Specifics of Open Access Publishing and Retrodigitization in Mathematics: An Experience from DML-CZ and EuDML Projects

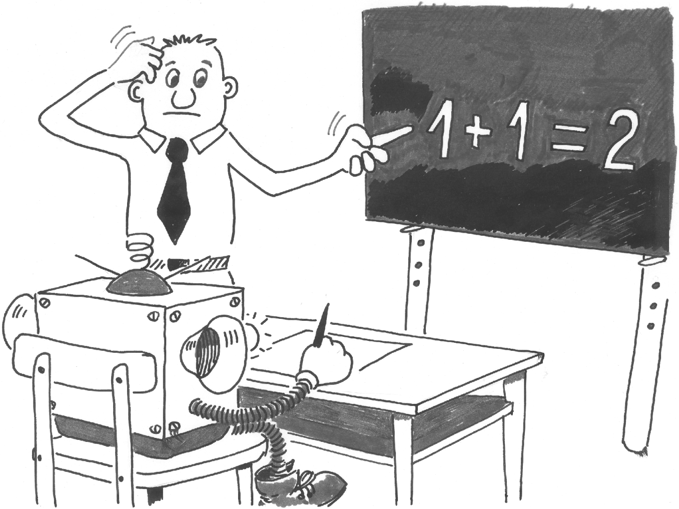
Petr Sojka
<sojka@fi.muni.cz> (Faculty of Informatics, Masaryk University, Brno)

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.

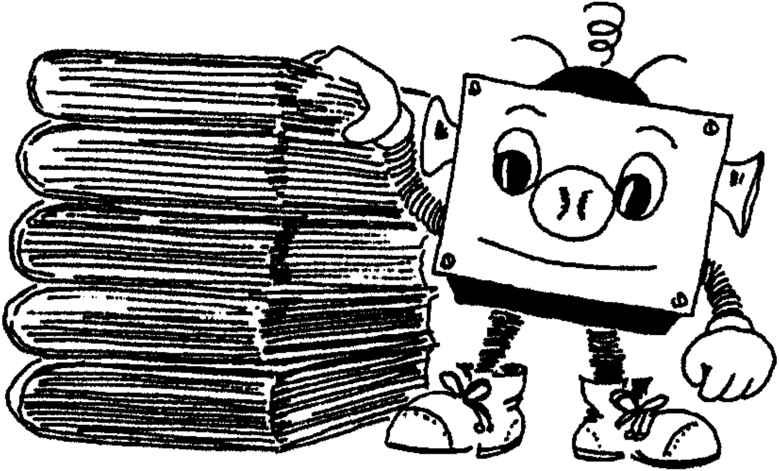
EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

Specifics of mathematics: Mathematics is the Queen of Science and Arithmetics is the Queen of Mathematics (C.F. Gauss)

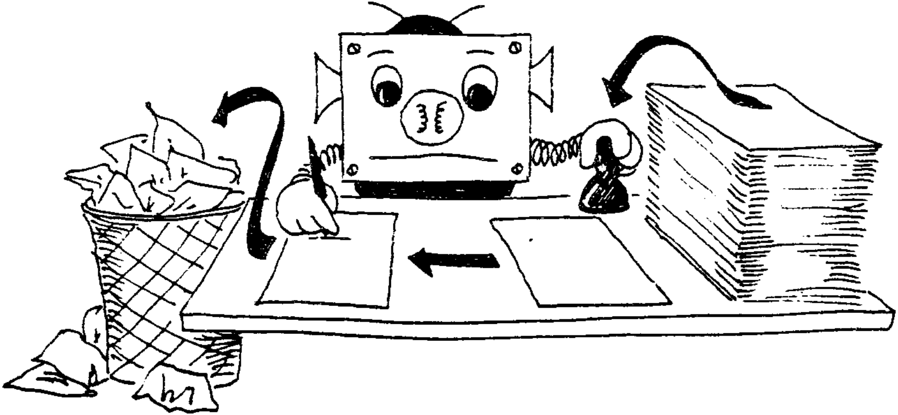


Digital processing is full of challenges



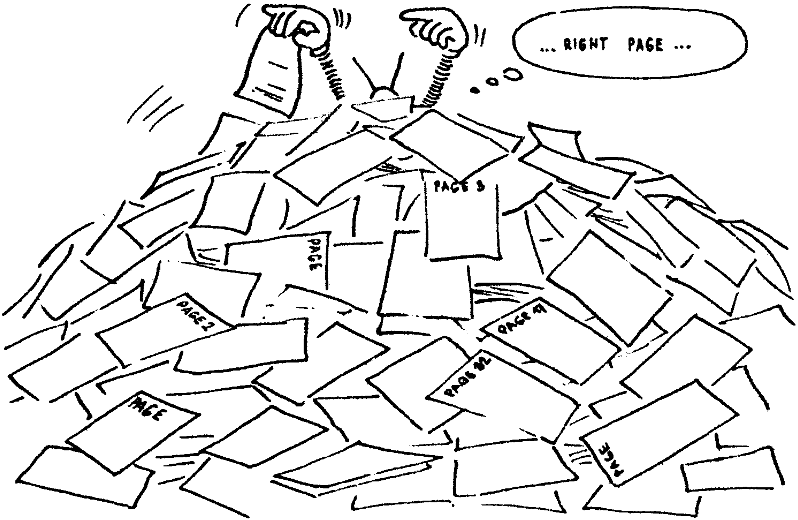
Amount of math ever published is about 100.000.000 pages

Specific publishing workflow and document engineering (tough reviewing, formulae handling)

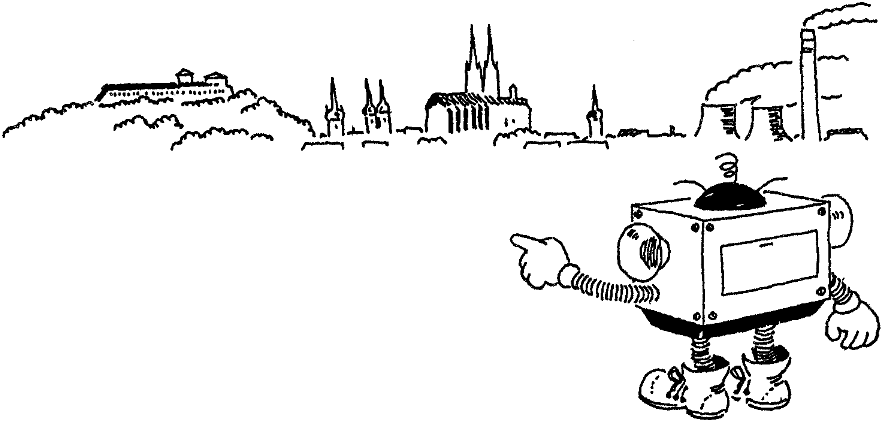


Specific problems of digitization (OCR, digital library development)

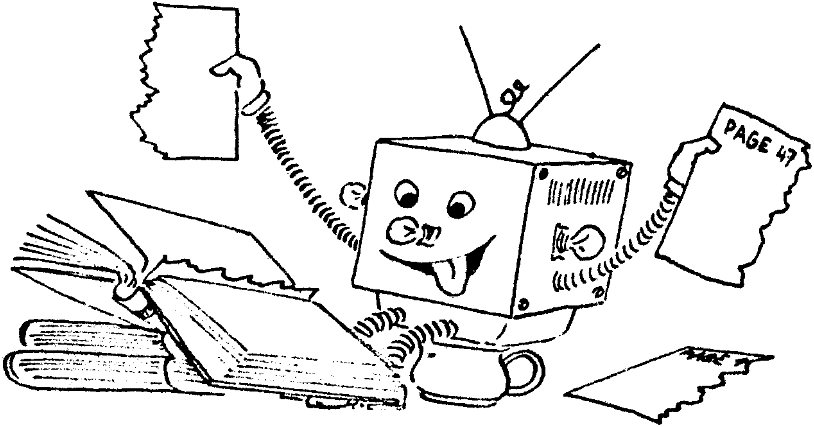
Challenges of math formulae search and indexing



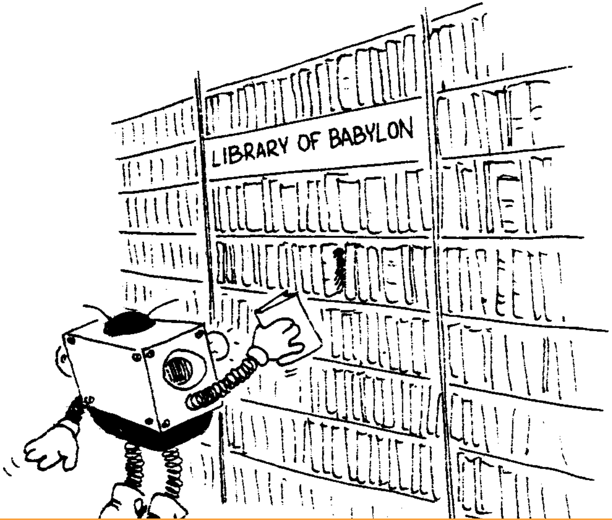
Challenges tackled during DML-CZ in Brno, CZ (pdfTEX's born town)



DML-CZ document engineering—best practices and tools

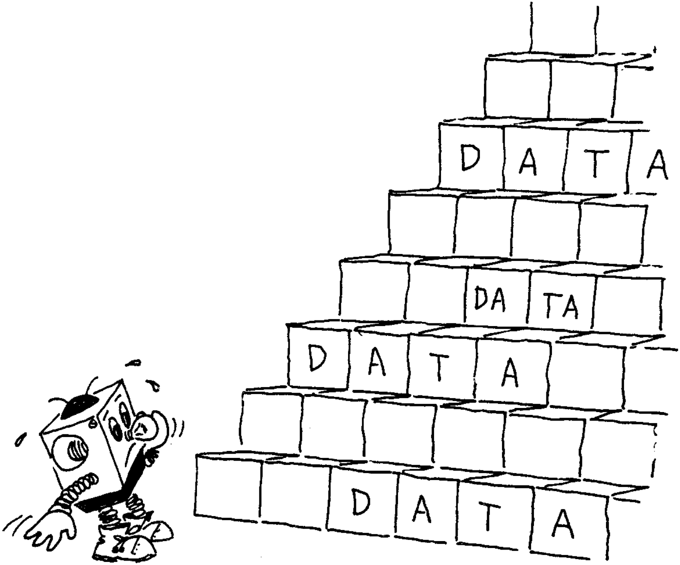


Towards European Virtual Library of Mathematics: bottom up DML processing towards EU (or worldwide) scale

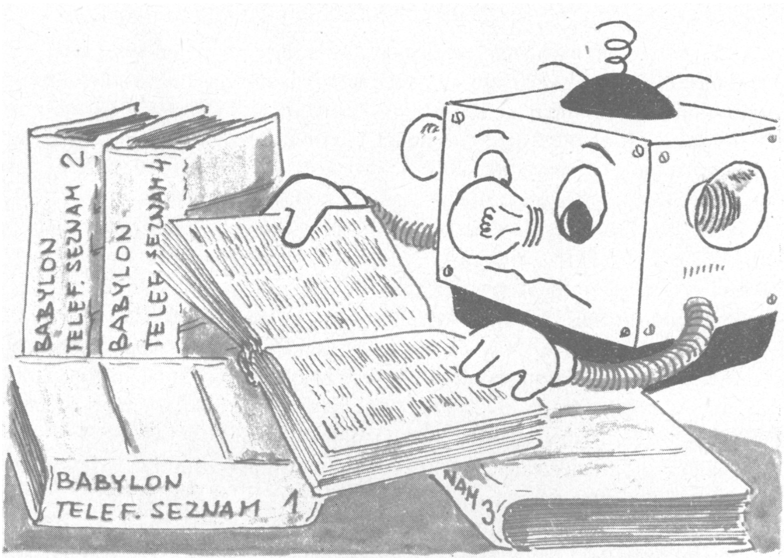


European Digital Mathematics Library

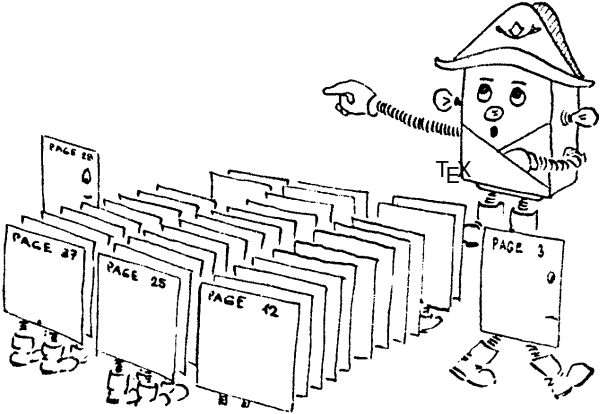
EuDML—from local data collections to the virtual digital library



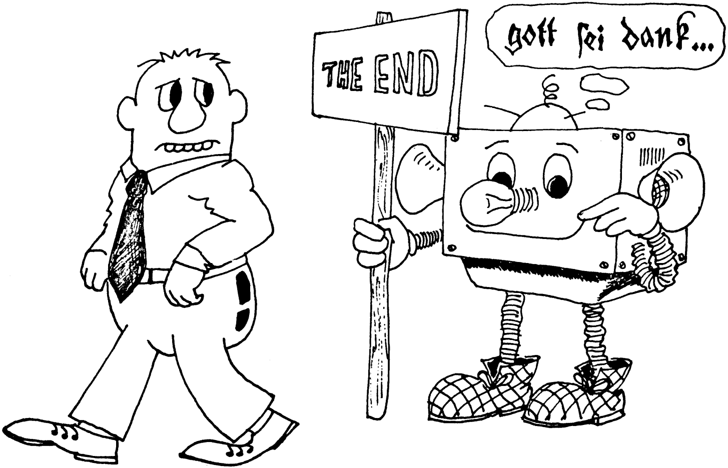
EuDML document engineering—scalable tools development



Yes, you can!



End of talk overview, quiz!



Open-ended story of a Vision

At the beginning there was a vision of all mathematical knowledge, *peer reviewed and verified* (100,000,000 pages) in ‘one stop shop’ (in the digital form), articulated at the beginning of millenium: reference mathematics library, grounding truths for not only mathematitions, but also other sciences.

Progress of IT, cheap disk space, 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 (some nine digit sum in $’s requested). Application was not successful–top down approach failed.

Vision of (European) Digital Mathematics Library

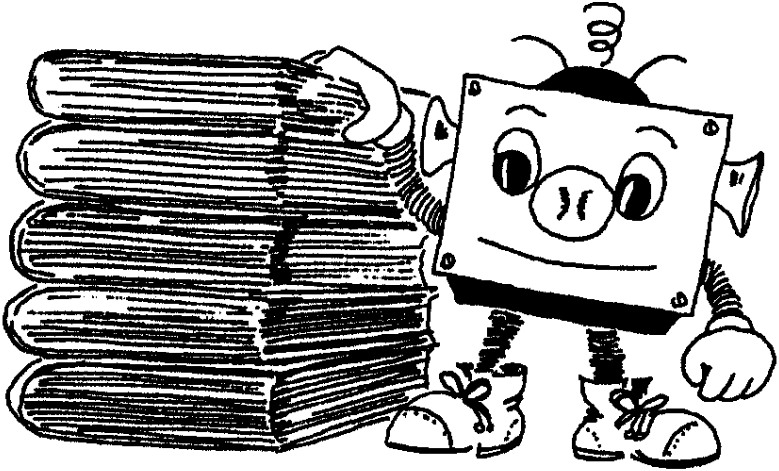
Several attempts to fund development of DML on European level (FP5, FP6) also was not successful.

Now, it starts to be realized: three year EU project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010 (MU and MU AV). EuDML The EUROPEAN DIGITAL MATHEMATICS LIBRARY

The strategy:

- to master the technology, develop tools and offer them;
- concept of *moving wall* to motivate and engage commercial publishers.
- to collect (bottom up) [virtual] *digital library*, ‘one-stop shop’ and achieve critical mass in the domain ⟶ ‘me too’ effect then.

From paper to digital processing, from local to the whole



Bottom up—from building bricks of regional repositories

As building bricks current (regional or publisher's) Digital Mathematics Library (DML) repositories as DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,…(from local repositories bottom-up to build the final thing—to realize the vision).

Example of DML-CZ: up and running digital mathematic library with nearly 30,000 papers. For more, see (who, what, browse, browse similar, how to search).

Live project—all comments to DML-CZ welcome!

DML-CZ: main facts

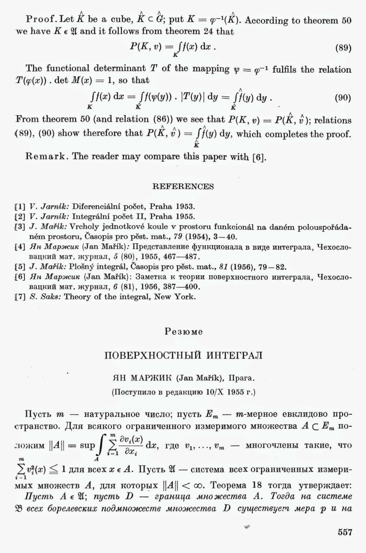
- Czech Academy of Sciences grant (program Information Society) 2005–2009, *full* (retro)digitization of 50,000 pages of mathematical literature per year, 8M CZK in total.
- Research part: **1)** gradual enhancement of the digital material by ‘knowledge enhancing’ filters on markup-rich XML data. **2)** 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).
- Production part: dig. center Jenštejn, overestimated costs.

DML-CZ data: what is there

- 12 mathematics journals, 7 OA, 5 with moving wall.
- 8 publishers of math in CZ and SK, close cooperation.
- 27,000+ articles with rich metadata (full bibliographic record, similarities, MSC classifications, translated titles).
- heavily multilingual (7+ languages).

Links on who, what, browse, browse similar, how to search).

DML-CZ – data: scientific math published in Czech and Slovak



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.)

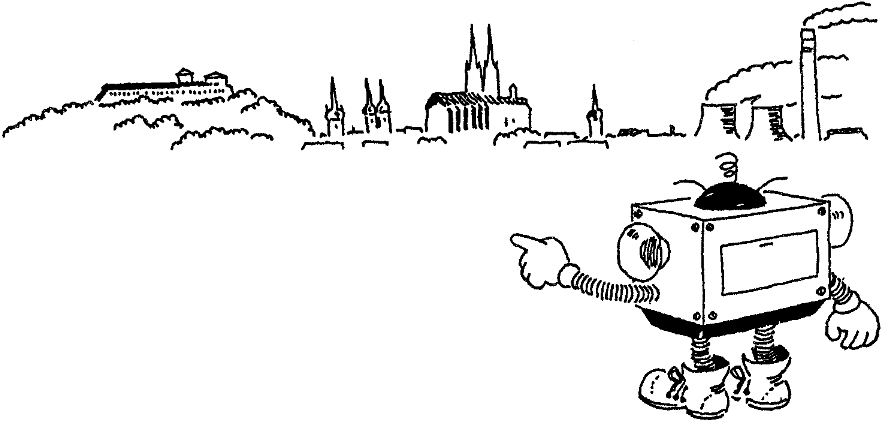
Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^{m} v_i^2(x)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

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



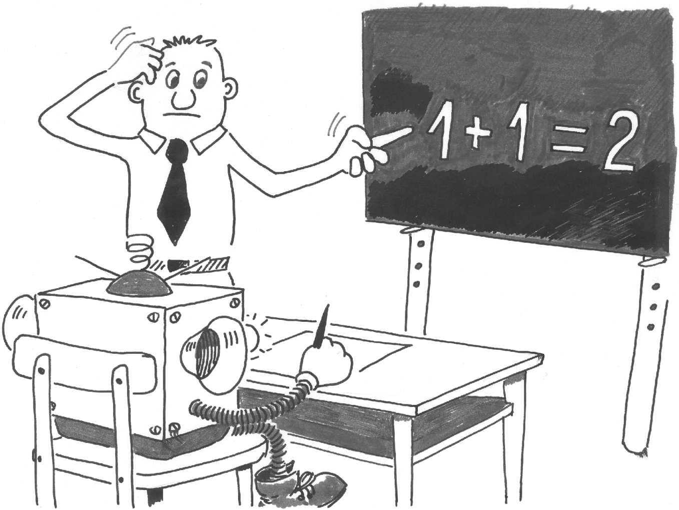
The approach used in DML-CZ

A succesfully 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'.

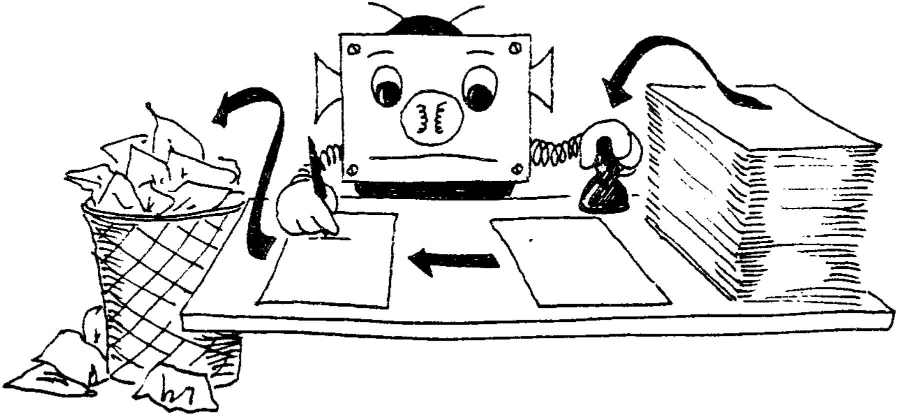
Design, technical and *political decisions* behind building the *Czech Digital Mathematics Library DML-CZ* (<http://dml.cz>) in the context of other thematical community projects (PubMed Central, ADS, INSPIRE, SCOAP3 and EuDML) have been solved. *No wheel reinvention.*

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. As much automated (using robots :-) as possible.

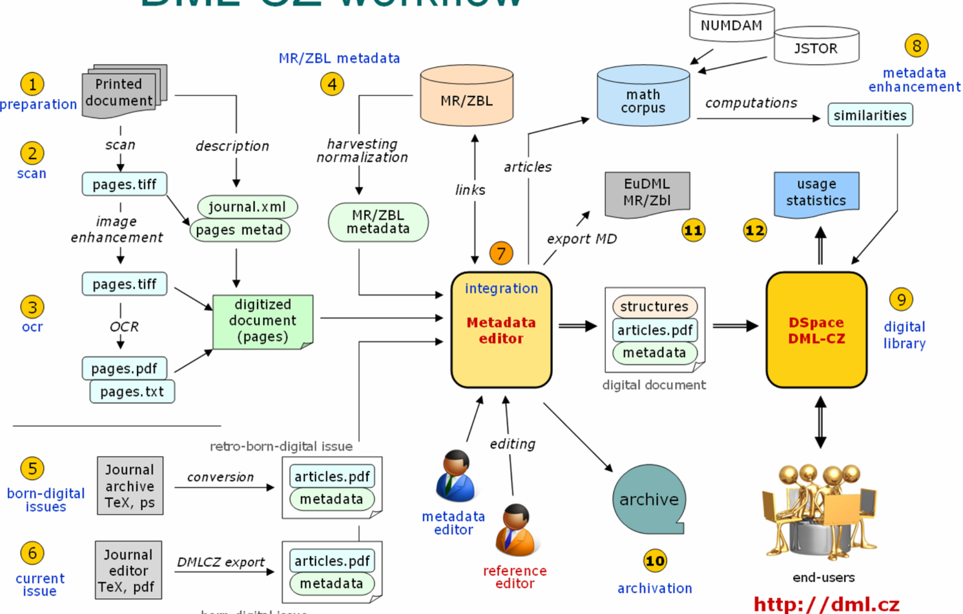
Math handling poses challenges—math OCR, math indexing,...



Document engineering—from paper to digital *workflow*



DML-CZ workflow



1 preparation
Printed document
description
2 scan
scan
pages.tiff
journal.xml
pages metad
image enhancement
pages.tiff
3 ocr
OCR
pages.pdf
pages.txt
digitized document (pages)

4 MR/ZBL metadata
MR/ZBL
harvesting normalization
MR/ZBL metadata
links
articles

5 born-digital issues
Journal archive TeX, ps
conversion
retro-born-digital issue
articles.pdf
metadata

6 current issue
Journal editor TeX, pdf
DMLCZ export
articles.pdf
metadata
born-digital issue

7
integration
Metadata editor
editing
metadata editor
reference editor
export MD

NUMDAM
JSTOR
math corpus
computations
similarities
8 metadata enhancement

EuDML MR/Zbl
11
usage statistics
12

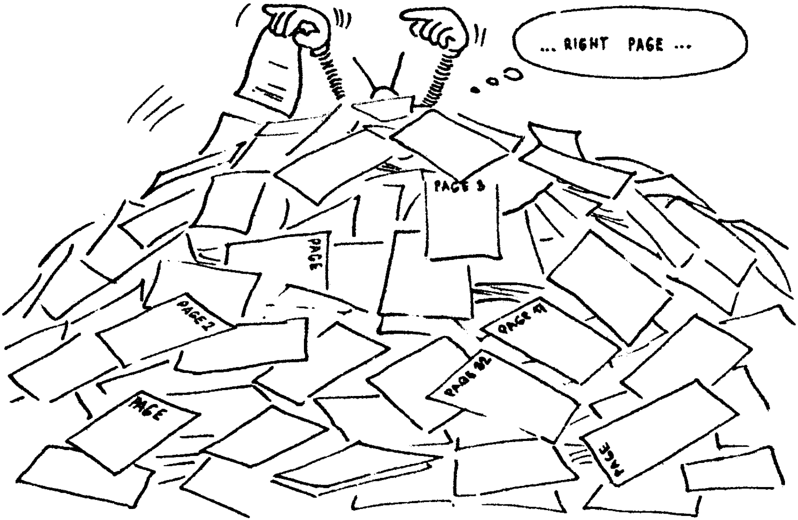
structures
articles.pdf
metadata
digital document

DSpace DML-CZ
9 digital library
end-users
http://dml.cz

archive
10 archivation

© DML-CZ, 2009

DML-CZ document engineering—data processing



DML-CZ now serves about *275,000 pages of math papers*.

Problems of *migration of existing workflows (born-digital, retro-digital) into the repository*. Negotiations with Google Scholar towards better visibility, indexing and search, and problems of copyright and sustainability issues, visualization, space and processing demands,….

Document engineering—digitization, digital library development

New tools and best practices for [meta]data processing

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

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

world of authors: LATEX, TEX notation of mathematics

world of applications/data exchange: XML, MathML

Verified and proven technologies (in DML-CZ)

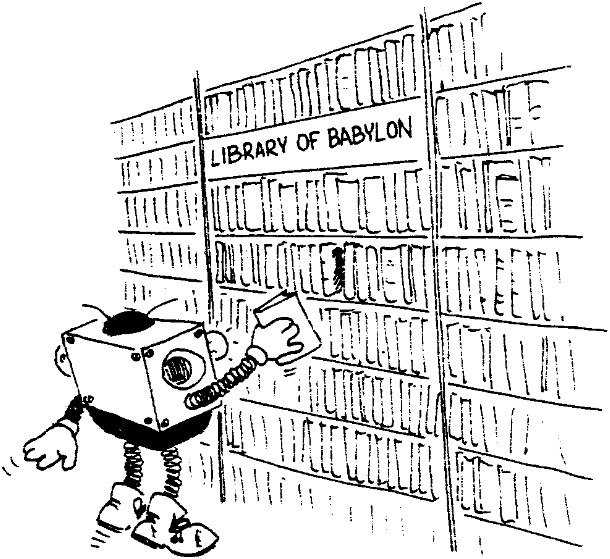
- scanned image processing and transformations, mathematical optical character recognition: OCR
- digital signature of PDF: pdfsign
- web-based long distance metadata editing: web application metadata editor
- optimization and recompression of PDF: downsizing of more than 50% without quality loss
- article similarity computations, demo.
- retroborn paper automated classification by MSC.

Verified and proven technologies (cont.)

- born-digital publishing system [for Archivum Mathematicum and other 4 journals] and conversions.
- retro-born-digital paper conversions and enhancements.
- data vizualization, browsing: adaptation of Visual Browser
- math retrieval: math formula indexing and search
- citation linking: CiteCrawl

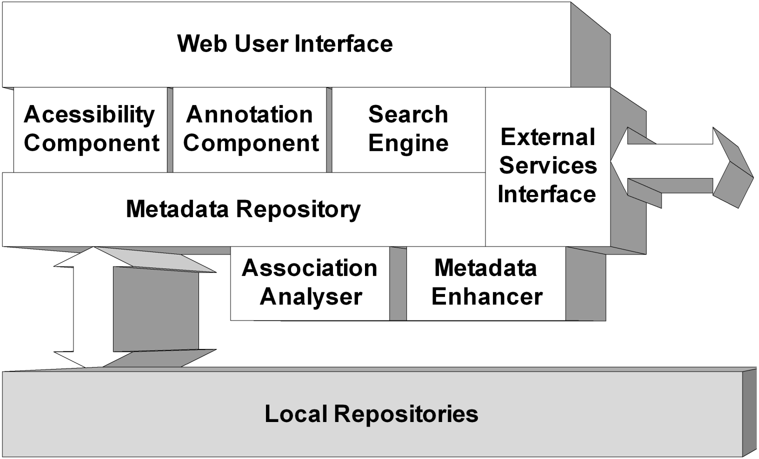
many open areas/challenges: multilingual retrieval?. MathML indexing,…

Bottom up processing towards EU or worldwide scale



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

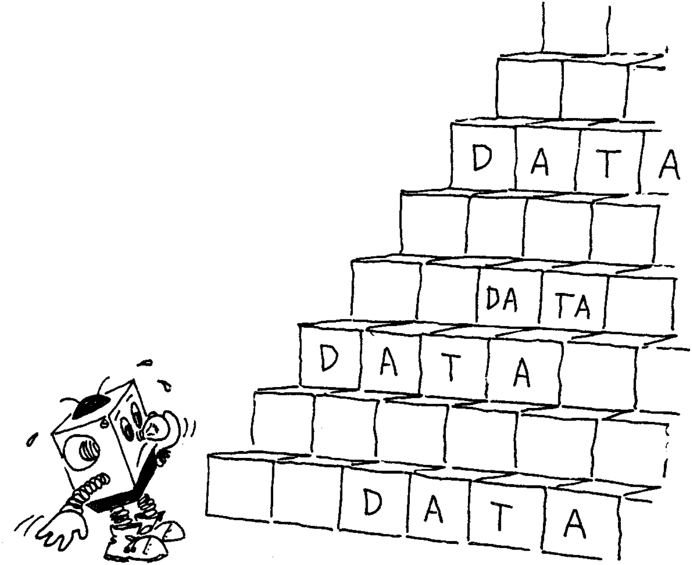
EuDML – data: legacy scientific math

- By 2013, EuDML should integrate *12 repositories*, have content from *200 integrated collections* (journals, book series, conference proceedings,…), more than *160,000 digital items* (papers, book chapters), *500,000 links between database objects*.
- It should be 'live' DL, having more than *1,000 users* contributing annotations, and more than *10,000 annotation* by 2013.
- Concept of *moving wall*: legacy data even from commercial publishers.

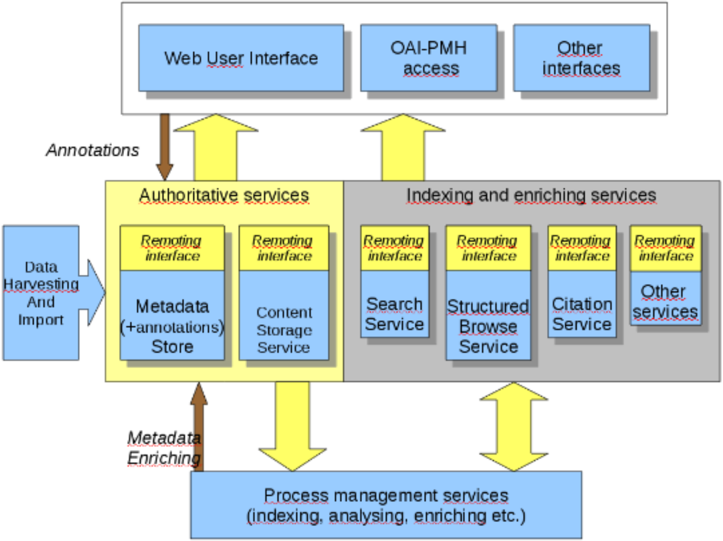
But how to actually implement it?

Experience from project partners from current digital library development.

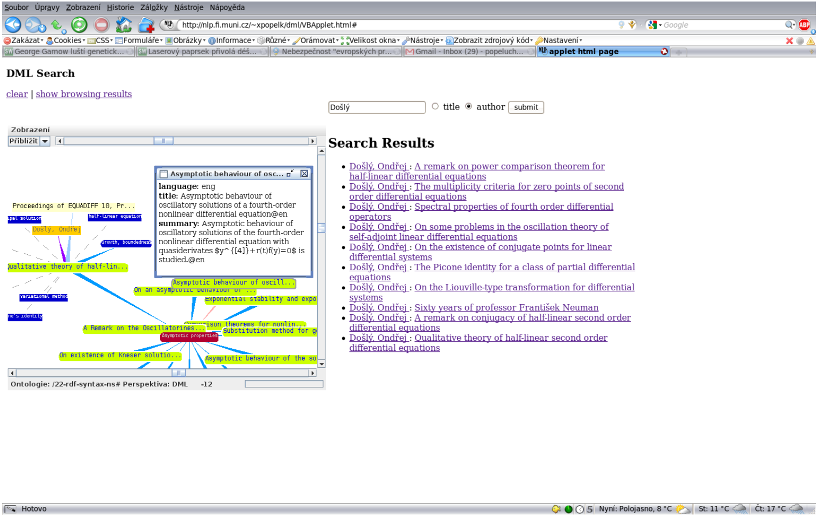
EuDML—from data collection to the virtual digital library



EuDML service based architecture



Alternative GUIs—visualization research


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 déš... Nebezpečnost "evropských pr... Gmail - Inbox (29) - popeluch... applet html page
Zobrazení
Přiblížit
Proceedings of EQUADIFF 10, Pr...
pal solution
half-linear equation
Došlý, Ondřej
growth, boundedness
ualitative 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...
Asymptotic properties
Substitution method for g
On existence of Kneser solutio...
Asymptotic behaviour of the so
Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -12
Hotovo
Nyní: Polojasno, 8 °C St: 11 °C Čt: 17 °C

DML Search

clear | show browsing results

Došlý ○ title ● author submit

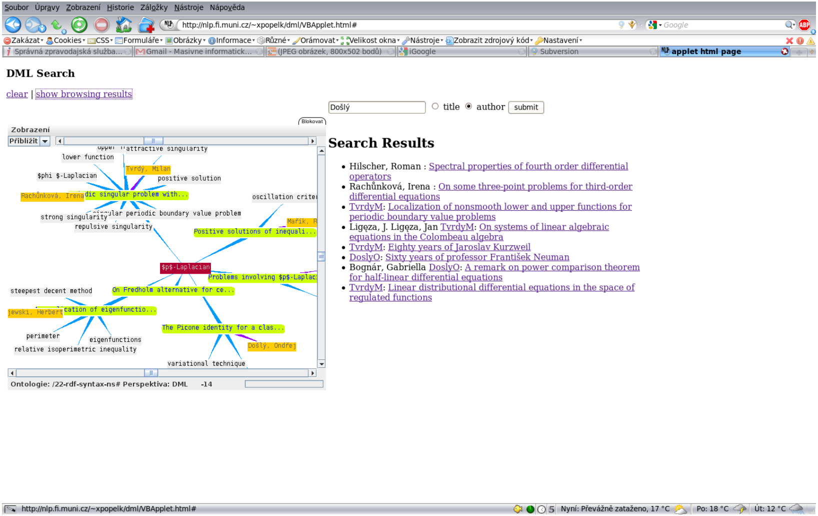
Asymptotic behaviour of osc...

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

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 development (DML-CZ)


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
singular periodic boundary value problem
strong singularity
Mařík, R
repulsive singularity
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

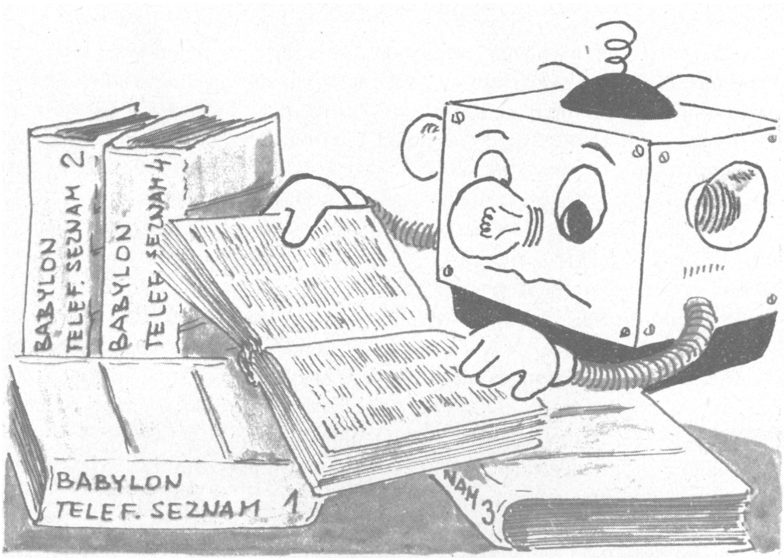
clear | show browsing results

Došlý ○ title ● author submit

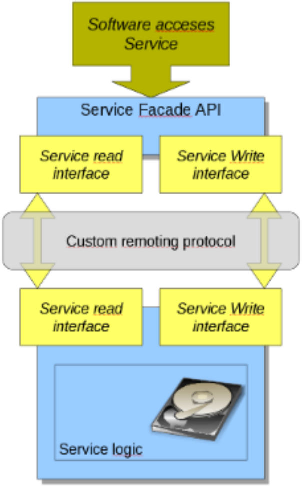
Search Results

- 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

EuDML document engineering—scalable tools development



EuDML service based architecture II



based on existing YADDA (used in Driver, Driver II) and REPOX (used in EuropeanaLocal, Telplus) projects – both are verified and mature platforms (implemented in Java)

math specifics needed to develop (TEX to MathML converter, math OCR, math in metadata support,…

MU offers: Metadata editor and other tools and expertize, mainly to be used in *WP7 Metadata Enhancements*

PDF Re-compression

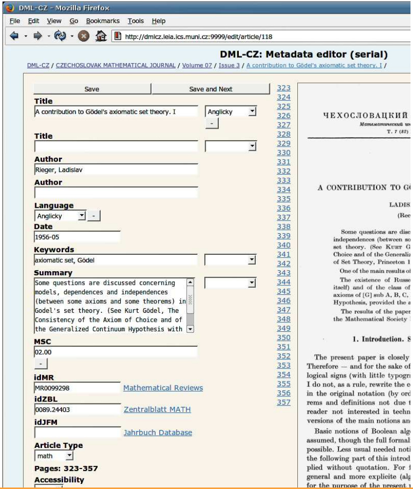
New tools developed (Radim Hatlapatka) to re-compress [bitonal] PDF files:

	Original PDF	After using PDF re-compressor	After using pdfsizeopt.py	After both
Size of whole PDF	100%	74.61%	50.02%	40.23%
Size of image and other objects	69.46%	37.14%	45.14%	35.36%

May be used for any PDF 1.4 (since Acrobat 5 released in 2001) file—JBIG2 compression.

Metadata Editor http://editor.dml.cz

Web-based client-server tool, developed (MU) for [meta]data import, editing, validation and checking. For testing, try <http://editor.dml.cz:9129>, admin/admin



DML-CZ - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://dmlcz.leia.ics.muni.cz:9999/edit/article/118

DML-CZ: Metadata editor (serial)

DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 07 / Issue 3 / A contribution to Gödel's axiomatic set theory. I /

Title
A contribution to Gödel's axiomatic set theory. I — Anglicky

Title

Author
Rieger, Ladislav

Author

Language
Anglicky

Date
1956-05

Keywords
axiomatic set, Gödel

Summary
Some questions are discussed concerning models, dependences and independences (between some axioms and some theorems) in Godel's set theory. (See Kurt Gödel, The Consistency of the Axiom of Choice and of the Generalized Continuum Hypothesis with

MSC
02.00

idMR
MR0099298 — Mathematical Reviews

idZBL
0089.24403 — Zentralblatt MATH

idJFM
Jahrbuch Database

Article Type
math

Pages: 323-357

Accessibility

323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357

ЧЕХОСЛОВАЦКИЙ
Математический ж
Т. 7 (82)

A CONTRIBUTION TO GÖ

LADIS
(Rec

Some questions are disc
independences (between so
set theory. (See Kurt G
Choice and of the Generaliz
of Set Theory, Princeton 1

One of the main results of

The existence of Russe
itself) and of the class of
axioms of [G] sub A, B, C,
Hypothesis, provided the a

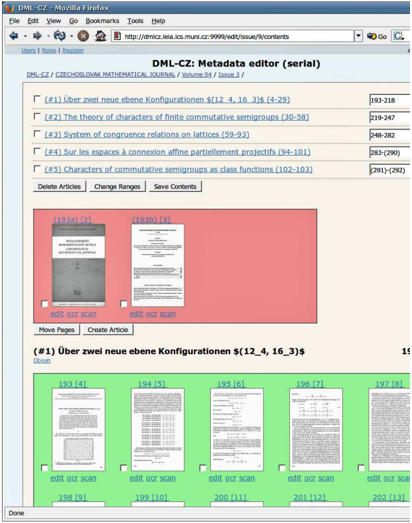
The results of the paper
the Mathematical Society

1. Introduction. S

The present paper is closely
Therefore — and for the sake of
logical signs (with little typogr
I do not, as a rule, rewrite the c
in the original notation (by ord
rems and definitions not due t
reader not interested in techn
versions of the main notions an

Basic notions of Boolean alg
assumed, though the full formal
possible. Less usual needed noti
the following part of this introd
plied without quotation. For f
general and more explicite (alg
for the purpose of the present

Metadata Editor localization (open source development)



DML-CZ - Mozilla Firefox

File Edit View Go Bookmarks Tools Help
http://dmlcz.leia.ics.muni.cz:9999/edit/issue/9/contents
Go
Users | Roles | Register

DML-CZ: Metadata editor (serial)

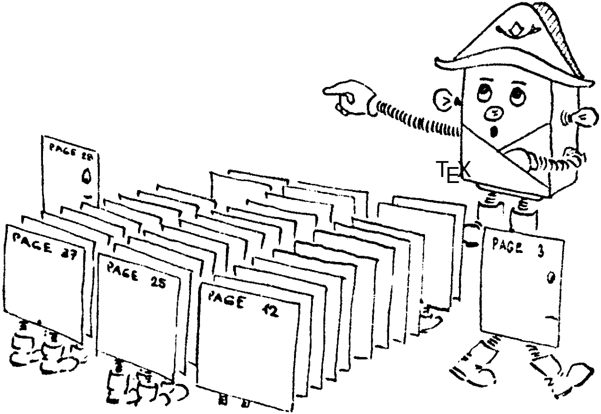
DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 04 / Issue 3 /

- ☐ (#1) Über zwei neue ebene Konfigurationen $(12_4, 16_3)$ (4-29) — 193-218
- ☐ (#2) The theory of characters of finite commutative semigroups (30-58) — 219-247
- ☐ (#3) System of congruence relations on lattices (59-93) — 248-282
- ☐ (#4) Sur les espaces à connexion affine partiellement projectifs (94-101) — 283-(290)
- ☐ (#5) Characters of commutative semigroups as class functions (102-103) — (291)-(292)

(#1) Über zwei neue ebene Konfigurationen $(12_4, 16_3)$

Obsah

Yes, you can!



Summary

Publishing and retrodigitization of mathematics poses many challenges, some yet to solve.

DML-CZ: up and running, most open access.

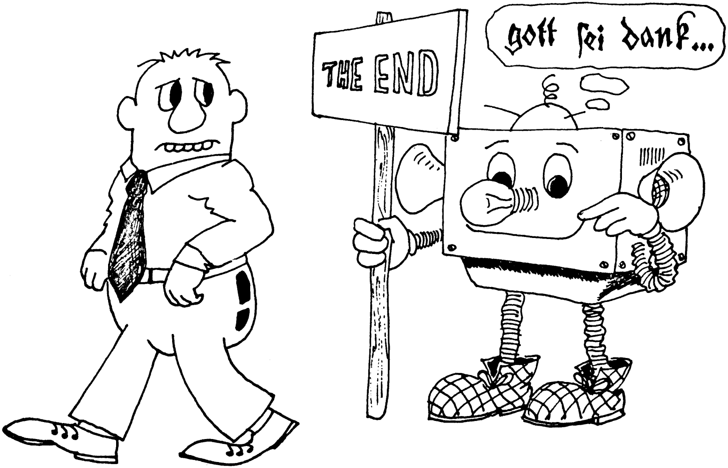
EuDML: work in progress, some DML-CZ experience and tools developed to be (re)used.

Invitation: EuDML meeting with content providers in Prague, October 15th: math publishers wanted to cooperate.

Further reading: DML workshop series proceedings (Openly accessed in DML-CZ).

Comments, cooperation offers welcome! We are Open and AND-minded (no black&white thinking :-).

End of the talk



Questions?

References, links

- DML-CZ team.
  Materials about DML-CZ, project publications [online, cit. 2010-08-24].
  <http://project.dml.cz/documents.html>.
- EuDML team.
  EuDML project info [online, cit. 2010-08-24].
  <http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503>
- EuDML team.
  EuDML webpage [online, cit. 2010-08-24].
  <http://eudml.eu/>.
- EuDML at MU team.
  EuDML at MU project info [online, cit. 2010-08-24].
  <http://nlp.fi.muni.cz/projekty/eudml/> or <http://www.muni.cz/research/projects/10067>.