

DML-CZ: From Scanned Image to Mathematical Knowledge Sharing

Petr Sojka

Faculty of Informatics
Masaryk University in Brno

June 29th, 2005

DML-CZ – Why?
(W)DML Initiatives

DML-CZ–What, Who
What?
Who?

DML-CZ–How?
Preparatory and Scanning Phases
Image and Metadata Processing
Indexing, Presentation, Visualization

Summary and Conclusions
Summary
Bibliography

WDML vision

- ▶ Vision of World Digital Math Library that will bring the enduring mathematical legacy to researchers 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.
- ▶ How – by creating partnership among scholars, libraries and publishers and sustaining these in the future. Digitize, give it to the publishers for free in the case of free access: Utopia? Probably yes, but.

(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.

- ▶ Failure of global funding of DML-EU within FP6.
- ▶ Funding plans (\$75.000.000) 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 government funding: DML-CZ,
 - 2) from scanned images to full text marked pages.

Specifics of Mathematical Publications

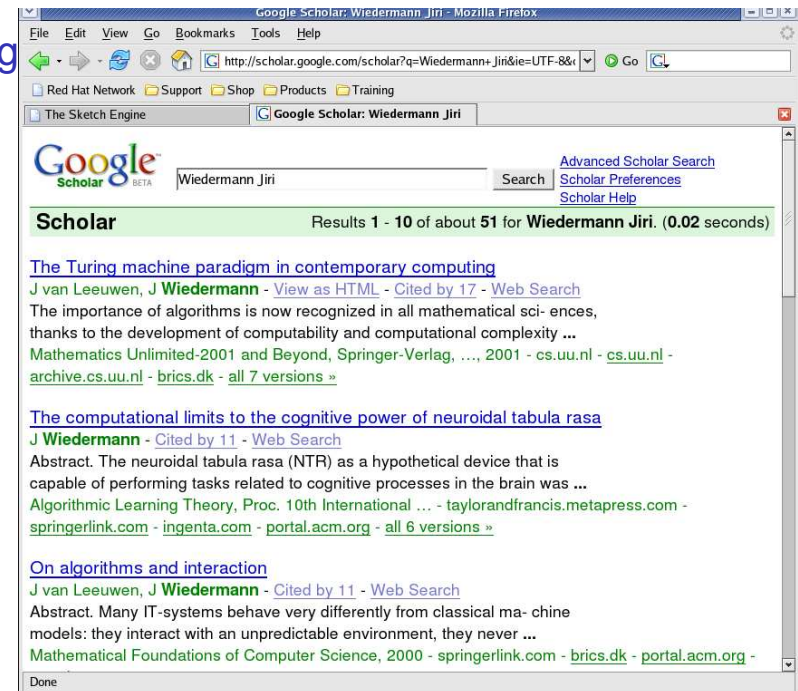
- ① review databases where entries are *classified* according to the Mathematics Subject Classification Scheme (MSC 2000).
- ② *Zentralblatt MATH* (more than 2.000.000 entries drawn from more than 2300 serial and journals) Jahrbuch über die Fortschritte der Mathematik (JFM) covering the period 1868–1942 (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.

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

On the ‘opposite’: CiteSeer, Google Scholar.

- ▶ 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
- ▶ Hardest part: sharing/delivery (economic models for knowledge sharing due to interests of content owners/publishers). More questions than answers in this crucial area.

Goog



What to digitize in DML-CZ?

Final selection not yet finished: 7–8 journals, 100–200 monographs and textbooks and conference proceedings. In total 200–300.000 pages. Refereed journals to start 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.
- ③ *Archivum Mathematicum* (2.000/4.000) Masaryk Uni in Brno.

Mathematica Bohemica already digitized in Göttingen, . . . Copyright issues crucial in negotiations.

On the way from 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.

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. Mirek Bartošek (Institute of Computer Science), content management system, metadata harvesting, long-term archiving.
- ③ **Charles University in Prague** Jiří Veselý, Oldřich Ulrych, selection and preparation of materials for digitization, metadata.
- ④ **Library of Academy of Sciences, Prague** Martin Lhoták, document scanning.

Preparation

document selection criteria?, 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 process of defrosting 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);

Metadata and Image Enhancements/Processing

metadata standards choice of standards (MODS, METS).

metadata acquisition 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?

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, 98%+ for mathematics

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

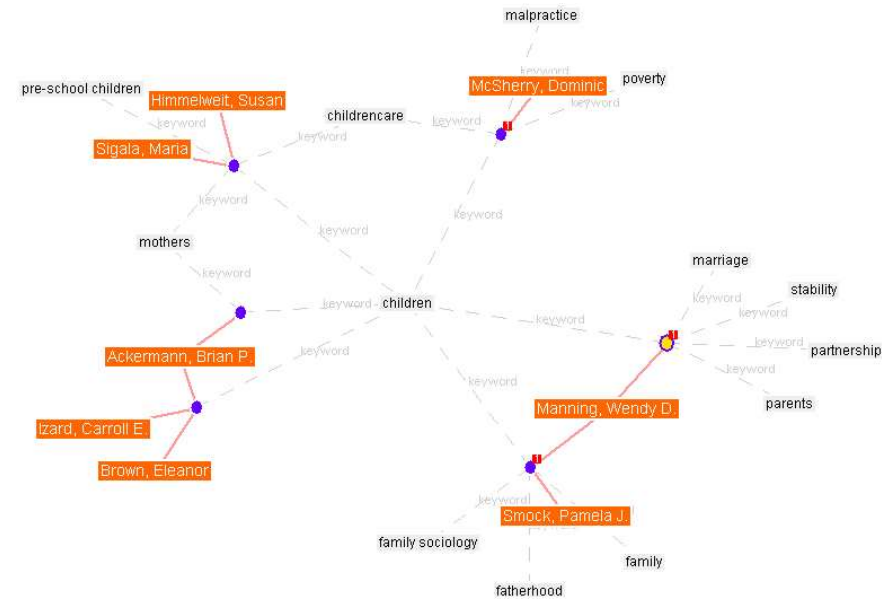
- ① context dependent mapping from visual to logical markup
- ② algorithms of language identification (bi-gram, tri-gram based, par or even sentence level)
- ③ document classification, metrics, ontology construction, comparison with AMS 2000 classification
- ④ semiautomatic bibliography markup and metrics, *global mathematics* citation index, "MathRank"
- ⑤ document clustering (for visualization, . . .), identification of near duplicates

Presentation

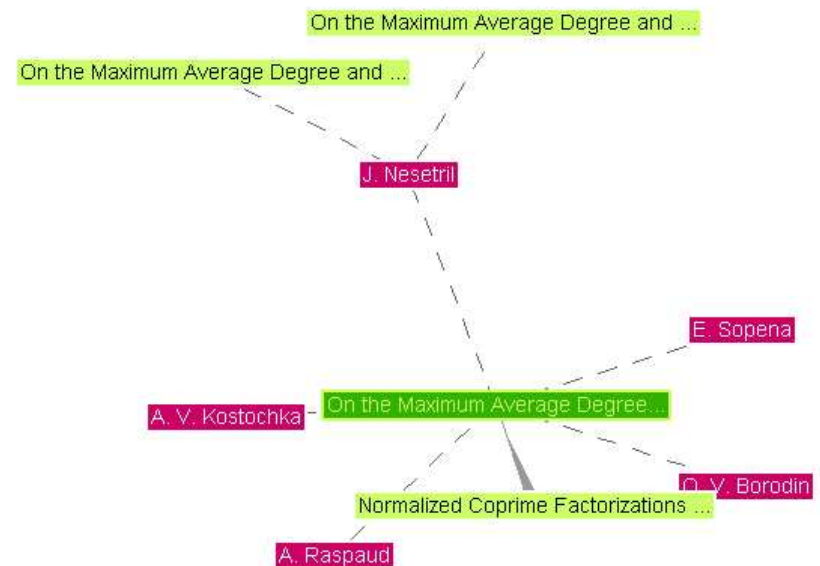
visualization techniques 'lost in hyperspace fear', visualization of document clustering, Visual Browser (different user's eyes).

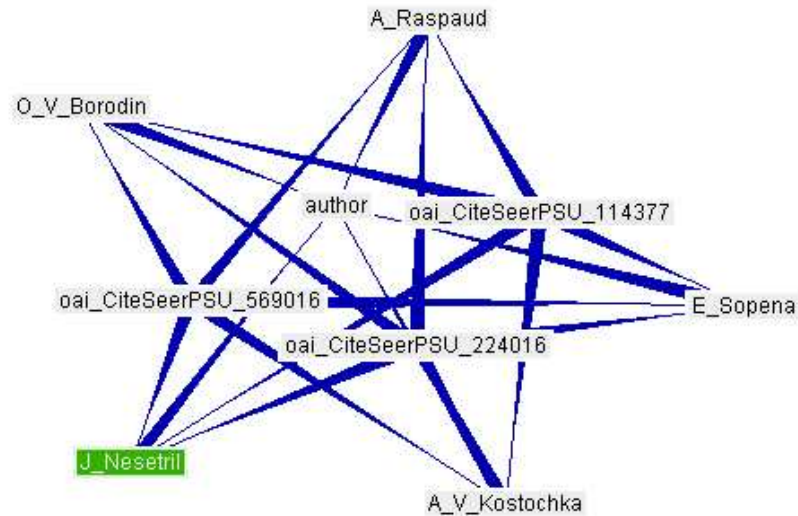
delivery system Kramerius (open source, created under contract) for scanned documents delivery.

Visualization



Visualization in Visual Browser





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?

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 prerequisites 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.