Optical Character Recognition of Mathematical Texts in the DML-CZ Project[1]

Petr Sojka, Radovan Panák, Tomáš Mudrák

Faculty of Informatics
Masaryk University, Brno

October 14th, 2008

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

What is this story about?

- **important problem:** from pixel sets to the information (when awake, your brain spends almost half of its capacity for this task)
- **important application:** how to have all the math papers published in a digital searchable for: imagine all mathematical information/knowledge available at your fingertips!
- **pleasant surprises** (unexpected connections, difficulties, solutions and beauty): it actually works reasonably well!
- **No sex and violence,** sorry.

Motivation, Goals

① DML-CZ **http://dml.cz**, **http://project.dml.cz**

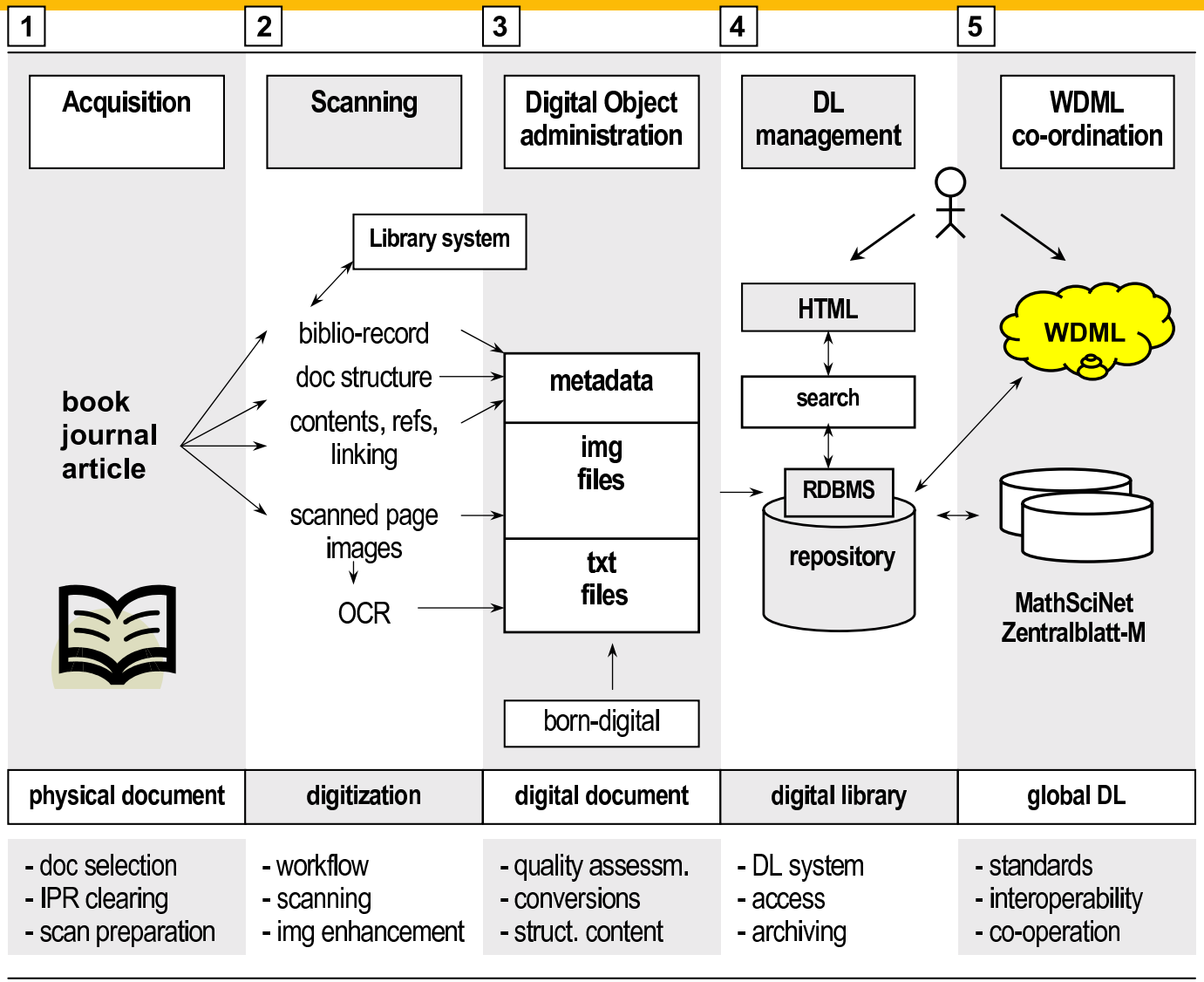
② not only page images, but added value wrt. Google Scholar

③ full text indexing, good searching (and ranking),

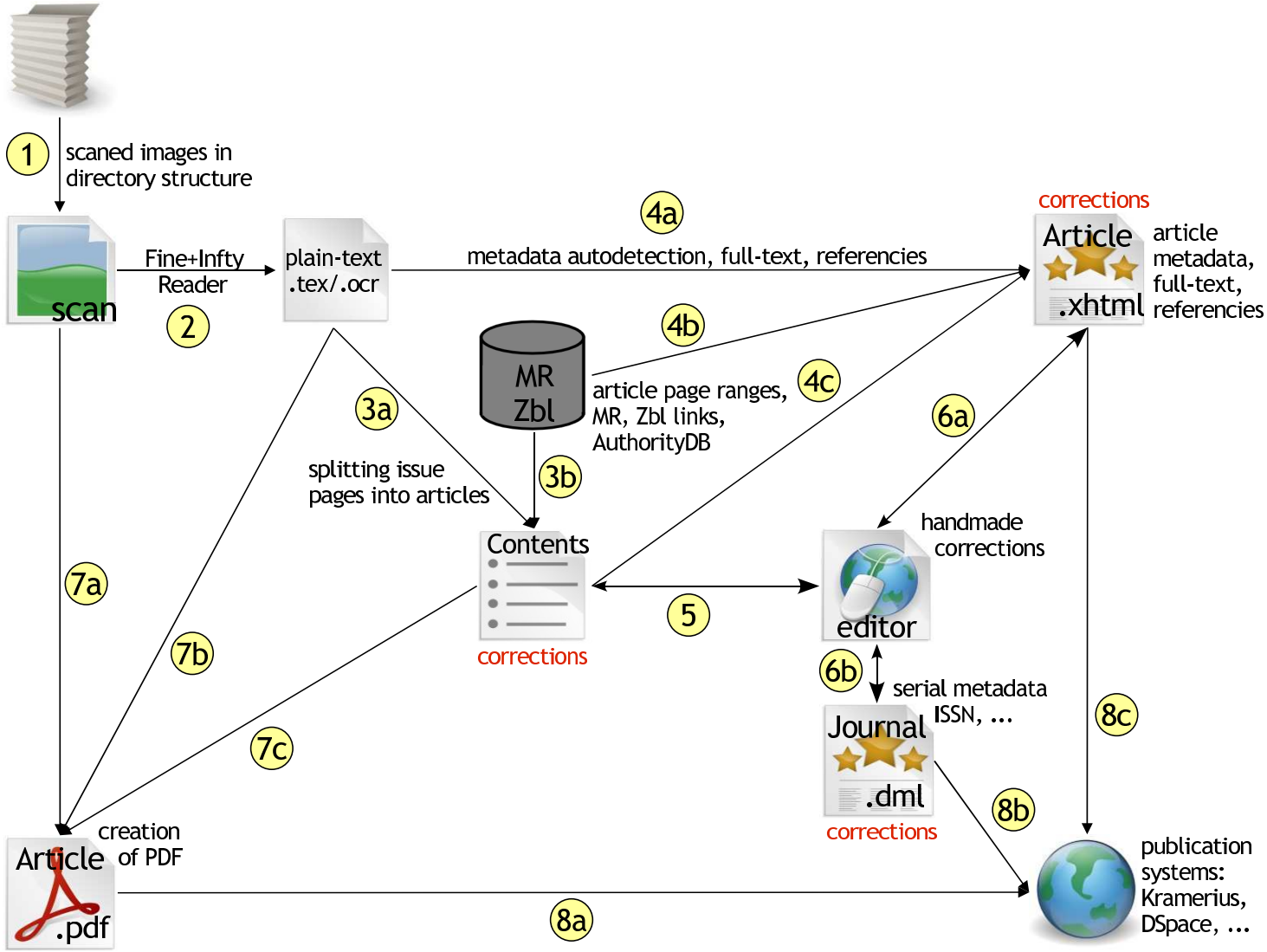
④ well clasified papers, with hypertext links between them and referee databases (ZentralBlatt MATH and Math Reviews)

⑤ persistent and stable access, aimed at full (text) visibility in the global information space (Google Scholar, OAI-PMH servers, . . .)

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

How to Find? Search!

① an entry gate to the digitized papers is **search**

② full text searching

③ searching for intext references

④ search and exchange of **mathematical formulas**: MathML, OpenMath

⑤ due to the massive size of digitized material, the only way is very good OCR, **including math**.

Existing OCR Systems

① Not to reinvent the wheel: trial of several OCR engines.

② No single OCR system with acceptable results: high error rate, working only for specific purposes (plain English text), direct use was not possible.

③ Fine Reader by ABBYY gave good results for (even multilingual) text, and allows for typeface learning.

④ InftyReader by **www.inftyproject.org** the only available solution for structural math recognition.

⑤ No out-of-the-shelf solution.

Our OCR Solution

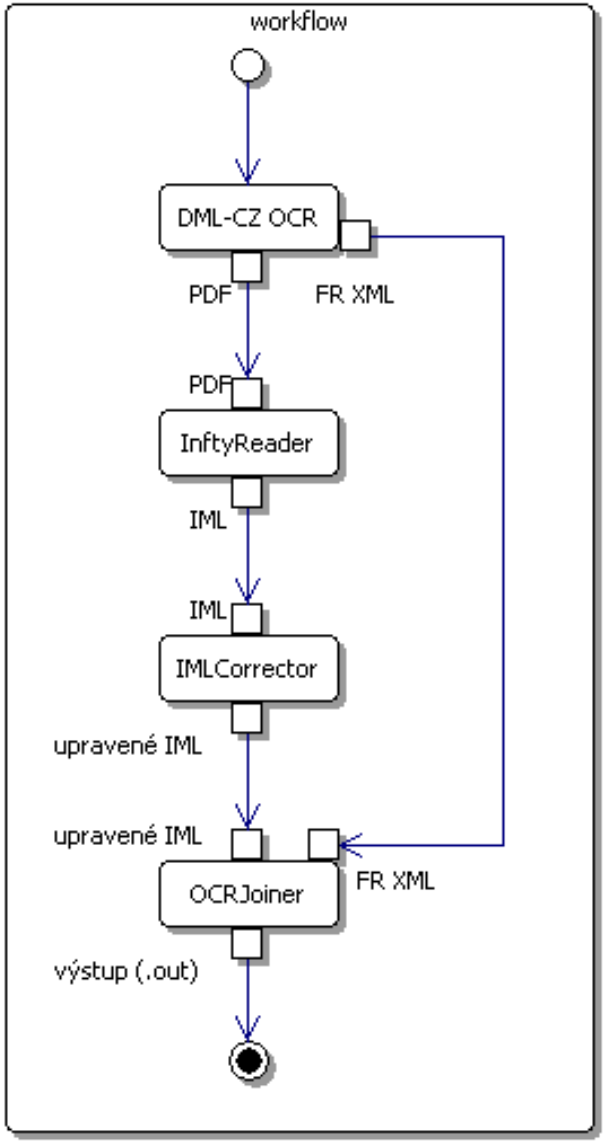
① combining both, using FineReader and InftyReader in a pipe to let every system to do what it is good for, then 'vote'

② top-level (Java) program to **automate** the process **and fix** some indeficiencies

③ instant setup unusable: **fine-tuning** and **gradually enhancing** the OCR procedure and program parameters so that OCR results would be acceptable for DML-CZ purposes

④ trying to improve the results further by close cooperation with the team of prof. Masakazu Suzuki (Infty Project leader, Kyushu University, Japan), and hopefully with other (retrodigitization) projects efforts.

DML-CZ OCR Workflow Diagram



DML-CZ OCR Workflow – middle level of details I

① Choosing the testbed data (30.000 pages of CMJ since 1951).

② Scanning 600 DPI, 4-bit depth (soft binarization advantage).

③ Lookup for hot typefaces used in CMJ.

④ Training the Fine Reader (FR) 8.0 OCR engine for the fonts used.

⑤ Training the Lingua::Ident Perl module for language identification of languages used in CMJ (EN, RU, F, GE, CZ, SK): very reliable statistical method based on character bigrams and trigram counts.

⑥ FR scanning using general setup profile (no specific language vocabulary used).

⑦ Evaluating the language of the scanned block.

⑧ Calling FR to scan for the 2nd time with profile appropriate to the recognized language(s).

DML-CZ OCR Workflow – middle level of details II

1. Export the result as layered PDF (+FineReader XML).
2. Importing this PDF by InftyReader.
3. InftyReader recognition and storing the result Infty Markup Language IML (XML+MathML) and LaTeX.
4. Running (our Java) program OMLCorrector to fix some Infty Reader indeficiencies in IML.
5. Running (our Java) program OCRJoiner to compare characters in bounding boxes by FR and InftyReader and store the final result in IML.
6. Use the resulted files in further DML-CZ workflow.

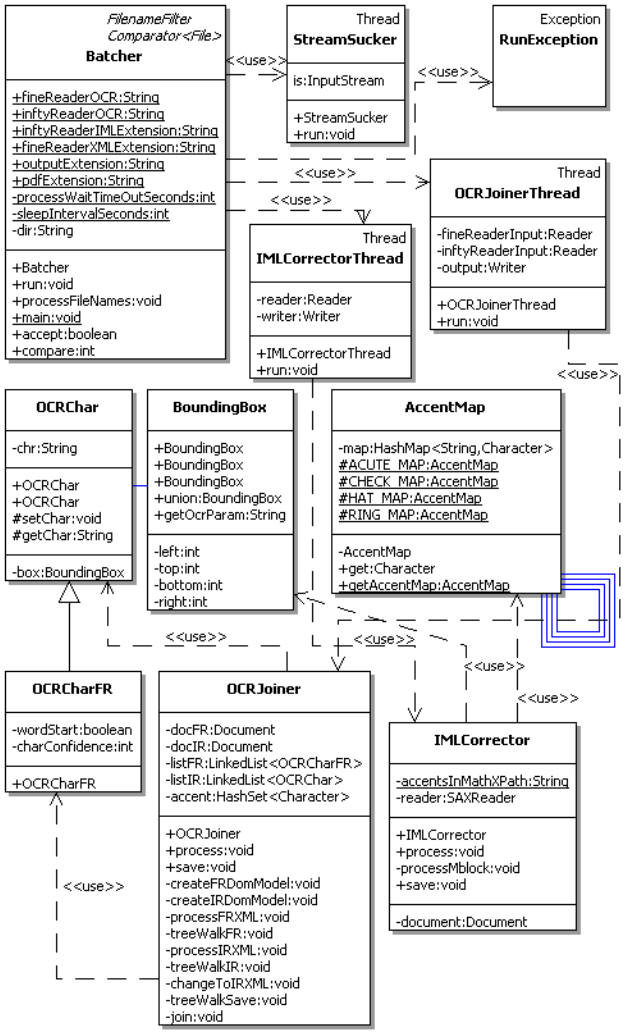
OCR XML Postprocessing

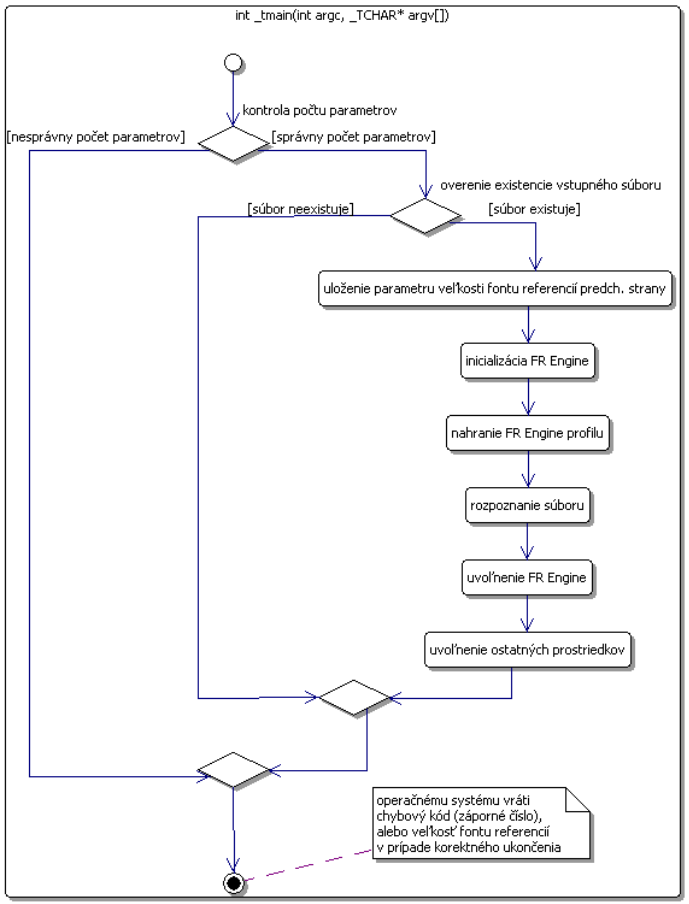
```
<mblock>
...
<munit entity="1" ocrparam="685,1746,704,1758,0">
check
<mlink type="under">
<munit ocrparam="684,1761,707,1794,0">s</munit>
</mlink>
</munit>
...
<mblock>
```

is transformed to

```
...
<char ocrparam"684,1746,707,1794" entity="1">š</char>
...
```

DML-CZ OCR Workflow Implementation Gory Details





Evaluation

Type of errors: T (text), D (diacritics), M (mathematics), L (layout)
Steps: 1 (FR1), 2 (FR2), 3 (Infty), 4 (OCRJoiner), 5 (IMLCorrector)

Step	T	D	M	L
1	10	0	224	82
2	4	0	170	78
3	4	0	168	71
4	14	0	24	15
5	14	0	24	15

DML-CZ OCR Results

Picture	FR 1	FR 2	FR8.0 PE	IR	IR fixed
1	84,99%	88,03%	88,46%	97,48%	97,48%
2	86,93%	88,76%	88,07%	98,97%	98,97%
3	89,19%	92,35%	91,53%	99,18%	99,18%
4	93,40%	93,52%	95,78%	99,15%	99,19%
5	91,09%	91,62%	92,15%	99,87%	99,87%
6	79,46%	80,05%	82,25%	99,61%	99,61%
7	92,59%	93,39%	93,71%	99,09%	99,09%
8	91,33%	91,33%	98,30%	98,18%	98,61%
Average	88,65%	89,90%	91,23%	98,97%	99,02%

Conclusions

- less than 1% error rate (counting **all** types of errors).
- still space for improvements (better text/math separation and Unicode support in InftyReader)
- possible merging of DML-CZ OCR (FR based) with Infty into one application?
- problems during FR SDK recognition of low quality images solved in the latest version of SDK

That's it!

Thanks for all contributions we build upon

The end of the story

Questions?