

DML-CZ: The Objectives and the First Steps

Jiří Rákosník, Petr Sojka, Martin Šárky

CMDE2006, Aveiro, 15-18 August 2006

Project



The Czech Digital Mathematics Library

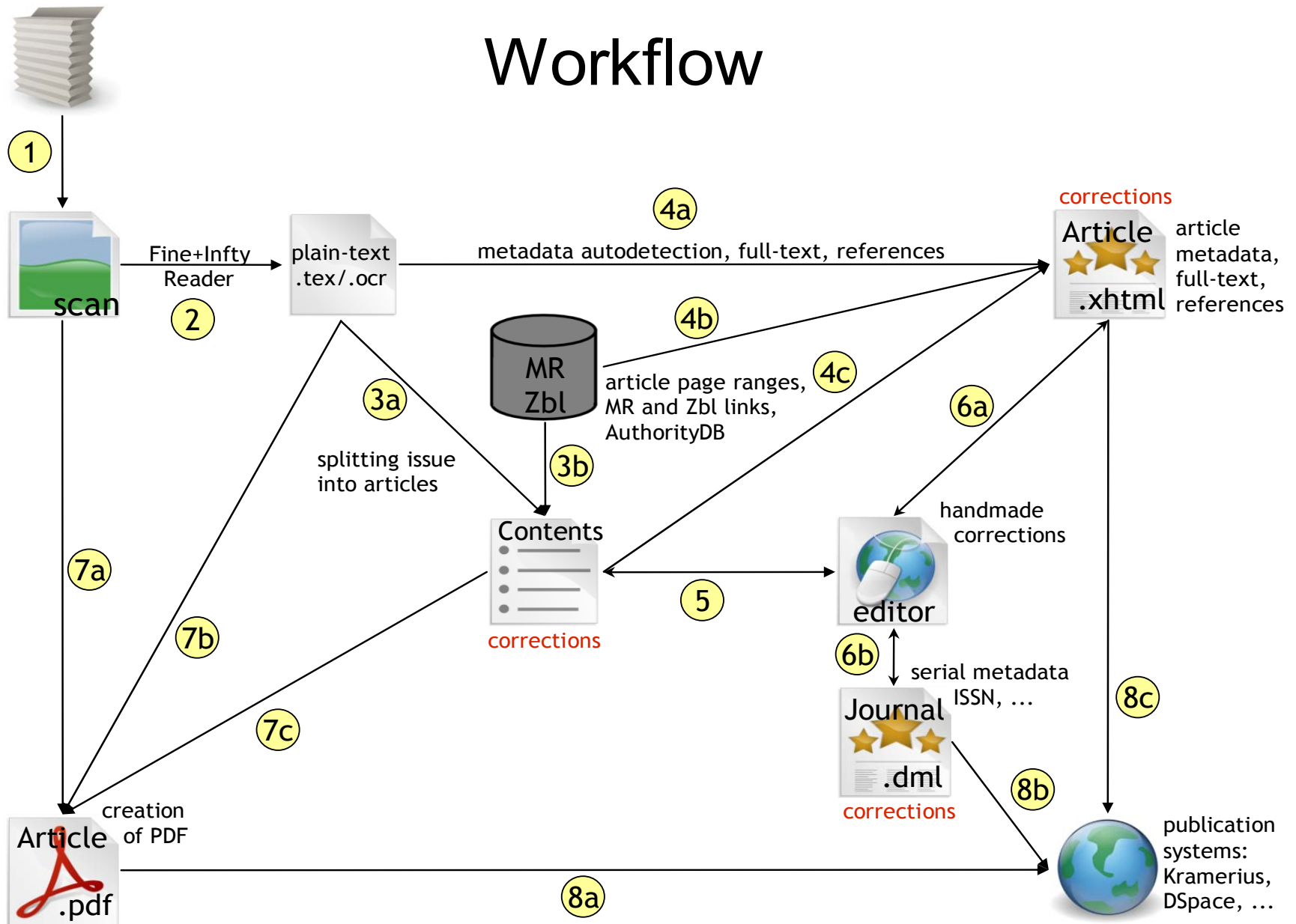
- The aim: to ensure availability of mathematical literature which has been published throughout history in the Czech lands, in digital archival form
- Funded by: the Academy of Sciences of the Czech Republic within the national R&D programme “Information Society”
- Period: 2005–2009
- Estimated extent: 150–200 thousand [digitized] pages

Partners



- **Mathematical Institute AS CR (Prague)**
project co-ordinator, selection & preparation of materials for digitization, IPR and copyright issues, operation and maintenance of the developed DML-CZ
- **Institute of Computer Science, Masaryk University (Brno)**
technical integration, development of the digital library, coordination of metadata provision and incorporation of the DML-CZ into the WDML
- **Faculty of Informatics, Masaryk University (Brno)**
OCR processing, techniques for searching and presenting digital documents, presentation formats and relevant technology development and testing
- **Faculty of Mathematics and Physics, Charles University (Prague)**
user requirements, metadata specifications and linkage to Zentralblatt MATH and Mathematical Reviews
- **Library AS CR (Prague)**
digitization, OCR, storage and presentation of the digitized content within the Academy of Science framework

Workflow



Test bed: Czechoslovak Math. J.



- 1951–1991 classical typesetting (almost 30 000 pages), since 1992 TeX
- multilingualism: Czech, Slovak, Russian, English, German, French, Italian
- the first two volumes published simultaneously in Czech, Russian and multilingual versions
- free-hand drawings, graphic figures, tables and photographs

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

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

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 funkcional 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 $\|A\| = \sup \int_A \sum_{i=1}^m \frac{\partial v_i(x)}{\partial x_i} dx$, 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{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 ρ и на

Доказательство. Если $S - L^*$ имеет только один элемент, то этим элементом является неизбежно e_r . Этот элемент сам по себе образует группу. Следовательно, можно ограничиться случаем, что $S - L^*$ имеет больше чем один элемент. В этом случае $S - L^*$ является полугруппой в силу леммы 5.

По лемме 8 существует L^* и M^* , и имеет место $M^* \subseteq L^*$. По теореме 5,2 есть $L^* = M^*$, и $S - L^*$ — слева простая полугруппа; так как $e_r \in S - L^*$, имеет эта слева простая полугруппа идемпотент. Значит, в силу теоремы 1,6, $S - L^*$ является соединением непересекающихся изоморфных групп.

Теорема 6,2. Пусть полугруппа S имеет двустороннюю единицу e . Пусть имеет хотя один двусторонний идеал $\neq S$. Пусть S/M^* имеет хотя один минимальный левый идеал. Тогда $S - M^*$ является группой.

Доказательство. Согласно лемме 8, существуют при наших предположениях идеалы M^* , R^* , L^* и имеет место $M^* \subseteq R^* \cap L^*$. Разностная полугруппа S/M^* является простой полугруппой с нулем и имеет двустороннюю единицу e . По предположению существует хотя один минимальный левый идеал из S/M^* . Подобно, как в начале доказательства теоремы 4,1, покажем, что никакой левый идеал из S/M^* не может быть нильпотентным. Следовательно, по теореме 1,4, S/M^* покрыто суммой своих минимальных левых идеалов. Но e не может входить ни в какой левый идеал $\neq S/M^*$. Значит, в S/M^* не содержится никакой левый идеал $\neq [0]$ и $\neq S/M^*$. Но далее имеет место: простая полугруппа, имеющая хотя один минимальный левый идеал и хотя один идемпотент $e \neq 0$, имеет также по меньшей мере один (не нильпотентный) минимальный правый идеал (см. [3], теорема 7,2). Значит, S/M^* является также соединением минимальных правых идеалов. Так как e не может принадлежать никакому правому идеалу $\neq S/M^*$, то и в S/M^* не содержится никакой правый идеал $\neq [0]$ и $\neq S/M^*$. Поэтому, в силу следствия 1,2в, S/M^* является группой с нулем. Значит, $S - M^*$ есть группа, чтд.

ЛИТЕРАТУРА

- Clifford A. H.: [1] Semigroups containing minimal ideals, Amer. J. Math. 70 (1948), 521—526. — [2] Semigroups without nilpotent ideals, Amer. J. Math. 71 (1949), 834—844.
- Green J. A.: [1] On the structure of semigroups, Ann. of Math. 54 (1951), 163—172.
- Schwarz Št.: [1] О максимальных идеалах в теории полугрупп, I, Чехословацкий математический журнал, т. 3 (78), 1953, 139—153. — [2] Структура простых полугрупп без нуля, Чехословацкий математический журнал, т. 1 (76), 1951, 51—65. — [3] О полугруппах, имеющих ядро, Чехословацкий математический журнал, т. 1 (76), 1951, 259—301.
- Воробьев Н. Н.: [1] Об идеалах ассоциативных систем, ДАН, т. 83 (1952), 641—643.

La torsion $\tau = \Delta/H^2$ peut être discontinue en un point (t, λ) où $A = B = C = 0$, donc $\varrho = 0$. C'est ce qui arrive pour $\lambda = \frac{1}{4}$, $t = 0$. D'ailleurs, pour $\lambda = \frac{1}{4}$ on retrouve la courbe non-orientable d'ordre 2 que nous avons déjà présentée au No. 3. On doit remarquer que la torsion τ reste bornée sur cette courbe $\lambda = \frac{1}{4}$, mais $\tau(t, \lambda)$ n'est pas bornée au voisinage du point $t = 0$, $\lambda = \frac{1}{4}$. En effet; on trouve

$$\tau = \frac{6\lambda^2[(4\lambda - 1) + (\cos t - 1)(4\lambda - 2\cos t - 2)]}{(4\lambda - 1)^2(5\lambda^2 - 4\lambda + 1) + (\cos t - 1)[12\lambda^2(\lambda^2 + 3)(\cos t + 1) - 4\lambda(3 + 26\lambda^2)]}$$

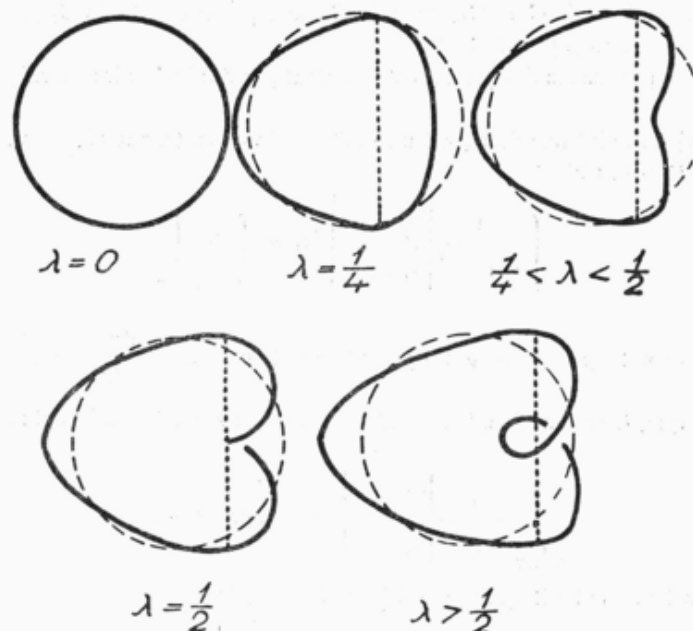


Fig. 9.

et l'on voit que pour $t \rightarrow 0$, $\lambda \rightarrow \frac{1}{4}$ la limite de ce rapport dépend de $\lim_{t \rightarrow 0} \frac{\cos t - 1}{4\lambda - 1}$;

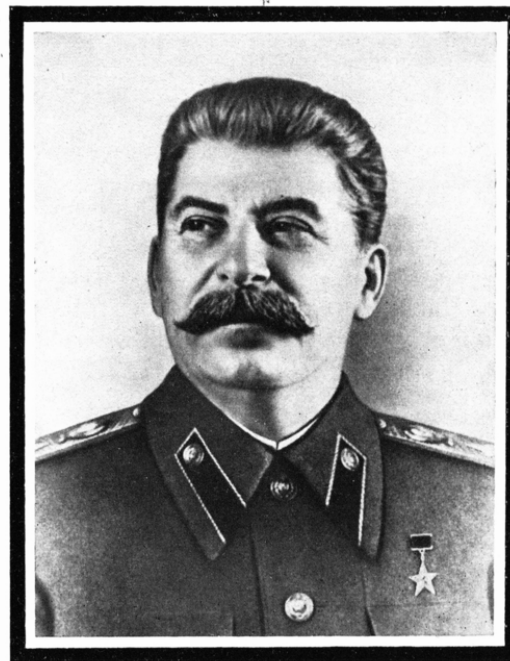
cette limite est infinie si $\lim_{t \rightarrow 0} \frac{\cos t - 1}{4\lambda - 1} = 0$.

On a $K = 0$ pour $0 \leq \lambda < \frac{1}{4}$ et $K = 1$ pour $\frac{1}{4} < \lambda \leq 1$, et ce saut est dû à la discontinuité de τ pour $\lambda = \frac{1}{4}$. Ce résultat est général:

Si C_1 et C_2 sont deux noeuds pour lesquels K prend des valeurs de parité différente ($K_2 - K_1 = 2h + 1$), pendant toute déformation continue (les dérivées des coordonnées jusqu'à l'ordre 3 étant continues) de C_1 en C_2 la torsion $\tau(s, \lambda)$ passe nécessairement par un point de discontinuité.



КЛЕМЕНТ ГОТВАЛЬД
1896—1953



ИОСИФ ВИССАРИОНОВИЧ СТАЛИН
1879—1953

Scanning



- Digitization Center of the Library AS CR
- Zeutschel scanners OS 7000
(90 A4 pages per hour at 600 DPI)
- grey scale, 600 (644) DPI, 4~bit in TIFF
- BookRestorer (i2S, France) for the graphical improvements of the scanned pages – mainly cropping, binarization and straightening lines
- the first OCR (all but mathematics): ABBYY FineReader engine integrated in the production system Sirius (Elsyst Engineering, CR)
- automated creation of minimal metadata with the aid of pre-defined models

From scanned images to articles



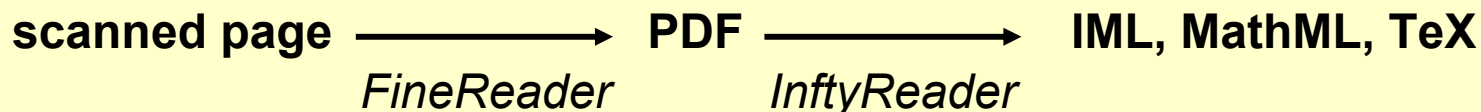
1. Sorting of the scanned page pictures into the hierarchical Journal-Volume-Issue directory structure
 - done implicitly during the scanning process and storage of the files in an appropriate directory structure

From scanned images to articles



2. Application of advanced OCR techniques based on the InftyReader software for processing of maths

- two runs:



- details will be given by Petr Sojka in his talk

Problems:

- language detection on the paragraph level
- page number detection (sometimes page number missing)
- if OCR does not detect page number, the scan is not matched automatically – has to be done manually

From scanned images to articles



3. Creation of the initial list of papers in the journal issue

- exploitation of metadata from existing databases
- location of beginnings and ends of papers by means of OCR
- identification of contents page and its items by means of OCR

This is very essential for further steps.

Problems:

- OCR may cause a false division of a paper into parts
- databases are not absolutely reliable (e.g. vacates)
- identification of contents page may be difficult, so far not solved

From scanned images to articles



4. Auto detection of descriptive metadata from external databases and/or from OCR

- some of them are obtained by OCR already during the scanning phase
- used in the next step for control

Problems (particularly with references):

- identification of the item beginning, separation of items and of their elements, finding of the URL, ...
- multilingualism (References, Bibliography, Bibliographie, Literaturverzeichnis, Littérature, Literatura, Литература)

From scanned images to articles



5. Manual revision of the list of papers and articles content

- crucial for the further steps in the workflow
- done manually with the help of the Metadata Editor that enables
 - visual control of page images
 - paper preview
 - shuffling pages within a paper and between papers
 - cancellation of a badly identified paper and constitution of a new one
 - identification of non-matematical “papers” (editorial, contents)
 - removal of pages without content
 - ...

▲

[3 / 19]

[395](#)
[396](#)
[397](#)
[398](#)
[399](#)
[400](#)
[401](#)
[402](#)
[403](#)
[404](#)

1133

61

•

1

10

DML-CZ - Mozilla Firefox

File Edit View Go Bookmarks Tools Help


http://dkf.ics.muni.cz:9999/edit/issue/17/contents

Go

<input type="checkbox"/> Transformation ou \$p\$-algèbre suffisante et minimum de la probabilité d'erreur. (118-126)	115-123
<input type="checkbox"/> --- Název článku nebyl zadán --- (127-132)	124-129
<input type="checkbox"/> The asymptotic behaviour of branching stochastic processes. (133-156)	130-153
<input type="checkbox"/> Reduzierende zufällige Transformationen. (157-161)	154-158
<input type="checkbox"/> --- Název článku nebyl zadán --- (162-169)	159-(166)

Zrušit články Aplikovat rozsahy Uložit obsah

(1a) [2]




[edit ocr scan](#)

Přesunout stránky Vytvořit článek

--- Název článku nebyl zadán --- (1b)

[Obsah](#)

(1b) [3]



Done

From scanned images to articles



6. Manual revision of descriptive metadata

- important for the quality of the DML, not for the workflow
- done also with the Metadata Editor

Problems:

- authors' names
 - transliteration
 - who decides – Zbl, MR, authority basis?
 - examples: Zakharov/Zaharov, Solomencev/Solomentsev, Nikolskiĭ/Nikol'skiĭ/Nikol'skij, ...
- missing MSC
- ...

From scanned images to articles



7. Generating PDF files on the article level

- double-layer PDF enabling search
- generated using the list of papers and corresponding page numbers
- we do not consider DjVu (PDF from 6.0 supports compression algorithm JBIG2)

From scanned images to articles



8. Export of papers and metadata into publication systems

Problems:

- to decide what system to use
 - Kramerius
 - DSpace
 - a specially developed one?
- to incorporate the DML-CZ into the WDML

IPR issues



Problems:

- according to the Czech law the electronic version of a printed document is considered a truly new document → special author's permission is required
- therefore, the electronic versions should not be displayed on a public net
- to negotiate with publishers/distributors (moving window, presentation of documents within this window, ...)

Further steps

- to solve the problems
- to digitize further literature
 - Applications of Mathematics, Kybernetika, couple of others
 - conference proceedings, textbooks, theses
- to handle the born-digital material
- to process the material digitized in Göttingen
- to process the Slovak journals
- to cooperate with other digitization initiatives
 - OCR maths
 - indexation and search in maths
 - classification
 - reference linking
 - we will be happy to share our experience and developed tools

Optimism



A: Things can't go any worse.

B: They still can.

Question: Who is optimist and who is pessimist?

In any way, I am optimist.

<http://dml.muni.cz/>