

Towards a Digital Mathematics Library: from DML-CZ to EuDML

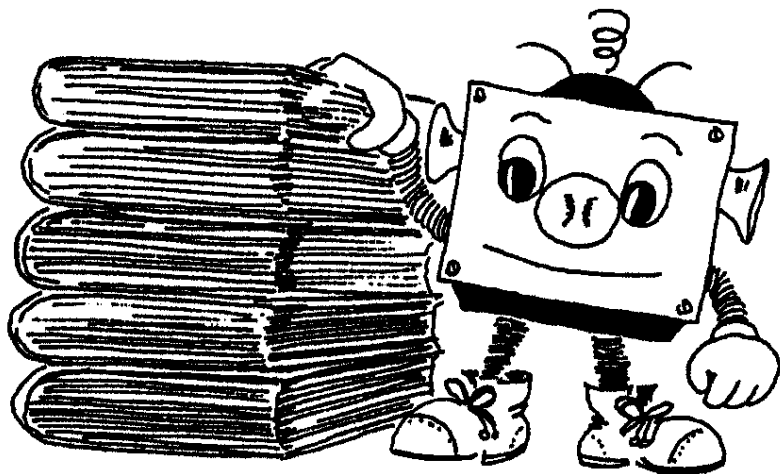
Petr Sojka

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

CEEDI 2010, Sarajevo, BiH, May 19th, 2010, 9:30 a.m.



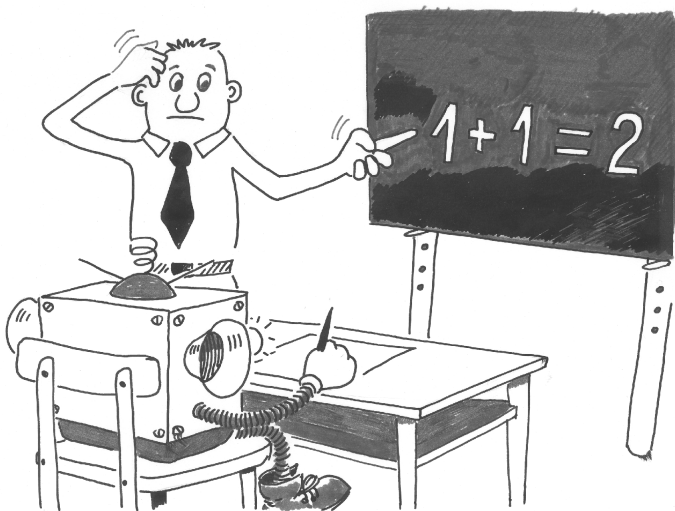
From paper to digital processing



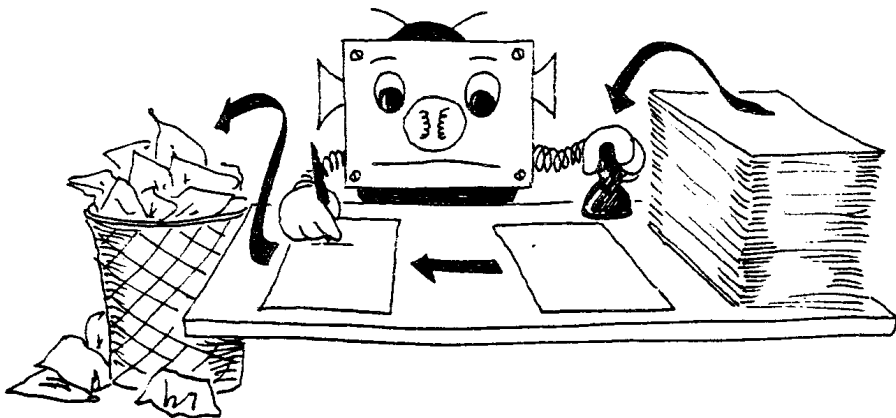
Information overload



Information overload in mathematics



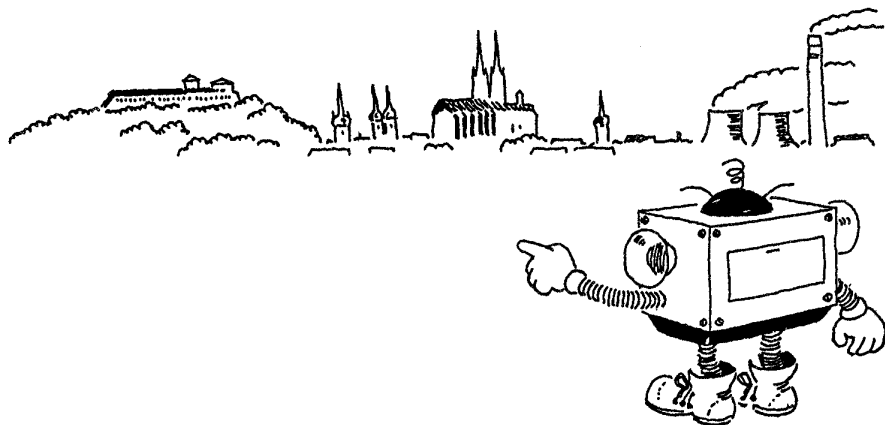
Document engineering—from paper to digital workflow



Document engineering—digitization, digital library development



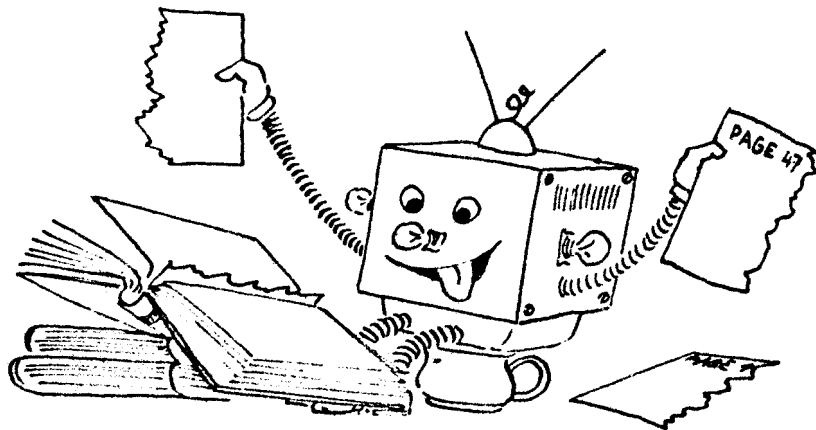
Bottom up processing—local (Brno, CZ) document engineering



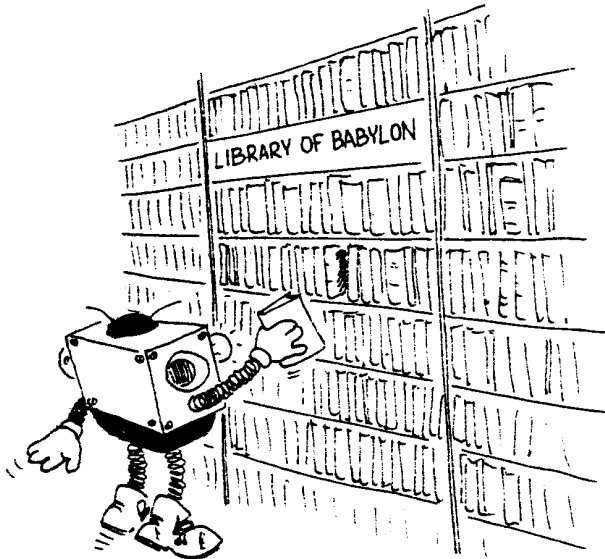
DML-CZ document engineering—data processing



DML-CZ document engineering—tools



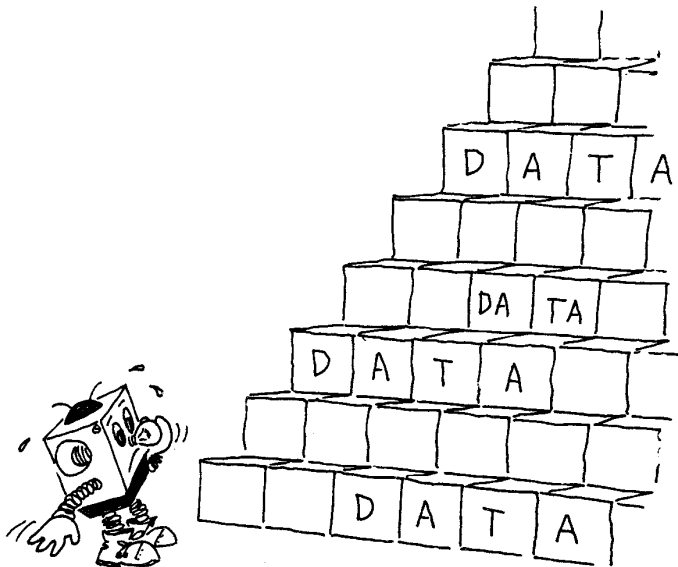
Bottom up DML processing towards EU or worldwide scale



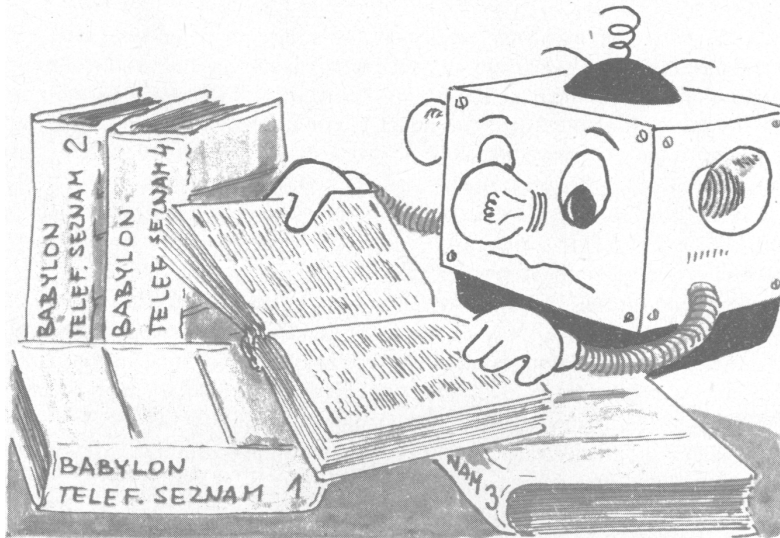
European Digital Mathematics Library



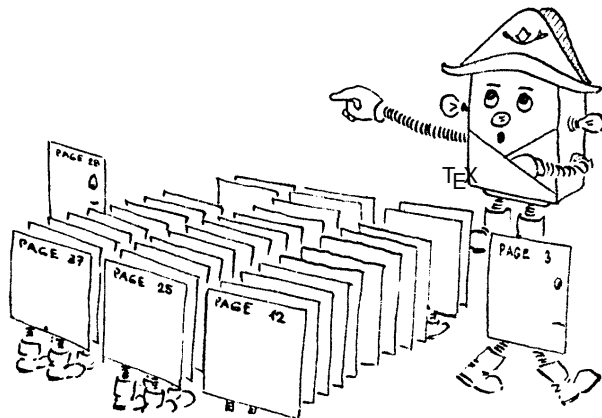
EuDML—from local data collections to the virtual digital library



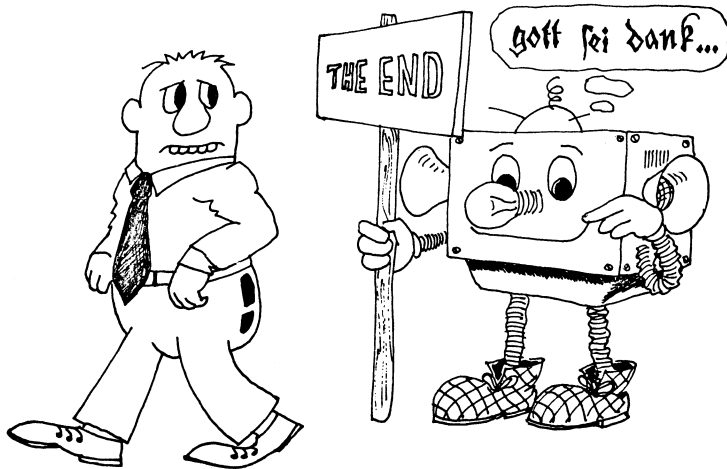
EuDML document engineering—scalable tools development



Yes, you can!



End of talk overview



Vision of WDML

At the beginning there was a vision of all mathematical knowledge, *peer reviewed and verified* (100,000,000 pages) on one spot and in the digital form.

Progress of IT, cheap 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 (100,000,000 \$ requested). Application was not successful.

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

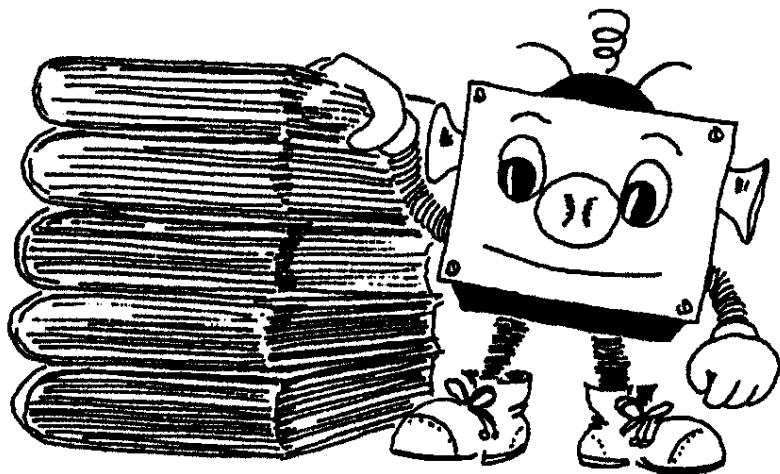
EuDML

(MU and MU AV). **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 a basis serve current DML repositories as DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,...(from local repositories bottom-up to build the final thing).

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

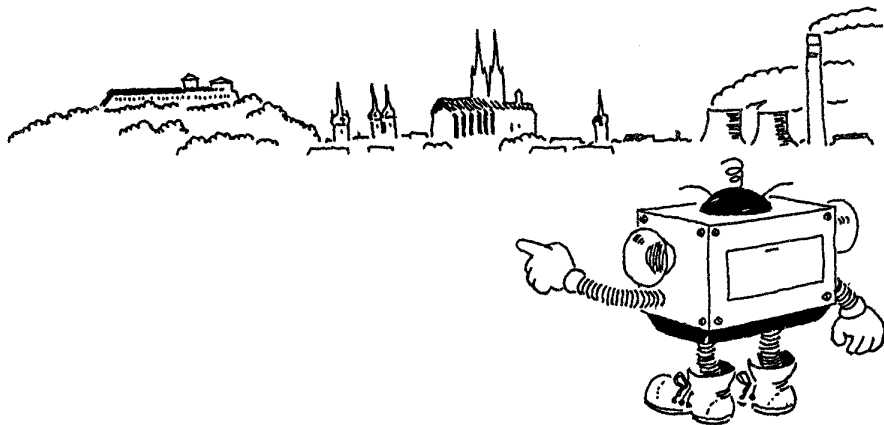
- Czech Academy of Sciences grant (program Information Society) 2005–2009, *full* (retro)digitization of 50,000 pages of mathematical literature per year, 8M CZK in total.
- Research part: **1)** gradual enhancement of the digital material by ‘knowledge enhancing’ filters on markup-rich XML data. **2)**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

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



Take care!



Information overload—what is there in DML-CZ?



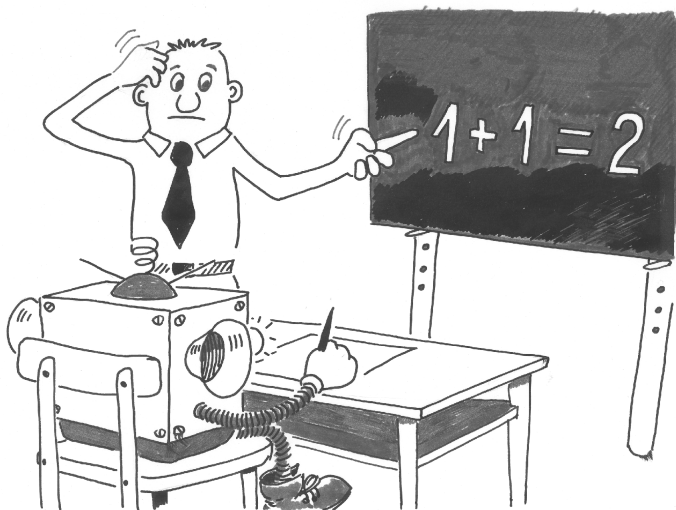
The approach used in DML-CZ

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

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

Our framework integrates workflow for the articles scanned from a paper (*math OCR*), for documents from retro-born digital period (data available in some type of electronic form) and for born-digital ones.

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



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 $K \in \mathfrak{U}$ and it follows from theorem 24 that

$$P(K, v) = \int_K f(x) \, dx. \quad (89)$$

The functional determinant T of the mapping $\varphi = \varphi^{-1}$ fulfils the relation $T(\varphi(x)) \cdot \det M(x) = 1$, so that

$$\int_K f(x) \, dx = \int_{\hat{K}} f(\varphi(y)) \cdot |T(y)| \, dy = \int_{\hat{K}} \hat{f}(y) \, dy. \quad (90)$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y) \, dy$, 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 poluspořádaném prostoru, *Casopis pro řést. 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, *Casopis pro řést. mat.*, 41 (1956), 79–82.
- [6] Ян Маржик (Jan Mařík): Заметки 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.)

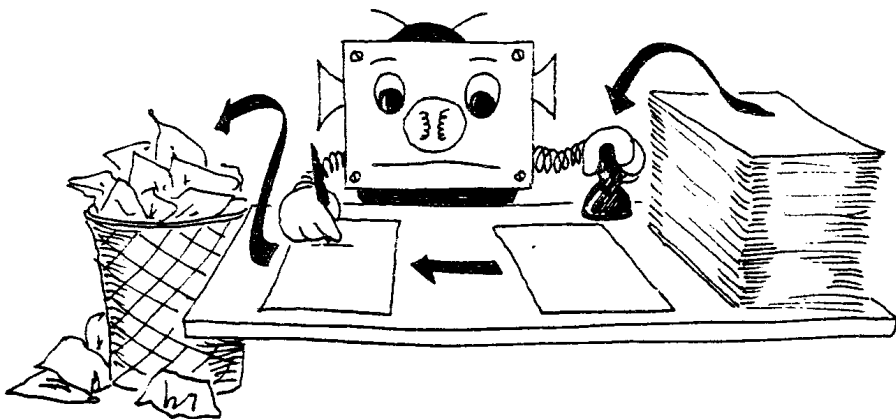
Пусть m — натуральное число; пусть E_m — m -мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^m \frac{\partial v_i(x)}{\partial x_i} \, dx$, где v_1, \dots, v_m — многочлены такие, что $\sum_{i=1}^m v_i^2(x) \leq 1$ для всех $x \in A$. Пусть \mathfrak{U} — система всех ограниченных измеримых множеств A , для которых $\|A\| < \infty$. Теорема 18 тогда утверждает: Пусть $A \in \mathfrak{U}$; пусть D — граница множества A . Тогда на системе \mathfrak{B} всех борелевских подмножеств множества D существует мера μ и на



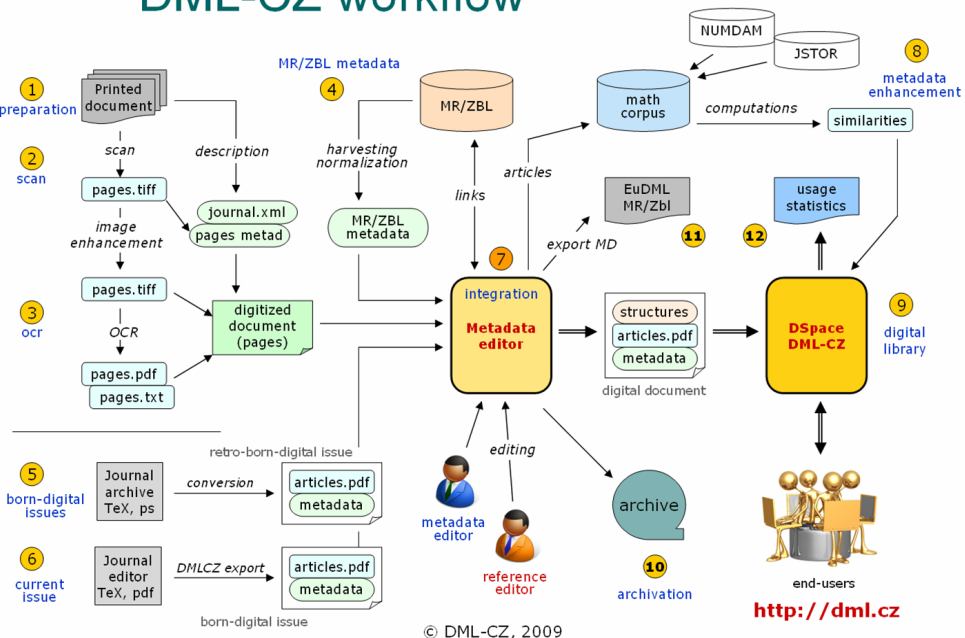
ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

Document engineering—from paper to digital *workflow*



DML-CZ workflow



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



MU expertize in [meta]data processing

Data heterogeneity, plethora of formats, validation and conversions:

retro-digital period: scanning, geometrical transformations (BookRestorer),
OCR (FineReader, InftyReader), two-layer PDF

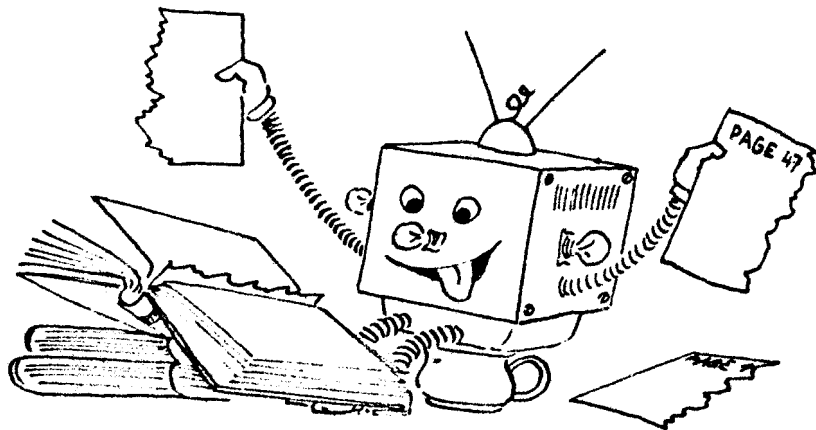
retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap
fonts of low resolution

born-digital period: typesetting by $\text{T}_{\text{E}}\text{X}$ with export of [meta]data into digital
library

world of authors: $\text{L}_{\text{A}}\text{T}_{\text{E}}\text{X}$, $\text{T}_{\text{E}}\text{X}$ notation of mathematics

world of applications/data exchange: XML, MathML

DML-CZ document engineering—tools and challenges



Typesetting of papers and cover pages

- Xe \LaTeX , Charis SIL (many alphabets and characters in author names, cyrillic,...)
- `\usepackage{pdfpages}` or `pdftk` (annotations).
- \TeX source generated from XML metadata (XSLT a perl), after validation of metadata full regeneration automatic (pipe of 7+ steps) `meta.xml` \rightarrow `item.xml` \rightarrow `item.tex` \rightarrow `item.pdf` \rightarrow ...

meta.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
  <number>1</number>
  <status>completed</status>
  <title lang="fre">Sur quelques applications des dispersions
  <title lang="eng">On some applications of cetral dispersions
  <author id="BoruvO" order="1">Borůvka, Otakar</author>
  <language>fre</language>
  <msc>34C10</msc>
  <idMR>MR0197823</idMR>
  <idZBL>Zbl 0151.10804</idZBL>
  <idUlrych>19650001</idUlrych>
  <category>math</category>
  <range>7-26</range>
  <range_pages>1-20</range_pages>
  <access>true</access>
</article>
```

item.tex

```

\newlength{\vsx} \vsx=148mm
\newlength{\vsy} \vsy=205mm
\newcommand\toptitle{Archivum Mathematicum}
\newcommand\maintitle{Sur quelques applications des dispersion
\newcommand\mainauthors{Otakar Borůvka}
\newcommand\PURL{http://dml.cz/dmlcz/104576}
\documentclass{dmlcz}
\begin{document}
\copyrightholders{$\copyright$ Masaryk University, 1965}
\bibtoks{\textit{Archivum Mathematicum},
Vol. 1 (1965), No. 1, 1--20}

\dmltitlepage
\dmlpage{../page/0007}{121mm}{193mm}
\dmlpage{../page/0008}{118mm}{189mm}
\dmlpage{../page/0009}{118mm}{186mm}
...
\end{document}

```

Verified and proven technologies (in DML-CZ)

- scanned image processing and transformations (with BookRestorer) (BT Pulkrábek).
- mathematical optical character recognition: OCR (MT Panák, Mudrák, BT Vystrčil).
- digital signature of PDF: pdfsign (BT Peter Bočák).
- web-based long distance metadata editing: web application metadata editor (ÚVT MU Mirek Bartošek, Martin Šárfy, Vlasta Krejčíř, Petr Kovář); to be localized by Miha Filej for EuDML.
- optimization of PDF: pdfopt (from ghostscript), pdfsizeopt.py (by Peter Szabó).
- similarity article computations (research with Radim Řehůřek), demo.

Verified and proven technologies (cont.)

- retroborn paper automated classification by MSC (Radim Řehůřek).
- data vizualization, browsing: adaptation of Visual Browser (MT Zuzana Nevěřilová), will be offered for EuDML GUI.
- PDF recompression using JBIG2: application based on jbig2enc/leptonica (BT Radim Hatlapatka), offered for EuDML.
- math retrieval: math formula indexing and search (MT Vítězslav Dostál, BT Martin Liška, BT Peter Mravec) – possibly to offer for EuDML.
- citation linking: CiteCrawl (BT Lukáš Lalinský)

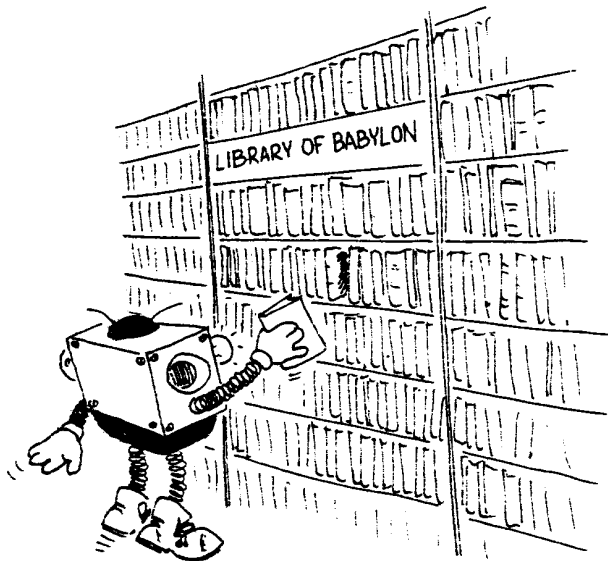
Verified and proven technologies (cont.)

- born-digital publishing system [for Archivum Mathematicum and other 4 journals] and conversions (BT&MT Michal Růžička), offered for EuDML.
- retro-born-digital paper conversions and enhancements (Michal Růžička), dtto.

open areas/challenges: multilingual retrieval?. MathML indexing using manatee/bonito?, math common sense?

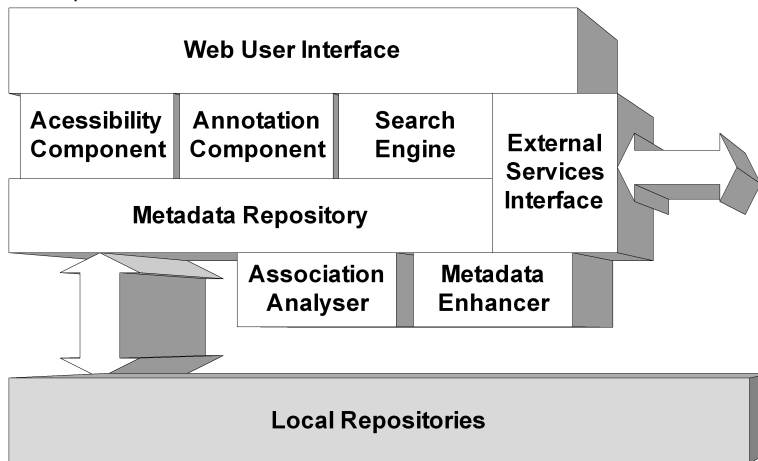
cooperation [problems above, `fixfont`, citation crawling, math OCR a indexing] “wanted!” to develop, enhance and offer for EUDML.

Bottom up processing towards EU or worldwide scale

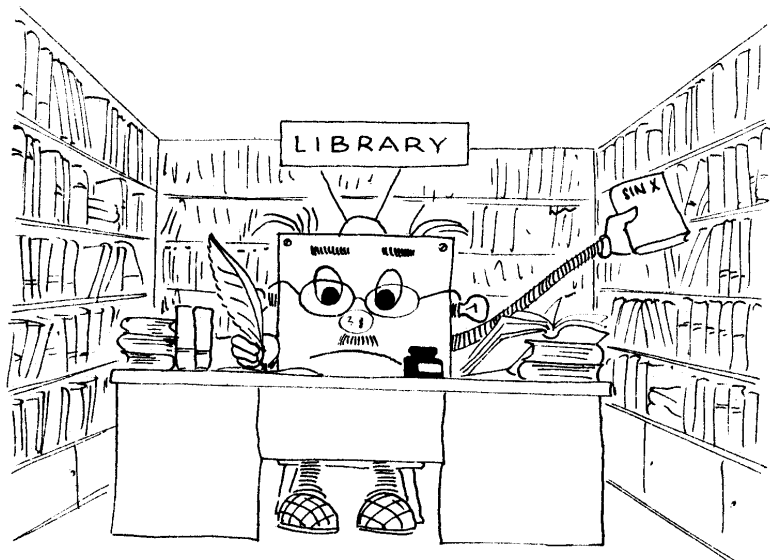


EuDML as a virtual library portal

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



European Digital Mathematics Library



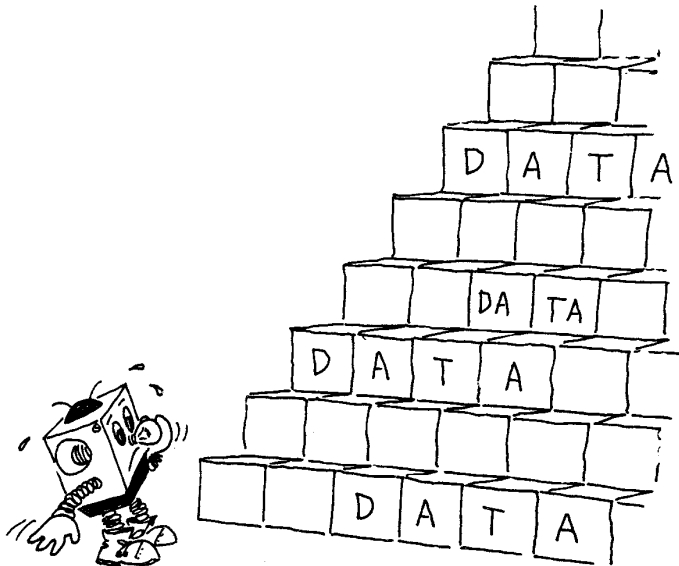
EuDML – data: legacy scientific math

- 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.

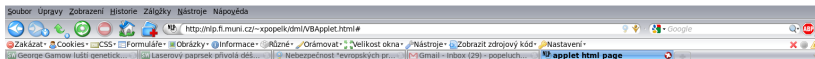
But how to actually implement it?

Experience from project partners from current digital library development.

EuDML—from data collection to the virtual digital library

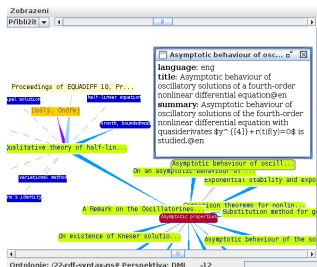


MU: alternative GUI—vizualisation research



DML Search

[clear](#) | [show browsing results](#)

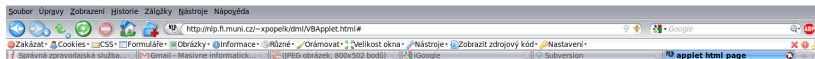


Došlý ☐ title ☒ author ☐ submit

Search Results

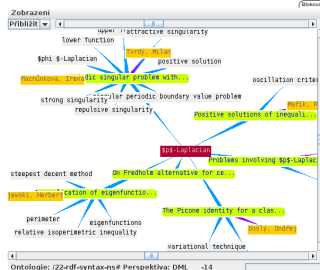
- Došlý Ondřej : A remark on power comparison theorem for half-linear differential equations
- Došlý Ondřej : The multiplicity criteria for zero points of second order differential equations
- Došlý Ondřej : Spectral properties of fourth order differential operators
- Došlý Ondřej : On some problems 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

MU: Visual Browser development (DML-CZ)



DML Search

```
clear | show browsing results
```

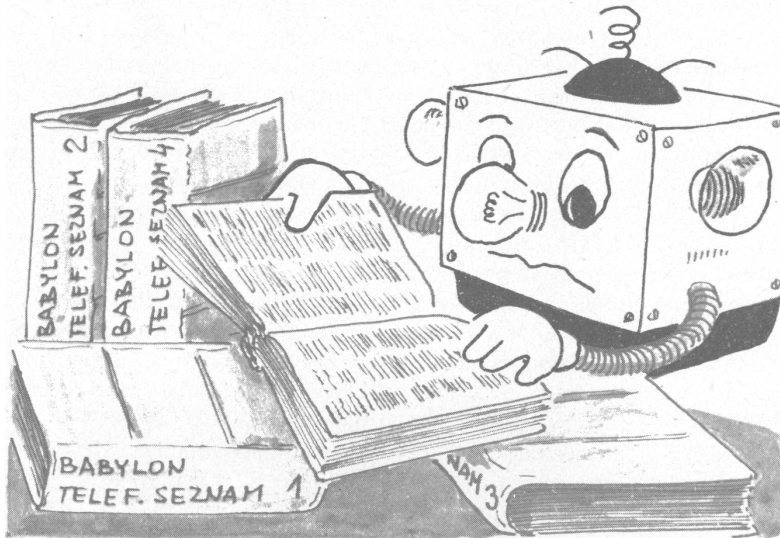
 ☐ title ☒ author 

Search Results

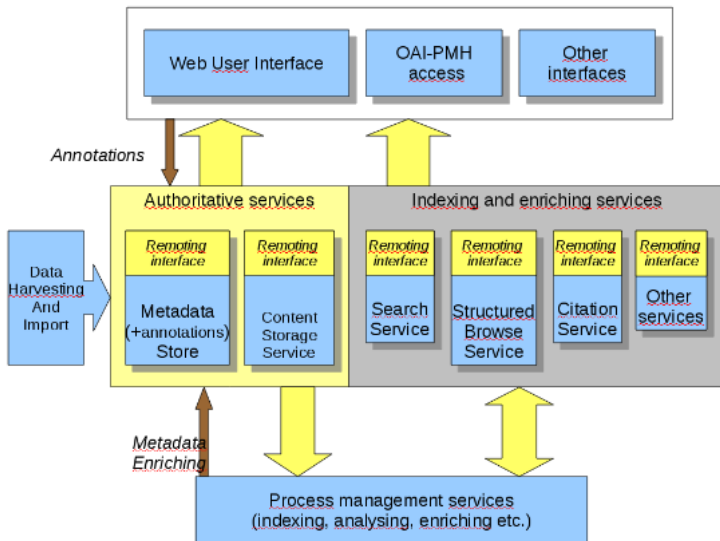
- Hilscher, Roman : **Spectral properties of fourth order differential operators**
- Rachůnková, Irena : **On some three-point problems for third-order differential equations**
- Tsyrdyk : **Localization of nonsmooth lower and upper functions for periodic boundary value problems**
- Ligeza, J. Ligeza, Jan Tsyrdyk : **On systems of linear algebraic equations in the Colombeau algebra**
- Tsyrdyk : **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**
- Bognár, Gabriella : **On the asymptotic behavior of solutions of nonlinear differential equations in the space of regulated functions**



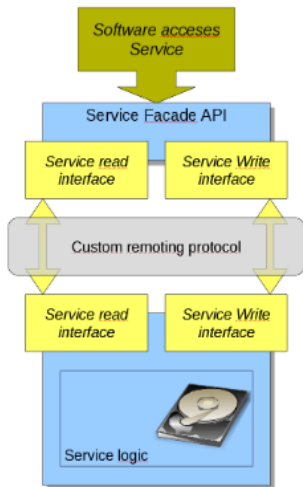
EuDML document engineering—scalable tools development



EuDML service based architecture



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 (T_EX to MathML converter, math OCR, math in metadata support,...

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

PDF Re-compression

New tools developed (with 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 (ICS MU) from scratch (Ruby) 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://editor.dml.cz:9129/>

DML-CZ: Metadata editor (serial)

DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 52 / Issue 3 / A Contribution to Gödel's axiomatic set theory. I /

Save Save and Next

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

Author
Friger, Ladislav

Language
Anglicky

Date
1955-05

Keywords
axiomatic set, Gödel

Summary
Some questions are discussed concerning models, dependences and independences (between some axioms and some theorems) in Gödel'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
j089.24403 [Zentralblatt MATH](#)

IDJFM
Jahrbuch Database

Article Type

ЧЕХОСЛОВАККИЙ МАТЕМАТИЧЕСКИЙ ЖУРНАЛ
Математический журнал
Т. 52 (1955)

A CONTRIBUTION TO GÖDEL'S AXIOMATIC SET THEORY. I
LADISLAV FRIGER
(1955-05)

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

1. Introduction.

The present paper is closely connected with the paper of Gödel [1] and for the sake of logical signs (with little typographical errors) I do not, as a rule, rewrite the original notation (by order) and definitions not due to Gödel are not interested in technical details of the main notions and basic notions of Boolean algebra are assumed, though the full formalism is possible. Less usual needed notations are given in the appendix.

Metadata Editor localization (Miha Filej)

DML-CZ - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://dmicr.iesl.icm.muni.cz:9999/edit/issue/9/contents

Users | Basic | Register


DML-CZ: Metadata editor (serial)

DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 04 / Issue 2 /

<input type="checkbox"/> (#1) Über zwei neue ebene Konfigurationen \$(12_4, 16_3)\$ (4-29)	193-218
<input type="checkbox"/> (#2) The theory of characters of finite commutative semigroups (30-38)	219-247
<input type="checkbox"/> (#3) System of congruence relations on lattices (59-93)	248-282
<input type="checkbox"/> (#4) Sur les espaces à connexion affine partiellement projectifs (94-101)	283-(290)
<input type="checkbox"/> (#5) Characters of commutative semigroups as class functions (102-103)	(291)-(292)


[Delete Articles](#) [Change Ranges](#) [Save Contents](#)

(193a) [2]



[edit ocr scan](#)

(193b) [3]




[edit ocr scan](#)

[Move Pages](#) [Create Article](#)

(#1) Über zwei neue ebene Konfigurationen \$(12_4, 16_3)\$ 11


[Group](#)

193 [4]




[edit ocr scan](#)

198 [9]




[edit ocr scan](#)

194 [5]




[edit ocr scan](#)

199 [10]




[edit ocr scan](#)

195 [6]




[edit ocr scan](#)

200 [11]




[edit ocr scan](#)

196 [7]




[edit ocr scan](#)

201 [12]




[edit ocr scan](#)

197 [8]



[edit ocr scan](#)

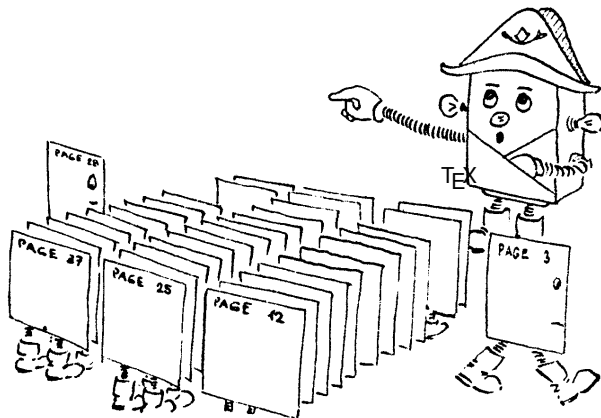
202 [13]



[edit ocr scan](#)

Done

Yes, you can!



Summary

EuDML: work in progress, based on DML-CZ experience and tools developed at FI and ICS during last 6 years.

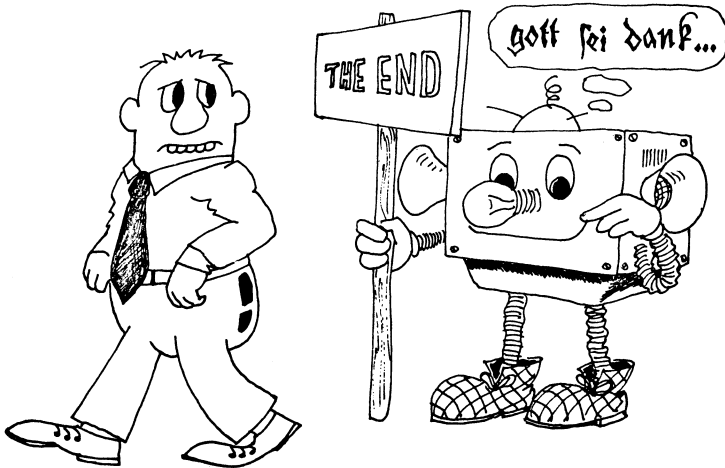
Current activities: WP4&5 meeting in Warsaw, 4 papers for forthcoming DML 2010 workshop, accepted paper at LREC 2010 (with Radim Řehůřek)...

Next activities: EuDML general meeting in Paris in July (c/o CICM 2010, DML 2010, July 7th–8th), WP7 technical meeting, tool implementation.

DML 2010 organization: <<http://www.fi.muni.cz/~sojka/dml-2010.html>>

Comments, cooperation offers welcome!

End of the talk



Questions?

References, links



DML-CZ team.

Materials about DML-CZ, project publications [online, cit. 2010-05-19].

<<http://project.dml.cz/documents.html>>.



EuDML team.

EuDML project info [online, cit. 2010-05-19].

<http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503>



EuDML team.

EuDML webpage [online, cit. 2010-05-19].

<<http://eudml.eu/>>.



EuDML at MU team.

EuDML at MU project info [online, cit. 2010-05-19].

<<http://nlp.fi.muni.cz/projekty/eudml/>> or <<http://www.muni.cz/research/projects/10067>>.