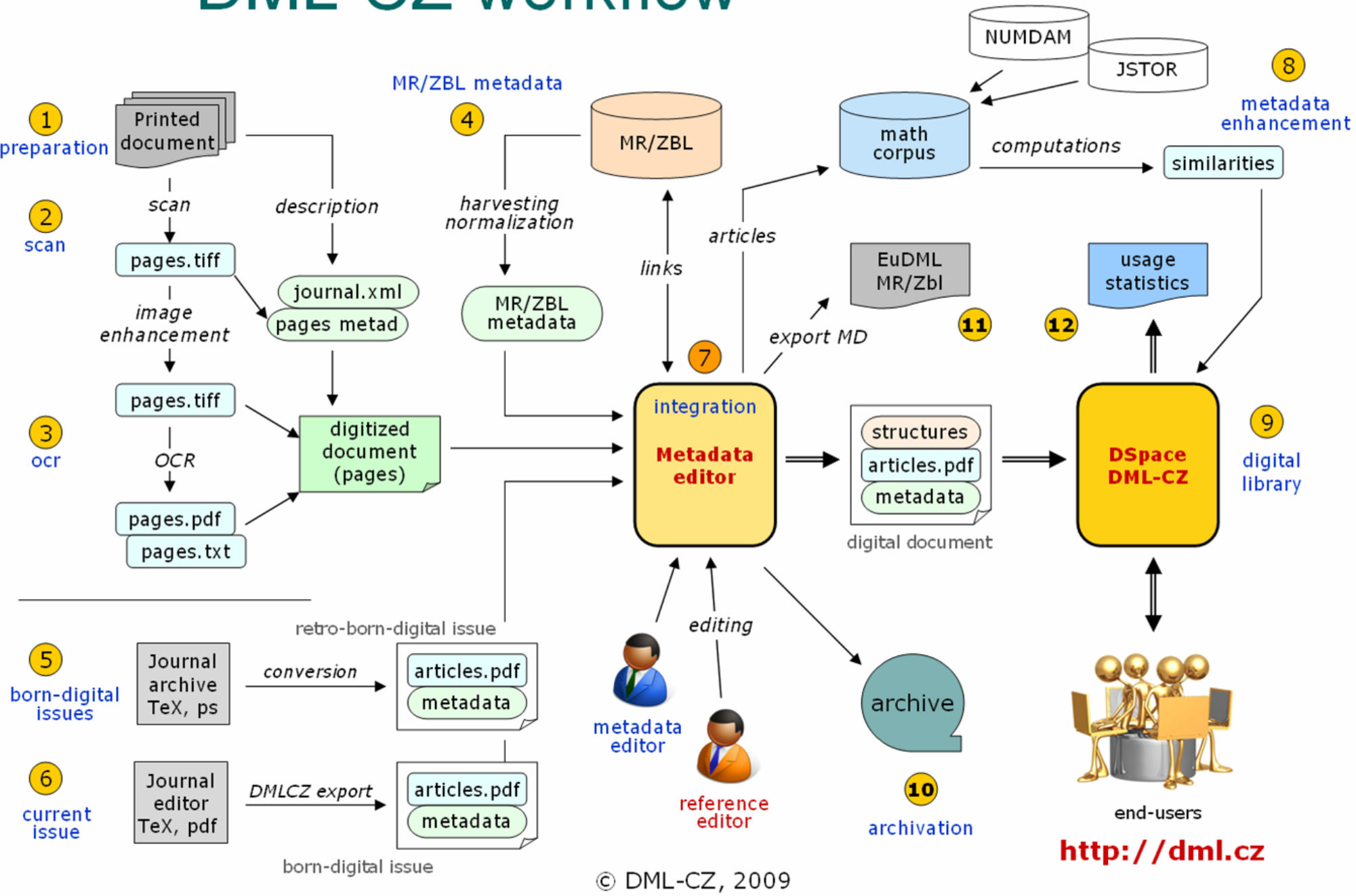
① *Czech Academy of Sciences, Prague* Jiří Rákosník, head of the project, responsibility for material selection, copyright negotiations.

② *Masaryk University, Brno* Petr Sojka (FI) formats and tools, technical coordination, information retrieval, indexing.
Mirek Bartošek (Institute of Computer Science), content management system, metadata Q/A, long-term sustainability.

③ *Charles University, Prague* Jiří Veselý, Oldřich Ulrych, selection and preparation of materials for digitization, metadata cleanup.

④ *Library of Academy of Sciences, Prague* Martin Lhoták, document scanning in Jenštejn.

DML-CZ workflow



Some of the verified and proven technologies (in DML-CZ)

- Scanned image processing and transformations (with BookRestorer) (BP Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (DP Panák, Mudrák, BP Vystrčil)
- Pre-MSC era papers' automated classification by MSC (with Radim Řehůřek)
- gensim framework: similarity article computations aka document clustering (Ph.D. research by Radim Řehůřek)

Similar articles to article

Borůvka, Otakar

Замечания к рецензии М. И. Ельшина моей статьи „О колеблющихся интегралах дифференциальных линейных уравнений 2-го порядка". (Russian). Czechoslovak Mathematical Journal, vol. 6 (1956), issue 3, pp. 431-433

-> Back to article

Method LSI	Method RP	Method TFIDF
An example of the tran...	An example of the tran...	An example of the tran...
Сообщения. Член-коррес...	Сообщения. Член-коррес...	Сообщения. Член-коррес...
О колеблющихся	О колеблющихся	О колеблющихся интегра

DML-CZ challenges and lessons learned

DML-CZ, the Czech Digital Mathematics Library, now serves more than *300,000 pages of more than 34,000 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
- semantic similarity metrics developments (Radim Řehůřek's Ph.D.)

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

For more, see (who, what, browse, browse similar, how to search).

Nature 454, 263 (2008) | doi:10.1038/454263b

Starting small but adding up: a free maths archive

A small group of researchers is meeting in Birmingham, UK, later this month to plan a free digital library of mathematics.

All the mathematical literature ever published runs to more han 50 million pages, with around 75,000 articles added each year. Over the past decade there have been several attempts to make this prodigious body of work accessible in a single digital archive, but so far none has succeeded.

A group of mathematicians intends to change this. They have started small, with a handful of digitization projects in Poland, Russia, Serbia and the Czech Republic. In a few years they hope to unite these repositories with their western European counterparts in an archive to be hosted by the European Union, according to the organizer, Petr Sojka, an informatics scientist at Masaryk University in Brno in the Czech Republic. Eventually this pan-European archive could be expanded globally, he says.

To make such an archive easier to search, researchers have found ways to guess the subject of a paper on the basis of the frequency of symbols in it. But there will be many more-practical challenges, such as finding the funds to scan millions of old papers and striking deals with publishers who hold rights to them.

It may already be too late to build a single free mathematical archive, according to John Ewing, head of the American Mathematical Society, which maintains a list of more than 1,500 journals whose archives have already been digitized. "A few years ago, this model had the potential to change the mathematics journal literature in profound ways," he says. But most publishers have rushed to scan their own archives in order to lock them up and sell them to libraries.

"While the effort to digitize the smaller collections is admirable, and it's certainly worthwhile, it's unlikely to effect a larger change," says Ewing. ■

Jascha Hoffman

Workshop series *Towards a Digital Mathematics Library* founded to tackle numerous challenges identified during DML-CZ project.

DML workshop series archived in DML-CZ

dml.cz/handle/10338.dmlcz/702563

Czech Digital Mathematics Library

About DML-CZ | FAQ | News | Conditions of Use | Math Archives | Contact Us

Search

Go

Advanced Search

Browse

- Collections
- Titles
- Authors
- MSC

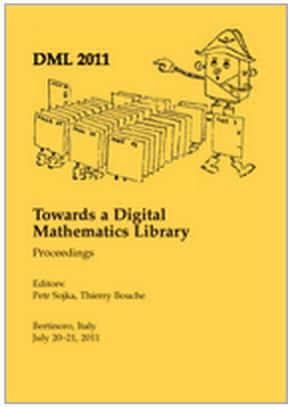
About DML-CZ

Partner of

EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY

DML-CZ Home >
DML >

DML



Description

Mathematicians dream of a digital archive containing all peer-reviewed mathematical literature ever published, properly linked and validated/verified. The objectives of DML workshops were to formulate the strategy and goals of a global mathematical digital library and to summarize the current successes and failures of ongoing technologies and related projects.

Archive:

DML 2008: **Proceedings of the 1st workshop, Birmingham, 2008**
DML 2009: **Proceedings of the 2nd workshop, Grand Bend, 2009**
DML 2010: **Proceedings of the 3rd workshop, Paris, 2010**
DML 2011: **Proceedings of the 4th workshop, Bertinoro, 2011**

Browse by Titles or Authors

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 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.

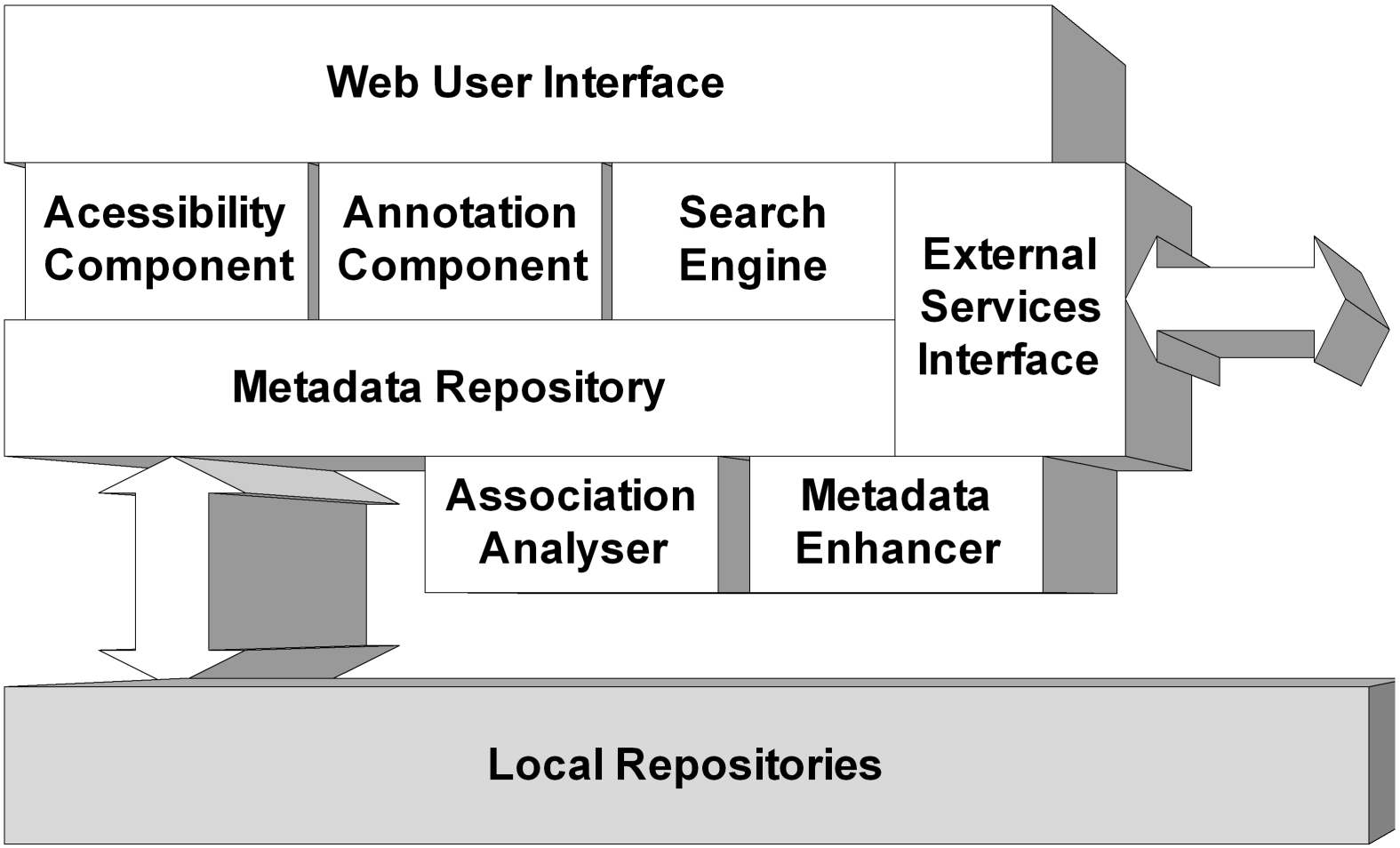
Vision of European Digital Mathematics Library

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 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 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: EuDML

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).

<http://eudml.org>

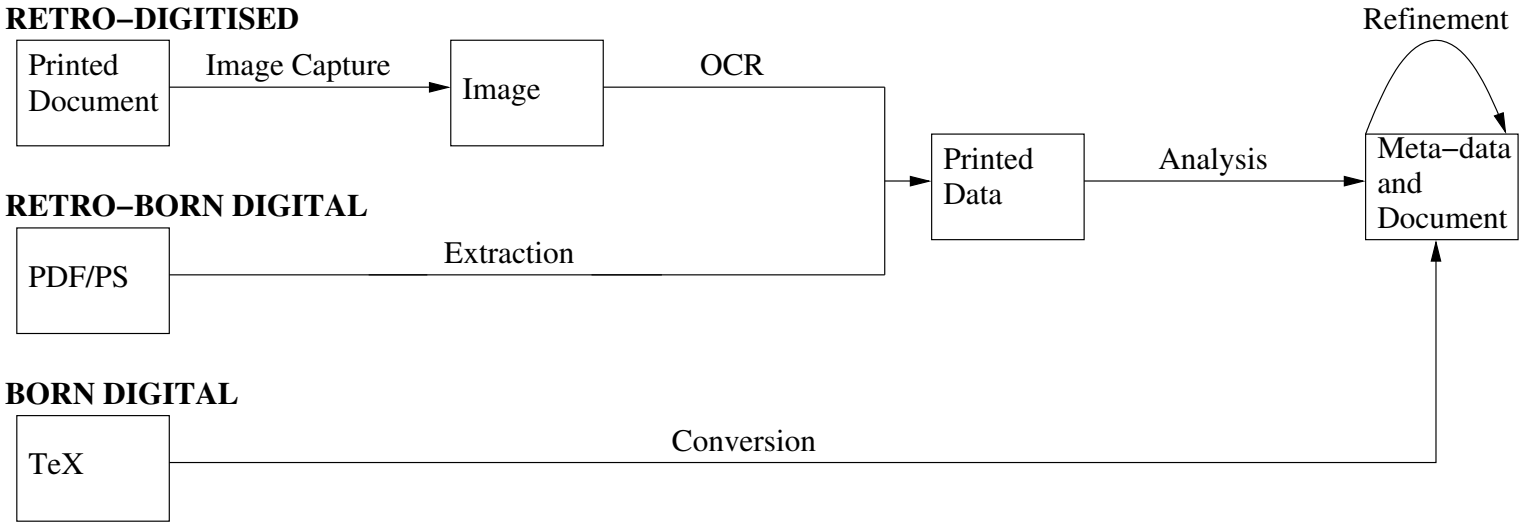
DML processing challenges: document accessibility

Conversions (inversion of authoring+typesetting) needed from:

born-digital period: typesetting by TEX with export of [meta]data into digital library: maxTract otherwise on PDFBox (plain text)

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader + InftyReader), otherwise on Tesseract (no math)

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution: finally Tesseract



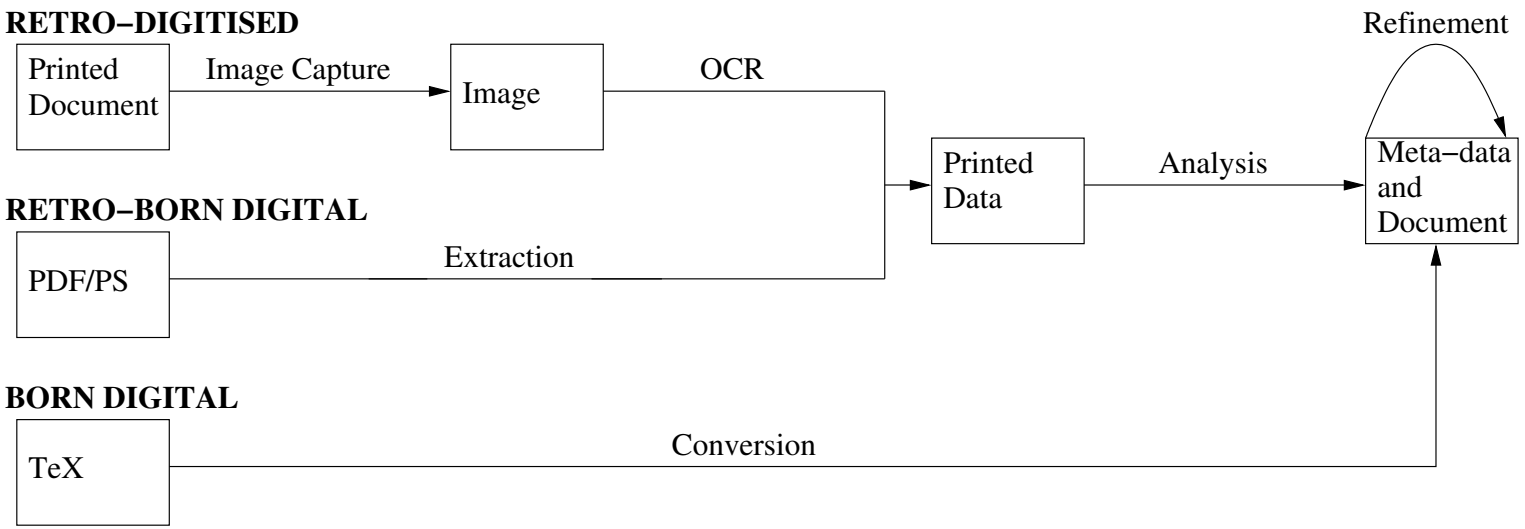
DML processing challenges: document accessibility

Conversions (inversion of authoring+typesetting) needed from:

born-digital period: typesetting by TEX with export of [meta]data into digital library: maxTract otherwise on PDFBox (plain text)

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader + InftyReader), otherwise on Tesseract (no math)

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution: finally Tesseract



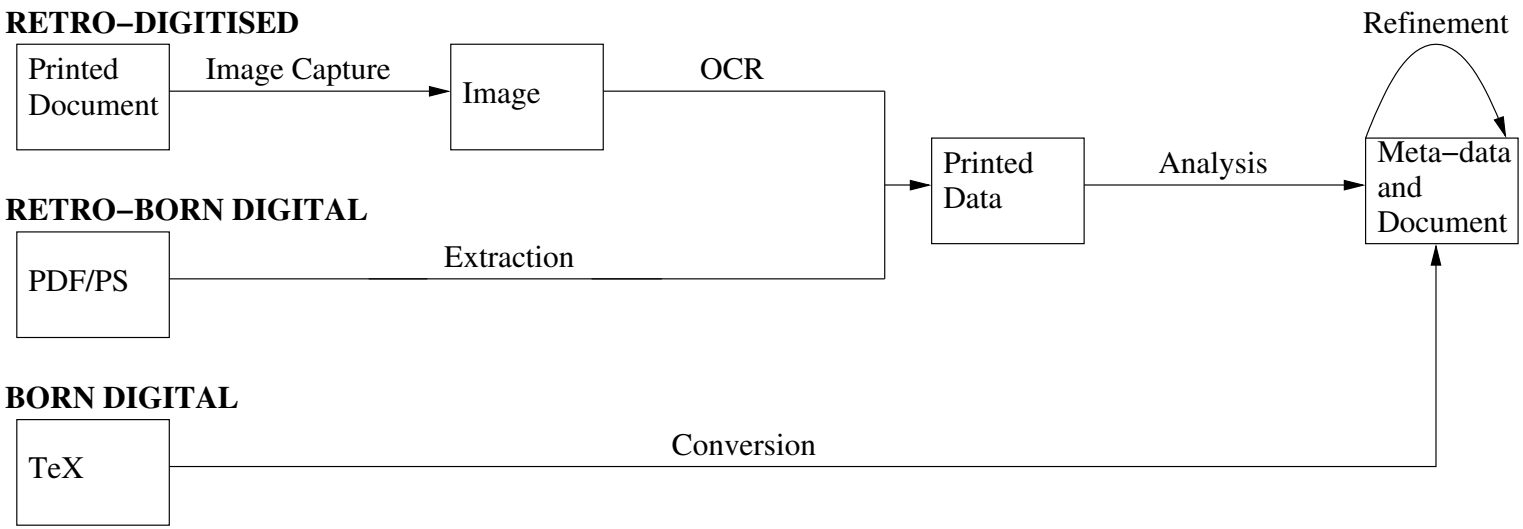
DML processing challenges: document accessibility

Conversions (inversion of authoring+typesetting) needed from:

born-digital period: typesetting by TEX with export of [meta]data into digital library: maxTract otherwise on PDFBox (plain text)

retro-digital period: scanning, geometrical transformations (BookRestorer), OCR (FineReader + InftyReader), otherwise on Tesseract (no math)

retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution: finally Tesseract



Getting fulltexts with math from bitmaps

Infty Reader from Fukuoka: working with prof. Suzuki to improve further (automation, support for Russian, LaTeX driver,...).

Automated only, no time (and money) to fix OCR errors.

Run in parallel in Brno, Grenoble and Lisbon to speed up. Almost 200K papers (more than 1M pages).

MathML output used for [internal] indexing and similarity computations only, not for metadata or export.

Getting fulltexts with math from PDF

Born-digital PDF: maxTract developments 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>
```

Adding accessibility

Use of linear grammars by Anderson (1968)

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 (Chromevox plugin) and/or Adobe Acrobat Reader (as multiple-layer PDFs, [no tagged PDFs yet]).

Content Similarity Results in EuDML: <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.

EuDML | The EUROPEAN DIGITAL MATHEMATICS LIBRARY

English (en) Jane Doe | Log Out

Title, Author, Keyword, Citation, Date... Search

Home Advanced Search Browse by Subject Browse by Journals Refs Lookup

Displaying similar documents to "On oscillation criteria for third order nonlinear delay differential equations"

On the solution of the differential equation $f\left(x, y, y^{(1)}, \ldots, y^{(n)}\right) = 0.$

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

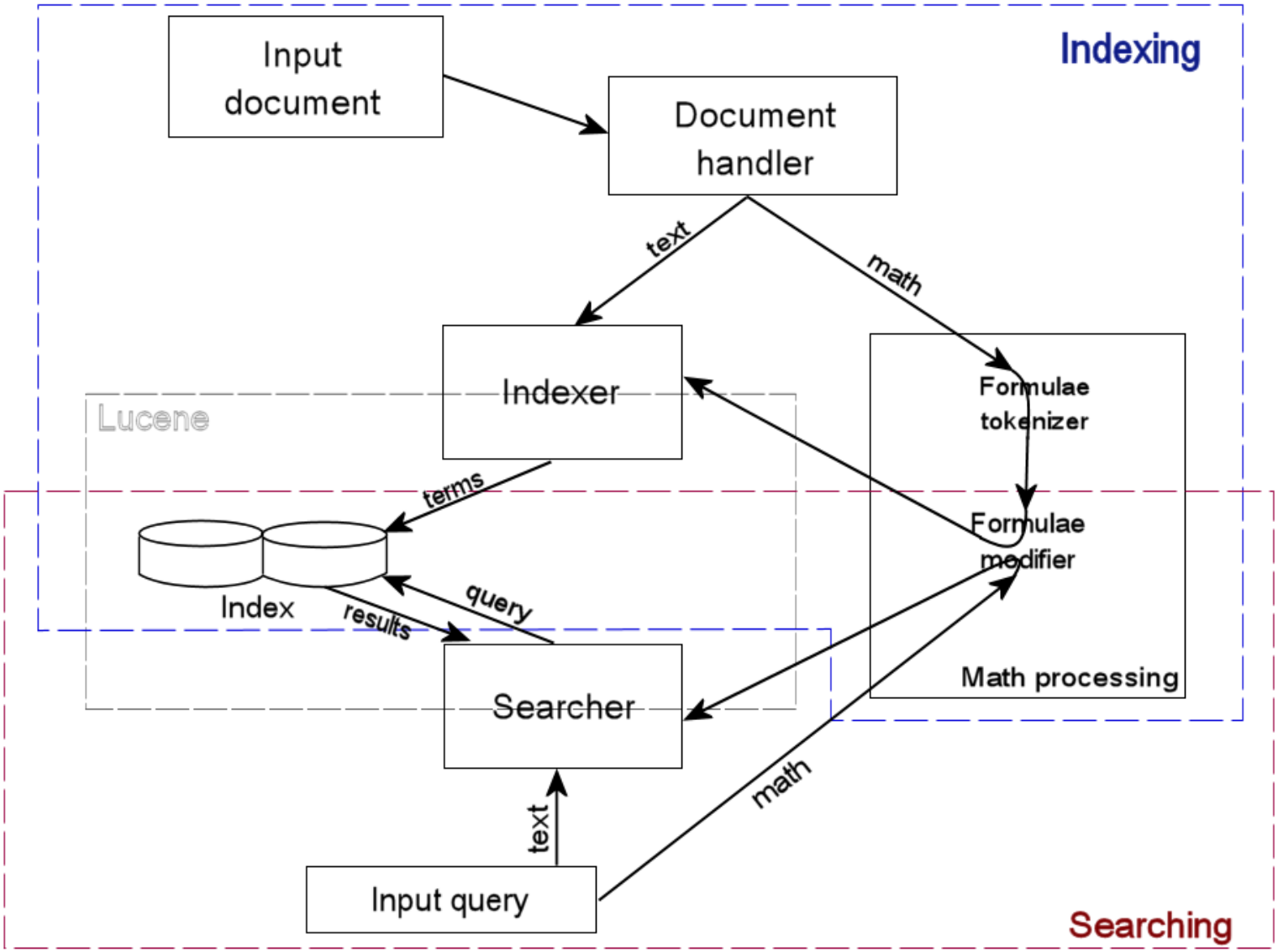
Math aware Search and Indexing

- Usual way of seeking information via [Google] search
- Conventional searching approaches are not applicable for math
- Usage of existing mathematical search engines (MathDex, EgoMath, LaTeXSearch, LeActiveMath, MathWebSearch) problematic
- new Math Indexer and Searcher (MIaS) developed at MU

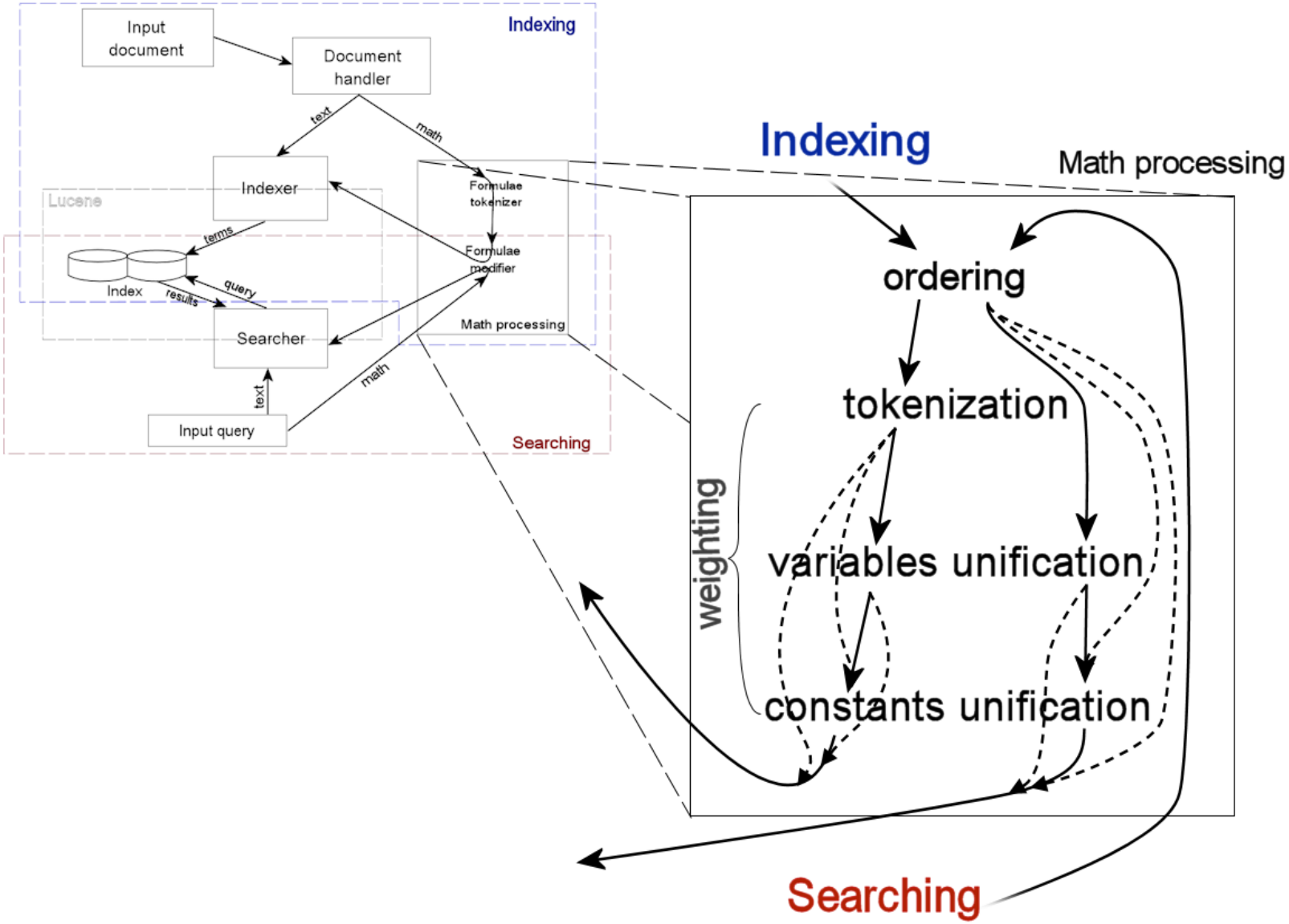
Math Indexer and Searcher — Features

- Inspired mostly by MathDex and EgoMath
- Based on full text core Apache Lucene
- Presentation MathML
- Allows similarity (not only exact match) between query and matched term
    - Commutativity
    - Unification of variables and constants
    - Subformulae matching
- Level of similarity calculation for expressions
- Mixed mathematical-textual queries

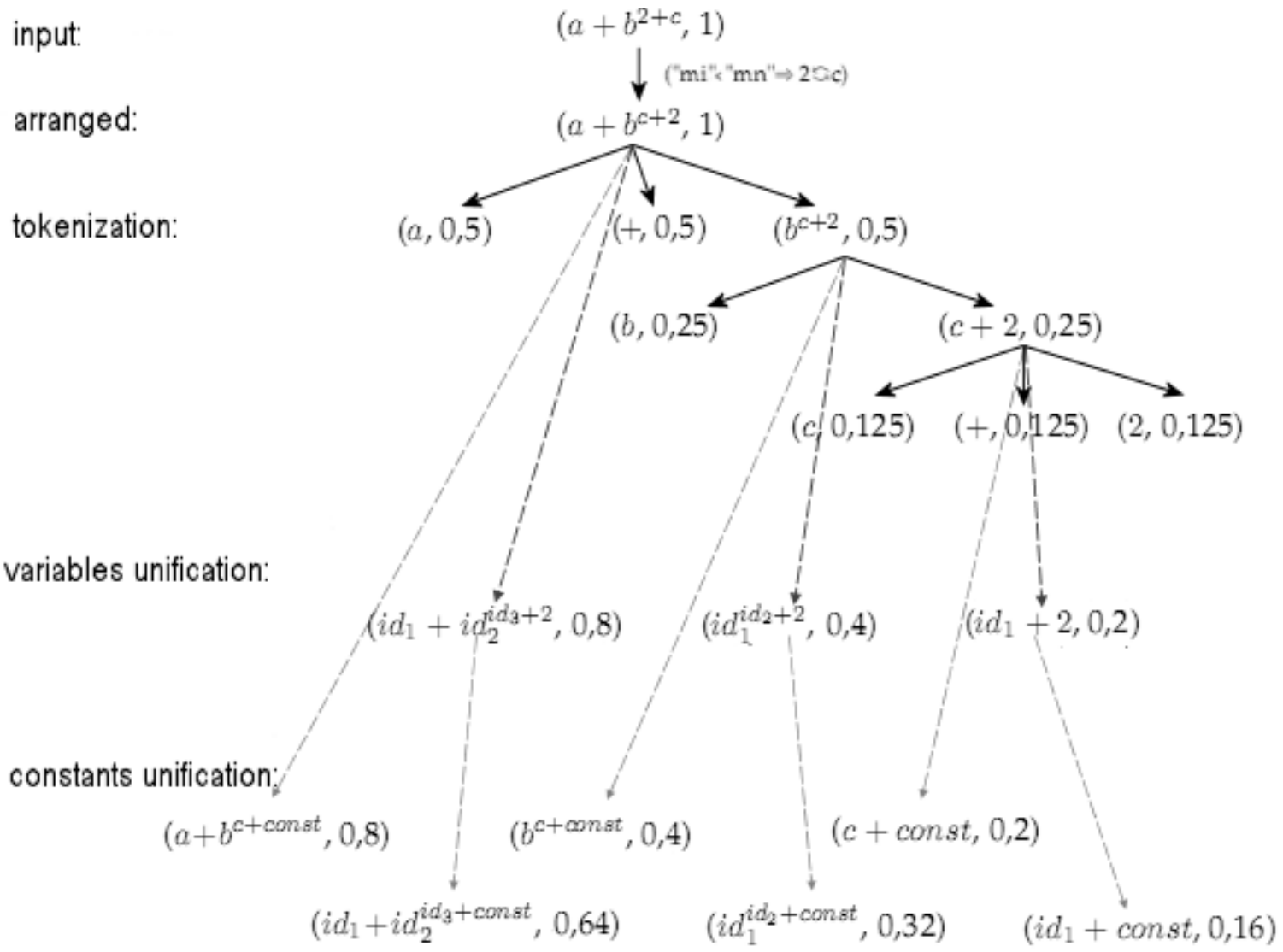
Math Indexer and Searcher — Design



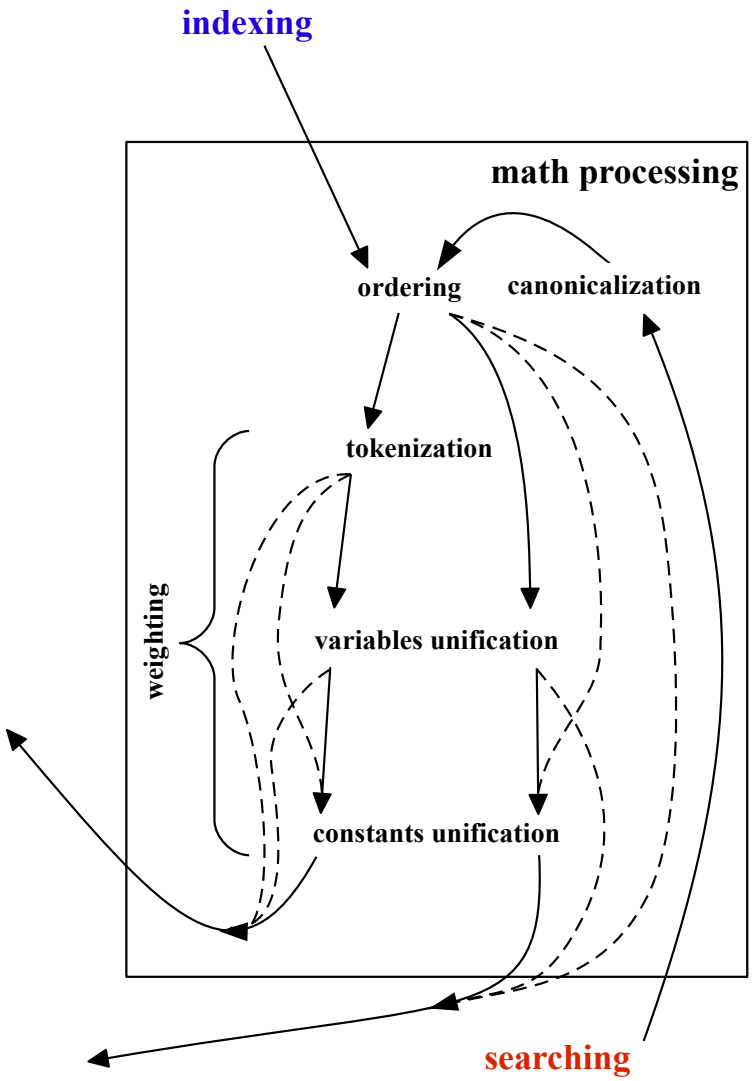
Math Indexer and Searcher — Design II



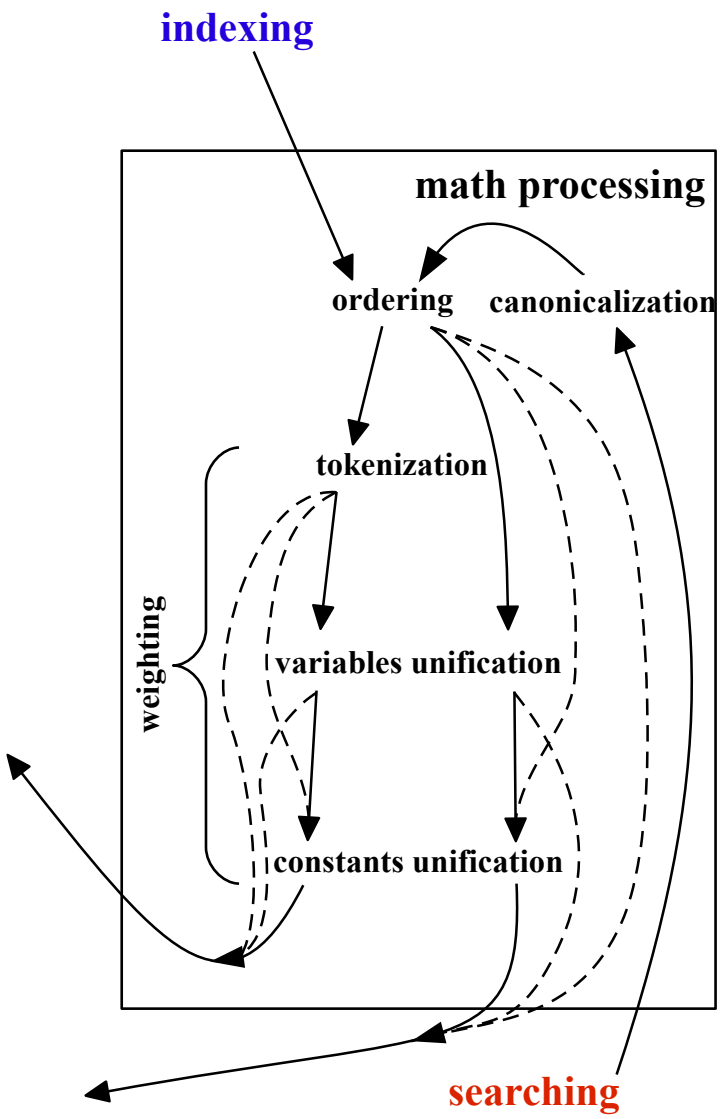
Formula Processing Weighting Example

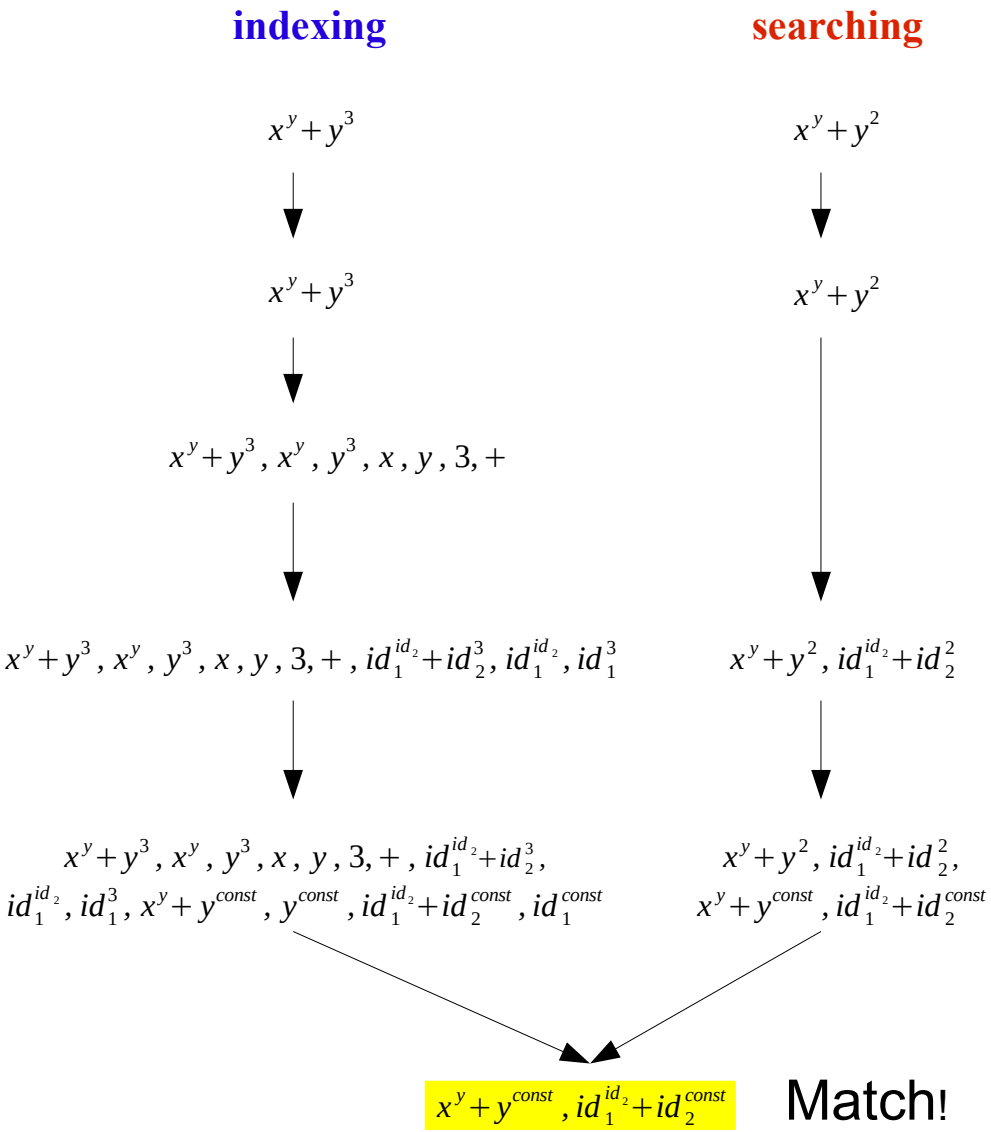


Math formulae indexing processing



Example


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


indexing
searching
x^y+y^3
x^y+y^2
x^y+y^3
x^y+y^2
$x^y+y^3, x^y, y^3, x, y, 3, +$
$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3$
$x^y+y^2, id_1^{id_2}+id_2^2$
$x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3,$
$id_1^{id_2}, id_1^3, x^y+y^{const}, y^{const}, id_1^{id_2}+id_2^{const}, id_1^{const}$
$x^y+y^2, id_1^{id_2}+id_2^2,$
$x^y+y^{const}, id_1^{id_2}+id_2^{const}$
$x^y+y^{const}, id_1^{id_2}+id_2^{const}$
Match!

Implementation

- Java
- Solr + Lucene
- scalable (10^9 formulae indexed from arXiv)
- jTidy for text extraction
- Mathematical part implements Lucene's interface Tokenizer — able to integrate to any Solr/Lucene based system as DSpace, many web pages…

Search demonstration

Help About

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

How to write query

```
<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: 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

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 = 1.6912912

Formulae search demonstration comments

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

- MathML/TEX input (Tralics [2] for conversion to MathML [9])
- 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: <http://eudml.org/search/>

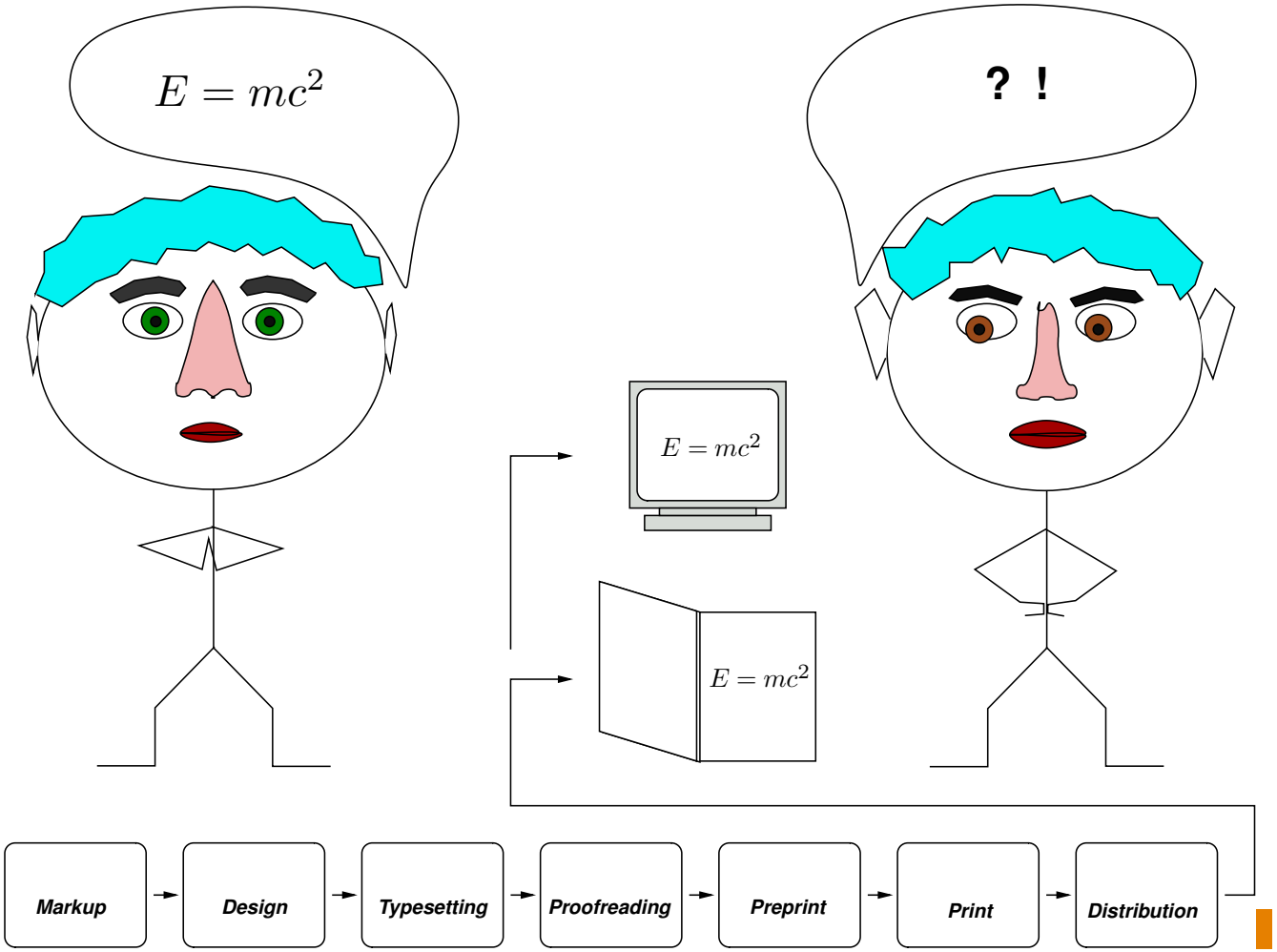
Open Access in DMLs

- controversial issue, huge misunderstandings, domain specific (PubMed Central example)
- publication quality is orthogonal to openess: PLOS example
- example of Impact factor boost
- serial crisis: Elsevier's margin 40% (e.g. billions of $s)
- library bundles, consortia
- controlled profit in OA vs. uncontrolled profit in commercial closed access?
- moving wall policy towards open access to DL heritage
- is Open Access future of DMLs?

Global DML

- EuDML, ‘success’ on European level, sustainability problems
- Sloan funded preparation project for WDML/GDML
- working group paving the way to GDML
- will we became the part of the history?

The Future: Scholarly Communication *via DMLs using rich KB*



Towards higher level content representations – knowledge bases

NLP processing from strings via words to meaning, including *math-awareness* math specifics: structures and abstractions

- to allow searching (semantically) similar papers, precise [semantic] indexing: search as a gate to knowledge
- to allow exploration of a DML by intelligent browsing of (semantically) similar papers: distributional semantics topic modeling as Latent Semantic Indexing, Latent Dirichlet Allocation
- to allow personalization and domain specifics, e.g. semantic faceted search (formulae,…)
- to track ‘train of thought’ – narrative qualities of papers, proofs (Mizar type of paper)
- finally have even math “knowledge at your fingertips”

Motivation for example I

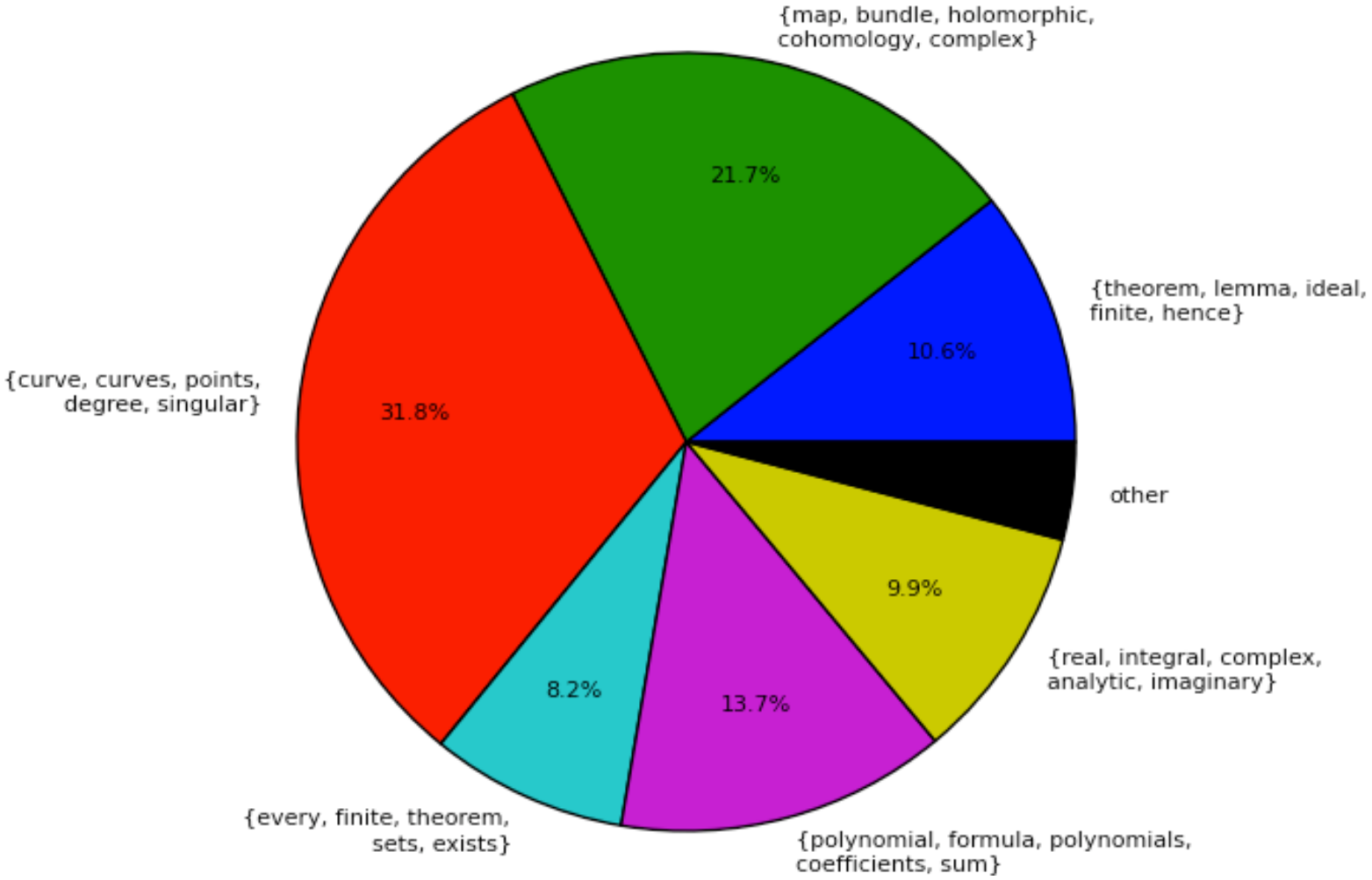
```
From: Shayan A Tabrizi <shayantabrizi@gmail.com>
Subject: [Corpora-List] Dataset for Different Research Areas

I want to find the relevance of each of the research papers of
my dataset to each of the research areas such as Physics, CS,
Math, Social Sciences, etc.
Thus, I need a dataset consisting of all research areas and
some sample texts (preferably papers) in that area, to estimate
the similarity of each of my papers to each of the areas.
              *Is there any such dataset?*
Some points:
   1. It is much much better if the dataset has areas in different
   granularities. e.g. in one level: Mathematics, Physics, CS, etc.
   and in a more fine-grained level divides CS to Networks,
   Artificial Intelligence, etc.
   2. Even if the dataset only consists of a specific domain
   (especially CS) and its sub-domains it is still usable.
```

Example I: Automated Meaning Picking from Texts

LDA Topics Pie Chart for math.0406240:

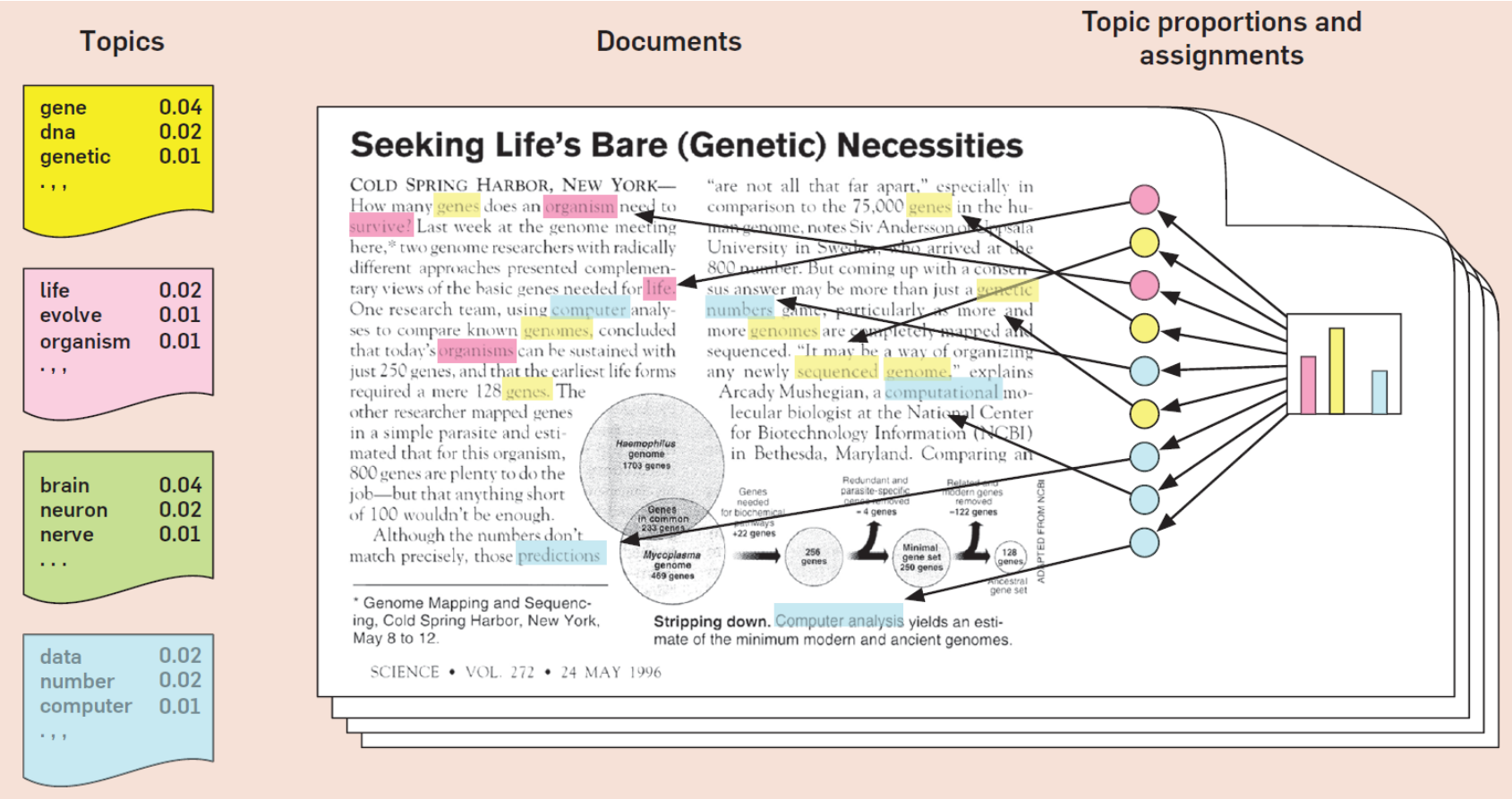
Each slice represents a different topic. The size of the slice corresponds to "how much is the article about this topic?". Topics which contribute <6% to the above document are aggregated under "other".

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



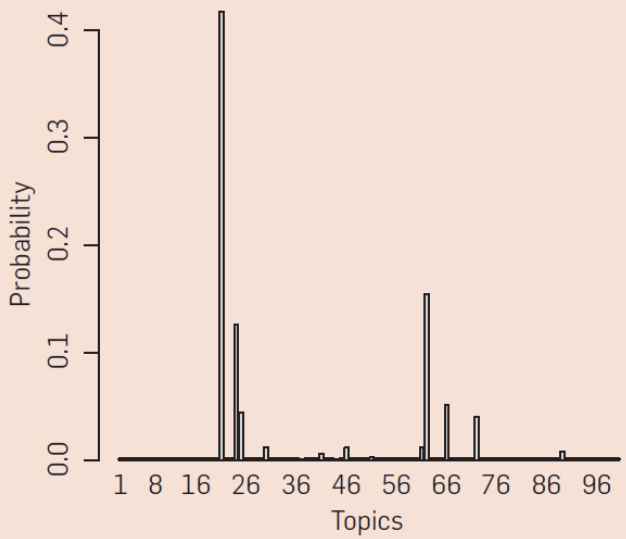
Probabilistic Topical Modeling: Latent Dirichlet Allocation

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



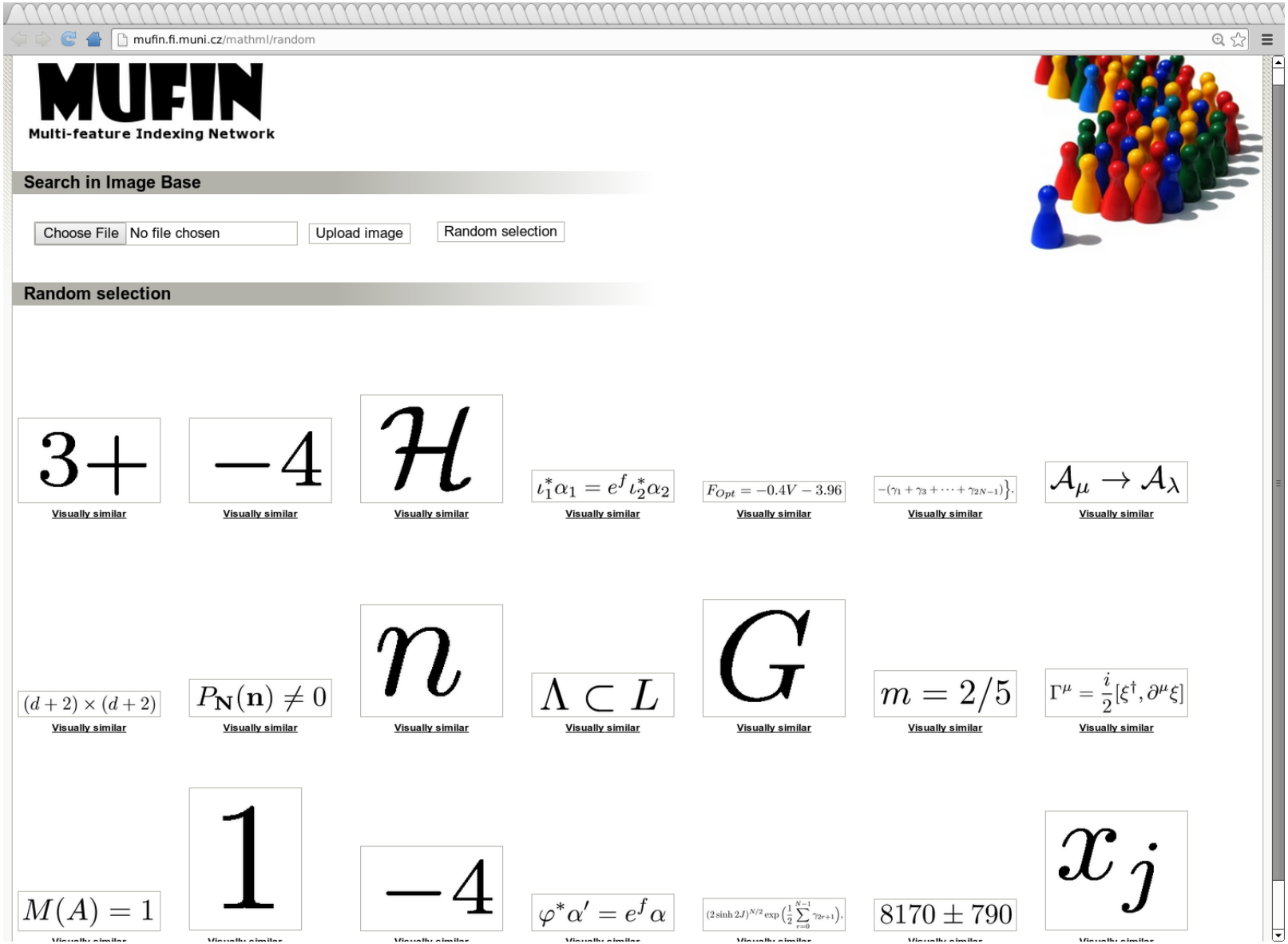
Topical Modeling: Latent Dirichlet Allocation II

- 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

Example II: metric-based indexing of math (+text)



Example III: text 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>
```

Future of DMLs?

- full text mining in semantic direction (typesetting^{-1}), higher level NLP
- open access in the long term
- author's direct publishing (arXiv, Perelman), peer review later?
- increase of automation and precision on [multiple] ways from author's head to the reader's one
- globalization (Google Scholar), automated suggestions
- personalization (up to the individual's preferences)

Future challenges

- Math-aware knowledge representation
- Math entailment (Partha Pakray), ‘flexiformat’ processing, ‘canonicalization’ of math formulae
- Math-aware corpora processing
- robust Math OCR is necessary
- robust born-digital PDF2Math conversion is needed as well
- only then challenges as: multilingual math retrieval, MathML indexing and search, math common sense, text and math disambiguation and understanding, mathematical document classification, document similarity could be possible

Challenge of MKM

- Math-aware knowledge representation: handling abstractions?
- math2vec? ‘smooth’ representation
- Canonicalization of math formulae processing

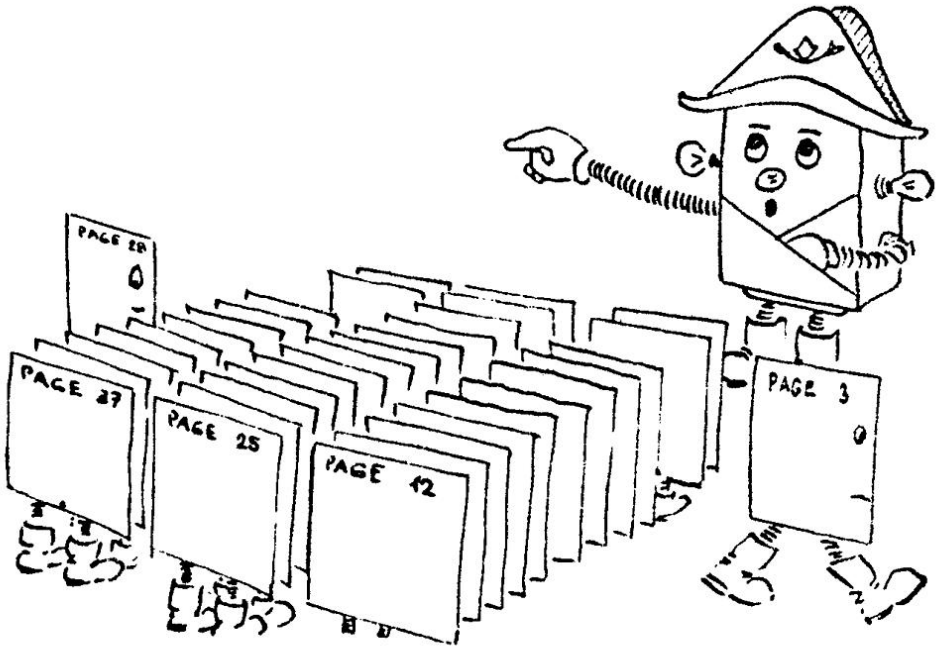
Challenge of math-aware corpora processing and tools

- Switching between different levels of structured data
- math2vec?
- Canonicalization of math formulae processing
- tools adaptation (handling trees and abstractions), ideally without supervision

Challenge of Evaluation of Math Information Retrieval

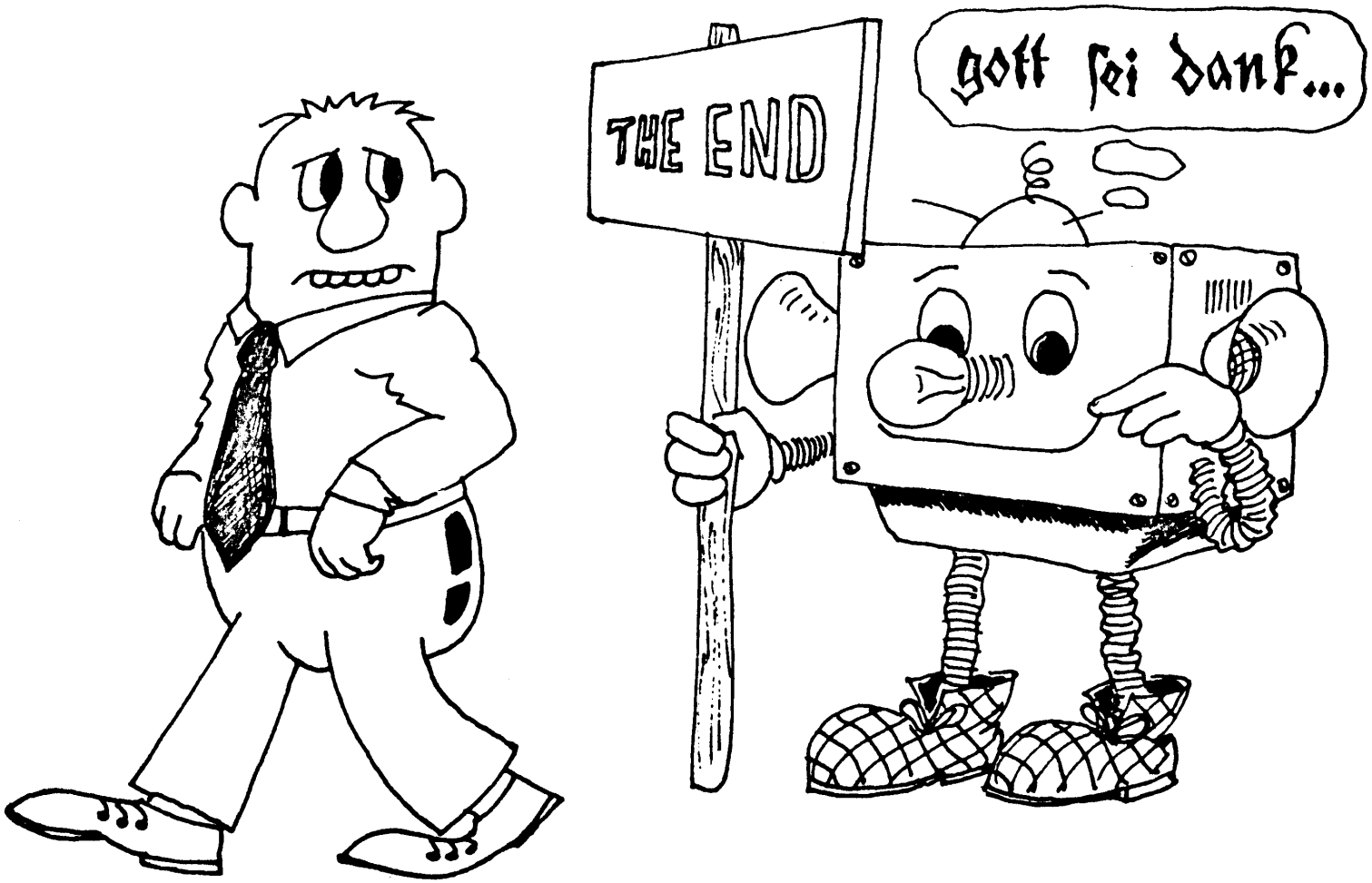
- What works in math-aware IR, UI, pragmatics
- first MIR happening in 2012, Math Tasks at NTCIR-10, NTCIR-11
- MIaS, our engine ‘won’ NTCIR-11 in Tokyo, December 2014

Acknowledgments and questions?



Acknowledgements: EuDML and DML-CZ projects (funding), EuDML and DML-CZ colleagues, Martin Líška, *Michal Růžička*, Radim Řehůřek, Radim Hatlapatka, Martin Jarmar, Maroš Kucbel, Zuzana Nevěřilová, Mirek Bartošek, Martin Šárfy, Vlastík Krejčíř, Petr Kovář, Vlastimil Dohnal, and many, many other authors and contributors of tools used.

That's it!



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 Task. 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>

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>.

Credits for LDA pictures goes to David M. Blei.

Credits for illustrations goes to Jiří Franek.