Towards Digital Mathematics Library (Optical Character Recognition of Mathematical Texts)

Petr Sojka, Radovan Panák, Tomáš Mudrák

Faculty of Informatics
Masaryk University in Brno

October 5th, 2006

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 Vision of World Digital Mathematics Library (WDML) in digital form

② DML-CZ project **`dml.muni.cz`**

③ not only page images, but added value wrt. Google Scholar

④ full text indexing, good searching (and ranking),

⑤ well clasified papers, with hypertext links between them and referee databases (ZentralBlatt and Math Reviews)

⑥ persistent and stable access, aimed at full (text) visibility in the global information space (Google Scholar, OAI-MPH servers, . . .)

How to Find? Search!

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

② full text searching (MathFind)

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

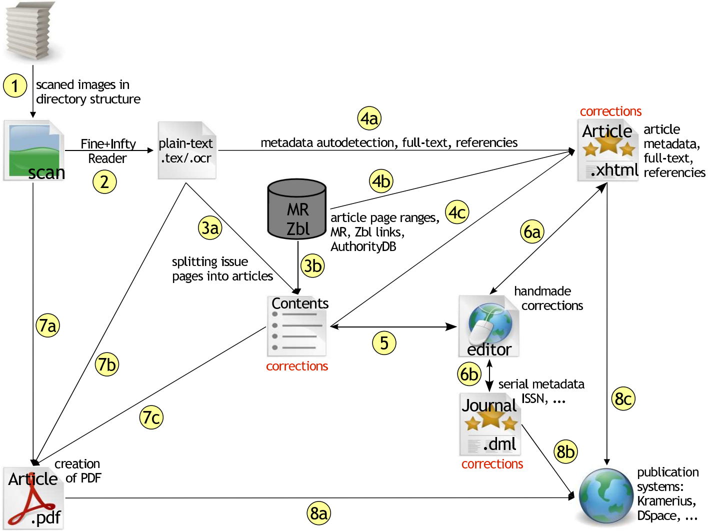
① combining both, using FineReader and InftyReader in a pipe to let every system to do what it is good for, then ‘vote’

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

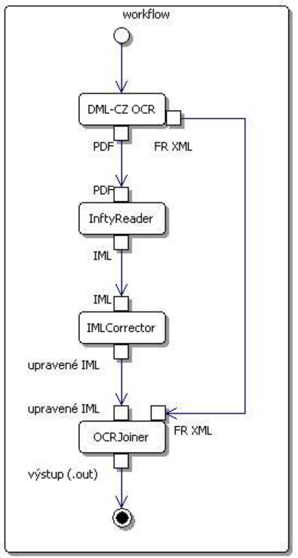
③ instant setup unusable: **fine-tuning** and **gradually enhancing** the OCR procedure and program parameters so that OCR results would be acceptable for DML-CZ purposes

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

Top-level DML-CZ Workflow



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

❶ Export the result as layered PDF (+FineReader XML).

❷ Importing this PDF by InftyReader.

❸ InftyReader recognition and storing the result Infty Markup Language IML (XML+MathML) and LaTeX.

❹ Running (our Java) program OMLCorrector to fix some Infty Reader indeficiencies in IML.

❺ Running (our Java) program OCRJoiner to compare characters in bounding boxes by FR and InftyReader and store the final result in IML.

❻ Use the resulted files in further DML-CZ workflow.

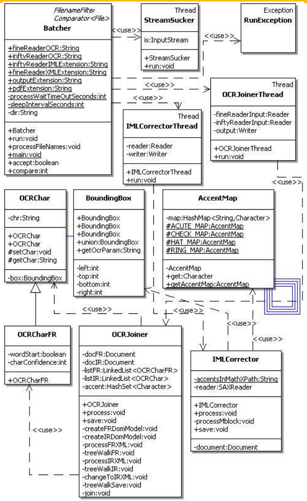
OCR XML Postprocessing

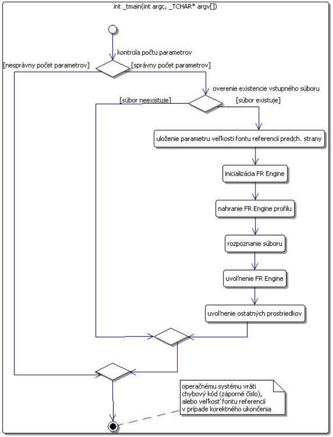
```
<mblock>
...
<munit entity="1" ocrparam="685,1746,704,1758,0">
check
<mlink type="under">
<munit ocrparam="684,1761,707,1794,0">s</munit>
</mlink>
</munit>
...
<mblock>
```

is transformed to

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

DML-CZ OCR Workflow Implementation Gory Details




int _tmain(int argc, _TCHAR* argv[])
kontrola počtu parametrov
[nesprávny počet parametrov]
[správny počet parametrov]
overenie existencie vstupného súboru
[súbor neexistuje]
[súbor existuje]
uloženie parametru veľkosti fontu referencií predch. strany
inicializácia FR Engine
nahranie FR Engine profilu
rozpoznanie súboru
uvoľnenie FR Engine
uvoľnenie ostatných prostriedkov
operačnému systému vráti chybový kód (záporné číslo), alebo veľkosť fontu referencií v prípade korektného ukončenia

Contact me, no secrets, no patents!

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%

Summary

- ☞ less than 1% character error rate (counting **all** types of errors).
- ☞ quality for math search, relevance ranking sufficient
- ☞ still space for improvements (better text/math separation and Unicode support in InftyReader)
- ☞ still space for better robustness
- ☞ still space for joining the efforts

Further Work

- storing the text in Manatee with MathML markup
- train classifiers on the papers with AMS classification tags
- compute AMS classification guesser and test it on the DML CZ papers
- experiment with math search and paper clustering
- configure and test Bonito interface for math search

That’s it!

Thanks for all contributions we build upon

The end of the story

Questions?