Digital Matematics Libraries: from DML-CZ towards EuDML

Petr Sojka

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

FI MU, NLP seminar, March 17th, 2010

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

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 projekt EuDML (programme EU CIP-ICT-PSP, type Pilot B) from February 2010 (MU and MU AV).

As a basis serve current DML repositories as DML-CZ or NUMDAM (bottom-up build up).

Example of DML-CZ (who, what, browse, browse similar, how to search).

All comments welcome, especially to similarity algorithms and results.

Data

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.

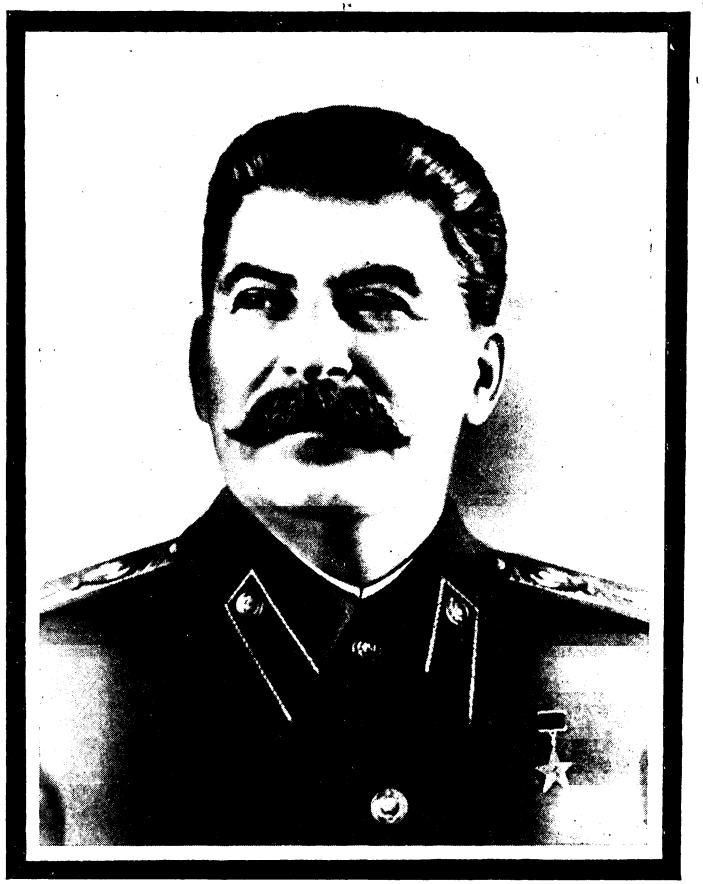
Резюме

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*



ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

Data

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.

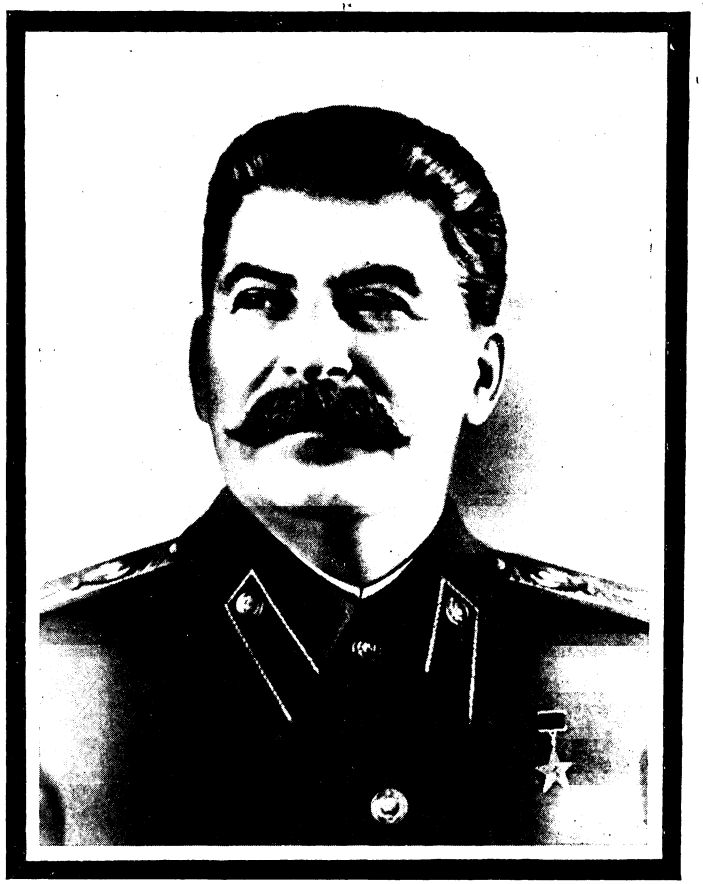
Резюме

ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЖИК (Jan Mařík), Прага.

(Поступило в редакцию 10/X 1955 г.)

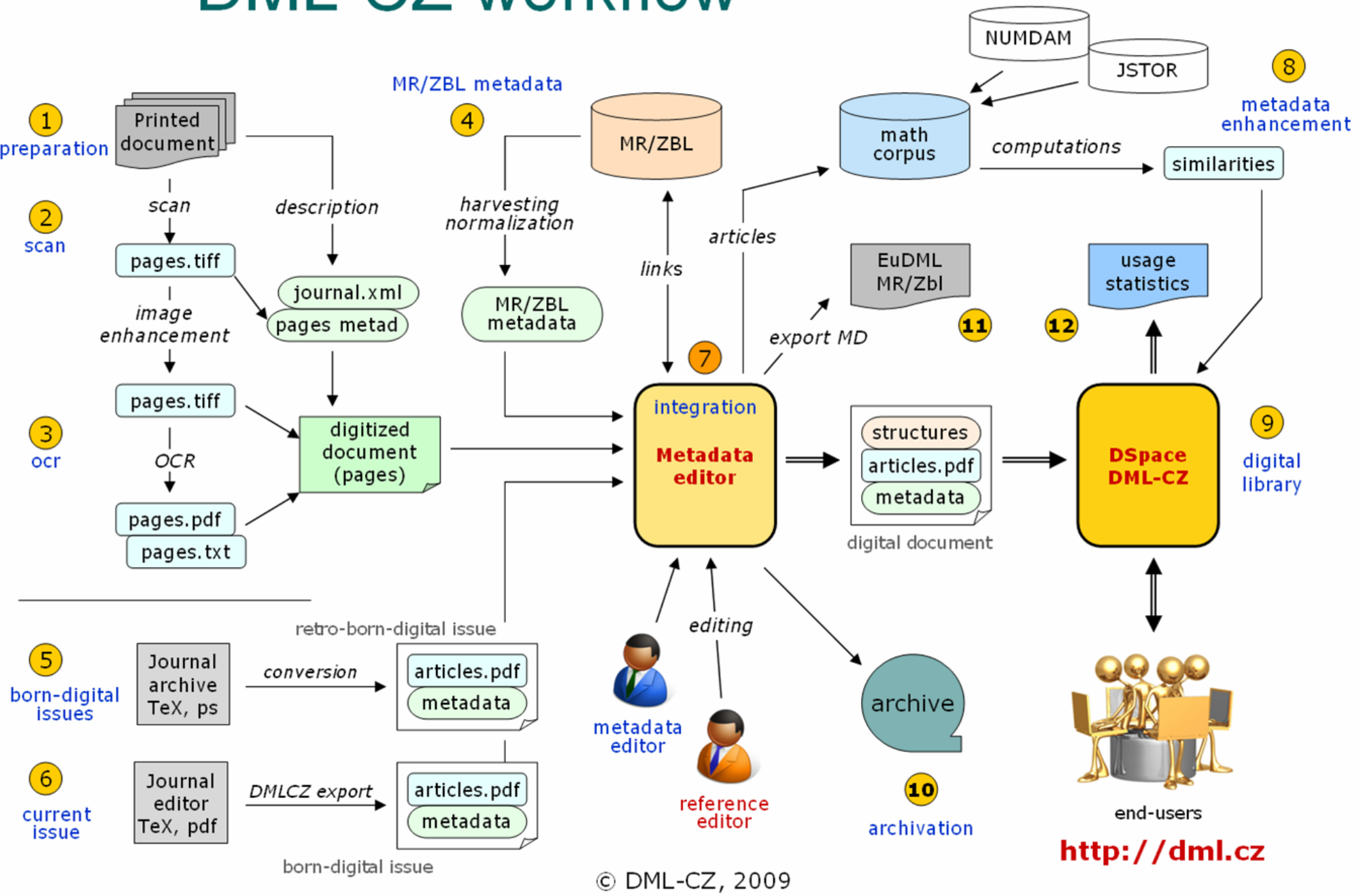
Пусть m — натуральное число; пусть E_m — m-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\,\mathrm{d}x$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^{m} v_i^2(x)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



ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

DML-CZ workflow



Take care!

Data heterogenity, different formats

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

Data heterogenity, different formats

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

Data heterogenity, different formats

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

Data heterogenity, different formats

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

Typesetting of papers and cover pages

- XeLaTeX, 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` → `item.xml` → `item.tex` → `item.pdf` → ...

Typesetting of papers and cover pages

- XeLaTeX, 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` → `item.xml` → `item.tex` → `item.pdf` → ...

Typesetting of papers and cover pages

- XeLaTeX, 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` → `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 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}
```

Other (verified and proven) technologies I

- image transformations (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).
- optimization of PDF: pdfopt (from ghostscript).
- similarity article computations (research with Radim Řehůřek), demo.

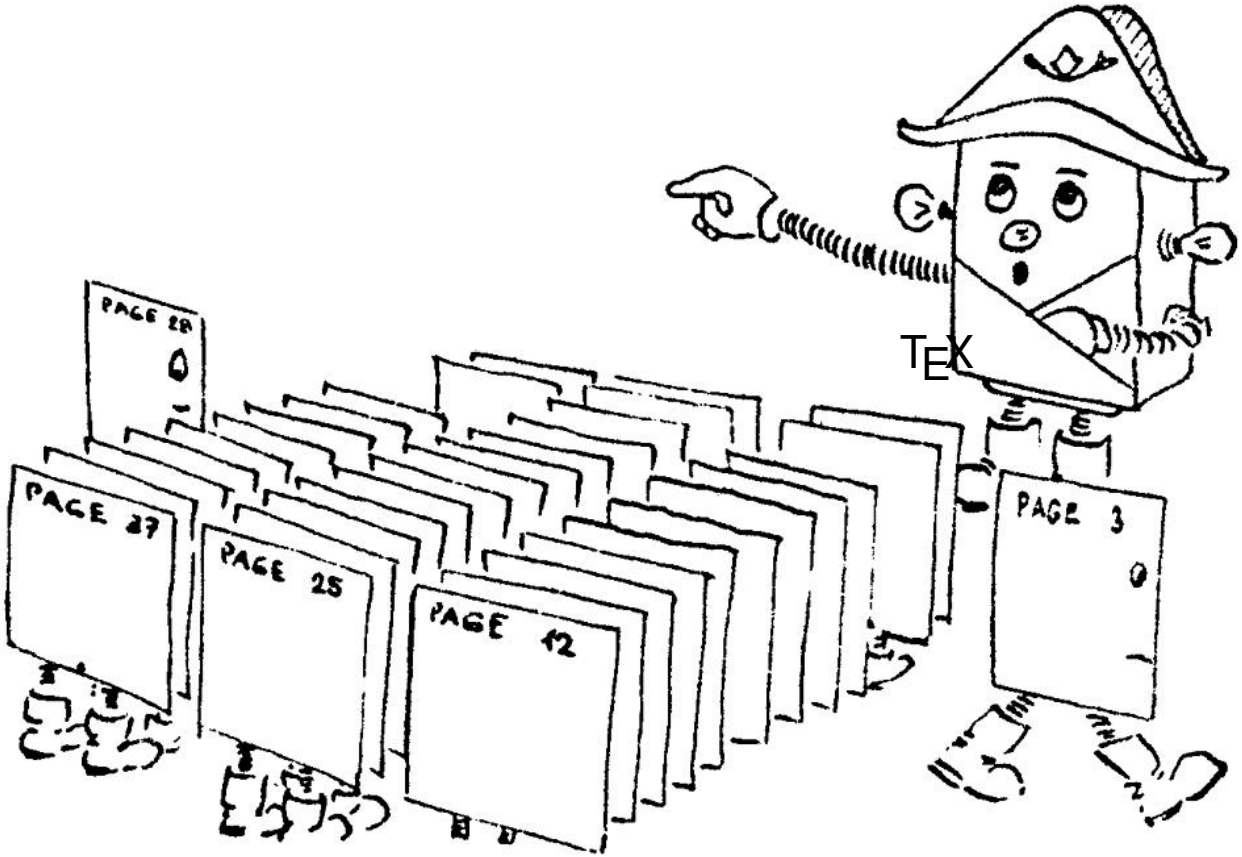
Other (verified) technologies II

- paper classification by MSC (Radim Řehůřek).
- data vizualization, browsing Visual Browser (DP Zuzana Nevěřilová).
- PDF recompression using JBIG2: application based on jbig2enc/leptonica (BP Radim Hatlapatka).
- born-digital publishing system [for Archivum Mathematicum] (BP Michal Růžička).
- retro-born-digital conversions (Michal Růžička).
- … [fixfont, math OCR a indexing (BP/DP/PhD “wanted!”)].

Other (verified) technologies II

- paper classification by MSC (Radim Řehůřek).
- data vizualization, browsing Visual Browser (DP Zuzana Nevěřilová).
- PDF recompression using JBIG2: application based on jbig2enc/leptonica (BP Radim Hatlapatka).
- born-digital publishing system [for Archivum Mathematicum] (BP Michal Růžička).
- retro-born-digital conversions (Michal Růžička).
- … [fixfont, math OCR a indexing (BP/DP/PhD “wanted!”)].

Comments? Cooperation? Questions?



References, links

DML-CZ team.
Materials about DML-CZ, project publications [online, cit. 2010-03-16].
<http://project.dml.cz/documents.html>.

EuDML team.
EuDML project info [online, cit. 2010-03-16].
<http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250503>.

EuDML at MU team.
EuDML at MU project info [online, cit. 2010-03-16].
<http://nlp.fi.muni.cz/projekty/eudml/> or <http://www.muni.cz/research/projects/10067>.