Report on DML-CZ project[1]

Petr Sojka et al.

DML-CZ
Faculty of Informatics, Masaryk University, Brno

Jul 8th, 2009

[1] Supported by the Academy of Sciences of Czech Republic grant #1ET200190513

Bottom-up way to WDML—DML-CZ

- Failure of global funding of DML-EU within FP6.
- Niche "markets" for Google Print or similar general digitization projects, mathematical literature published in CE not covered.
- Making WDML (bottom up)^2 by creation of "microclima": 1) with the help of the local goverment funding: DML-CZ, 2) from scanned images to full text marked pages.

The Goal

- Czech Academy of Sciences grant (program Information Society) 2005–2009, **full** (retro)digitization of 50,000 pages of mathematical literature per year.
- We do not want to reinvent the wheel (scanning, text OCR).
- 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
- IPR part:sharing/delivery (economic models for knowledge sharing due to interests of content owners/publishers).

What to digitize in DML-CZ?

7–8 Czech and Slovak math journals, 100–200 monographs and textbooks and conference proceedings, in total about 250,000 pages:

① *Czechoslovak Mathematical Journal* (30,000 scanned, 7,000 are already born digital). Published by Academy of Sciences of CR, distributed partially by Springer. Founded as *Časopis pro pěstování matematiky* in 1872, under current name since 1951. 272 pages quarterly.

② *Applications of Mathematics* (20,000/5,000). Published by Academy of Sciences of CR. Founded in 1956 (as *Aplikace matematiky*). 80 pages bimonthly.

③ *Archivum Mathematicum* (2,000/4,000) Masaryk Uni in Brno.

Mathematica Bohemica and *Archivum Mathematicum* already partially digitized in Göttingen,... Copyright issues crucial.

Who is in the project?

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.

On the way from digital image to knowledge

acquisition preparation, document acquisition, copyright issues handling;

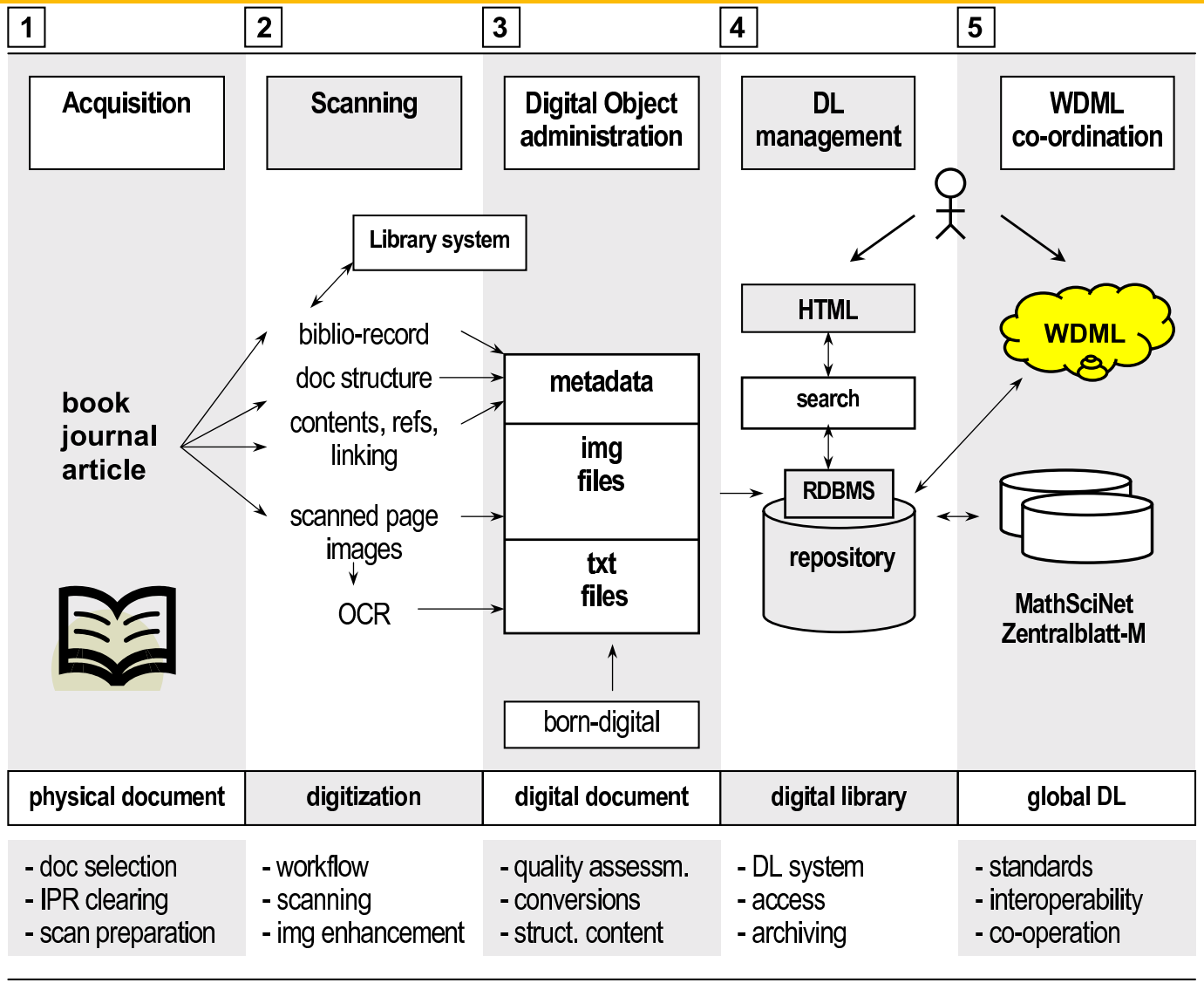
scanning document scanning (1/5 of the budget only) main metadata entering, scanning checks;

image processing main OCR, image enhancements.

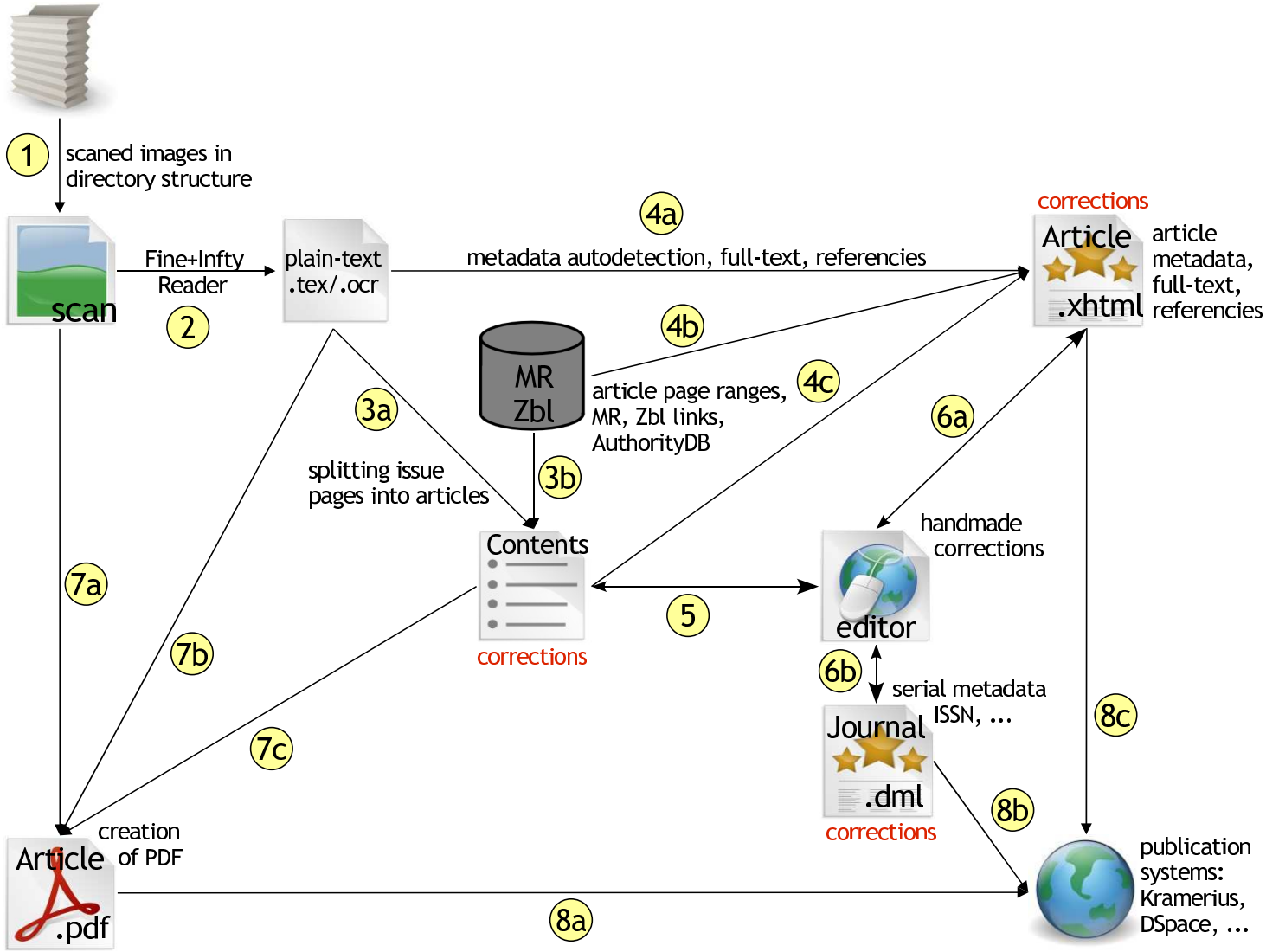
semantic processing document markup enhancement, semantic processing, document classification, citation linking, document clustering, [math] indexing;

delivery and presentation visualization techniques of document repository, digital library web portal, interfaces to other services and search engines for the semantic based document processing/delivery.

DML-CZ workflow steps



Top-level DML-CZ workflow overview (simplified)



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.

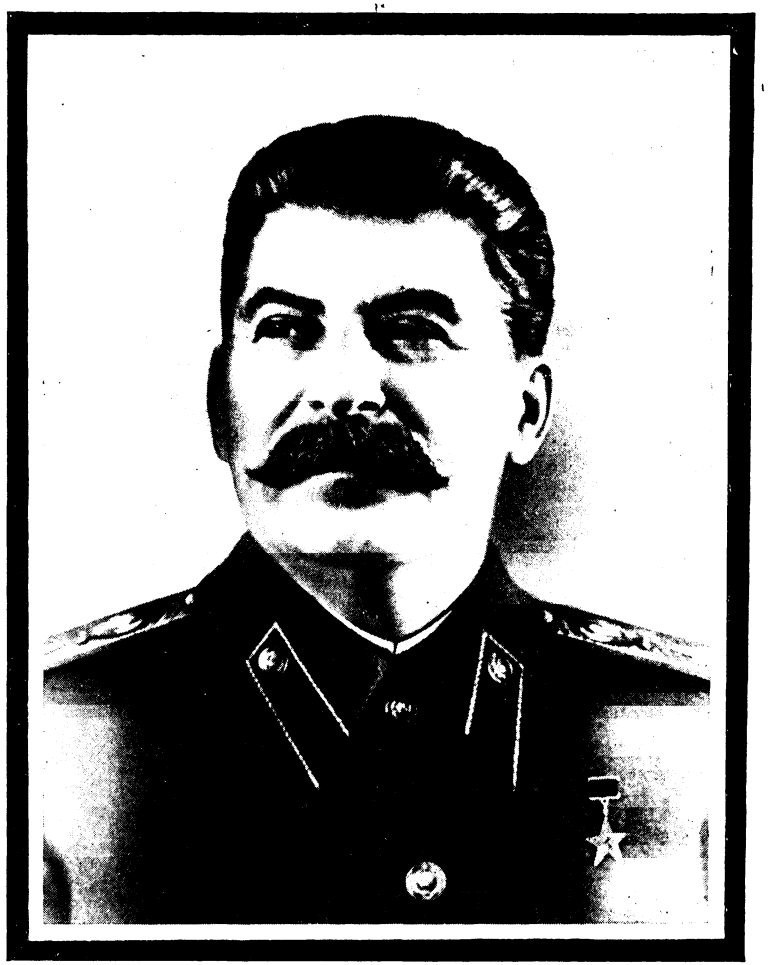
Резюме

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

Summary of current status quo

- **http://dml.cz** up (11,000 articles) and running (as beta)
- new papers (both born-digital (5 journals), and retro-digitized), and new features added continuously: metadata exports, similar papers by LSI and other methods
- project ends by the end of 2009, then hopefully EuDML or EVLM.