# From Pixels and Minds to the Mathematical Knowledge in Digital Library[1]

Petr Sojka, Jiří Rákosník

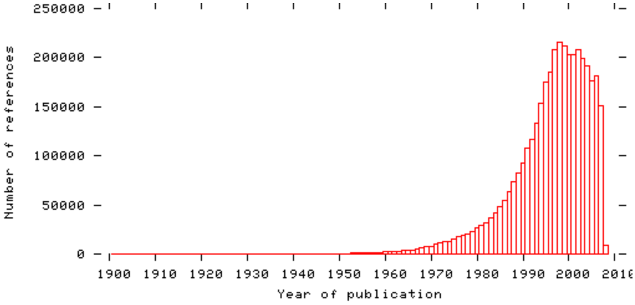
DML-CZ
Faculty of Informatics, Masaryk University, Brno

July 28th, 2008

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

# Digital Mathematics Library — motivations

- All math knowledge at your fingertips (*text or code*)!
- Using bibliographical **global** citation analysis and ranking to tackle information overload (# of references in The Collection of Computer Science bibliographies):



# Publish or perish – publication growth

*"If [in 2600] you stacked all the new books being published next to each other, you would have to move at ninety miles an hour just to keep up with the end of the line. Of course, by 2600 new artistic and scientific work will come in electronic forms, rather than as physical books and paper. Nevertheless, if the exponential growth continued, there would be ten papers a second in my kind of theoretical physics, and no time to read them."*

*Stephen Hawking*

# Publish or perish – publication growth

*"If [in 2600] you stacked all the new books being published next to each other, you would have to move at ninety miles an hour just to keep up with the end of the line. Of course, by 2600 new artistic and scientific work will come in electronic forms, rather than as physical books and paper. Nevertheless, if the exponential growth continued, there would be ten papers a second in my kind of theoretical physics, and no time to read them."*

*Stephen Hawking*

- problems with reviewing (author/reviewer discrepancy)

# From Minds to **Digital** Mathematics Library

- Going digital increases impact (citation scores) [Giles 1999]
- authors put preprints on the web, publishers eager to be indexed by search engines (75% traffic from there) ⟶ Google Scholar, Citeseer.
- – persistence of author's information on the web
- + ad surrogate ⟶ ad fondes
- + implications of digital access: from factography ⟶ **art of posing questions**.

# From Minds to **Digital** Mathematics Library

- Going digital increases impact (citation scores) [Giles 1999]
- authors put preprints on the web, publishers eager to be indexed by search engines (75% traffic from there) $\longrightarrow$ Google Scholar, Citeseer.
- – persistence of author's information on the web
- + ad surrogate $\longrightarrow$ ad fondes
- + implications of digital access: from factography $\longrightarrow$ **art of posing questions**.
- $\longrightarrow$ (W)DML!

# From [old] minds to **Library** (via pixels): (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. |

# 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**: 2,329,742 publications (May 20th, 2008), 80,000 new items and 60,000 reviews added each year; 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.

# Mathematical Knowledge Library – **who** should care?

1. publishers?

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)?

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. *Google Scholar?*

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. Google Scholar? have IT, money, no interest
4. Librarians?

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. Google Scholar? have IT, money, no interest
4. Librarians? have money (sponsors, culture heritage,. . . ), little interest (in math)
5. **D**igital ⟶

# Mathematical Knowledge Library – **who** should care?

① publishers? have money, have IT, but no interest and sometimes continuity

② mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT

③ *Google Scholar*? have IT, money, no interest

④ Librarians? have money (sponsors, culture heritage,. . . ), little interest (in math)

⑤ **D**igital ⟶Computer Scientist

⑥ **M**athematical ⟶

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. Google Scholar? have IT, money, no interest
4. Librarians? have money (sponsors, culture heritage,...), little interest (in math)
5. **D**igital ⟶Computer Scientist
6. **M**athematical ⟶Mathematicians
7. **L**ibrary ⟶

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. Google Scholar? have IT, money, no interest
4. Librarians? have money (sponsors, culture heritage,...), little interest (in math)
5. **D**igital ⟶Computer Scientist
6. **M**athematical ⟶Mathematicians
7. **L**ibrary ⟶Librarians (sustainability)

# Mathematical Knowledge Library – **who** should care?

1. publishers? have money, have IT, but no interest and sometimes continuity
2. mathematical institutions (EMS, AMS, CEIC)? have interest, but no money, a little IT
3. Google Scholar? have IT, money, no interest
4. Librarians? have money (sponsors, culture heritage,...), little interest (in math)
5. **D**igital ⟶Computer Scientist
6. **M**athematical ⟶Mathematicians
7. **L**ibrary ⟶Librarians (sustainability)
8. ⟶ **all together**: NUMDAM+CEDRAM example

# Minds to D(M)L support

Better publishing support:

- institutional (Göttingen paying Springer flat fee for open access for all scientists affiliated with the university)
- making publishing easier (publishing platforms [CEDRAM] and tools [biblio servers, arXiv, YADDA]
- better capture of semantics (formalized systems or supporting semantic features of formats as MathML, OpenMath,...)
- capturing semantics as easily as possible
- different minds $\rightarrow$ different (meaning) representations $\rightarrow$ never perfect unification
- `{X \over Y\egroup` $\longrightarrow$ `\frac{X}{Y}`

# Bottom-up way to WDML—DML-CZ

- Failure of global funding of DML-EU within FP6.

# Bottom-up way to WDML—DML-CZ

- 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).

# Bottom-up way to WDML—DML-CZ

- 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.

# Bottom-up way to WDML—DML-CZ

- 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 goverment funding: DML-CZ, 2) from scanned images to full text marked pages.

# The Goal

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

# The Goal

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

# What to digitize in DML-CZ?

7–8 Czech and Slovak math journals, 100–200 monographs and textbooks and conference proceedings, in total about 250,000 pages:

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.

# What to digitize in DML-CZ?

7–8 Czech and Slovak math journals, 100–200 monographs and textbooks and conference proceedings, in total about 250,000 pages:

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* and *Archivum Mathematicum* already partially 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, Brno** Petr Sojka (FI) 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, 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;

# 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;

# 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.

# 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, [math] indexing;

# On the way from digital image to knowledge

**acquisition** preparation, document acquisition, copyright issues handling;

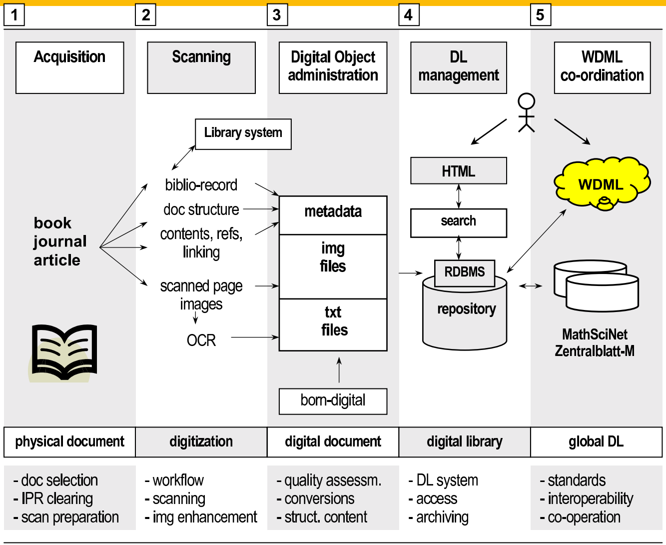
**scanning** document scanning (1/5 of the budget only) main metadata entering, scanning checks;

**image processing** main OCR, image enhancements.

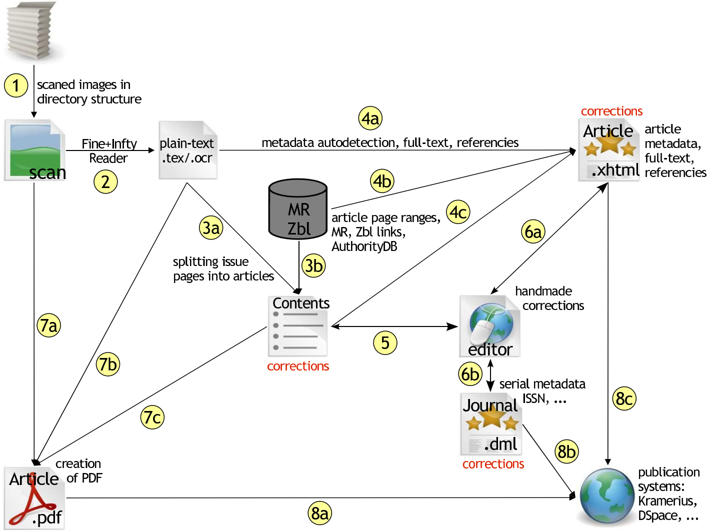
**semantic processing** document markup enhancement, semantic processing, document classification, citation linking, document clustering, [math] indexing;

**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.

# DML-CZ workflow steps



# Top-level DML-CZ workflow overview (simplified)



## Project DML-CZ, partners, goals, methods, workflow overview

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

$$P(K, v) = \int_K f(x)\,dx\,. \tag{89}$$

The functional determinant $T$ of the mapping $\psi = \varphi^{-1}$ fulfils the relation $T(\varphi(x))\,.\,\det M(x) = 1$, so that

$$\int_K f(x)\,dx = \int_{\hat{K}} f(\psi(y))\,.\,|T(y)|\,dy = \int_{\hat{K}} \hat{f}(y)\,dy\,. \tag{90}$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y)\,dy$, 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 $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\,dx$, 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*

557

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

$$P(K, v) = \int_K f(x)\, dx\,. \tag{89}$$

The functional determinant $T$ of the mapping $\psi = \varphi^{-1}$ fulfils the relation $T(\varphi(x))\,.\,\det M(x) = 1$, so that

$$\int_K f(x)\, dx = \int_{\hat{K}} f(\psi(y))\,.\,|T(y)|\, dy = \int_{\hat{K}} \hat{f}(y)\, dy\,. \tag{90}$$

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{v})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{v}) = \int_{\hat{K}} \hat{f}(y)\, dy$, 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 $\|A\| = \sup \int_A \sum_{i=1}^{m} \frac{\partial v_i(x)}{\partial x_i}\, dx$, где $v_1, \ldots, v_m$ — многочлены такие, что $\sum_{i=1}^{m} v_i^2(x) \leq 1$ 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*

557

ИОСИФ ВИССАРИОНОВИЧ СТАЛИН

1879—1953

# Preparation

**document selection** by quality, but 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 :-(. Following NUMDAM :-).

"I have worked for the digital math library in different committees since 1992, and now I am tired of this topic. The main obstacles are of legal nature (misuse of copyright laws by big commercial publishers), and we missed some opportunities along the way." Peter Michor

# 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 as TIFFs);

# Optical Character Recognition

- Text OCR by two phase DML-OCR implemented with ABBYY FineReader SDK 8.1.

# Optical Character Recognition

- Text OCR by two phase DML-OCR implemented with ABBYY FineReader SDK 8.1.
- Errors in math $\longrightarrow$ 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

# Optical Character Recognition

- Text OCR by two phase DML-OCR implemented with ABBYY FineReader SDK 8.1.
- Errors in math ⟶ 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
- Multilayer PDF with several OCR layers (text, math in TeX, math in MathML or OMDoc)
- 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** Zbl/MR, OCR tagging, [retyping]

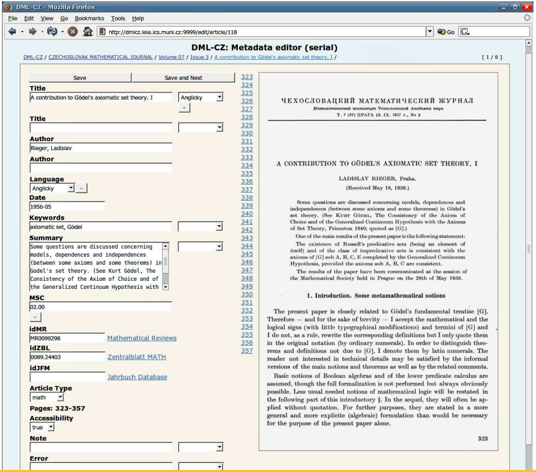
**image enhancements** TIFF, PDF, jbig2 compression as a measure of quality

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

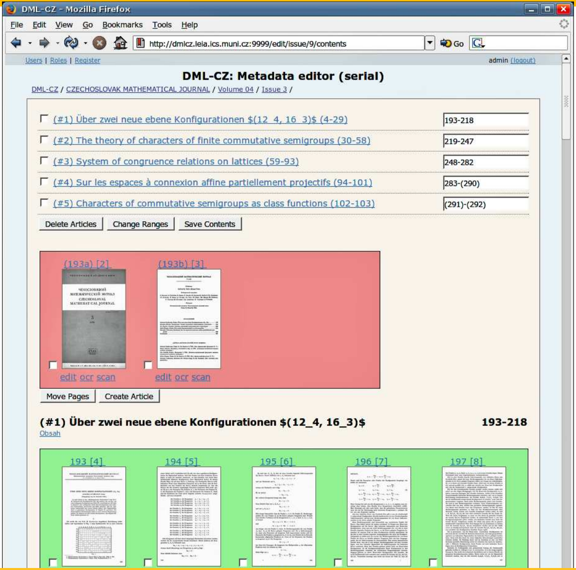
References and fulltexts are metadata as well, English titles and MSC mandatory. OAI-MPH export.

# Metadata Editor http://editor.dml.cz

Web-based client-server tool, developed (ICS MU) from scratch (Python) for metadata import, editing and checking.



# Metadata Editor



# 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

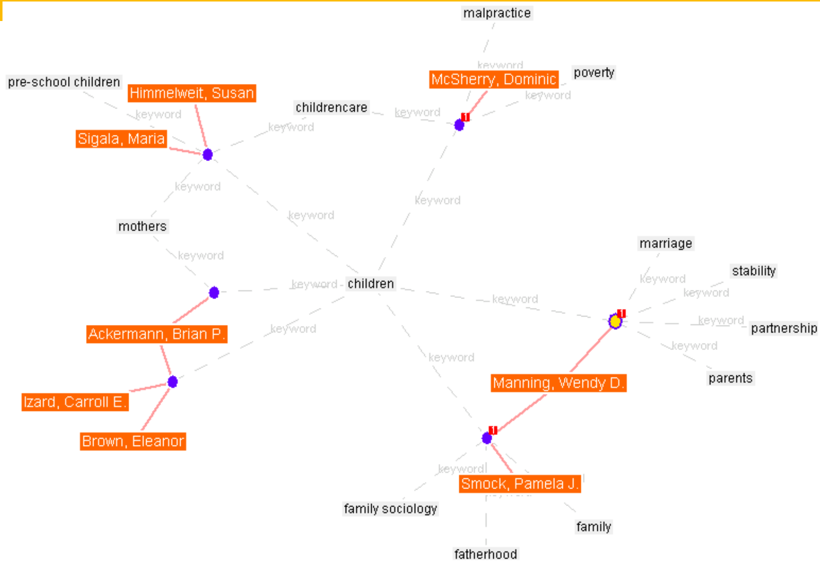
② algorithms of language identification (bi-gram, tri-gram based, par or even sentence level)

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

# Visualization

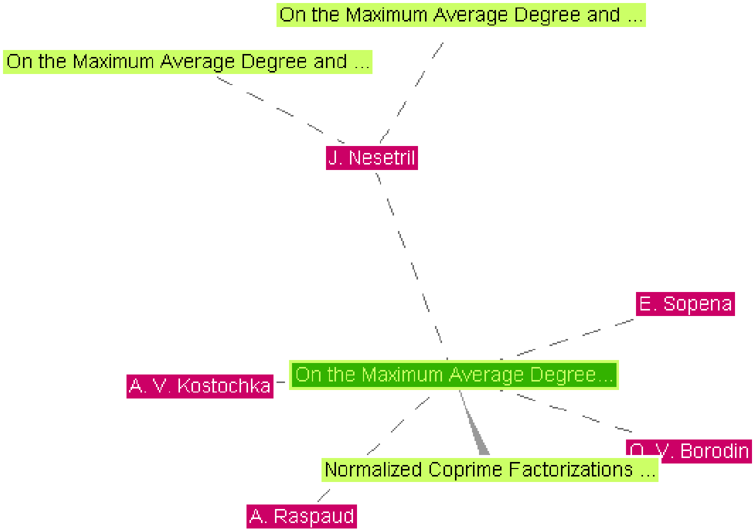


# Presentation

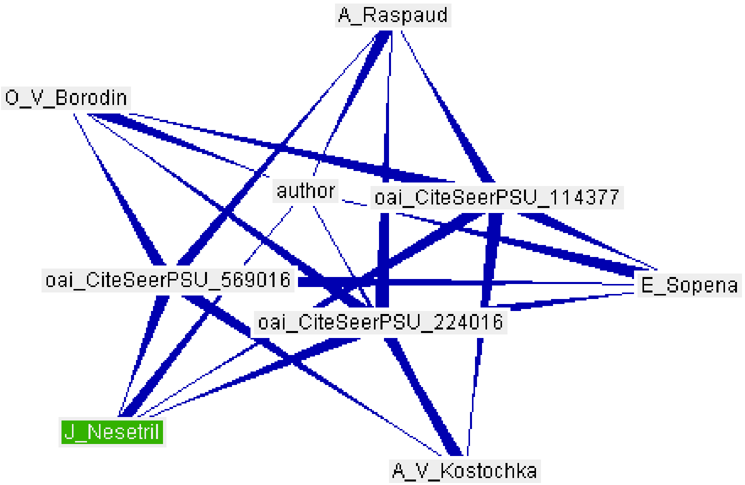
**visualization techniques** 'lost in hyperspace fear', vizualization of document clustering, Visual Browser (different user's eyes).

**delivery** customised digital library system DSpace (open source, created at MIT) for final articles delivery, search. Manakin interface.

# Visualization in Visual Browser



# Visualization in Visual Browser



# 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** Lucene, customized for math. (Experiments with Manatee and EDBM-2 (Zbl, NUMDAM))?

# Delivery: Thierry's CMUC example

**GDZ** : Goettingen

**DML-CZ** : Brno/Prague

# Paper Classification

1. every math journal paper today classified by MSC (five alphanumerical letter code) taxonomy
2. one primary, several secondary MSC
3. useful for search narrowing, clustering, documenent distance basis
4. old papers were not classified when published or reviewed

# Mathematical Paper Classification and Categorization

We thrive in information-thick worlds because of our marvelous and everyday capacity to select, edit, single out, structure, highlight, group, pair, merge, harmonize, synthesize, focus, organize, condense, reduce, boil down, choose, **categorize**, catalog, **classify**, list, abstract, scan, look into, idealize, isolate, discriminate, distinguish, screen, pigeonhole, pick over, sort, integrate, blend, inspect, filter, lump, skip, smooth, chunk, average, approximate, cluster, aggregate, outline, summarize, itemize, review, dip into, flip through, browse, glance into, leaf through, skim, refine, enumerate, glean, synopsize, winnow the wheat from the chaff and separate the sheep from the goats.
Edward R. Tufte

1. every math journal paper today classified by MSC (five alphanumerical letter code) taxonomy (tree)
2. one primary, several secondary
3. useful for search narrowing, MSC 1991, MSC 2000, MSC 2010

# Automated MSC Classification Experiment

To date (March 2008), in the digitized part there are 369 volumes of 14 journals and book collections: 1,493 issues, 11,742 articles on 177,615 pages. From NUMDAM, we got another 15,767 full texts of articles (in simple XML format) for an experiment.

1. several different languages
2. trained on papers with one primary MSC
3. NLP lab's GVP project code as basis

# Automated MSC Machine learning

**tokenization and lemmatization:** the first part of the preprocessing relates to how the text is split into tokens (words)—alphabetic, lowercase, Krovetz stemmer, lemmatization, bi-gram tokenization;

**feature selectors:** how to choose the tokens that discriminate best—$\chi^2$, mutual information (MI-score);

**feature amount:** how many features are needed to classify best—500, 2,000 or 20,000 features;

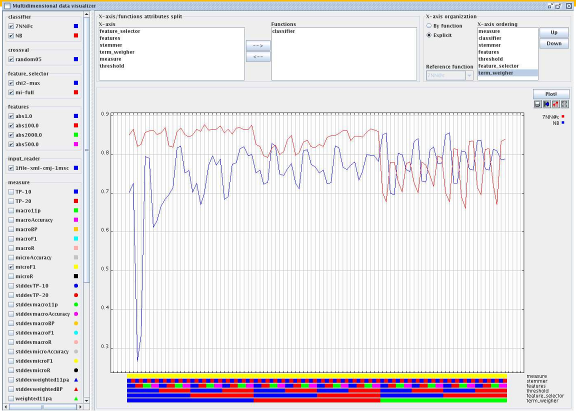
**term weighting:** how the features will be weighted (**tfidf** variants and weights normalizations (**atc** (augmented term frequency), **bnn** and **nnn**));

**classifiers:** Naïve Bayes (NB), $k$-Nearest Neighbours ($k$NN), Support Vector Machines (SVM), Artificial Neural Nets (ANN);

**threshold estimators:** how to choose the category status of the classifier based on a threshold—**fixed** or **s-cut** strategy for threshold setting;

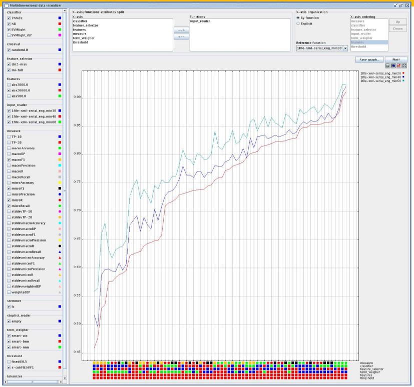
**evaluation and confidence estimation:** how results are measured and how the confidence is estimated in them—Receiver Operating Characteristic (), Normalized Cross Entropy (NCE).

# GVP Framework for comparing learning methods



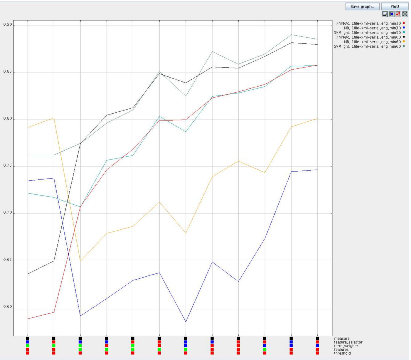

The two differently colored curves correspond to the chosen learning methods ($k$-NN, Naïve Bayes in the legend on the right). From the colors below chosen function values, one immediately sees which combination (at the bottom) of preprocessing methods

# Dependency of performance on the number of examples per class limit



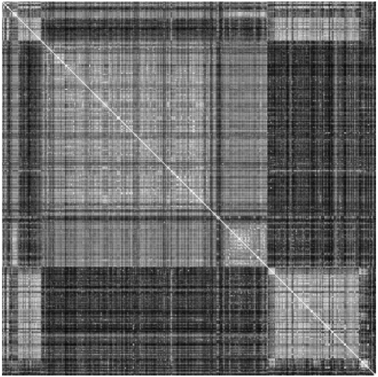

From the three curves one can see that by increasing the threshold of minimum category size one gets better results in every aspect (color square combination at the bottom).

# Classifiers' learning methods comparison by $F_1$ measure




SVM and *k*NN run hand in hand while NB lags behind. The major influence is due to the threshold on minimum category size.

# Detail of MSC-sorted documents' similarity matrix



Matrix computed by LSA for top-level MSC code 20-xx **Group theory and generalizations**. The white lower right square corresponds to the 20Mxx **Semigroups** subject papers. We can see strong similarity of 20Mxx to 20.92 **Semigroups, general theory** and 20.93 **Semigroups, structure and classification** (white lower left and upper right rectangles).

# Metadata from born-digital papers

1. main idea: metadata exported as a side-effect of publishing printed journal issues with only minimal additional costs (by requirement of proper tagging).
2. references, full text for searching
3. minimal changes in the workflow
4. *Archivum Mathematicum* pilot project.

# Pilot project of Archivum Mathematicum

1. inspired by CEDRAM
2. papers in LaTeX with AMS styles, references in BibTeX.
3. new styles files by Michal Růžička
4. automated typesetting, page numbering, EMIS web page generation,...
5. use of configurable Tralics converter to XML
6. high automation by program make
7. automated import to DML-CZ
8. first issue already available

# How to Find? Search!

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

# How to Find? Search!

1. an entry gate to the digitized papers is **search**
2. full text searching, searching for intext references

# How to Find? Search!

1. an entry gate to the digitized papers is **search**
2. full text searching, searching for intext references
3. search and exchange of **mathematical formulas** in MathML, OpenMath: project Mathdex

# How to Find? Search!

1. an entry gate to the digitized papers is **search**
2. full text searching, searching for intext references
3. search and exchange of **mathematical formulas** in MathML, OpenMath: project Mathdex
4. 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.

# 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.

# 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.

# 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.

# 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'

# 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

# 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

# Our OCR Solution

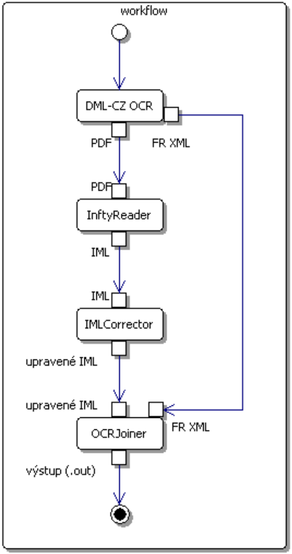
① combining both, using FineReader and InftyReader in a pipe to let every system to do what it is good for, then 'vote'

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, wait for next talk), 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.

# 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.

# 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.

# 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.

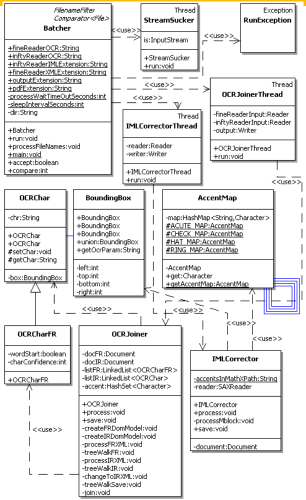
# OCR XML Postprocessing

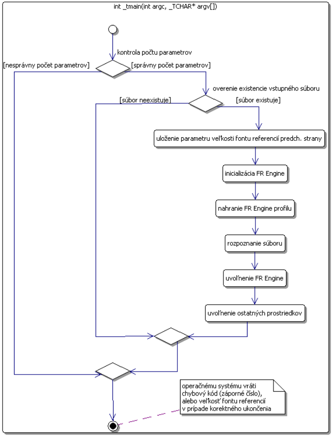
```
<mblock>
...
<munit entity="1" ocrparam="685,1746,704,1758,0">
check
<mlink type="under">
<munit ocrparam="684,1761,707,1794,0">s</munit>
</mlink>
</munit>
...
<mblock>
```

is transformed to

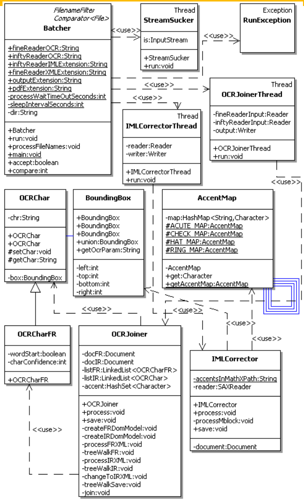
```
...
<char ocrparam"684,1746,707,1794" entity="1">š</char>
...
```

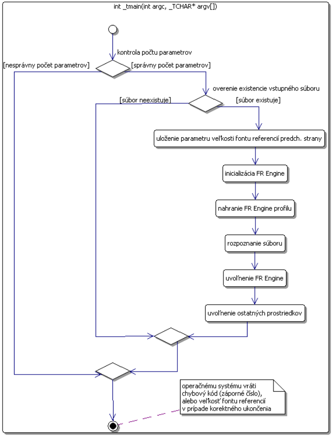
# DML-CZ OCR Workflow Implementation Gory Details



# 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% |

# OCR—Conclusions

- ☞ less than 1% error rate (counting **all** types of errors).

# OCR—Conclusions

- less than 1% error rate (counting **all** types of errors).
- still space for improvements (better text/math separation and Unicode support in InftyReader)

# OCR—Conclusions

- less than 1% error rate (counting **all** types of errors).
- still space for improvements (better text/math separation and Unicode support in InftyReader)
- still space for better robustness and precission
- several bachelor (Vystrčil) and diploma thesis (Panák, Mudrák) using FR SDK

# 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)**

Preliminary DML-CZ project web pages are at http://dspace.dml.cz/ and http://project.dml.cz/.

Metadata editor. MSC classification, math document similarity. New born-digital workflow (pilot project of Archivum Mathematicum).

TODO: Even better **math OCR**. **EuDML project** integration—real data are needed to explore methods (classification, similarity, OCR) further.

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

- S. Lawrence, C.L. Giles, and K. Bollacker, *Digital Libraries and Autonomous Citation Indexing*, Computer, June 1999, pp. 67–71.
- M. Bartošek, M. Lhoták, J. Rákosník, P. Sojka, M. Šárfy: *DML-CZ: The Objectives and the First Steps.* book chapter in a forthcoming book by A.K. Peters Ltd., 2008. pp. 69–79.
- Eisenbud: World Digital Mathematics Library.
  *A presentation to the Gordon and Betty Moore Foundation*, August 19, 2004.
- R. Řehůřek, P. Sojka: *Automated Classification and Categorization of Mathematical Knowledge* Intelligent Computer Mathematics [Proceedings of 7th International Conference on Mathematical Knowledge Management MKM 2008], LNCS, Springer, to appear, 15 p.
- P. Sojka: *DML-CZ: From Scanned Image to Knowledge Sharing.* In: Klaus Tochtermann, Hermann Maurer (Eds): Proceedings of KSR @ I-Know 2005 5th International Conference on Knowledge Management, pp. 664–672, June 29 - July 1, 2005, Graz.
- P. Sojka, J. Rákosník: *From Pixels and Minds to the Mathematical Knowledge in a Digital Library.* submitted to DML 2008.

P. Sojka, M. Růžička: *Single-source publishing in multiple formats for different output devices.* Tugboat, 29(1):118-124. ISSN 0896-3207. January 2008.

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.