Accessibility Issues in a Digital Mathematical Library

Examples and Experience of DML-CZ and EuDML

Petr Sojka

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

Universal Learning Design 2011, Brno, Czech Republic

February 10th, 2011, 9:50AM

FACULTAS ARTIS INFORMATICAE • UNIVERSITAS MASARYKIANA •

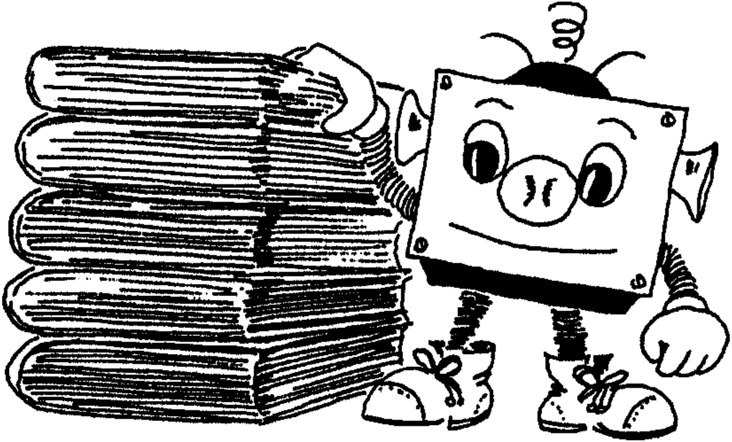
EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

DML-CZ

Outline and two take-home messages

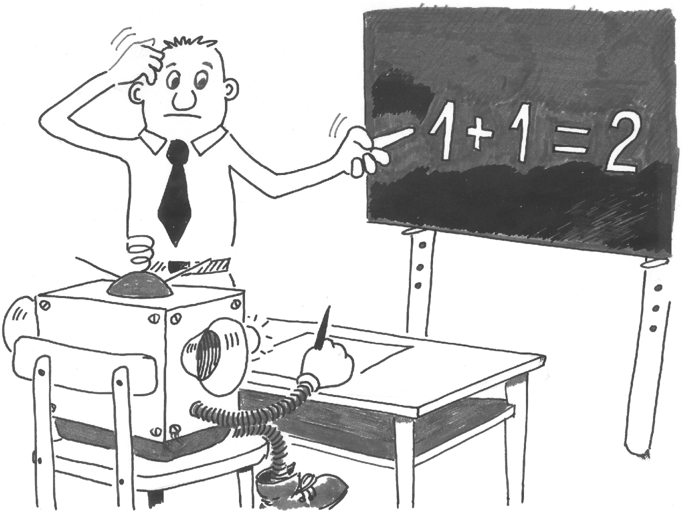
1. Pictorial overview
2. Motivation, vision of PubMed Central for Mathematics
3. Complexity of digitization workflow of The Czech Digital Mathematics Library DML-CZ
4. Math Indexer and Searcher
5. Accessibility improving technologies and tools for DML-CZ and EuDML
6. Tools developed
7. Summary, conclusions and future work

Accessibility issues in *digital* libraries

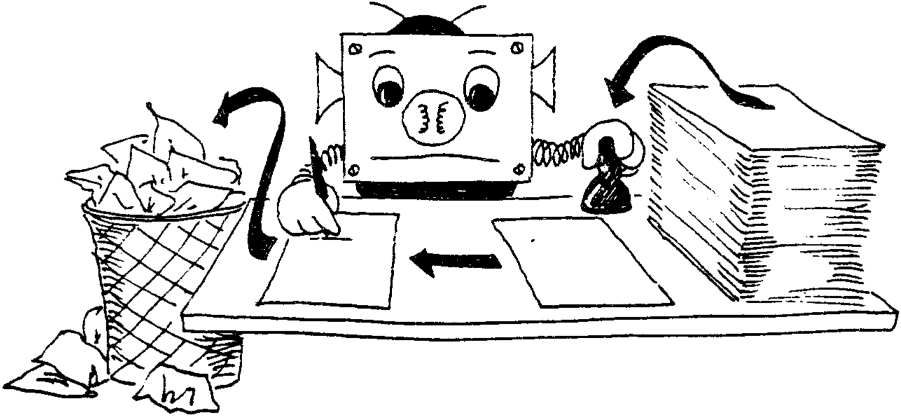


Information overload in globalized *scientific* world

Universal learning design using digital *mathematics*

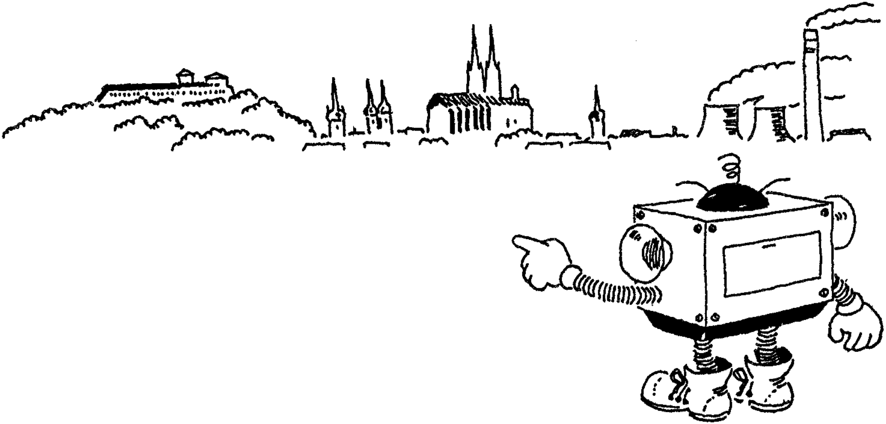


From paper to digital *workflow*

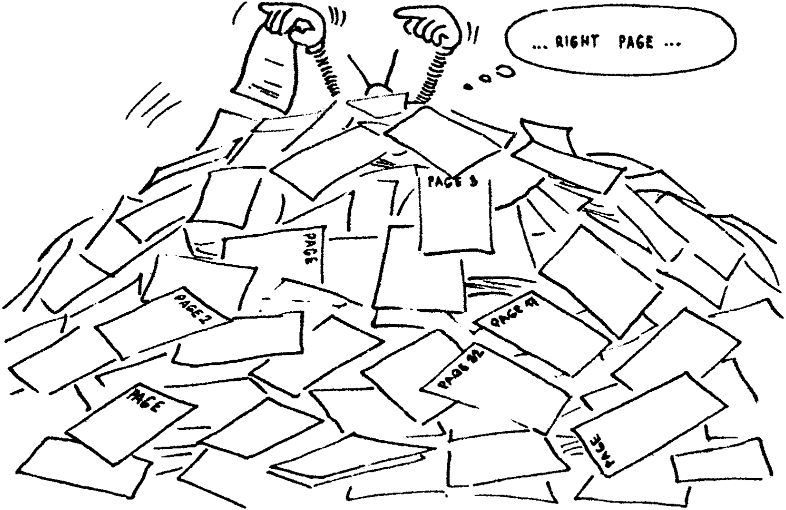


Retro-digitization, *accessible* digital library development

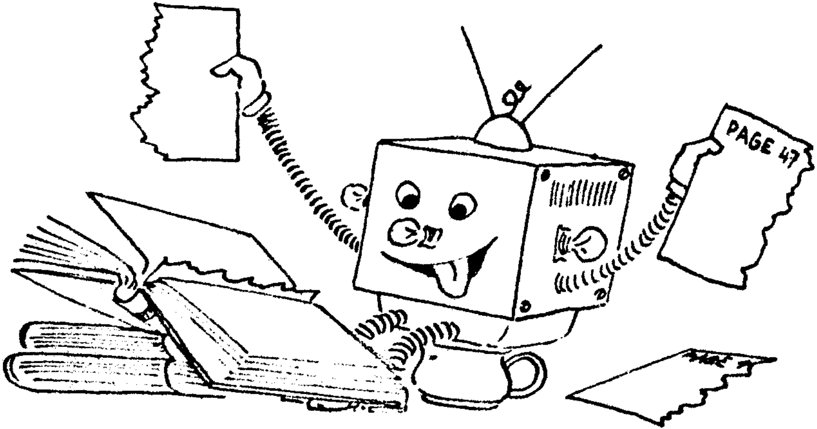
Experiences from project *DML-CZ* (Brno, CZ)



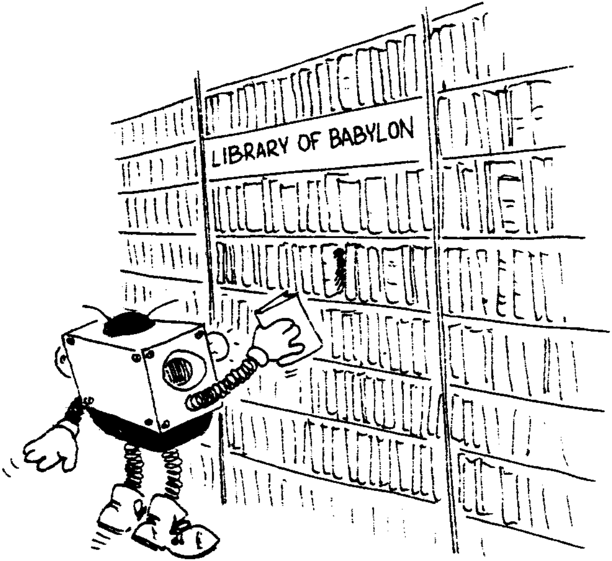
DML-CZ: new *workflows* and math data indexing



New approaches to *math document retrieval*

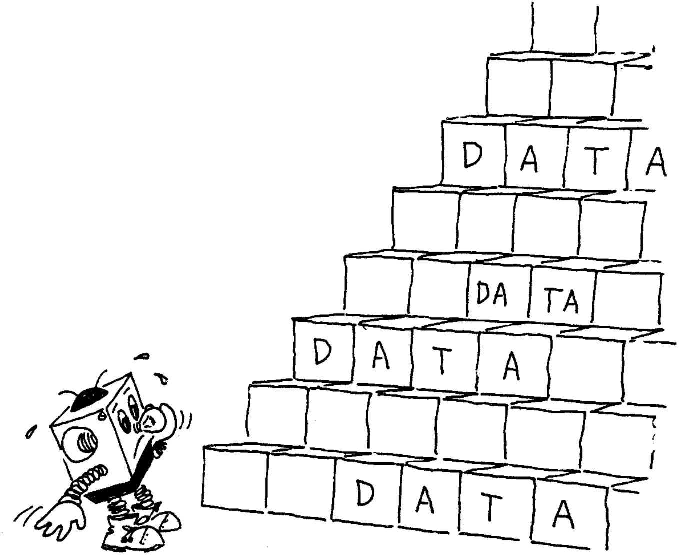


‘Bottom up’ deployment towards EU or *worldwide scale*

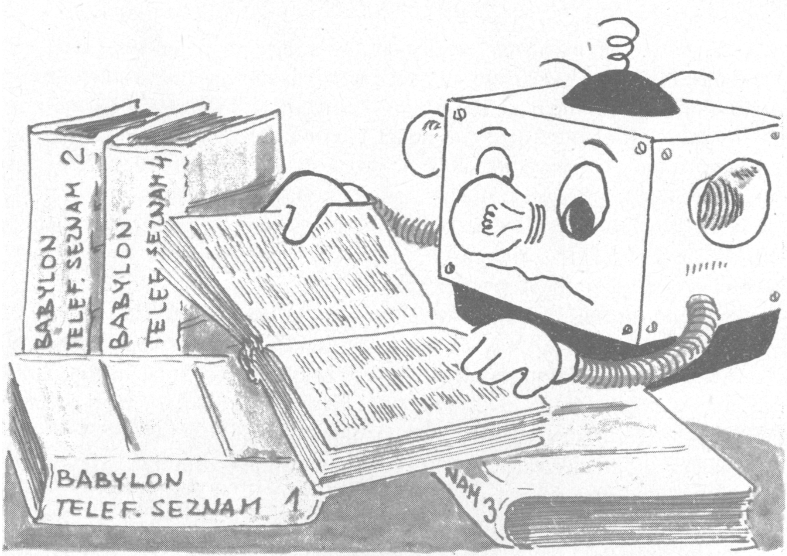


The European Digital Mathematics Library: *EuDML*

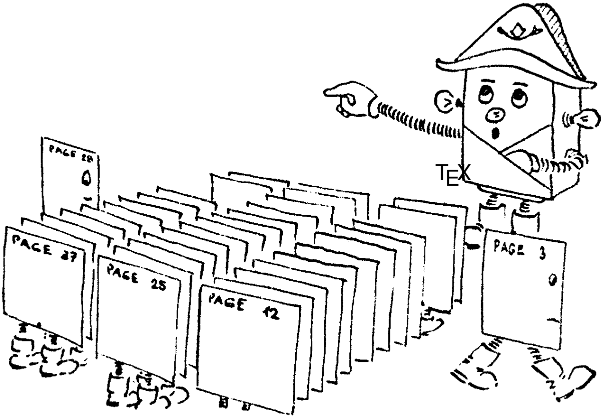
EuDML: from local data collections to the virtual DL



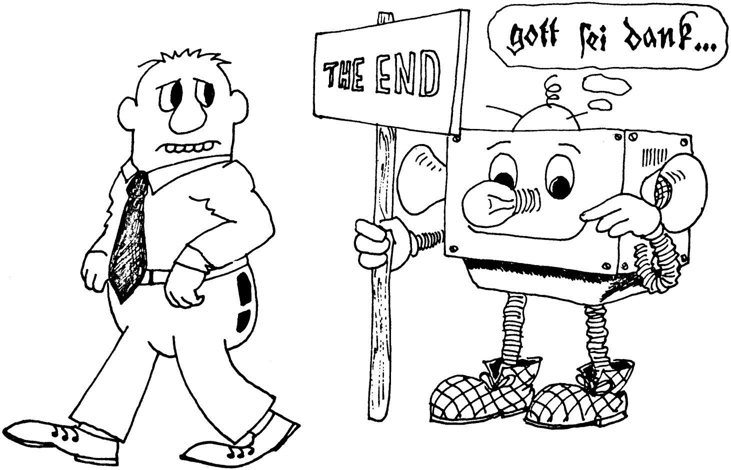
Tools for *automated math extraction* from PDF



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.

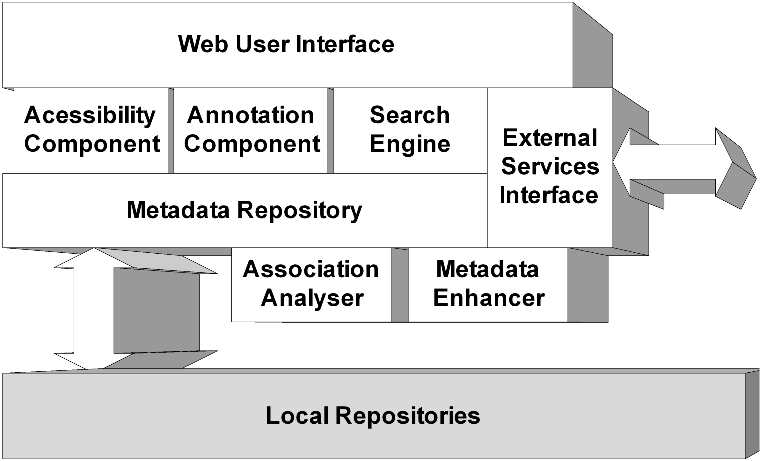
Vision of European Digital Mathematics Library

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 → 'a must/me too' effect then as with PubMed.

EuDML as a virtual library portal

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



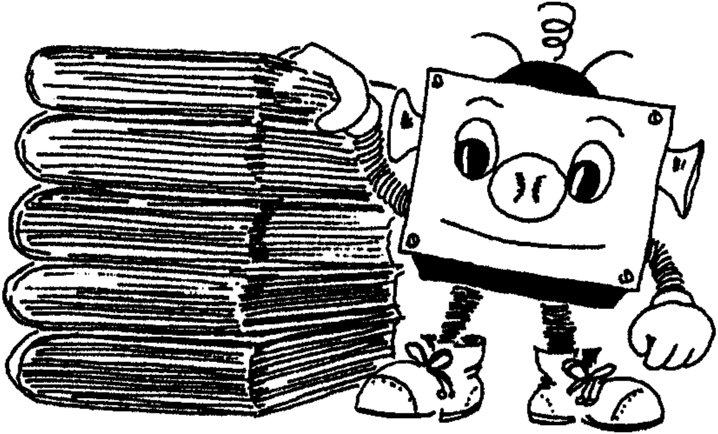
European Digital Mathematics Library

Bottom up—from building bricks of regional repositories

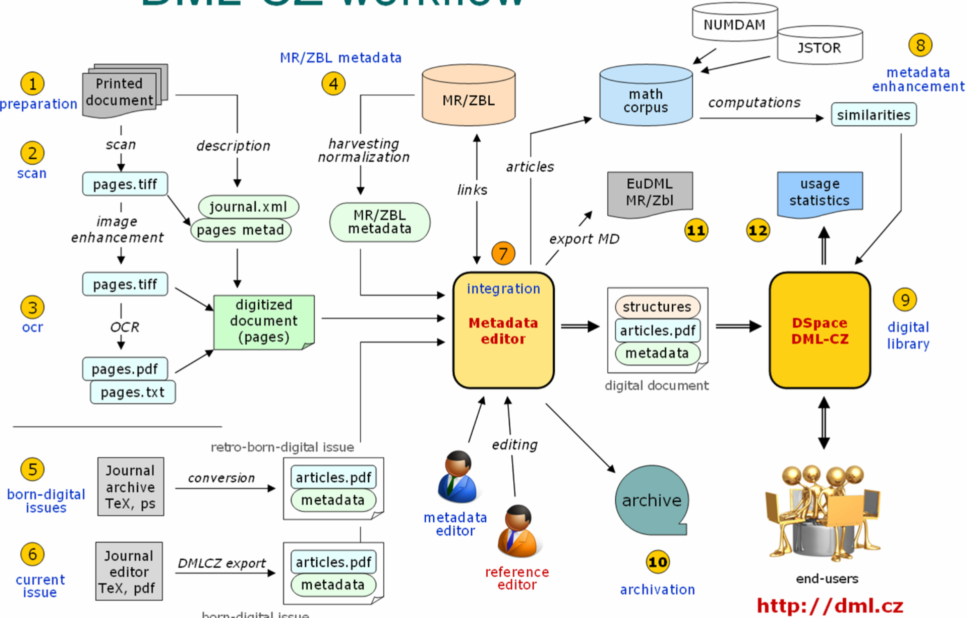
As DML content providers serve mostly publisher's or regional DML repositories as The Czech Digital Mathematics Library DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,…: aggregating content from local repositories to build the bigger (global?) DML.

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

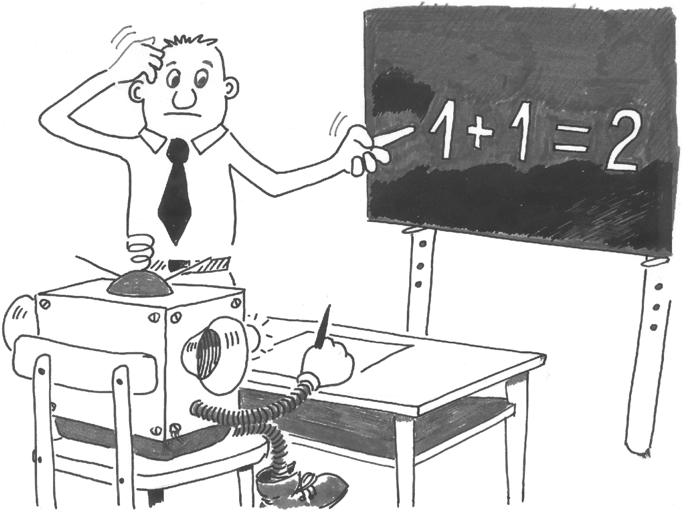


DML-CZ workflow

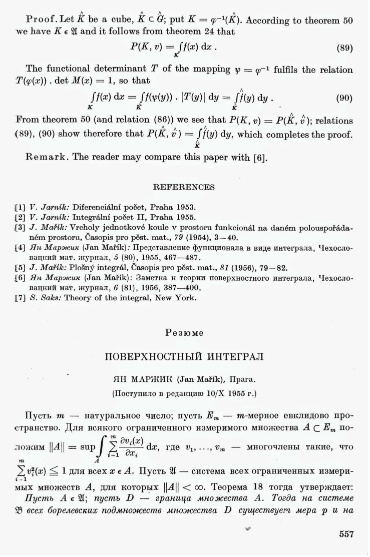


© DML-CZ, 2009

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

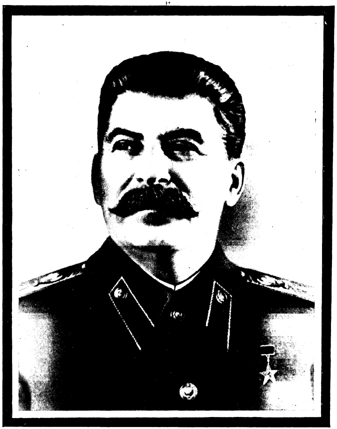
ЯН МАРЖИК (Jan Mařík), Прага.

(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

557



ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

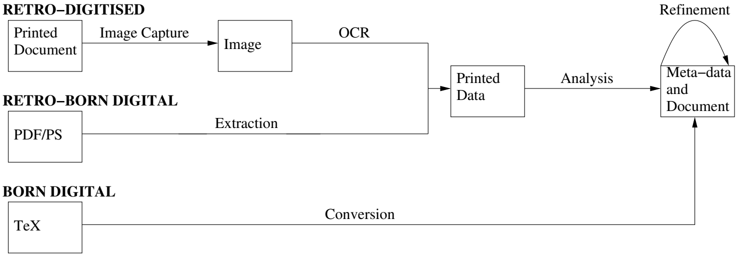
Document accessibility 4 DML processing challenges

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



RETRO–DIGITISED

Printed Document

Image Capture

Image

OCR

Printed Data

Analysis

Meta–data and Document

Refinement

RETRO–BORN DIGITAL

PDF/PS

Extraction

BORN DIGITAL

TeX

Conversion

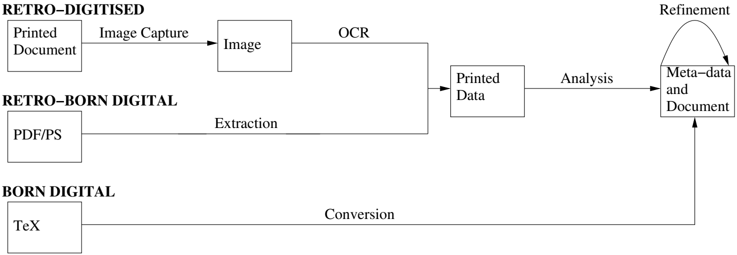
Document accessibility 4 DML processing challenges

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



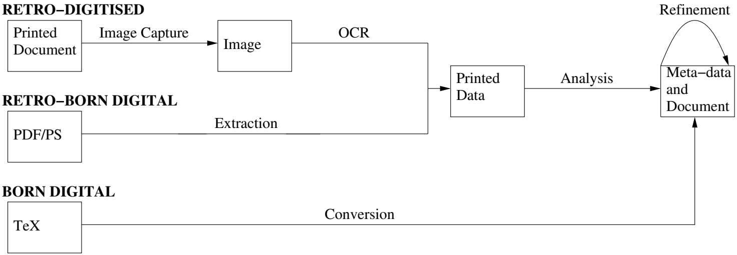
Document accessibility 4 DML processing challenges

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: → 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).

MathML or LaTeX? MathML and LaTeX!

Data heterogenity, plethora of formats, validation and conversions:

world of authors: LaTeX, TeX notation of mathematics

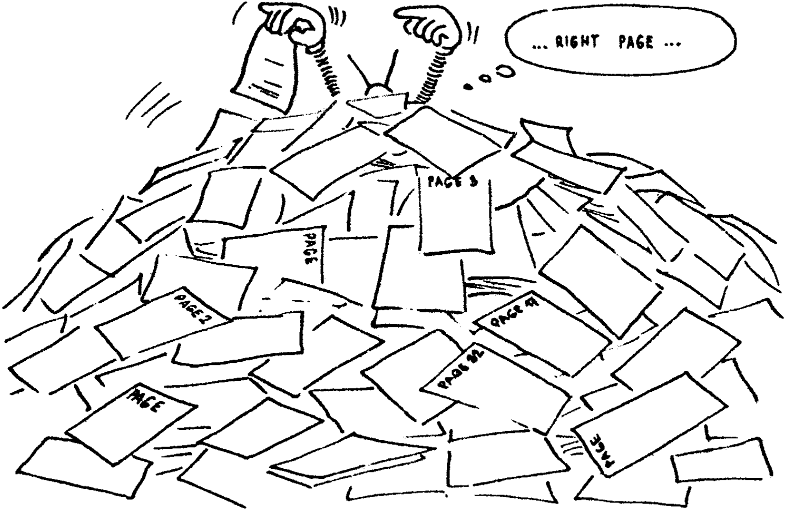
world of applications/data exchange: XML, *MathML*

Big volumes: → 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 almost *300,000 pages of 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.

Math 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) is being developed at MU

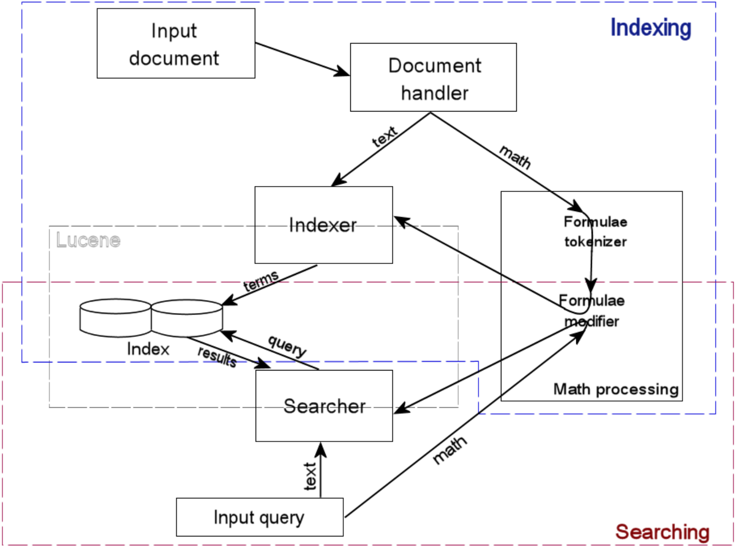
Comparison

	Input documents	Internal representation	Used converters	Approach	α-eq.	Query language	Queries	Indexing core
MathDex	HTML, TeX/LaTeX, Word, PDF	Presentation MathML (text)	jtidy, blahtex, LaTeXML, Hermes, Word+MathType, pdf2tiff->Infty	syntactic	✗	?	text, math, mixed	Apache Lucene
LeActiveMath	OMDoc, OpenMath	OpenMath (text)	-	syntactic	✗	OpenMath (palette editor)	text, math, mixed	Apache Lucene
LaTeXSearch	LaTeX	LaTeX(text)	-	syntactic	✗	LaTeX	titles, math, DOI	?
MathWeb Search	Presentation MathML, Content MathML, OpenMath	Content MathML, OpenMath (substitution trees)	-	semantic	✔	QMath, LaTeX, Mathematica, Maxima, Maple, Yacas styles (palette editor)	text, math, mixed	Apache Lucene (for text only)
EgoMath	Presentation MathML, Content MathML, PDF	Presentation MathML (text)	Infty	mixed	✗	LaTeX	text, math, mixed	EgoThor

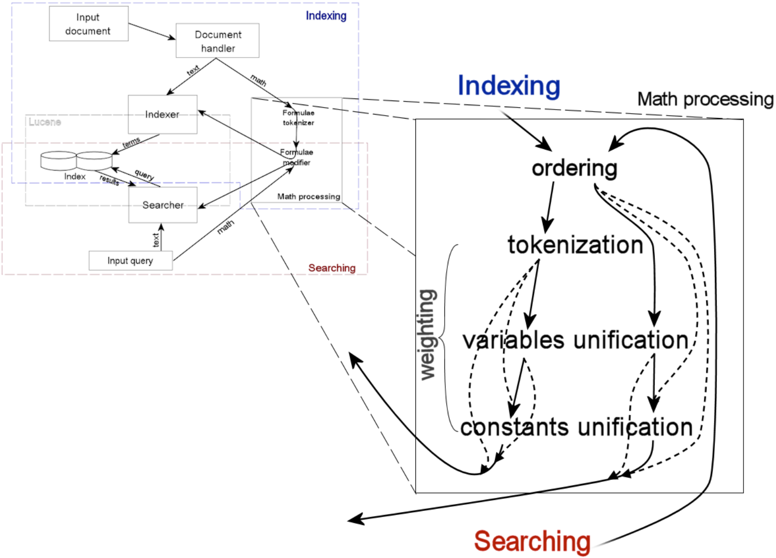
Math Indexer and Searcher — Features

- 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

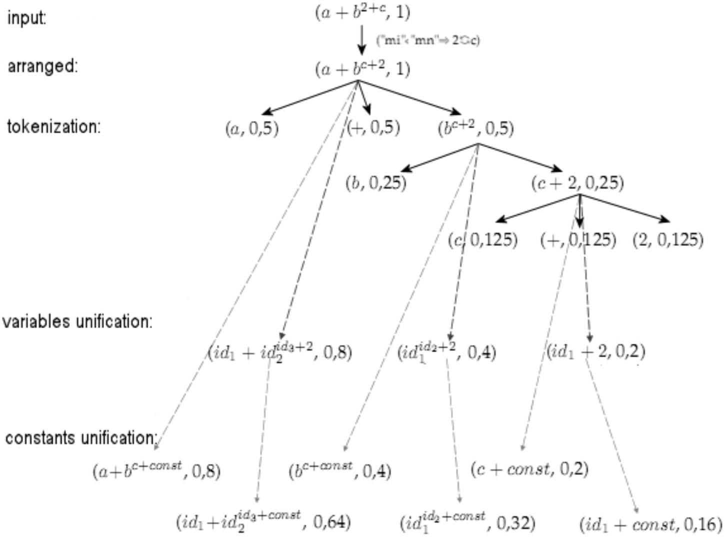
Math Indexer and Searcher — Design



Math Indexer and Searcher — Design II



Formula Processing Example



Implementation

- Java
- Lucene 3.0.1
- jTidy for text extraction
- Mathematical part implements Lucene’s interface Tokenizer — able to integrate to any Lucene based system as YADDA, DSpace,…

Evaluation

- Math corpus from arXMLiv in XHTML + MathML
    - 324,060 real math documents
    - Uncompressed corpus size 53 GB, ZIP compressed 6.7 GB
    - 112 million input formulae
- Indexed
    - Produced over 2 billion math index expressions
    - Index size 45 GB
- Simple demo web interface: WebMIaS

Math Search Conclusion

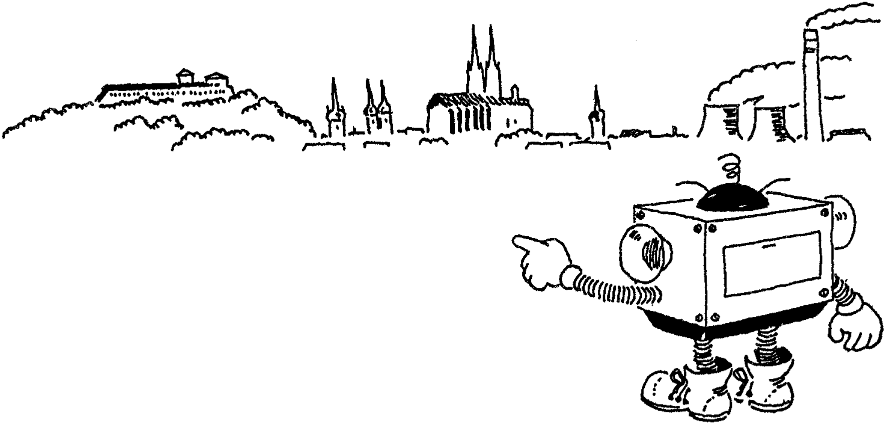
MIaS is

- *text+math IR compatible* (fits mathematician's needs),
- *scalable* (index with 2 billions formulae tested),
- *Lucene/SOLR compatible* (transformable into SOLR-usable plugin) system,
- *easily deployable/integrable into EuDML core* (written in Java).

MIaS is *ready to be used in EuDML*!

Accessibility improving technologies and tools

6+ years of local (Brno, CZ) document engineering



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

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT 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)

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

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT 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)

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

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT 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)

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

- Scanned image processing and transformations (with BookRestorer) (BT Pulkrábek)
- Mathematical optical character recognition: OCR by combining FineReader (SDK 8.1) and Infty by prof. Suzuki (MT Panák, Mudrák, BT 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)

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

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

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

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

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

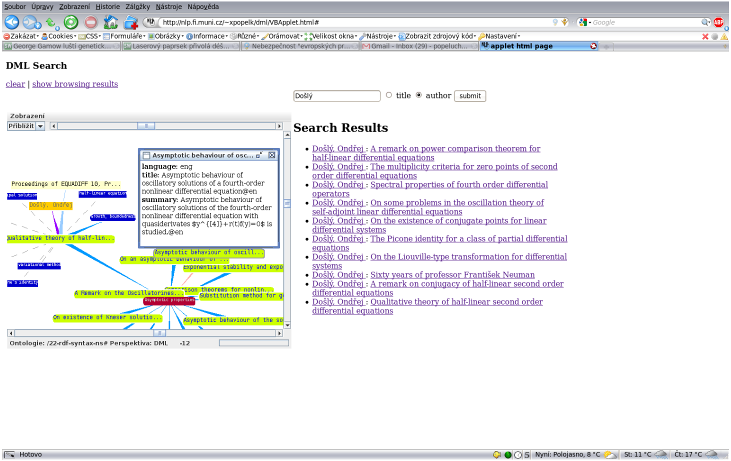
- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

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

- Google Scholar partnership: interface to use our metadata instead of those parsed from landing pages' HTML
- Math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec)
- Born-digital publishing system [for Archivum Mathematicum and for other 10 journals] and retro-born-digital paper conversions and enhancements (BT&MT Michal Růžička)
- Visualization and browsing interface (MT Zuzana Nevěřilová)

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

Metadata (in RDF) visualisation, browsing: Visual Browser tool (MT Zuzana Nevěřilová) for [Eu]DML GUI.


Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda
http://nlp.fi.muni.cz/~xpopelk/dml/VBApplet.html#
Google
Zakázat Cookies CSS Formuláře Obrázky Informace Různé Orámovat Velikost okna Nástroje Zobrazit zdrojový kód Nastavení
George Gamow luští genetick...
Laserový paprsek přivolá dés...
Nebezpečnost "evropských pr...
Gmail - Inbox (29) - popeluch...
applet html page
Zobrazení
Přiblížit
Proceedings of EQUADIFF 10, Pr...
ipal solution
half-linear equation
Došlý, Ondřej
Growth, boundedness
Qualitative theory of half-lin...
variational method
ne's identity
Asymptotic behaviour of oscill...
On an asymptotic behaviour of ...
Exponential stability and expo
Comparison theorems for nonlin...
A Remark on the Oscillatorines...
Substitution method for g
Asymptotic properties
On existence of Kneser solutio...
Asymptotic behaviour of the so
Hotovo
Nyní: Polojasno, 8 °C
St: 11 °C
Čt: 17 °C

DML Search

clear | show browsing results

Došlý ○ title ● author submit

Asymptotic behaviour of osc...

language: eng
title: Asymptotic behaviour of oscillatory solutions of a fourth-order nonlinear differential equation@en
summary: Asymptotic behaviour of oscillatory solutions of the fourth-order nonlinear differential equation with quasiderivates $y^{[4]}+r(t)f(y)=0$ is studied.@en

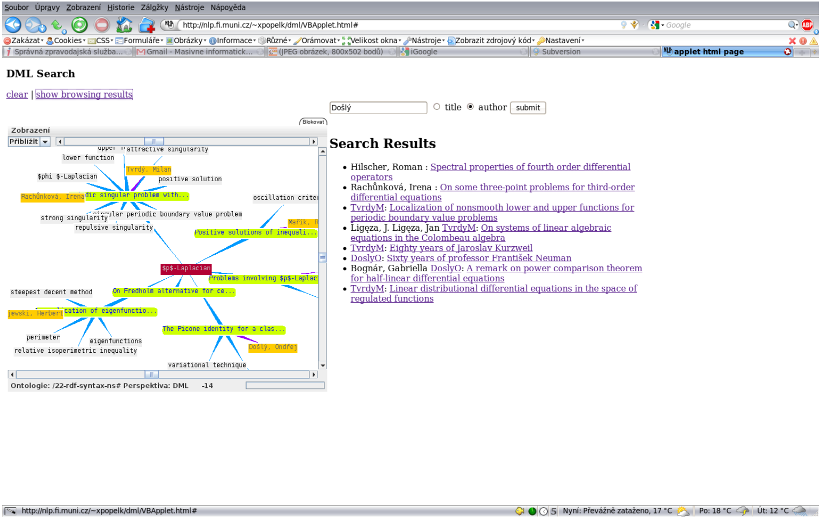
Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -12

Search Results

- Došlý, Ondřej : A remark on power comparison theorem for half-linear differential equations
- Došlý, Ondřej : The multiplicity criteria for zero points of second order differential equations
- Došlý, Ondřej : Spectral properties of fourth order differential operators
- Došlý, Ondřej : On some problems in the oscillation theory of self-adjoint linear differential equations
- Došlý, Ondřej : On the existence of conjugate points for linear differential systems
- Došlý, Ondřej : The Picone identity for a class of partial differential equations
- Došlý, Ondřej : On the Liouville-type transformation for differential systems
- Došlý, Ondřej : Sixty years of professor František Neuman
- Došlý, Ondřej : A remark on conjugacy of half-linear second order differential equations
- Došlý, Ondřej : Qualitative theory of half-linear second order differential equations

Visual Browser to meet Universal Learning Design

One of ULD principles is to present information and content in different ways:


Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda
http://nlp.fi.muni.cz/~xpopelk/dml/VBApplet.html#
Google
Zakázat Cookies CSS Formuláře Obrázky Informace Různé Orámovat Velikost okna Nástroje Zobrazit zdrojový kód Nastavení
Správná zpravodajská služba...
Gmail - Masivne informatick...
(JPEG obrázek, 800x502 bodů)
iGoogle
Subversion
applet html page
Blokovat
Zobrazení
Přiblížit
upper
attractive singularity
lower function
Tvrdý, Milan
$phi $-Laplacian
positive solution
Rachůnková, Irena
dic singular problem with...
oscillation criter
strong singularity
singular periodic boundary value problem
repulsive singularity
Mařík, R
Positive solutions of inequali...
p-Laplacian
Problems involving p-Laplac
steepest decent method
On Fredholm alternative for ce...
jewski, Herbert
cation of eigenfunctio...
The Picone identity for a clas...
perimeter
eigenfunctions
Došlý, Ondřej
relative isoperimetric inequality
variational technique
Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -14
http://nlp.fi.muni.cz/~xpopelk/dml/VBApplet.html#
5 Nyní: Převážně zataženo, 17 °C Po: 18 °C Út: 12 °C

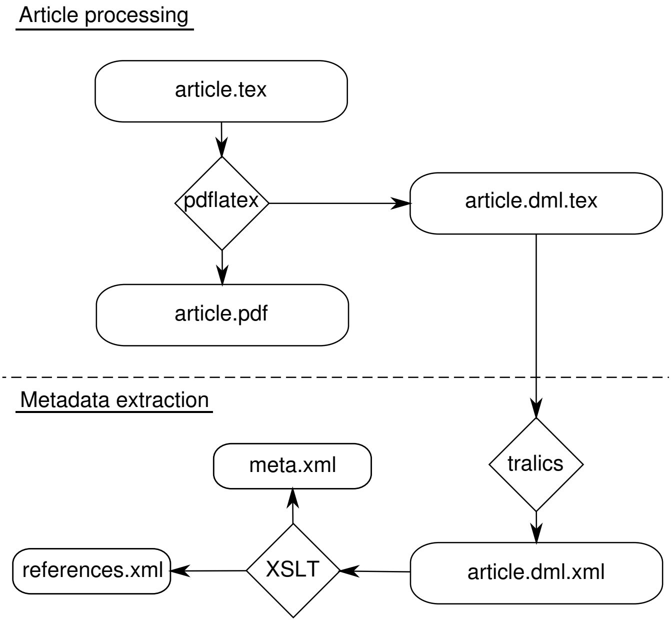
DML Search

clear | show browsing results

Došlý ○ title ● author submit

Search Results

- Hilscher, Roman : Spectral properties of fourth order differential operators
- Rachůnková, Irena : On some three-point problems for third-order differential equations
- TvrdyM: Localization of nonsmooth lower and upper functions for periodic boundary value problems
- Ligęza, J. Ligęza, Jan TvrdyM: On systems of linear algebraic equations in the Colombeau algebra
- TvrdyM: Eighty years of Jaroslav Kurzweil
- DoslyO: Sixty years of professor František Neuman
- Bognár, Gabriella DoslyO: A remark on power comparison theorem for half-linear differential equations
- TvrdyM: Linear distributional differential equations in the space of regulated functions



Article processing

article.tex

pdflatex

article.dml.tex

article.pdf

Metadata extraction

meta.xml

tralics

references.xml

XSLT

article.dml.xml

How Does It Work

- A lightweight set of LaTeX macros in the form of a LaTeX macro package.
  - Can be easily customized to meet needs of a particular journal document class / style file.
  - The LaTeX macro package itself does not transform the LaTeX source code to XML.
  - Literally exports selected parts of the LaTeX document to an external file.
  - This file is subsequently processed by a journal-independent Tralics-based procedure.

How Does It Work (cont.)

```
\documentclass[runningheads]{llncs}
\usepackage{dmlcommon}
\usepackage{dmlcz}

\begin{document}

\author{Petr Sojka}
\dmlaindex{Sojka}{Petr}
\dmltitle{Towards a Digital Mathematical Library}
...
\maketitle

\begin{dmlabstract}
The workshop's objectives were to formulate the strategy
and goals of a global mathematical digital library...
\end{dmlabstract}
...
```

How Does It Work (cont.)

```
\documentclass{dmlczmeta}\begin{document}

\begin{xmlelement}{author}{Sojka, Petr
\XMLaddatt{order}{1}}\end{xmlelement}

\begin{xmlelement}{title}{Towards a Digital Mathematical
Library\XMLaddatt{lang}{eng}}\end{xmlelement}

\begin{xmlelement}{abstract}\XMLaddatt{lang}{eng}\bgroup
The workshop’s objectives were to formulate the strategy
and goals of a global mathematical digital library...
\egroup\end{xmlelement}

\begin{xmlelement}{keyword}{OCR\XMLaddatt{lang}{eng}}
\end{xmlelement}

...
\end{document}
```

How Does It Work (cont.)

- Tralics is a LaTeX to XML translator.
  - The most indispensable part of the system.
  - Its engine is able to process regular LaTeX code.
  - It is not necessary to
    - convert the LaTeX code to plain text directly,
    - nor deal with the LaTeX macro expansion or the complexity of its syntax.
- Tralics outputs a UTF-8 encoded XML file.
- This output is finally processed by the XLST processor furnishing DML-CZ metadata in its final form.

How Does It Work (cont.)

```
<?xml version=’1.0’ encoding=’UTF-8’?>
<!DOCTYPE std SYSTEM ’classes.dtd’>
<!-- Translated from latex by tralics 2.13.5,
     date: 2010/07/03-->
<std><p>
<author order=’1’>Sojka, Petr</author>
<title lang=’eng’>Towards a Digital Mathematical
Library</title>

<abstract lang=’eng’>The workshop’s objectives were to
formulate the strategy...</abstract>
<keyword lang=’eng’>OCR</keyword>
<keyword lang=’eng’>OpenMath</keyword>

<language>eng</language>
<abstractlanguage>eng</abstractlanguage>
...
</p></std>
```

How Does It Work (cont.)

```
<?xml version=”1.0” encoding=”UTF-8”?>
<article>
  <title lang=”eng”>Towards a Digital
    Mathematical Library</title>

  <author order=”1”>Sojka, Petr</author>

  <language>eng</language>

  <keyword lang=”eng”>OCR</keyword>
  <keyword lang=”eng”>OpenMath</keyword>

  <summary lang=”eng”>The workshop’s objectives
    were to formulate the strategy...</summary>
  ...
</article>
```

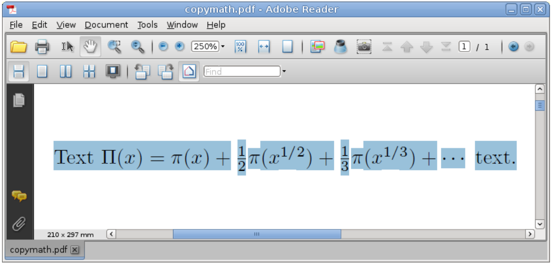
Why It Is Useful

- It is easy to integrate this procedure to an existing journal processing workflow. It is thus acceptable to all the involved editors.
  - Current TEX processing is used.
  - Platform independent.
    - The TEX itself produces the source file.
    - XML generated using Tralics and XSLT.
  - No need for BibTEX.
- It is safe.
  - At the same time as the final PDF document is created, the metadata is automatically generated based on the same source code.
- Since Tralics supports MathML we are able to translate mathematical expressions from the input LATEX notation to this XML language.

Maths, TEX, PDF

- PDF is widely adopted and very often used for electronic publications.
  - The DML-CZ project stores full texts of the articles as PDF files as do many other digital libraries.
- Thanks to pdfTEX, PDF is also the *de facto* standard output format of the modern TEX distributions.
- LATEX mathematical notation is well known and effective.
  - Used not only in LATEX documents, but also in a variety of other projects, such as Wikipedia.
- LATEX source code is usually a good choice for plain text representation of mathematical expressions.

Standard PDF document

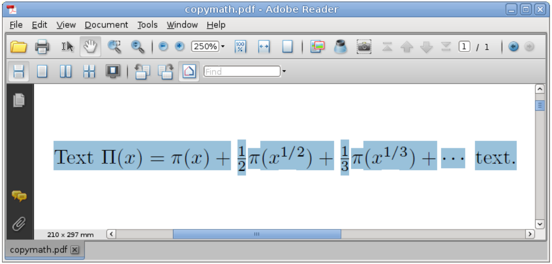

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

LaTeX source code:

```
Text $\Pi(x) = \pi(x) +
\frac{1}{2}\pi(x^{1/2}) +
\frac{1}{3}\pi(x^{1/3}) + \cdots$
text.
```

Standard PDF document

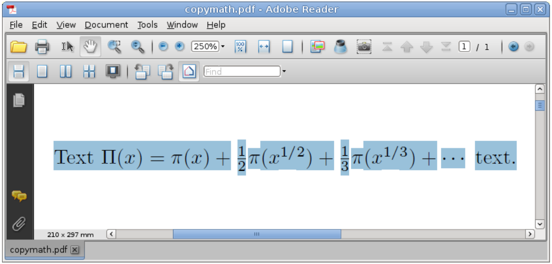

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

PDF code:

```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ/F17 9.9626 Tf 23.247 0 Td
[(\005\050)]TJ/F20 9.9626 Tf 11.346 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[(\051)-278(=)]TJ/F20 9.9626 Tf 17.158 0 Td [(\031)]TJ/F17 9.9626 Tf 6.036 0 Td
[(\050)]TJ/F20 9.9626 Tf 3.875 0 Td [(x)]TJ/F17 9.9626 Tf 5.694 0 Td
[(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923 Td [(1)]TJ
ET
```

Standard PDF document

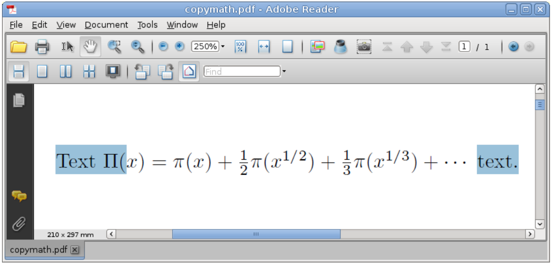

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

Text obtained using Copy & Paste function of PDF reader:

```
Text  ( ) =  ( ) + 1
2 ( 1/2) + 1
3 ( 1/3) + · · · text.
```

copymath-enabled PDF document

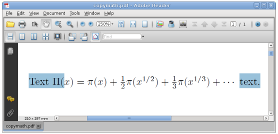

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

LaTeX source code:

```
Text $\Pi(x) = \pi(x) +
\frac{1}{2}\pi(x^{1/2}) +
\frac{1}{3}\pi(x^{1/3}) + \cdots$
text.
```

copymath-enabled PDF document

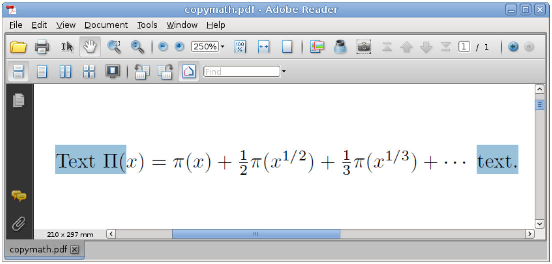

copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

PDF code:

```
BT
/F16 9.9626 Tf 148.712 707.125 Td [(T)83(ext)]TJ
ET
1 0 0 1 171.959 707.125 cm
/Span <<
/ActualText<245C506920287829203D205C706920287829202B205C66726163207B317D7B32
7D5C70692028785E7B312F327D29202B205C66726163207B317D7B337D5C70692028785E7B31
2F337D29202B205C63646F74732024> >> BDC
1 0 0 1 -171.959 -707.125 cm
BT
/F17 9.9626 Tf 171.959 707.125 Td [(\005\050)]TJ/F20 9.9626 Tf 11.346 0 Td
[(x)]TJ/F17 9.9626 Tf 5.694 0 Td [(\051)-278(=)]TJ/F20 9.9626 Tf 17.158 0 Td
[(\031)]TJ/F17 9.9626 Tf 6.036 0 Td [(\050)]TJ/F20 9.9626 Tf 3.875 0 Td
[(x)]TJ/F17 9.9626 Tf 5.694 0 Td [(\051)-222(+)]TJ/F18 6.9738 Tf 17.247 3.923
Td [(1)]TJ
ET
```

copymath-enabled PDF document


copymath.pdf - Adobe Reader
File Edit View Document Tools Window Help
250%
1 / 1
Find
210 x 297 mm
copymath.pdf

Text $\Pi(x) = \pi(x) + \frac{1}{2}\pi(x^{1/2}) + \frac{1}{3}\pi(x^{1/3}) + \cdots$ text.

Text obtained using Copy & Paste function of PDF reader:

```
Text $\Pi (x) = \pi (x) +
      \frac {1}{2}\pi (x^{1/2}) +
      \frac {1}{3}\pi (x^{1/3}) + \cdots $
text.
```

Implementation

- The `ActualText` command of the PDF language is used to mark the region of the mathematical expression inside the PDF document.
- We want the package to be as user friendly as possible – users should not be forced to modify their mathematical expressions in any way, `\usepackage{copymath}` should cater for all their needs.
  - The implementation is not easy.
  - This requires nonstandard modifications of the LaTeX mathematical environments.

Implementation (cont.)

- We need to add `\pdfliteral` at the beginning and end of every mathematical environment.
- The dollar sign ($) is activated and redefined.
- It is necessary to keep track of nested mathematical environments.
- Simple redefinition of $\mathcal{AMS}$-LaTeX mathematical environments is not possible.
- It seems that not all PDF viewers respect contents of the `ActualText` command.
- Adobe Reader ignores the "_" sign inside `ActualText` provided another character is present.
- Possibility to be misused.

Summary

- 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 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

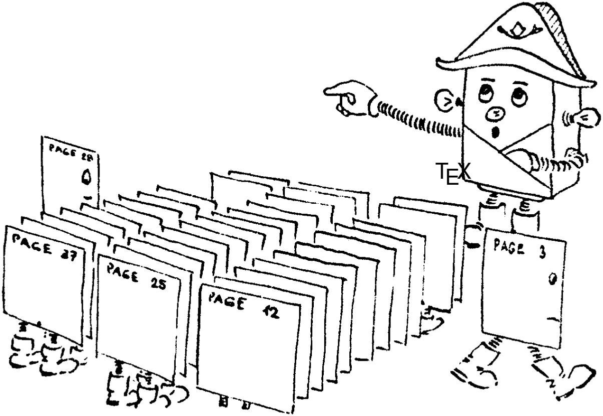
Summary

- 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 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

Summary

- 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 2011 c/o CICM Bertinoro, Italy, July 18th–23rd, 2011

Yes, you can!



Future work

- Robust Math OCR
- Robust PDF2Math conversion
- Design alternative, novel and accessible user interfaces for the digital library
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense, mathematical document classification, document similarity

Future work

- Robust Math OCR
- Robust PDF2Math conversion
- Design alternative, novel and accessible user interfaces for the digital library
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense, mathematical document classification, document similarity

Future work

- Robust Math OCR
- Robust PDF2Math conversion
- Design alternative, novel and accessible user interfaces for the digital library
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense, mathematical document classification, document similarity

Future work

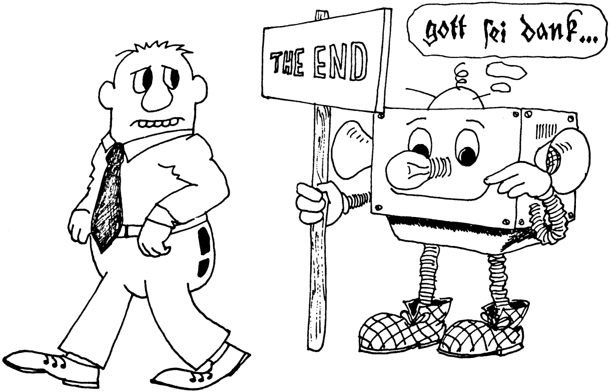
- Robust Math OCR
- Robust PDF2Math conversion
- Design alternative, novel and accessible user interfaces for the digital library
- Other challenges: multilingual math retrieval, MathML indexing and search, math common sense, mathematical document classification, document similarity

Take-home messages and acknowledgements

- Even though some problems are not resolved and complexity of dealing with math is high,
- there seems to be light at the end of the tunnel!

Many thanks to all members of DML-CZ and EuDML team, especially to those whose tools and work had been reported: Michal Růžička, Martin Líška, Martin Šárfy, Radim Řehůřek, Volker Sorge, Thierry Bouche et al.

End of the talk



Questions? Comments? Cooperation offers?

Czech Digital Mathematics Library [online].
[online, cit. 2011-02-10].
Available from WWW: <http://dml.cz/>.

EuDML: The European Digital Mathematics Library [online].
This page was last modified on 20 January 2010, at 08:09. [online, cit. 2011-02-10].
Available from WWW: <http://www.eudml.eu/>.

Bouche, T.:
A pdfLaTeX-based automated journal production system.
In Proceedings of EuroTeX 2006, TUGboat **27**(1) (2006) 45–50.

Centre de diffusion de revues académiques mathématiques [Center for diffusion of mathematic journals] [online].
[online, cit. 2011-02-10].
Available from WWW: <http://www.cedram.org/>.

Růžička, M.:
Automated Processing of TeX-Typeset Articles for a Digital Library.
In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, 167–176.

Archivum Mathematicum [online].
Masaryk University, Brno, Czech Republic.
Last modified December 18, 2009 [online, cit. 2011-02-10].
Available from WWW: <http://www.emis.de/journals/AM/>.

Grimm, J.:
Tralics, a LaTeX to XML Translator.
In Proceedings of EuroTeX, TUGboat **24**(3) (2003) 377–388.

Tralics: a LaTeX to XML translator [online].
Last modified $Date: 2009/11/24 17:17:03 $ [online, cit. 2011-02-10].
Available from WWW: <http://www-sop.inria.fr/apics/tralics/>.

Infty Project: Research Project on Mathematical Information Processing [online].

[online, cit. 2011-02-10].
Available from WWW: <http://www.inftyproject.org/en/>.

Suzuki, M.; Kanahori, T.; Ohtake, N.; Yamaguchi, K.:
An Integrated OCR Software for mathematical Documents and Its Output with Accessibility.
Computers Helping people with Special Needs, 9th International Conference ICCHP 2004, Paris, July 2004, Lecture Notes in Computer Sciences 3119, Springer (2004) 648–655.

EuDML at MU team.
EuDML at MU project info [online, cit. 2011-02-10].
<http://nlp.fi.muni.cz/projekty/eudml/> or
<http://www.muni.cz/research/projects/10067>.