# DML-CZ: From Scanned Image to Mathematical Knowledge Sharing

Petr Sojka

Faculty of Informatics
Masaryk University in Brno

April 8th, 2005





## The Goal

- Czech Academy of Sciences grant (program Information Society).
- Five years, 2005–2009.
- Digitization of 50.000 pages of mathematical literature per year is planned.
- Contract (approx EUR 50.000 per year) signed two months ago.
- Design issues to discuss: gradual enhancement of the digital material by 'knowledge enhancing' filters on markup-rich XML data.

Many questions remain—that is why I am here.





## Bottom-up way to WDML – DML-CZ

- Vision of WDML for years.
- Failure of global funding of DML-EU within FP6.
- Google Print project: digitization of Harvard, Stanford, Oxford, U. Michigan and New York Public libraries ($150.000.000).
- Niche "markets", grey literature, mathematical literature published in CE not covered.
- Making WDML bottom up (with the help of the goverment funding): DML-CZ





## (W)DML Initiatives

We do not want to reinvent the wheel:

NUMDAM Numérisation de documents anciens mathématiques.

German digital research library at Göttingen.

EMANI electronic mathematical archiving network

... JSTOR, Cornell, Russian DML,...

DML-CZ Digital Mathematical Library of mathematical literature published in Czech.

## What to digitize?

Selection not yet finished: 5–8 journals, 100–200 conference proceedings, monographs and textbooks. In total 200–300.000 pages. First 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,...





## Who is in the project?

Four contractors:

① *Czech Academy of Sciences* 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* Martin Lhoták, document scanning.





## Phases planned

acquisition preparation, document acquisition, copyright issues handling;

scanning document scanning, 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;

presentation visualization techniques of document repository, digital library web portal, interfaces to other services and search engines for the semantic database document.





## 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?

## 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 bitonal according to BPS

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





## Scanning (cont.)—Questions

- What is really necessary to do *during* the scanning?
- Any metadata entering at this stage? retyping or OCR correction?
- OCR? Software choice?
- Quality assurance—what to check and when?
- What to outsource and what not?





## Metadata

**OCR** ABBYY FineReader? XDOC? Several OCR layers? storage of references to unresolved images (math) for future processing (AutoTag)?

**metadata** choice of, retyping or OCR tagging?

**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?





## Presentation, Visualization

**visualization techniques** 'lost in hyperspace fear', vizualization of document clustering, Visual Browser.

**web portal** unique and persistent URLs (DOI? URN? PURL?,...)

**interfaces to other services** OAI-PMH harvesting, bibitem export

**indexing, search relevance** EDBM-2?, mirroring? Google Scholar?

Any experience with the results of EU project SciX (EUR 1.000.000)? Conditions of the use of EDBM?

# 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 unresolved.
Will Google Print take over before (W)DML is finished?
Preliminary project web pages are at http://dml.muni.cz/.
Thank you for comments, hints, suggestions, possibility to make this presentation, or even for software tools and possible cooperation in the future.