Towards Digital Mathematics Library DML-CZ (OCR of Mathematical Texts)

Petr Sojka

Faculty of Informatics
Masaryk University in Brno

December 8th, 2006

Digital Library of Mathematics – why?

- Information/knowledge at your fingertips!
- Vision of World Digital Math Library that will bring the enduring mathematical legacy to researchers (and students) worldwide.
- Estimation of 75.000.000 pages in total only (able to be cleverly stored in one portable 500 GB disc (EUR 200) these days).
- New science depends critically on old mathematics: 50% of current references are to pre-1990 papers, 25% to pre-1980 (Eisenbud 2004).
- 250.000 distinct authors sent papers for a review in the last decade in mathematical sciences.

Digital Library – why?

- Information overload? Yes, but we to face it anyway: Google PageRank chooses what is worth (by bibliographical global citation analysis).
- Google Scholar, Citeseer.
- ad surrogate $\longrightarrow$ ad fondes
- Implications: From factography to the art of posing **questions**.

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, ...)

Bottom-up way to WDML—DML-CZ

- Failure of global funding of DML-EU within FP6.
- Similar situation with WDML funding plans by the Gordon and Betty Moore Foundation.
- Google Print project: massive digitization of Harvard, Stanford, Oxford, University of Michigan and New York Public libraries ($150.000.000).
- Niche “markets”, grey literature, 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.

(W)DML Initiatives

NUMDAM	Numérisation de documents anciens mathématiques.
ERAM	The Jahrbuch Project—Electronic Research Archive for Mathematics (1868–1942): „Jahrbuch über die Fortschritte der Mathematik“
JSTOR	(AMS journals)
EMANI	*electronic mathematical archiving network* (Cornell, SUB Göttingen, MathDoc, Tsinghua University Library)
RusDML	Russian DML (2.000.000 pages of papers in Zbl *refereed journals*)
DML-CZ	Digital Mathematical Library of mathematical literature published in the Czech and Slovak Republics.

The Goal

- Czech Academy of Sciences grant (program Information Society) 2005–2009, **full** (retro)digitization of 50.000 pages of mathematical literature per year is planned.
- 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).

Specifics of Mathematical Publications

1. review databases where entries are **classified** according to the Math Subject Classification Scheme (MSC 2000).
2. **Zentralblatt MATH** (more than 2.000.000 entries drawn from more than 2300 serial and journals) Jahrbuch über die Fortschritte der Mathematik (JFM) covering the period 1868–1942 (200.000 entries digitized in ERAM).
3. **MathSciNet**: 24.157 items added in 2005; 1799 journals covered; links to 501.123 original articles; 11.304 active reviewers; 428.680 authors indexed. Since 1940.
4. 50 year old or even older papers are frequently cited.

Limited search in review databases, only things as collaboration distances.

On the ‘opposite’: CiteSeer, Google Scholar.

What to digitize in DML-CZ?

Final selection not yet finished: 7–8 journals, 100–200 monographs and textbooks and conference proceedings. In total 250.000 pages. Refereed journals we started with:

1. *Czechoslovak Mathematical Journal* (30.000 pages to scan, 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.

Mathematica Bohemica and *Archivum Mathematicum* already 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 in Brno** Petr Sojka (Faculty of Informatics) 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 in 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, 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.

Preparation

document selection by quality; grey literature too.

preparation acquisition of documents for scanning.

copyright negotiation with publishers (or even authors?)

In what order? What is important when signing digitization contract? Current trends in EU: paying for the rights to digitize and to the authors rights organizations for everything not older than 70 years.

Scanning

Floods in Bohemia three years ago. Many manuscripts were under water, and frozen (put into the refrigerator). Workflow for proces of defrosing includes scanning (Library of Academy of Sciences, Jenštejn near Prague, capacity of 40.000 pages per month or more!).

parameters 600 dpi 4bit depth.

scanning facilities Digibook RGB 10000, A1 color book scanner; two book scanners Zeutschel OS 7000, A2 B/W.

software Book Restorer to make the scanned pages uniform (white space around text body,...); system Sirius for archival storage of scanned materials (they are put on CDs in TIFF G4);

Optical Character Recognition

- Methods for separation of text OCR and mathematics OCR.
- Math: Infty system (Suzuki et al., Japan): 1) layout analysis, 2) character recognition, 3) structure analysis of math. expressions, and 4) manual error correction
- Text OCR by ABBYY FineReader API.
- Several OCR layers + comparison and machine learning techniques for logical tagging (AutoTag, XDOC systems).
- Quality assurance—quality matters most! 99%+ accuracy for text, 96%+ for mathematics

Metadata and Image Enhancements/Processing

metadata standards choice of standards (MODS, METS).

metadata acqusition retyping, OCR tagging, getting elsewhere?

image enhancements multiple format, PDF, DjVU conversions, software?

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

Dublin Core, miniDML or ZentralBlatt+MR? Or all? BibTeX or XML? software for digital repository? (DSpace?) bibitem handling, addition of ZBL, MR, JHR hypertext links in miniDML? Technology for doing the linking?

Storage, Indexing

space multiple OCR layers, multiple attribute layers (lemmas, reviewer comments, semantic classifications, etc.) no problems to store and index all of that for **all** mathematics literature so far.

software 1) client/server architecture, Bonito and Manatee developed at NLPLAB FI MU, used by OUP dictionary development (Oxford Thesaurus of English, 2004) based on corpora of 100.000.000 word positions, superior scaling qualities. 2) Lucene indexing software (OSS).

Document Markup Enhancement Methods

① context dependent mapping from visual to logical markup

② algorithms of language identification (bi-gram, tri-gram based, par or even sentence level)

③ document classification, metrics, ontology construction, comparison with AMS 2000 classification

④ semiautomatic bibliography markup and metrics, **global mathematics** citation index, "MathRank"

⑤ document clustering (for visualization, . . .), identification of near duplicates

Presentation

visualization techniques 'lost in hyperspace fear', vizualization of document clustering, Visual Browser (different user's eyes).

delivery system Kramerius (open source, created under contract) for scanned documents delivery.

Delivery

web portal unique and persistent URLs: Digital Object Identifier DOI (URN? PURL?,...)

interfaces to other services OAI-PMH harvesting, bibitem export, Googlebot optimization

indexing, search relevance Manatee, Lucene, or EDBM-2? mirroring? Google Scholar? Paid content model?

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**.

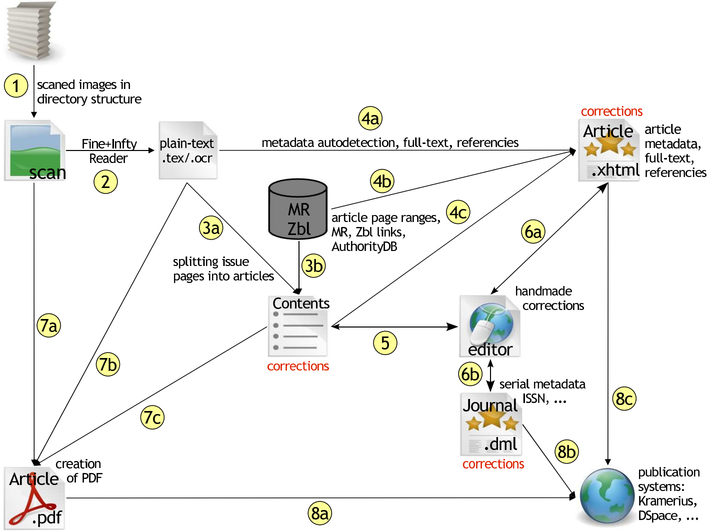
Existing OCR Systems

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.

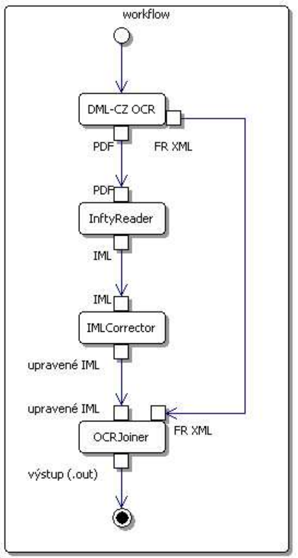
Our OCR Solution

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

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.

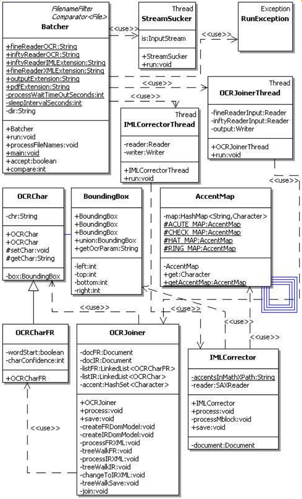
OCR XML Postprocessing

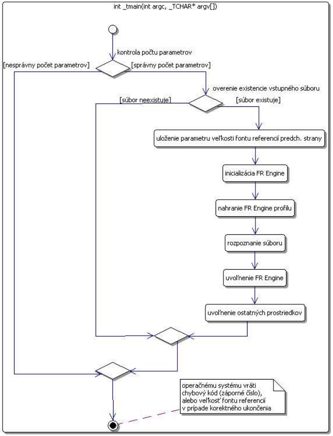
```
<mblock>
...
<munit entity="1" ocrparam="685,1746,704,1758,0">
check
<mlink type="under">
<munit ocrparam="684,1761,707,1794,0">s</munit>
</mlink>
</munit>
...
<mblock>
```

is transformed to

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

DML-CZ OCR Workflow Implementation Gory Details





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

Summary and Conclusions

> We should experiment; we should try out new things;
> we should tinker with technology and find better ways
> to communicate. **John Ewing (2002)**

We are at the start—many problems are unresolved. Preliminary project web pages are at http://dml.muni.cz/. Will Google Print and Google Scholar projects take over before (W)DML is finished (90:10% rule)?

Real data are needed to explore methods further.

Properly designed **visualization** may help to **reveal** enormous amounts of (textual) **data.** „Graphics reveal data." (Tufte)

Free access+intelligent retrieval are prerequisities of sharing knowledge.

Eisenbud: World Digital Mathematics Library.
A presentation to the Gordon and Betty Moore Foundation, August 19, 2004.

M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori.
INFTY—An integrated OCR system for mathematical documents.
Proceedings of DocEng 2003, Grenoble, France.

A. Shapiro.
TouchGraph LLC at SourceForge, 2004.
Available from: `http://touchgraph.sourceforge.net/`.

E. Tufte.
Envisioning Information.
Graphics Press, 1990.