

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

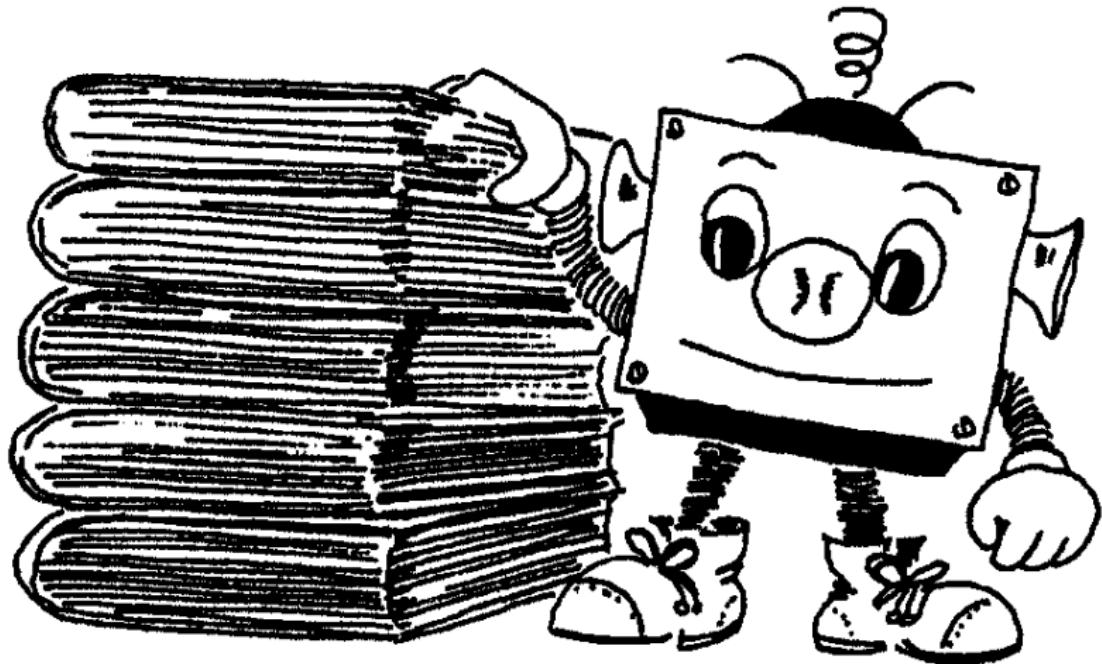
Petr Sojka

<sojka@fi.muni.cz> (FI MU, Brno)

FI MU, Brno, CZ, Informatics Colloquium, May 4th, 2010, 2 p.m.



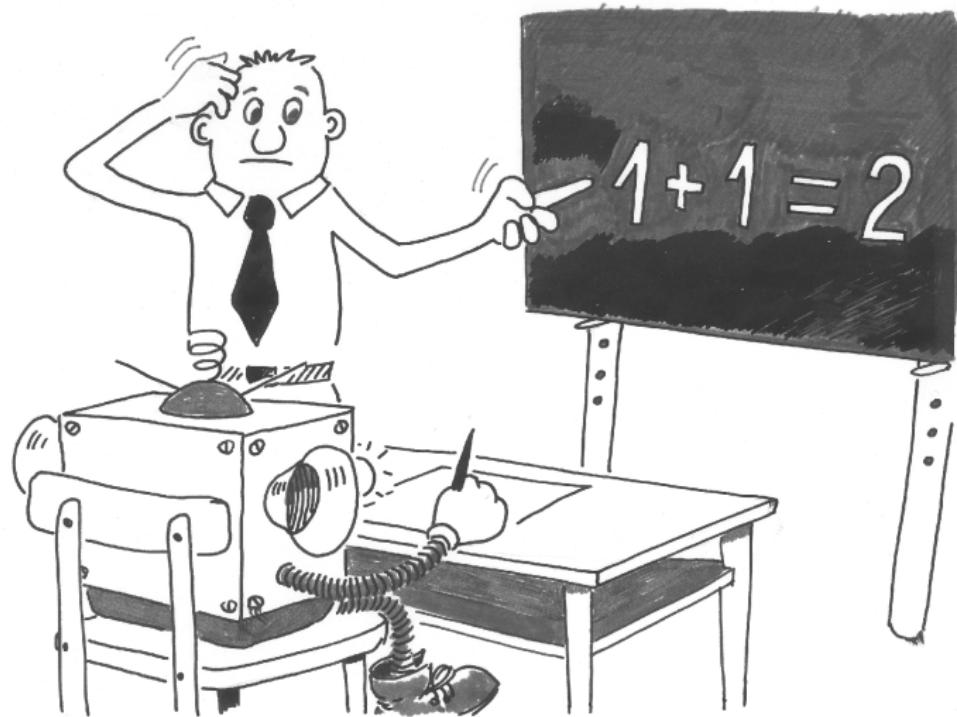
# From paper to digital processing



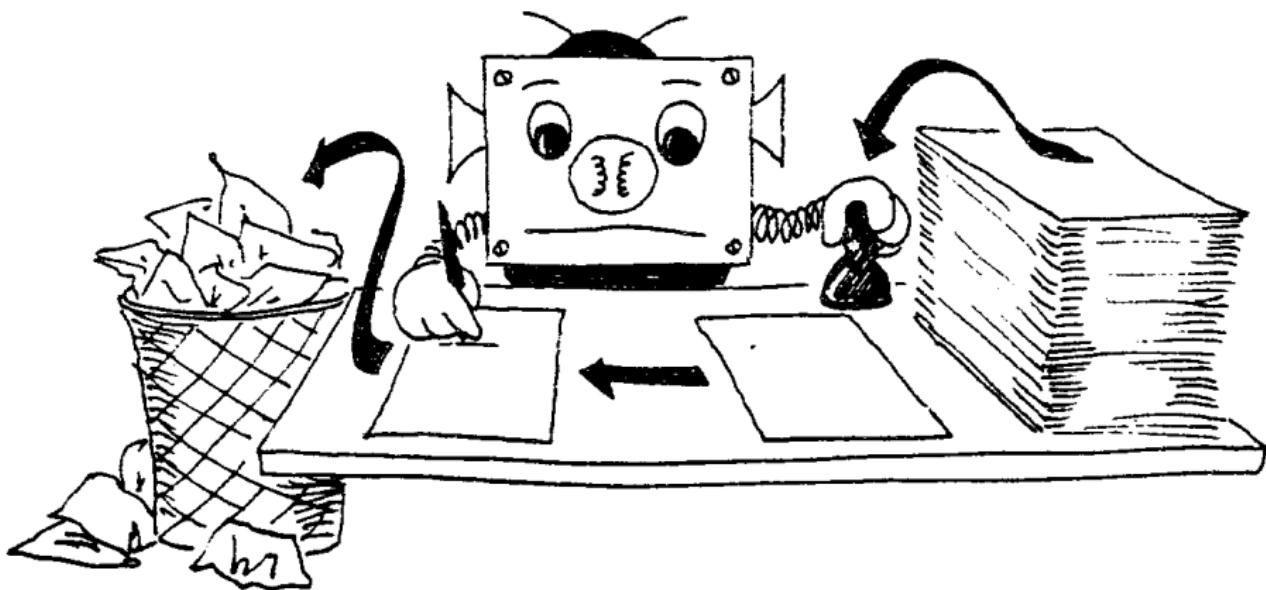
# Information overload



# Information overload in mathematics



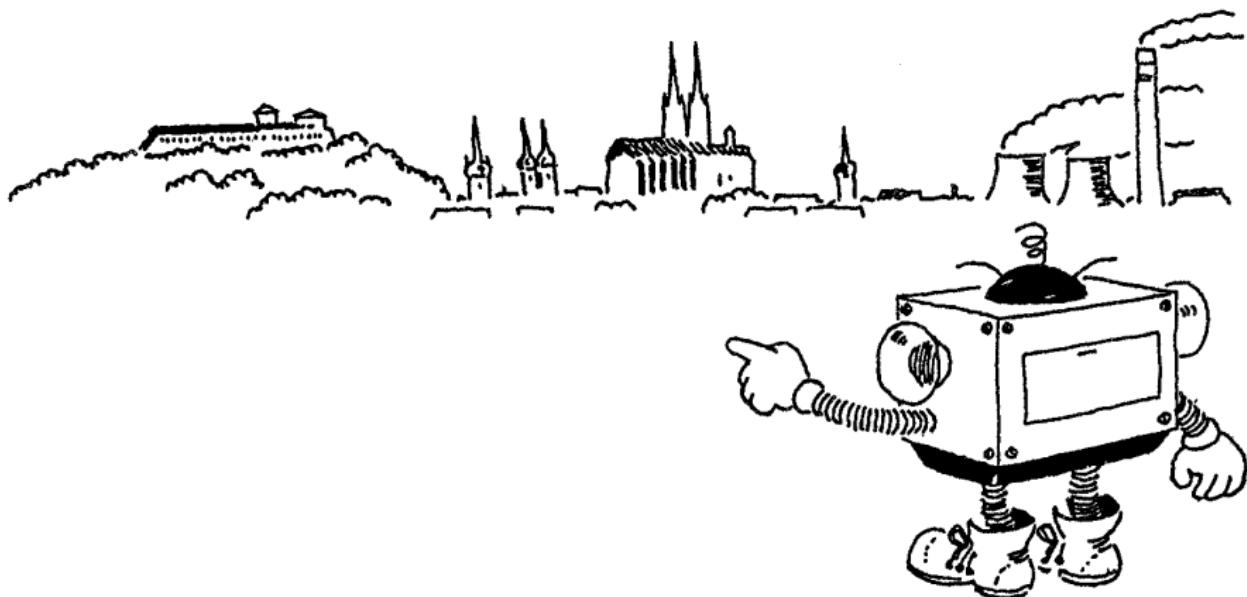
# Document engineering—from paper to digital workflow



# Document engineering—digitization, digital library development



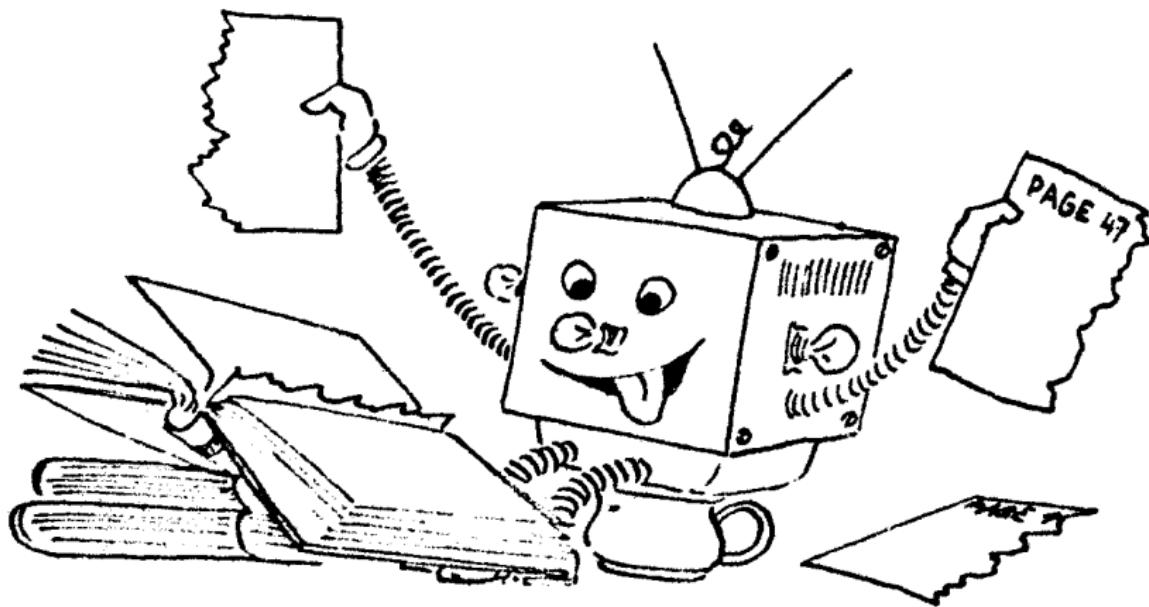
# Bottom up processing—local (Brno, CZ) document engineering



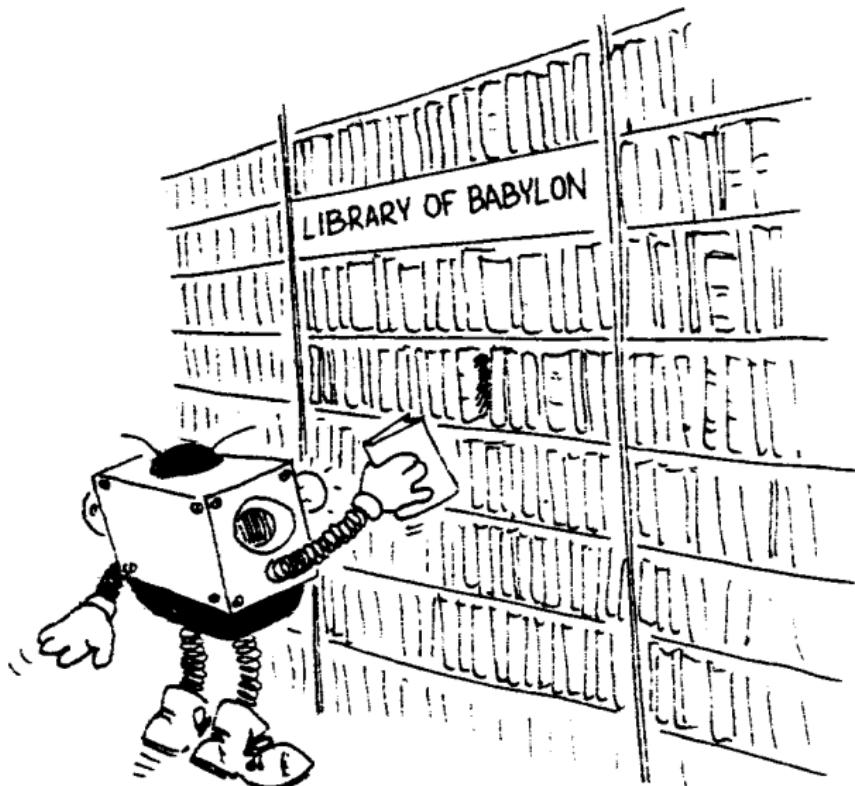
# DML-CZ document engineering—data processing



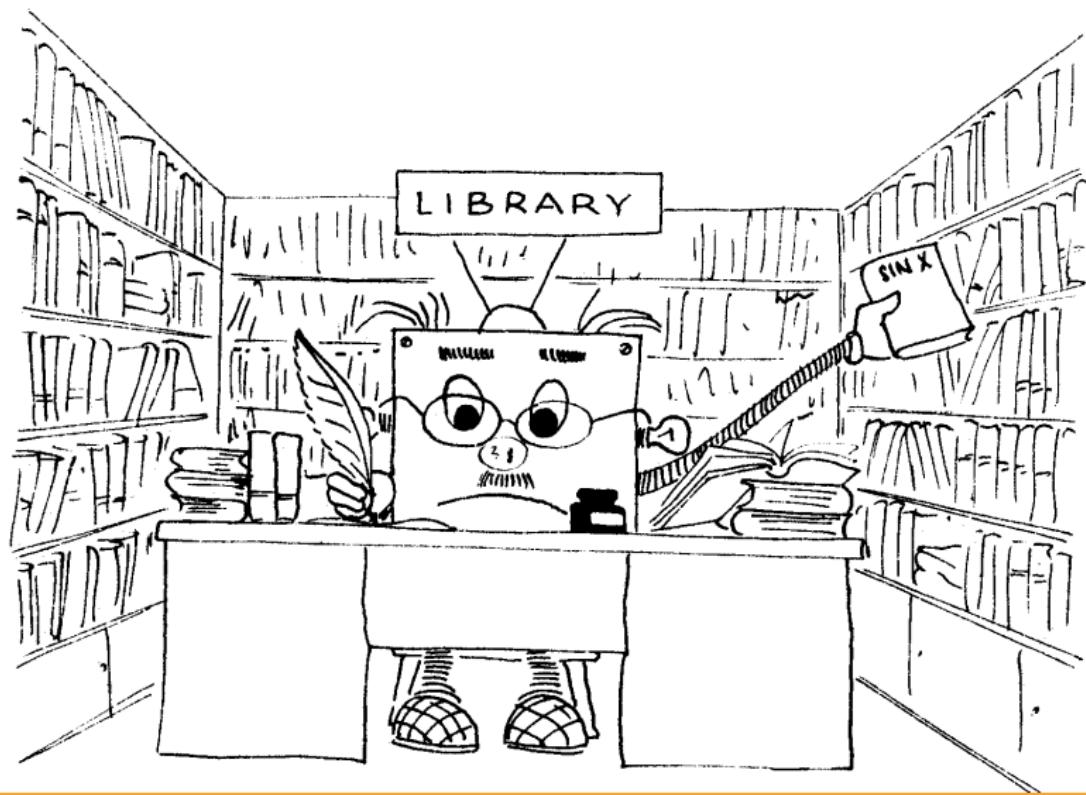
# DML-CZ document engineering—tools



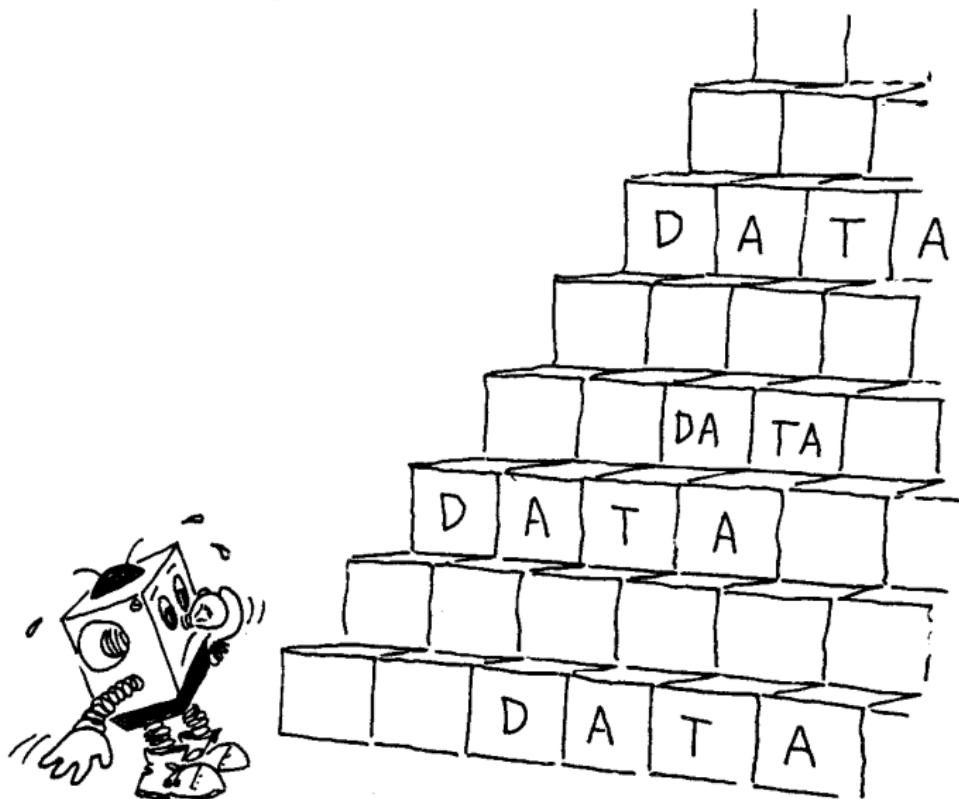
# Bottom up DML processing towards EU or worldwide scale



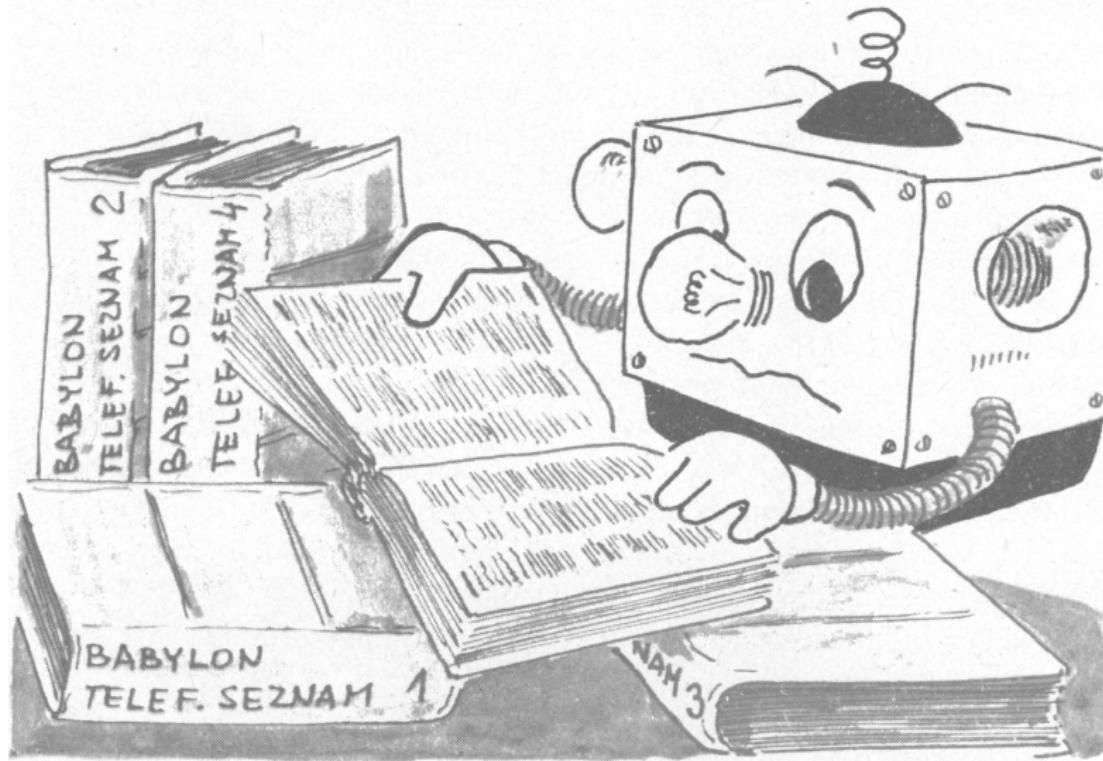
# European Digital Mathematics Library



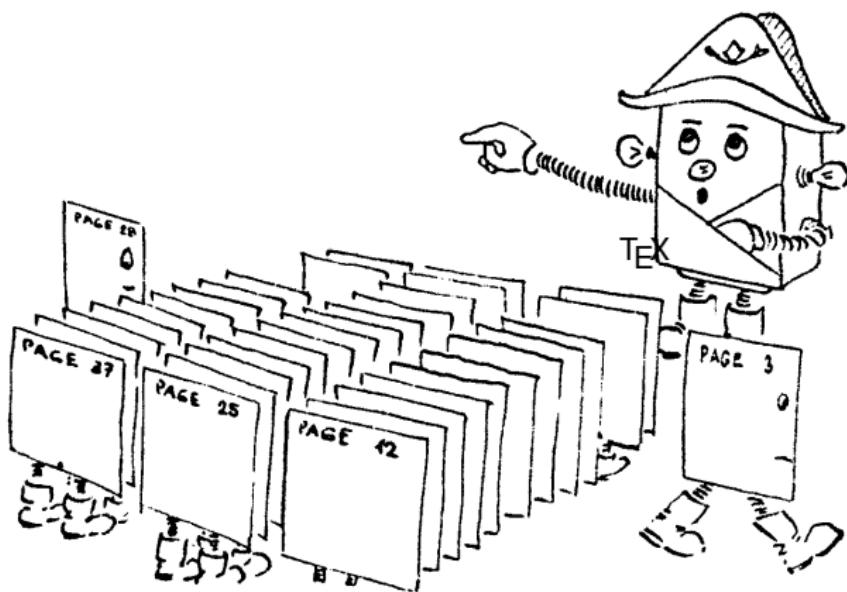
# EuDML—from data collection to virtual digital library



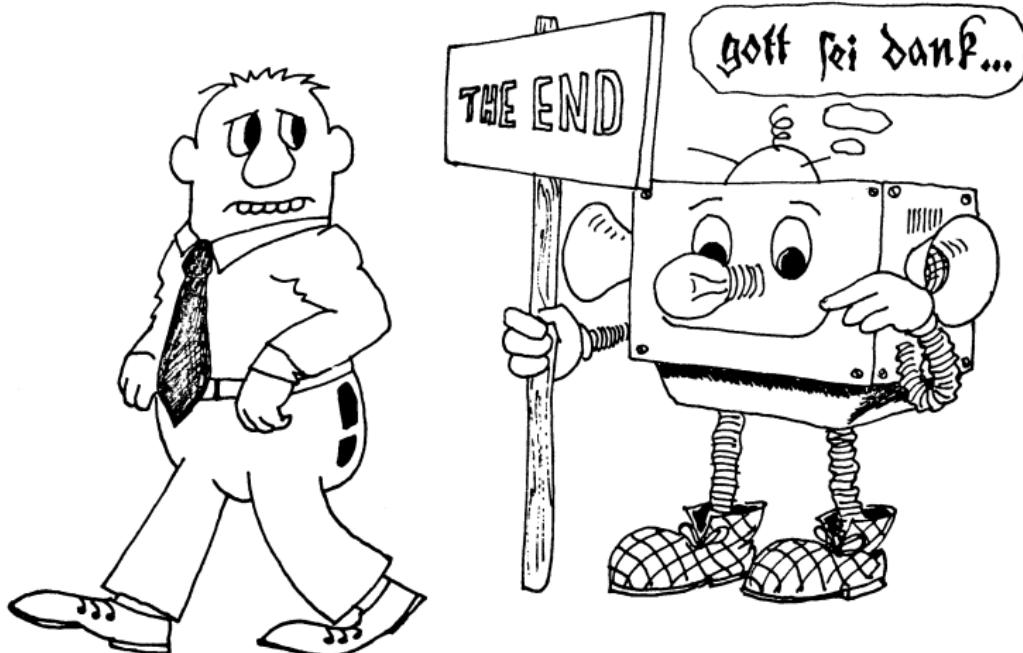
# EuDML document engineering—scalable tools development



# Yes, you can!



## End of talk overview



# Vision of WDML/EuDML

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.

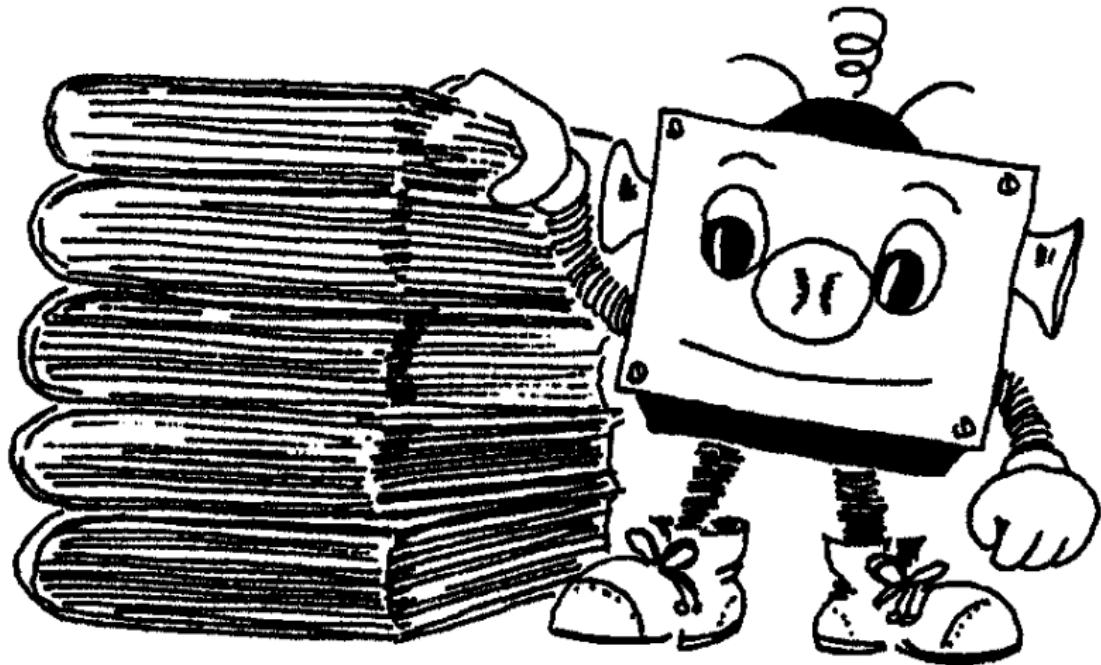
It starts to happen, but slowly: three year EU project EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).



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

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).

# 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?



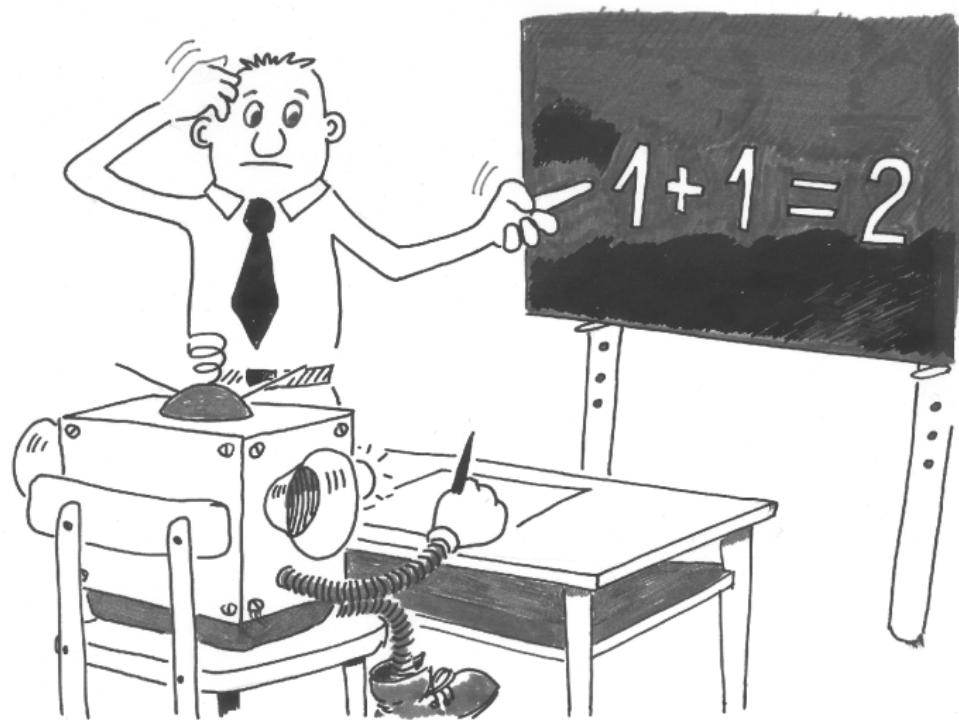
## The approach used in DML-CZ

A successfully built repository (e.g. set of *workflows*) needs a *coordinated* effort of *librarians*, *IT specialists* and representatives of users – *content specialists*.

*Design, technical and political decisions* behind building the *Czech Digital Mathematics Library DML-CZ* (<<http://dml.cz>>) in the context of other successful thematic community projects (PubMed Central, ADS, SCOAP3 and planned 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  $\varphi = \varphi^{-1}$  fulfills 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 polospásaděném prostoru, Časopis pro pěst. mat., 79 (1954), 3–40.
- (4) Илья Марцик (Jan Mařík): Представление функционала в виде интеграла, Чехословацкий мат. журнал, 5 (80), 1955, 467–487.
- (5) J. Mařík: Plný integrál, Časopis pro pěst. mat., 81 (1956), 79–82.
- (6) Илья Марцик (Jan Mařík): Заметка к теории поверхности интеграла, Чехословацкий мат. журнал, 6 (81), 1956, 387–400.
- (7) S. Saks: Theory of the integral, New York.

## Резюме

### ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЦИК (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

$$\sum_{i=1}^m v_i^2(x) \leq 1 \text{ для всех } x \in A.$$

Пусть  $\mathfrak{U}$  — система всех ограниченных измеримых множеств  $A$ , для которых  $\|A\| < \infty$ . Теорема 18 тогда утверждает:

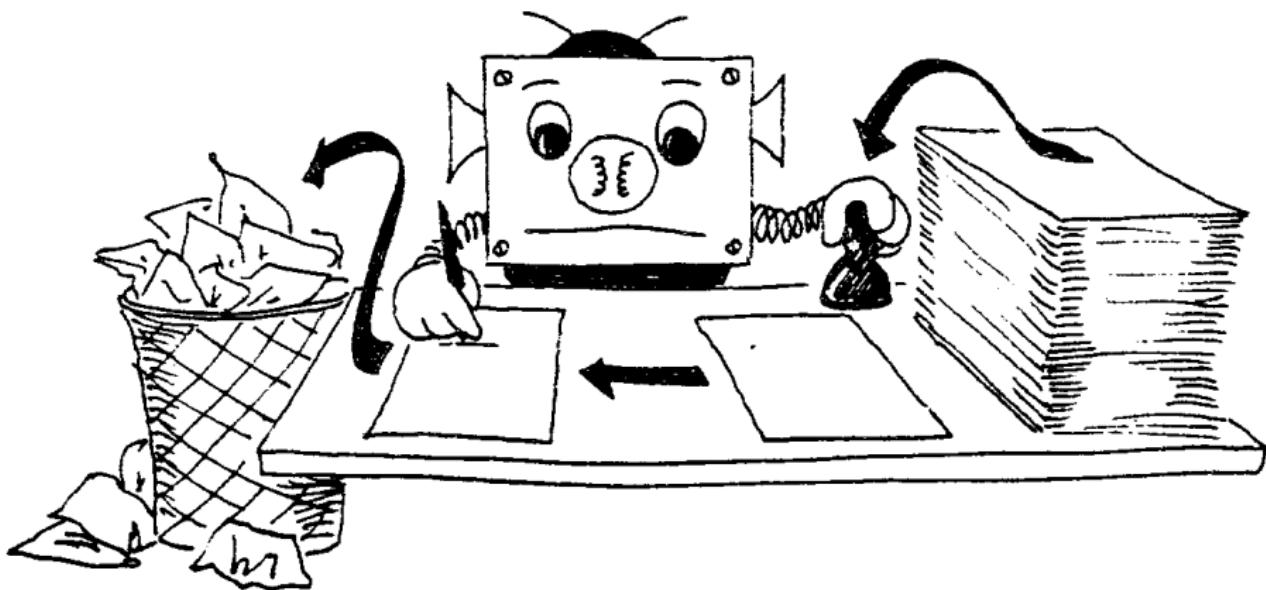
Пусть  $A \in \mathfrak{U}$ ; пусть  $D$  — граница множества  $A$ . Тогда на системе  $\mathfrak{U}$  всех борелевских подмножеств множества  $D$  существует мера  $r$  и на



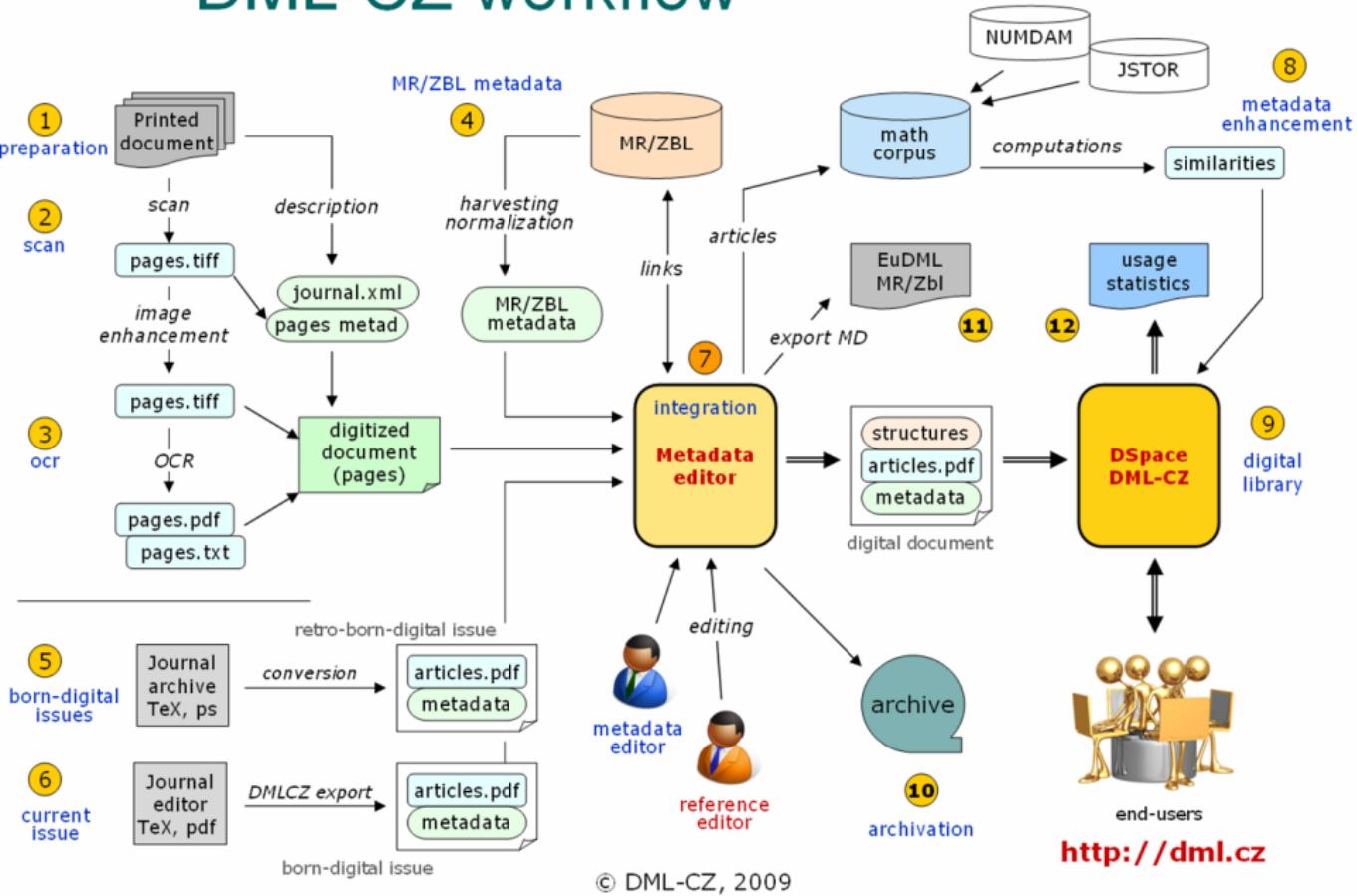
ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

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

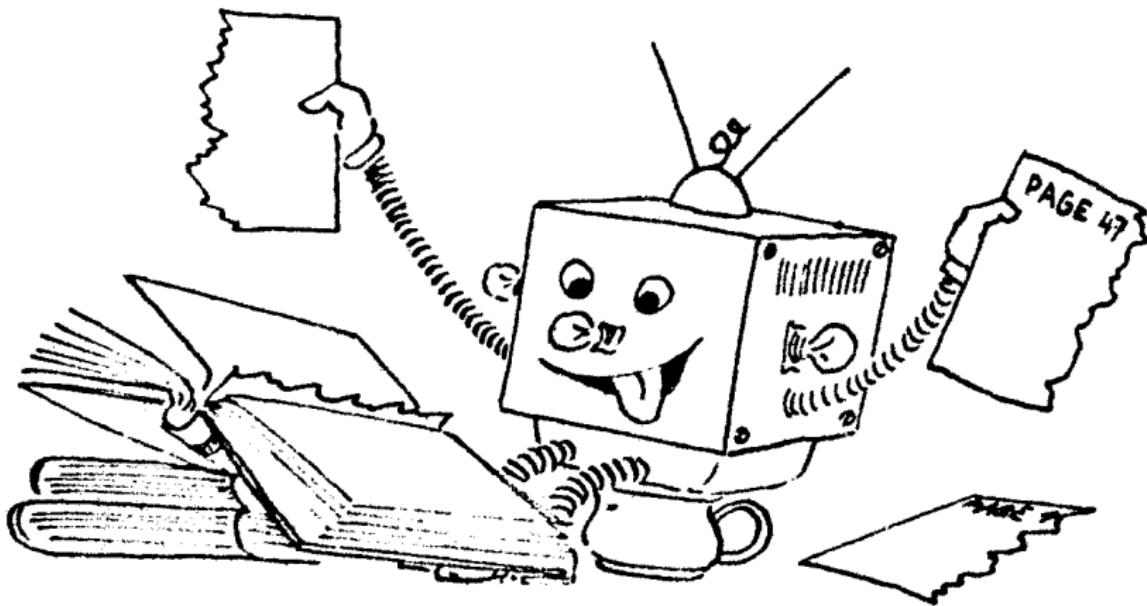
retro-born-digital period: not complete .tex or .dvi data, bad formats, bitmap  
fonts of low resolution

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

world of authors:  $\text{\LaTeX}$ ,  $\text{\TeX}$  notation of mathematics

world of applications/data exchange: XML, MathML

# DML-CZ document engineering—tools and challenges



# Typesetting of papers and cover pages

- Xe<sup>L</sup>A<sub>T</sub>E<sub>X</sub>, Charis SIL (many alphabets and characters in author names, cyrillic,...)
- \usepackage{pdfpages} or pdftk (annotations).
- T<sub>E</sub>X source generated from XML metadata (XSLT a perl), after validation of metadata full regeneration automatic (pipe of 7+ steps) meta.xml → item.xml → item.tex → item.pdf → ...

# 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 central 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}

\dmtitlepage
\dmtpage{..../page/0007}{121mm}{193mm}
\dmtpage{..../page/0008}{118mm}{189mm}
\dmtpage{..../page/0009}{118mm}{186mm}

...
\end{document}
```

## Verified and proven technologies (in DML-CZ)

- scanned image processing and transformations (with BookRestorer) (BP Pulkrábek).
- mathematical optical character recognition: OCR (DP Panák, Mudrák, BP Vystrčil).
- digital signature of PDF: pdfsign (BP 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 visualization, browsing: adaptation of Visual Browser (DP Zuzana Nevěřilová), will be offered for EuDML GUI.
- PDF recompression using JBIG2: application based on jbig2enc/leptonica (BP Radim Hatlapatka), offered for EuDML.
- math retrieval: math formula indexing and search (DP Vítězslav Dostál, BP Martin Liška, BP Peter Mravec) – possibly to offer for EuDML.
- citation linking: CiteCrawl (BP Lukáš Lalinský)

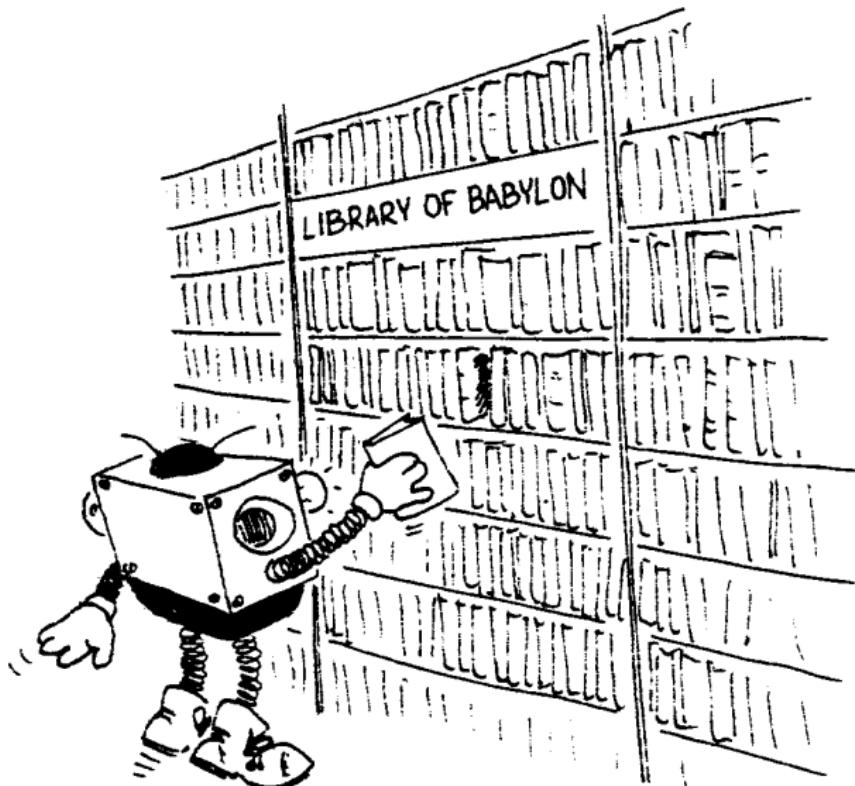
## Verified and proven technologies (cont.)

- born-digital publishing system [for Archivum Mathematicum and other 4 journals] and conversions (BP&DP 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?

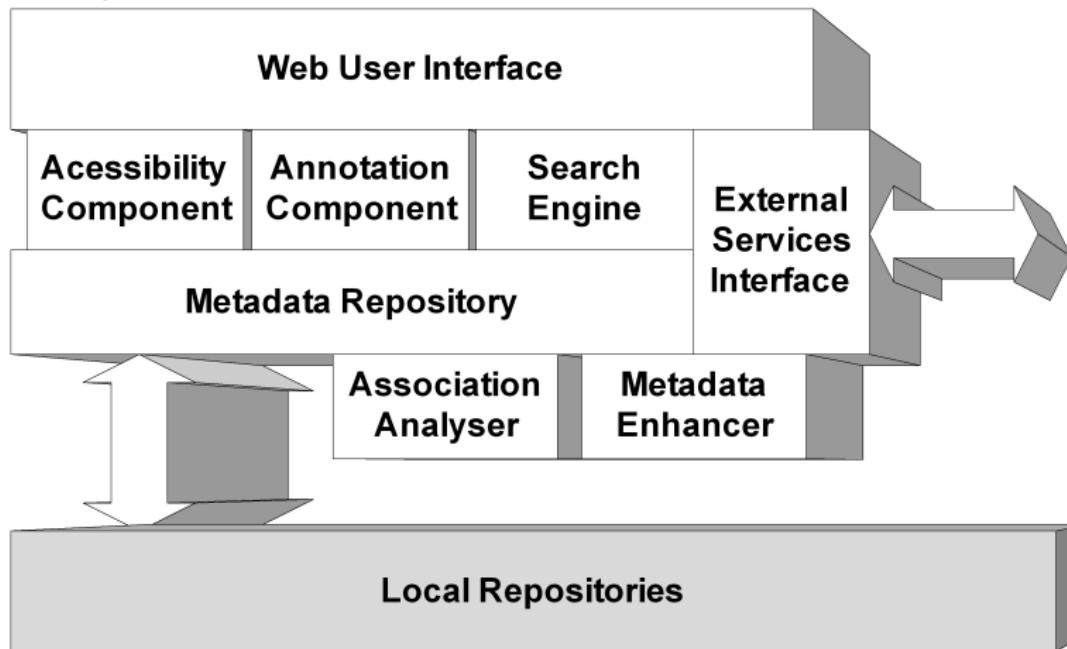
cooperation [problems above, fixfont, citation crawling, math OCR a indexing (BP/DP/PhD)] “wanted!” to develop, enhance and offer for EUDML.

# 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



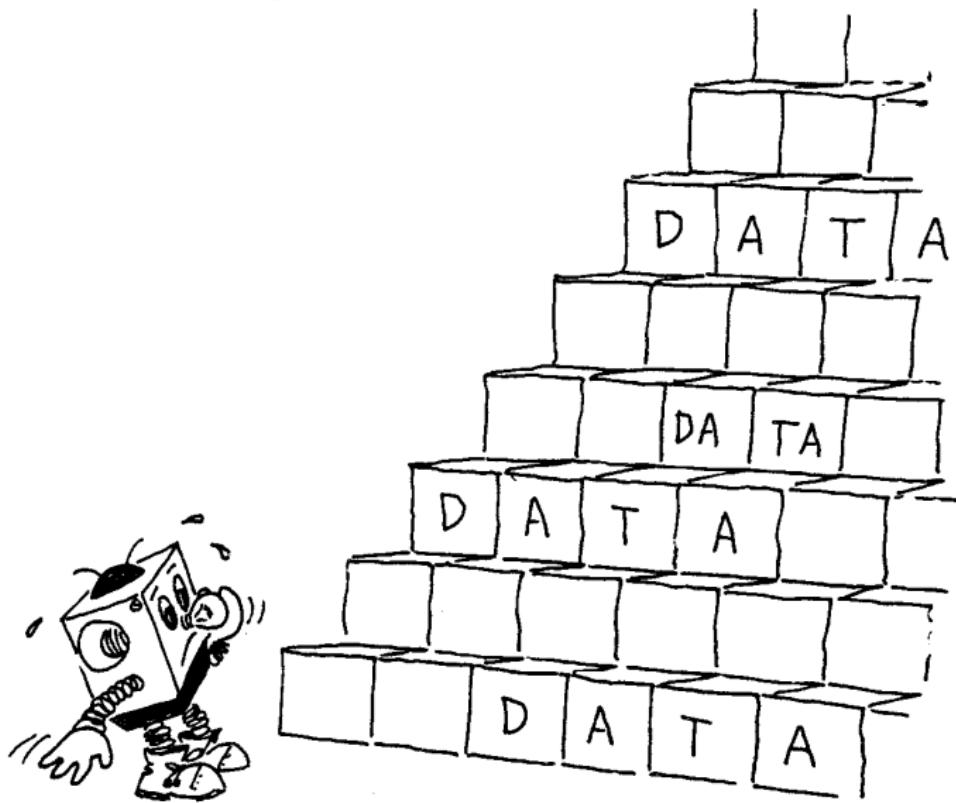
## EuDML – data: legacy scientific math

- 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 virtual digital library



# MU: alternative GUI—vizualisation research

Soubor   Úpravy   Zobrazení   Historie   Záložky   Nástroje   Nápověda  
<http://nlp.fi.muni.cz/~xpopek/dml/VBApplet.html#>

Zakázat\* Cookies CSS Formuláře Obrázky Informace Různé Orámovat\* Velikost okna Nástroje Zobrazit zdrojový kód Nastavení\* Správná zpravodajská služba... Gmail - Masivne informaticke... (JPEG obrázek, 800x502 bodů) Google Subversion applet html page

DML Search

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

Zobrazení  
Přibližit ▾

Bartušek  title  author

**Search Results**

- Nikitin, S. : Decoupling normalizing transformations and local stabilization of nonlinear systems
- Knoflíček, František : A combinatorial approach to the known projective planes of order nine
- Zelinka, Bohdan : Subtraction semigroups
- Novák, Jiří : Packings of pairs with a minimum known number of quadruples
- Neustupa, Jiří : A principle of linearization in theory of stability of solutions of variational inequalities
- Zelinka: Domination in graphs with few edges
- Pokorný, M. Pokorný, Milan Neustup: Axisymmetric flow of Navier-Stokes fluid in the whole space with non-zero angular velocity component
- Ewert, Janina : Quasicontinuity and related properties of functions and multivalued maps
- : Book reviews
- Vanžurová, A. Vanžurová, Alena : On torsion of a \$3\$-web
- Jakubík, J. Jakubík, Ján : On sequences in vector lattices
- Palumbiny, Oleg : On existence of Kneser solutions of a certain class of \$n\$-th order nonlinear differential equations
- Yanagi, Shigenori : Asymptotic behavior of the solutions to a one-dimensional motion of compressible viscous fluids
- Roubíček, Tomáš : Relaxation of vectorial variational problems

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

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

Exponential stability and exponential decay of oscillations in differential equations [en]  
 Existence of weak solutions of boundary value problems for nonlinear elliptic equations [en]  
 Qualitative theory of half-linear differential equations [en]  
 On existence of weaker solutions of boundary value problems for nonlinear elliptic equations [en]  
 Decoupling normalizing transformations and local stabilization of nonlinear systems [en]

Ontologie: /22-rdf-syntax-ns# Perspektiva: DML - 7

Hотово

Nyní: Převážné zataženo, 17 °C Po: 18 °C Ut: 12 °C

# MU: Visual Browser development (DML-CZ)

Soubor   Úpravy   Zobrazení   Historie   Záložky   Nástroje   Nápověda

<http://nlip.fi.muni.cz/~xpopek/dml/VBApplet.html#>

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) Google Subversion applet html page

DML Search

clear | show browsing results

Důležité   title   author   submit

Zobrazení Přiblížit 4

Search Results

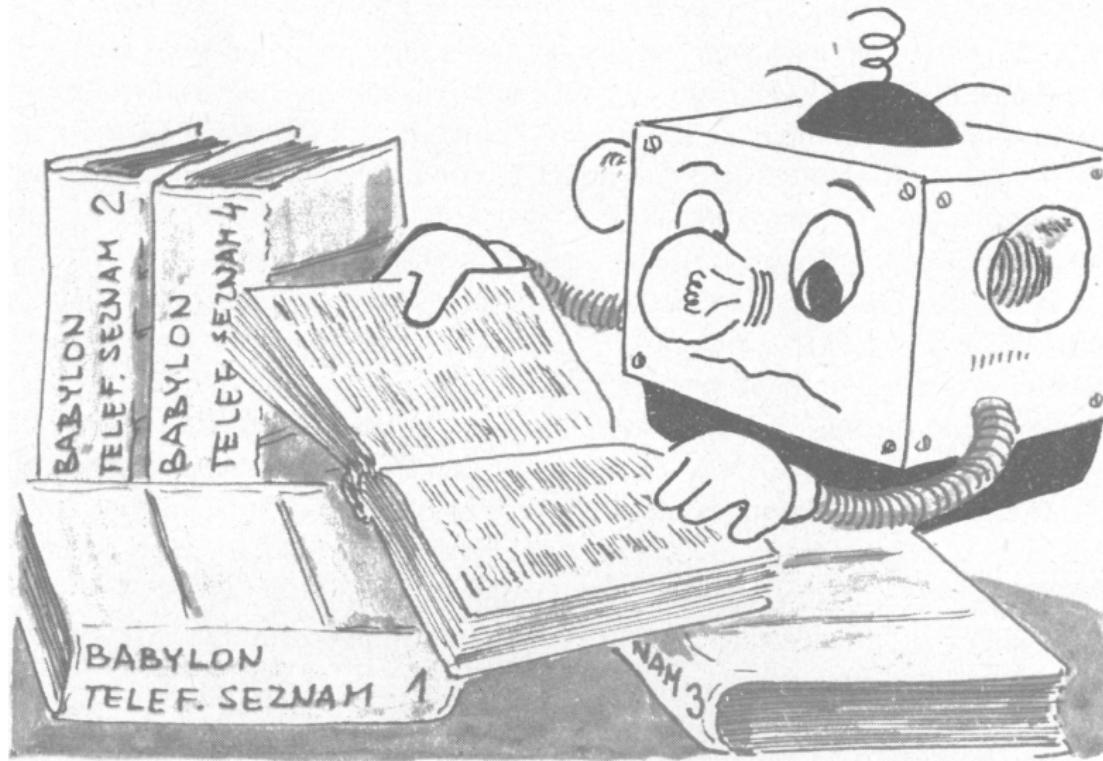
- Hilscher, Roman : [Spectral properties of fourth order differential operators](#)
- Rachdičková, Irena : [On some three-point problems for third-order differential equations](#)
- TvrdyM: Localization of nonsmooth lower and upper functions for periodic boundary value problems
- Ligęza, J. Ligęza, Jan 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

Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -14

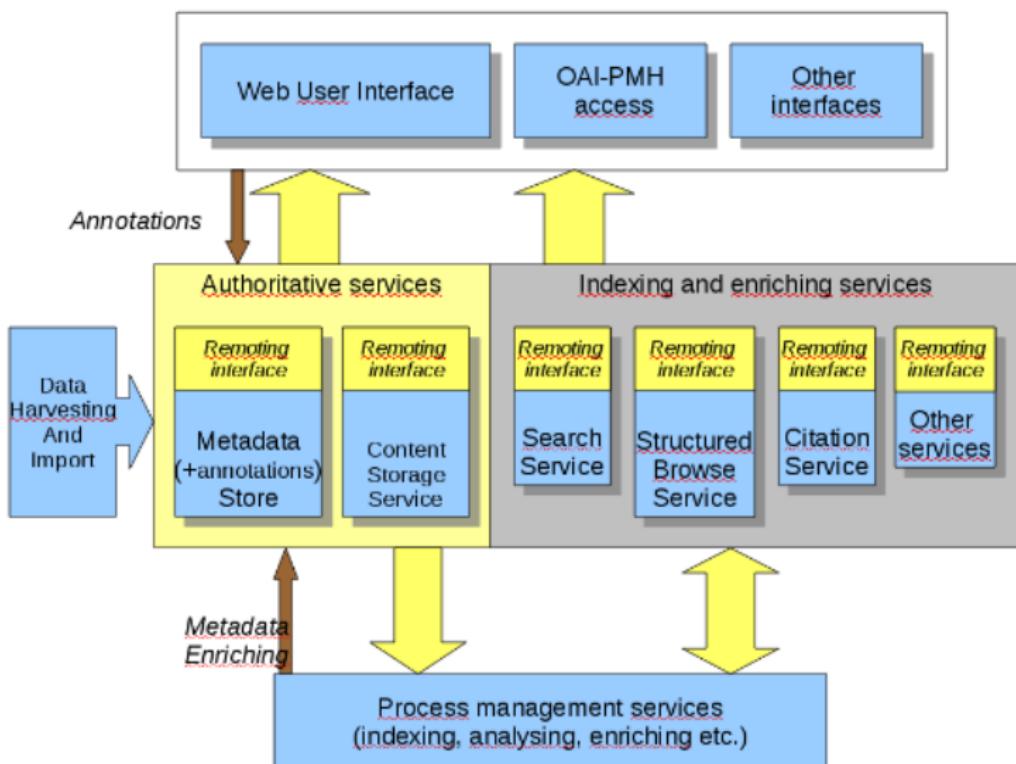
http://nlip.fi.muni.cz/~xpopek/dml/VBApplet.html#

Nyní: Převážně zataženo, 17 °C Po: 18 °C Ut: 12 °C

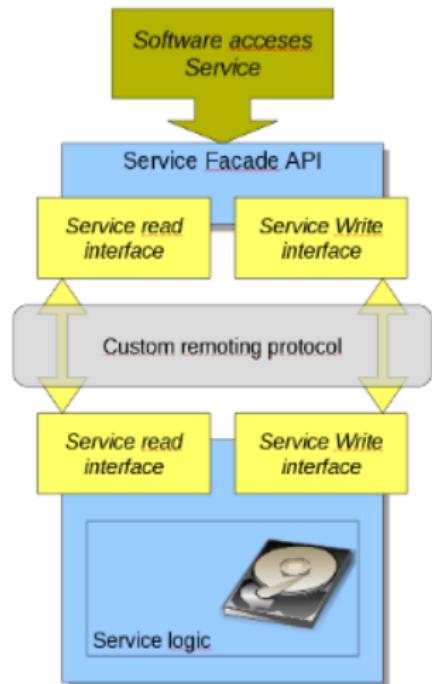
# EuDML document engineering—scalable tools development



# 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 ( $\text{\TeX}$  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.

The screenshot shows the DML-CZ Metadata editor interface for serials. The left side contains a form with dropdown menus and input fields for various metadata fields such as Title, Author, Date, Language, and MSC. The right side displays a preview of the document's content, which includes a title in Russian ('ЧЕХОСЛОВАЦКИЙ'), a summary, and sections like 'A CONTRIBUTION TO G' and '1. Introduction.' The preview also shows some internal document structure with page numbers (323-357).

Save	Save and Next	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

# Metadata Editor localization (Miha Filej)

The screenshot shows a Mozilla Firefox browser window displaying the DML-CZ Metadata editor for a serial issue. The URL is <http://dmlcz.kar.muni.cz:9999/edit/Issue/9/contents>. The page title is "DML-CZ: Metadata editor (serial)". The main content area lists five articles with their titles and page ranges:

Title	Page Range
(#1) Über zwei neue ebene Konfigurationen \$(12_4, 16_3)\$	[193-218]
(#2) The theory of characters of finite commutative semigroups	[219-247]
(#3) System of congruence relations on lattices	[248-282]
(#4) Sur les espaces à connexion affine partiellement projectifs	[283-290]
(#5) Characters of commutative semigroups as class functions	[291-292]

Below the list are buttons: "Delete Articles", "Change Ranges", and "Save Contents".

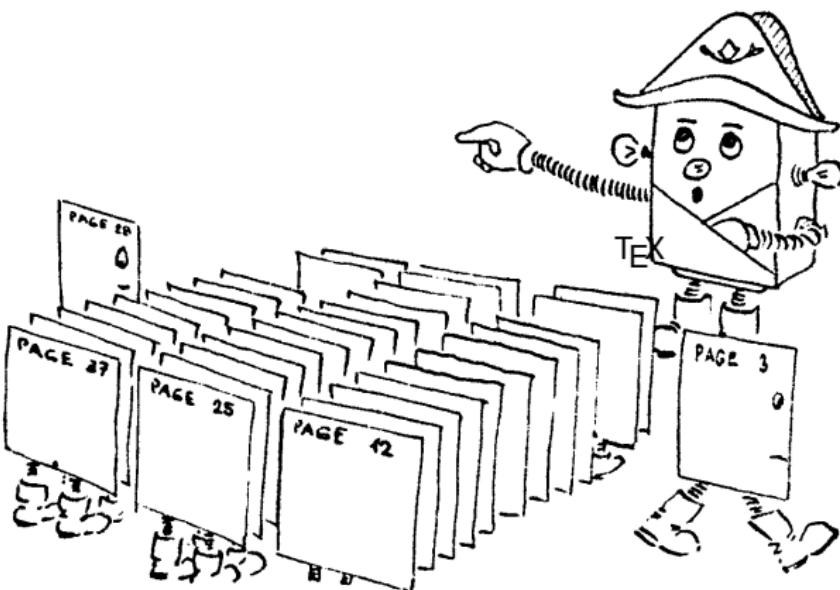
Under the article list, there is a red box containing two thumbnail images of scanned pages from the journal. The left image is labeled "(193a) [21]" and the right image is labeled "(193b) [3]". Below each image is a button: "edit ocr scan".

At the bottom of the red box, there are buttons: "Move Pages" and "Create Article".

The page content continues with a detailed view of article #1, showing its full text and several smaller thumbnails of scanned pages labeled 193 [4] through 202 [13]. Each thumbnail has a "edit ocr scan" button below it.

At the very bottom of the page is a "Done" button.

# 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 (deadline today ;-), accepted paper at LREC 2010 (with Radim Řehůřek)...

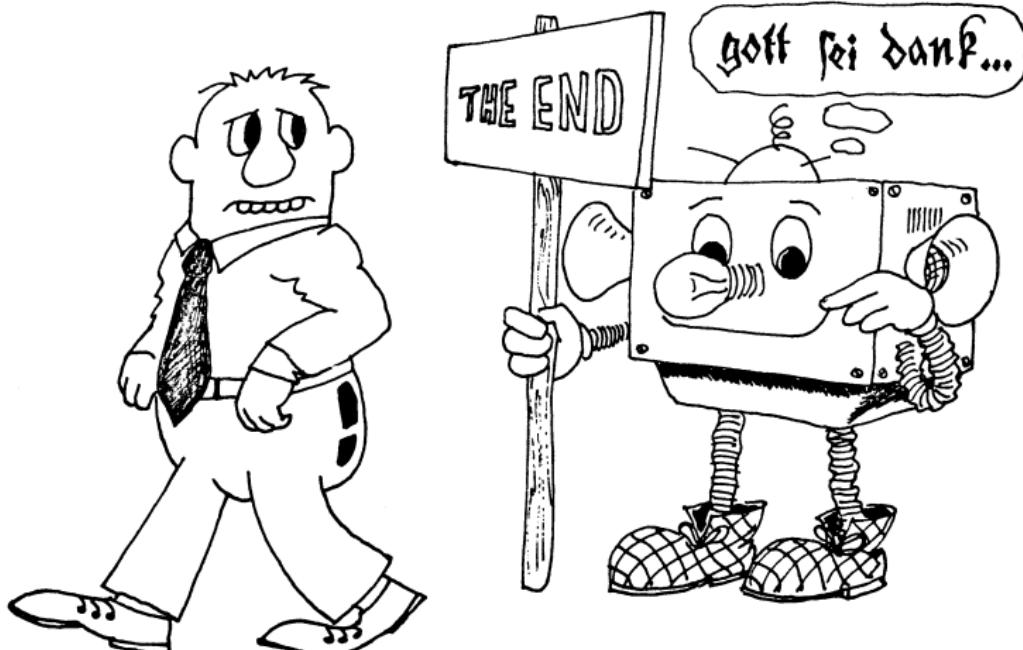
Next activities: EuDML general meeting in Paris in July (c/o CICM 2010, DML 2010), WP7 technical meeting, tool implementation.

DML 2010 organization: <[http://www.fi.muni.cz/ sojka/dml-2010.html](http://www.fi.muni.cz/sojka/dml-2010.html)>

Working meetings at MU every Wednesday, 2pm.

Comments, cooperation offers welcome!

# End of the talk



Questions?

# References, links



DML-CZ team.

*Materials about DML-CZ, project publications* [online, cit. 2010-05-4].

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



EuDML team.

*EuDML project info* [online, cit. 2010-05-4].

<[http://ec.europa.eu/information\\_society/apps/projects/factsheet/index.cfm?project\\_ref=250503](http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503)>



EuDML team.

*EuDML webpage* [online, cit. 2010-05-4].

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



EuDML at MU team.

*EuDML at MU project info* [online, cit. 2010-05-4].

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