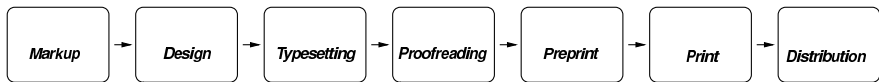
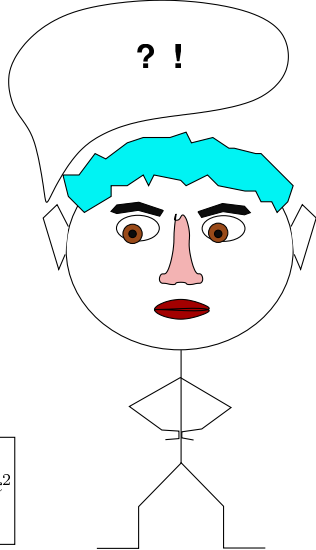
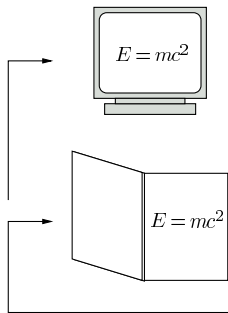


# From Minds to Pixels and Back

Petr Sojka

Faculty of Informatics, MU, Brno

October 21th, 2008



- ① Minds→Pixels: Publishing
  - Content and Form Separation

## 1 Minds→Pixels: Publishing

- Content and Form Separation

## 2 Competing Patterns

- Hyphenation Pattern Generation

## 1 Minds→Pixels: Publishing

- Content and Form Separation

## 2 Competing Patterns

- Hyphenation Pattern Generation

## 3 Thai Segmentation

# Part One: From Minds to Pixels

- 1 Minds→Pixels: Publishing
  - Content and Form Separation
- 2 Competing Patterns
  - Hyphenation Pattern Generation
- 3 Thai Segmentation
- 4 Summary of Contributions (Part One)

*Discover the outer logic of the typography in the inner logic of the text.*  
— Robert Bringhurst

- Document = **content** + **form**.
- Content should be **marked up** in author's terms and notions of domain language.

*Discover the outer logic of the typography in the inner logic of the text.*  
— Robert Bringhurst

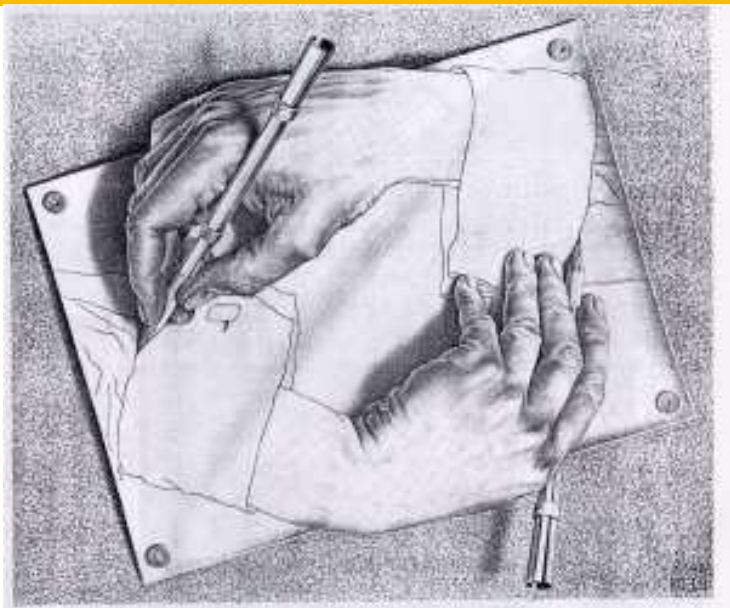
- Document = **content** + **form**.
- Content should be **marked up** in author's terms and notions of domain language.
- Form (appearance) should reflect the **design**, it should use the graphical means consistently (sameness).
- Possibilities of a form of a document are constrained by **output devices** (paper, LCD monitor, PDA).



*Discover the outer logic of the typography in the inner logic of the text.*  
—Robert Bringhurst

- Document = **content** + **form**.
- Content should be **marked up** in author's terms and notions of domain language.
- Form (appearance) should reflect the **design**, it should use the graphical means consistently (sameness).
- Possibilities of a form of a document are constrained by **output devices** (paper, LCD monitor, PDA).
- **Single-source publishing** allows structured aggregation of content and form markup and cost-effective maintenance.

# Single-source publishing from author's markup



# Content vs. visual markup

```
&\elevenit I\kern.7ptllustrations by\cr
&DU\kern-1ptANE BIBBY\cr
\noalign{\vfill}
&\setbox0=\hbox{\manual77}%
\setbox2=\hbox to\wd0{\hss\manual6\hss}%
\raise2.3mm\box2\kern-\wd0\box0\cr % A-W logo
&ADDISON\kern.1em--WESLEY\cr
&PUBLISHING COMP\kern-.13emANY\kern-1.5mm\cr
```

OK?

## Content vs. visual markup

```
&\elevenit I\kern.7ptllustrations by\cr
&DU\kern-1ptANE BIBBY\cr
\noalign{\vfill}
&\setbox0=\hbox{\manual77}%
\setbox2=\hbox to\wd0{\hss\manual6\hss}%
\raise2.3mm\box2\kern-\wd0\box0\cr % A-W logo
&ADDISON\kern.1em--WESLEY\cr
&PUBLISHING COMP\kern-.13emANY\kern-1.5mm\cr
```

OK?

NO! (at least for single-source publishing for multiple outputs)

# Single-source publishing example I: Math textbook

From one, properly marked source, multiple output versions

- optimized for print (PDF)
- optimized for LCD screen (PDF)
- optimized for web browser (portable HTML)
- optimized for web browser (scalable XML+MATHML)
- ...

From one, properly marked source, multiple output versions

- optimized for print (PDF)
- optimized for LCD screen (PDF)
- optimized for web delivery (searchable via [is.muni.cz](http://is.muni.cz))
- ...

From one, properly marked source, multiple output versions

- optimized for print (PDF)
- optimized for LCD screen (PDF)
- optimized for web delivery (searchable via [is.muni.cz](http://is.muni.cz))
- ...

# Hyphenation task

“**pattern** ORIGIN Middle English *patron* ‘something serving as a model’, from Old French. The change in sense is from the idea of patron giving an example to be copied. Metathesis in the second syllable occurred in the 16th cent. By 1700 *patron* ceased to be used of things, and the two forms became differentiated in sense.” (NODE, 1998 edition)

- Hyphenation separated from content.
- There are “**long-distance**” dependencies.
- **discreteness**: small change in input  $\Rightarrow$  fundamental change in output
- **ambiguity**: *o-blít, ob-lít; na-rval, nar-val; po-drobit, pod-robit; wach-stube, wachs-tube; ...*
- **hard generalization, exceptions, exceptions of exceptions, ...**



# The method (competing patterns)

the way to perfection (space & time minimization): instead of one big set of patterns, decomposition into several layered approximations (subpatterns)  $p_1$  (positive subpatterns),  $p_2$  (negative subpatterns—exceptions for  $p_1$ ),  $p_3$  (positive subpatterns to cover what has not been covered by “ $p_1 \wedge \neg p_2$ ”), ...

```
h y p h e n a t i o n
p1          1n a
p1          1t i o
p2          n2a t
p2          2i o
p2          h e2n
p3 h y3p h
p4          h e n a4
p5          h e n5a t
h0y3p0h0e2n5a4t2i0o0n
h y-p h e n-a t i o n
```

How to generate  
the patterns

?

# Techniques of pattern generation

- ☞ **Stratification technique:** elimination of “not necessary” training examples speeds up learning.
- ☞ **Bootstrapping technique:** Iterative bootstrapping technique for corpus tagging and error correction.
- ☞ Final parameters of patterns generation setting—**Fine tuning:** With parameters of learning process we can fine-tune size and quality of patterns (general problem of finding minimal full coverage patterns is NP optimization class of problems).

# Stratified sampling technique

“A large body of information can be comprehended reasonably well by studying more or less random portions of the data. The technical term for this approach is **stratified sampling.**” Knuth, 1991

Example of stratification rule for e.g. hyphenation task:

- 1 only every 7th (actually 17th worked as well) derived word form from the full list added to the PATGEN input list, with exceptions that:
- 2 every stem must be accompanied by at least 1 derived form, and
- 3 every derived form with overlapping prefixes has to be present in the PATGEN input list as well, and
- 4 only one word with prefixes **ne** (by which one can create negation to almost every word) and **nej** (by which one creates superlatives) is included.

# Parameters of pattern generation

- ☞ heuristics for pattern acceptance/addition in given level:  
 $good * good\_weight - bad * bad\_weight \geq threshold$

Table: Liang's patterns for English (hyphen.tex)

level	length	param	hyphens	% correct	% wrong	# patterns
1	2-3	1 2 20	67604 14156	76.6	16.0	+ 458
2	3-4	2 1 8	7407 11942	68.2	2.5	+ 509
3	4-5	1 4 7	13198 551	83.2	3.1	+ 985
4	5-6	3 2 1	1010 2730	82.0	0.0	+1647
5	5-8	1 ∞ 4	1320 6428	89.3	0.0	+1320

4447 patterns, 1 hour CPU (PDP-10), total size 27667 B

# PATGEN statistics for Czech hyphenation

Table: Standard Czech hyphenation with Liang's parameters for English

level	length	param	% correct	% wrong	# patterns	size
1	2-3	1 2 20	96.95	14.97	+ 855	
2	3-4	2 1 8	94.33	0.47	+1706	
3	4-5	1 4 7	98.28	0.56	+1033	
4	5-6	3 2 1	98.22	0.01	+2028	32 kB

Table: Standard Czech hyphenation with improved (size optimized) strategy

level	length	param	% correct	% wrong	# patterns	size
1	1-3	1 2 20	97.41	23.23	+ 605	
2	2-4	2 1 8	85.98	0.31	+ 904	
3	3-5	1 4 7	98.40	0.78	+1267	
4	4-6	3 2 1	98.26	0.01	+1665	23 kB

# Czech/Slovak hyphenation

# of words	# of hyphenation points		
	Correct	Wrong	Missed
Czech			
372562	1019686 (98.26 %)	39 (0.01 %)	18086 (1.74 %)
Slovak			
333139	1025450 (98.53 %)	34 (0.01 %)	15273 (1.47 %)

Table: Standard Czech hyphenation  
with improved (% of correct optimized) strategy

level	length	param	% correct	% wrong	# patterns	size
1	1-3	1 5 1	95.43	6.84	+2261	
2	1-3	1 5 1	95.84	1.17	+1051	
3	2-5	1 3 1	99.69	1.24	+3255	
4	2-5	1 3 1	99.63	0.09	+1672	40 kB

Table: Czech hyphenation of composed words  
(Liang but allowing 1-length patterns in level 1)

level	length	param	% correct	% wrong	# patterns	size
1	1-3	1 2 20	72.97	14.32	+ 300	
2	2-4	2 1 8	69.32	3.09	+ 450	
3	3-5	1 4 7	84.09	4.02	+ 870	
4	4-6	3 2 1	82.61	0.33	+2625	25 kB

# Czech hyphenation of compounds

Table: Czech hyphenation of compound words  
(% of correct slightly optimized)

level	length	param	% correct	% wrong	# patterns	size
1	1-3	1 2 20	72.97	14.32	+ 300	
2	2-4	2 1 8	69.32	3.09	+ 450	
3	3-5	1 4 3	90.82	4.24	+3014	
4	4-6	3 2 1	89.07	0.36	+2770	40 kB

Table: Czech hyphenation of compound words with parameters  
(% of correct optimized, but % of wrong and size increase)

level	length	param	% correct	% wrong	# patterns	size
1	1-3	1 5 1	64.35	5.34	+1415	
2	2-4	1 5 1	67.10	1.88	+1261	
3	3-5	1 3 1	97.94	5.39	+8239	
4	4-6	1 3 1	97.91	1.14	+2882	84 kB



# Compression with patterns

- Reaching full recall the method may be viewed as compression: 6,000,000 hyphenated Czech words ( $\approx 40$  MB) can be stored in patterns of 40 kB—1:1000 ratio.
- In addition, searching for a word hyphenation in **constant time** (patterns stored in packed trie data structure) with respect to the dictionary size.
- 100,000+ hyphenated words per second on modern PC in tens of kB of space.
- The key is representing the problem as competing patterns (longer patterns beat shorter patterns as exceptions): hierarchy of exceptions.

# Thai segmentation

- Our testbed for pattern application.
- Thai: 44 consonants, 28 vowels.
- No explicit syllable, word, and sentence boundaries in paragraphs.
- No punctuation.
- We need to know when typesetting
  - ▶ at least word (and sentence) boundaries to break lines
  - ▶ `<wbr>` tag for a web browser
- Even native Thai don't agree: is a compound word one word or more?

# Thai segmentation patterns development

- Training from available Orchid corpus.
- Evaluation measures:

$$\text{Precision} = \frac{\# \text{ found well}}{\# \text{ found well} + \# \text{ bad}}$$

$$\text{Recall} = \frac{\# \text{ found well}}{\# \text{ found well} + \# \text{ missed}}$$

- *segment is correct iff both the start and the end are correctly predicted*
- *In addition, combined into a single measure*

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## Results of Thai segmentation patterns generation (8000 paragraphs from Orchid corpus)

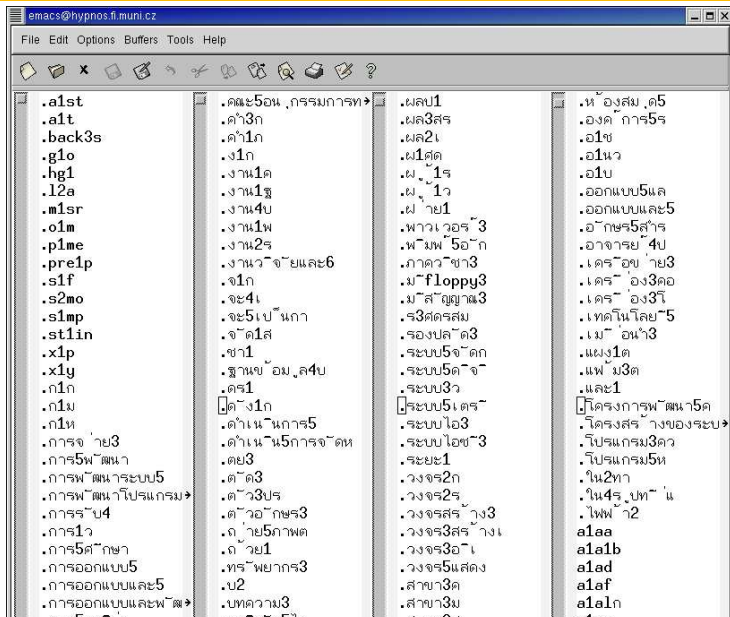
level	length	param	% correct	% wrong	# patterns	utf size (kB)
1	1-5	1 6 1	97.92	4.86	+15443	161
2	2-6	1 1 1	96.53	0.65	+ 2596	196
3	3-11	1 3 1	99.57	0.79	+ 3448	267
4	4-12	4 1 1	97.87	0.03	+ 953	286
5	9-19	1 3 1	99.68	0.12	+ 2468	364
6	10-20	1 1 1	99.67	0.04	+ 129	368

## Thai patterns generation results

- Nearly 100% precision on the training set
- Training sets 4,000, 6,000, and 8,000 paragraphs
- Tested on previously unseen text

# par.	# good	# bad	# missed	prec. %	recall %	F-score
4000	139788	11231	15529	92.56	90.00	91.26
6000	98243	7951	9432	92.51	91.24	91.87
8000	46361	3358	3703	93.25	92.60	92.92

# Thai segmentation patterns in Emacs



# Contributions summary

*Formal definition of competing patterns.* We have described and developed a new approach to language engineering based on the theory of covering and inhibiting patterns.

*New approaches to competing pattern generation.* We have verified the plausibility and usefulness of bootstrapping and stratification techniques for machine learning techniques of the pattern generation process. We have related our new techniques to those used so far—the results improve significantly with the new approach.

*Properties of the pattern generation process.* We have shown that reaching size-optimality of the pattern generation process is an NPO problem; however, it is possible to achieve full data recall and precision on the given data with the heuristics presented.

## Contributions summary (cont.)

New approach to the Thai text segmentation problem. We show that an algorithm using competing patterns learnt from segmented Thai text returns better results than current methods for this task.

Thai segmentation patterns. New patterns for the Thai segmentation problem were generated from data in the ORCHID corpus.

New Czech and Slovak hyphenation patterns. The new hyphenation patterns for Czech and Slovak give a much better performance than the previous ones, and are in practical use in distributions of text processing systems ranging from  $\text{T}_{\text{E}}\text{X}$ , SCRIBUS, OPENOFFICE.ORG to Microsoft Word.



## Contributions summary (cont.)

*New patterns for specific tasks.* Patterns for specific tasks demanded in the areas of computer typesetting and NLP were developed—phonetic hyphenation, universal syllabic hyphenation, and the possibility of using context-sensitive patterns for disambiguation tasks were shown.

*The foundation for new pattern generation algorithms.* The design of a program OPATGEN for pattern generation in an object oriented manner allows easy experimentation with new pattern generation heuristics.

*Usage of the methodology for partial morphological disambiguation.*

We have shown that the methodology of competing patterns can be used for partial disambiguation tasks. Experiments showed an improved performance for the partial morphological disambiguation of Czech.

and last but not least: ecological contribution—saving a lot of trees by better hyphenation patterns.

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

## 6 DML-CZ overview

- DML-CZ workflow: preparation, scanning, metadata, OCR, indexing, delivery

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

## 6 DML-CZ overview

- DML-CZ workflow: preparation, scanning, metadata, OCR, indexing, delivery

## 7 MSC

- Mathematical Subject Classification

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

## 6 DML-CZ overview

- DML-CZ workflow: preparation, scanning, metadata, OCR, indexing, delivery

## 7 MSC

- Mathematical Subject Classification

## 8 Publishing

- Born-digital (retro-born-digital) paper handling

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

## 6 DML-CZ overview

- DML-CZ workflow: preparation, scanning, metadata, OCR, indexing, delivery

## 7 MSC

- Mathematical Subject Classification

## 8 Publishing

- Born-digital (retro-born-digital) paper handling

## 9 OCR

- DML-CZ Optical Character Recognition: (Fine+Infty)Reader++

# From Pixels to Minds (Digitization, Tagging)

## 5 What and why?

- Better than *Google Scholar* for mathematical peer reviewed literature; both

## 6 DML-CZ overview

- DML-CZ workflow: preparation, scanning, metadata, OCR, indexing, delivery

## 7 MSC

- Mathematical Subject Classification

## 8 Publishing

- Born-digital (retro-born-digital) paper handling

## 9 OCR

- DML-CZ Optical Character Recognition: (Fine+Infty)Reader++

## 10 Summary

- Summary, Conclusions, Bibliography

# From pixels to minds? Digitization needed

- *The need to digitize.*



# From pixels to minds? Digitization needed

- *The need to digitize.*
- *Google Scholar*

# From pixels to minds? Digitization needed

- The need to digitize.
- Google Scholar `||_peer_reviewed_math`

# From pixels to minds? Digitization needed

- The need to digitize.
- Google Scholar || `peer_reviewed_math` but better!

# From pixels to minds? Digitization needed

- The need to digitize.
- Google Scholar || `peer_reviewed_math` but better!
- Vision of World Digital Math Library (WDML) that will bring the enduring mathematical legacy to researchers worldwide.

# From pixels to minds? Digitization needed

- The need to digitize.
- Google Scholar || `peer_reviewed_math` but better!
- Vision of World Digital Math Library (WDML) that will bring the enduring mathematical legacy to researchers worldwide.
- High quality, checked content, crosslinking via reviewing databases Zentralblatt MATH or Mathematical Reviews (more than 2,500,000 reviewed articles)

# From pixels to minds? Digitization needed

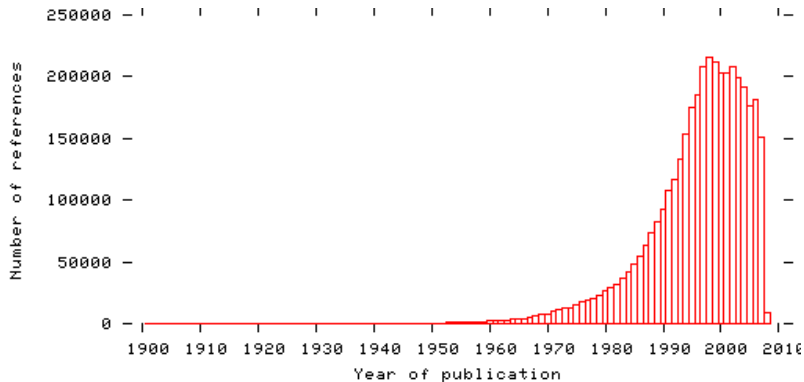
- The need to digitize.
- Google Scholar || `peer_reviewed_math` but better!
- Vision of World Digital Math Library (WDML) that will bring the enduring mathematical legacy to researchers worldwide.
- High quality, checked content, crosslinking via reviewing databases Zentralblatt MATH or Mathematical Reviews (more than 2,500,000 reviewed articles)
- Estimation of 100,000,000 pages in total only (able to be cleverly stored in one portable disc (EUR 200) today), but cannot be read in the entire (wo)man's life).

# From pixels to minds? Digitization needed

- The need to digitize.
- Google Scholar || `peer_reviewed_math` but better!
- Vision of World Digital Math Library (WDML) that will bring the enduring mathematical legacy to researchers worldwide.
- High quality, checked content, crosslinking via reviewing databases Zentralblatt MATH or Mathematical Reviews (more than 2,500,000 reviewed articles)
- Estimation of 100,000,000 pages in total only (able to be cleverly stored in one portable disc (EUR 200) today), but cannot be read in the entire (wo)man's life)..
- 250,000 distinct authors (**minds**) sent papers for a review in the last decade in mathematical sciences.

# Digital Mathematics Library – motivations

- Publish or perish – publication growth: but reviewers hard to found.
- Using bibliographical **global** citation analysis and ranking to tackle information overload (# of references in The Collection of Computer Science bibliographies):





- *Going digital increases impact (citation scores) [Giles 1999]*
- *authors put preprints on the web, publishers eager to be indexed by search engines (50% traffic from there) → Google Scholar, Citeseer.*
- *– persistence of author's information on the web*
- *+ ad surrogate → ad fontes*
- *+ implications of digital access: from factography → **art of posing questions.***

- Going digital increases impact (citation scores) [Giles 1999]
- authors put preprints on the web, publishers eager to be indexed by search engines (50% traffic from there) → Google Scholar, Citeseer.
- – persistence of author's information on the web
- + ad surrogate → ad fontes
- + implications of digital access: from factography → **art of posing questions.**
- → (W)DML!

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 years old or even older papers are frequently cited.

# Google Scholar vs. MR/Zbl

[http://scholar.google.com/scholar?q=Antonin Kucera](http://scholar.google.com/scholar?q=Antonin+Kucera)

<http://www.ams.org/mathscinet/search/publications.html?pg1=IID&s1=695584>

<http://www.ams.org/mathscinet/pdf/1992331.pdf?pg1=IID&s1=695584&r=16>

Author and institution disambiguation:

[http://www.ams.org/mathscinet/search/institution.html?code=CZ\\_MASC](http://www.ams.org/mathscinet/search/institution.html?code=CZ_MASC)

*See the difference? Hyperlinking needed for computing H-index, high quality metadata for its robustness etc.*

# The Goal: Bottom-up way to WDML—DML-CZ

- 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: Bottom-up way to WDML—DML-CZ

- 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.



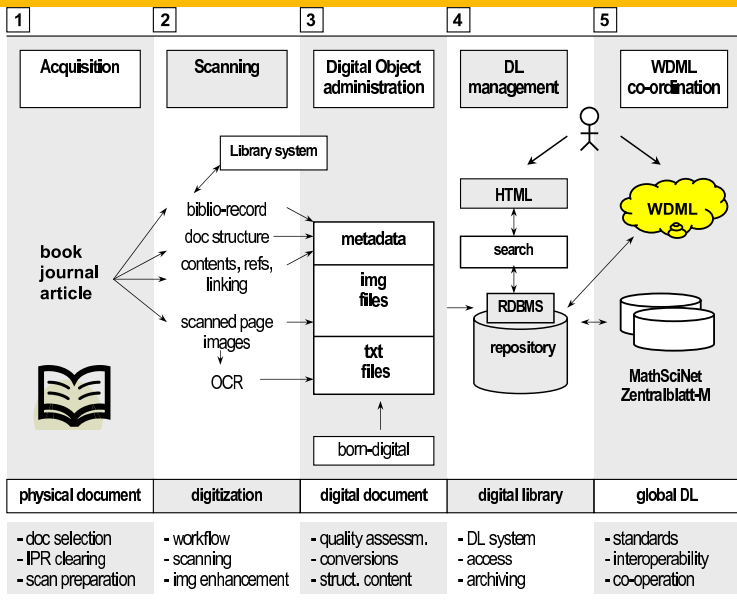
# What to digitize in DML-CZ?

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

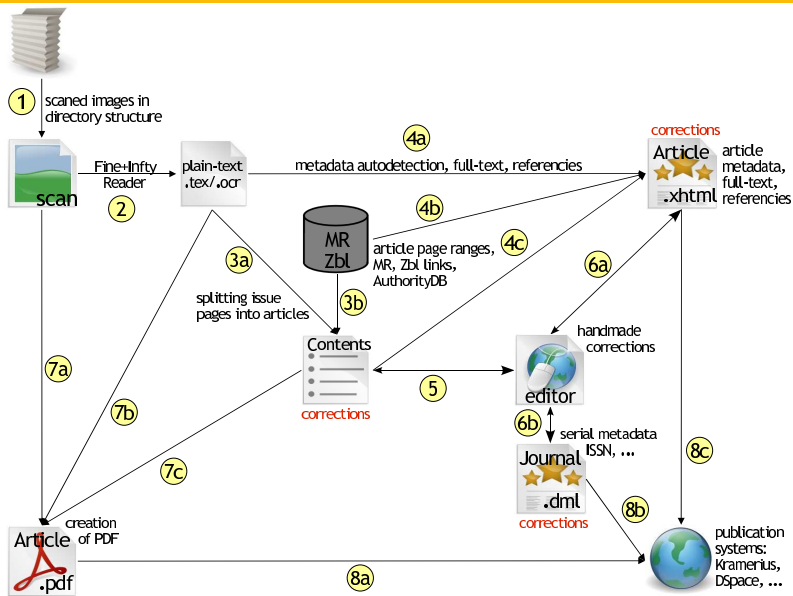
- ① *Czechoslovak Mathematical Journal* (30,000 pages to scan, 7,000 are already born digital). Published by Academy of Sciences of CR, distributed partially by Springer. Founded as *Časopis pro pěstování matematiky* in 1872, under current name since 1951. 272 pages quarterly.
- ② *Applications of Mathematics* (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.

# DML-CZ workflow steps



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



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

$$P(K, v) = \int_K f(x) dx. \quad (89)$$

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

$$\int_K f(x) dx = \int_{\hat{K}} f(\varphi(y)) \cdot |T(y)| dy = \int_{\hat{K}} \hat{f}(y) dy. \quad (90)$$

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 funkcí na daném polouspořádaném prostoru*, *Časopis pro řést. 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 řést. 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  $v_1, \dots, v_m$  — 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  $\mu$  и на

Proof. Let  $\hat{K}$  be a cube,  $\hat{K} \subset \hat{G}$ ; put  $K = \varphi^{-1}(\hat{K})$ . According to theorem 50 we have  $K \in \mathfrak{U}$  and it follows from theorem 24 that

$$P(K, v) = \int_K f(x) dx. \quad (89)$$

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

$$\int_K f(x) dx = \int_{\hat{K}} f(\varphi(y)) \cdot |T(y)| dy = \int_{\hat{K}} \hat{f}(y) dy. \quad (90)$$

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

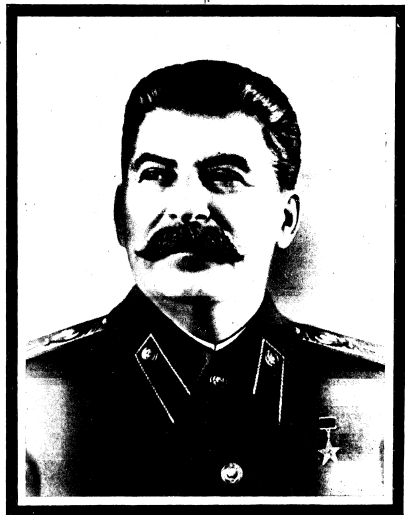
- [1] V. Jarník: *Diferenciální počet*, Praha 1953.
- [2] V. Jarník: *Integrální počet II*, Praha 1955.
- [3] J. Mařík: *Vrcholy jednotkové koule v prostoru funkcionál na daném poluospořádaném prostoru*, *Časopis pro řést. mat.*, 79 (1954), 3—40.
- [4] Ян Маржик (Jan Mařík): *Представление функционала в виде интеграла*, *Чехословацкий мат. журнал*, 5 (80), 1955, 467—487.
- [5] J. Mařík: *Plošný integrál*, *Časopis pro řést. 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  $v_1, \dots, v_m$  — многочлены такие, что  $\sum_{i=1}^m v_i^2(x) \leq 1$  для всех  $x \in A$ . Пусть  $\mathfrak{U}$  — система всех ограниченных измеримых множеств  $A$ , для которых  $\|A\| < \infty$ . Теорема 18 тогда утверждает: Пусть  $A \in \mathfrak{U}$ ; пусть  $D$  — граница множества  $A$ . Тогда на системе  $\mathfrak{B}$  всех борелевских подмножеств множества  $D$  существует мера  $r$  и на



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 defrozing 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);

- 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 → 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  $\text{T}_\text{E}\text{X}$ , 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 acquisition** 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.

**DML-CZ: Metadata editor (serial)**

Save | Save and Nest

Title: A contribution to Gödel's axiomatic set theory. I | Anglicky

Author: Finger, Ladislav

Language: Anglicky

Date: 1956-05

Keywords: axiomatic set, Gödel

Summary: Some questions are discussed concerning models, dependences and independences (between some axioms and some theorems) in Gödel's set theory. (See Kurt Gödel, The Consistency of the Axiom of Choice and of the Generalized Continuum Hypothesis with the Axioms of Set Theory, *Princeton PhD*, quoted as (G).)

MSC: 02.00

idMR: MR0099298 | [Mathematical Reviews](#)

idZBL: 0089.24403 | [Zentralblatt MATH](#)

idJFM: | [Jahrbuch Database](#)

Article Type: math

Pages: 323-357

Accessibility: true

Note:

Error:

**ЧЕХОСЛОВАЦКИЙ МАТЕМАТИЧЕСКИЙ ЖУРНАЛ**  
Математический институт Чехословацкой Академии наук  
Т. 7 (1956) СЛАВА 18. IX. 1957 г., № 4

**A CONTRIBUTION TO GÖDEL'S AXIOMATIC SET THEORY. I**

LADESLAV REIDIKER, Praha.  
(Received May 16, 1956.)

Some questions are discussed concerning models, dependences and independences (between some axioms and some theorems) in Gödel's set theory. (See Kurt Gödel, The Consistency of the Axiom of Choice and of the Generalized Continuum Hypothesis with the Axioms of Set Theory, *Princeton PhD*, quoted as (G).)

One of the main results of the present paper is the following statement:  
The existence of Russell's predicative sets (being an element of itself) and of the class of impredicative sets is consistent with the axioms of (I) sub A, B, C, E completed by the Generalized Continuum Hypothesis, provided the axioms sub A, B, C are consistent.

The results of the paper have been communicated at the session of the Mathematical Society held in Prague on the 28th of May 1956.

**1. Introduction. Some metamathematical notions**

The present paper is closely related to Gödel's fundamental treatise (G). Therefore — and for the sake of brevity — I accept the mathematical and the logical signs (with little typographical modifications) and terms of (G) and I do not, as a rule, rewrite the corresponding definitions but I only quote them in the original notation (by ordinary numerals). In order to distinguish theorems and definitions not due to (G), I denote them by latin numerals. The reader not interested in technical details may be satisfied by the informal versions of the main notions and theorems as well as by the related comments.

Basic notions of Boolean algebra and of the lower predicate calculus are assumed, though the full formalization is not performed but always obviously possible. Less usual needed notions of mathematical logic will be restated in the following part of this introductory §. In the sequel, they will often be applied without quotation. For further purposes, they are stated in a more general and more explicit (algebraic) formulation than would be necessary for the purpose of the present paper alone.

323

# Metadata Editor

DML-CZ - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://dmicz.leia.ics.muni.cz:9999/edit/issue/9/contents


Users | Roles | Register admin (logout)


## DML-CZ: Metadata editor (serial)

DML-CZ / CZECHOSLOVAK MATHEMATICAL JOURNAL / Volume 04 / Issue 3 /

<input type="checkbox"/> (#1) Über zwei neue ebene Konfigurationen $\$(12_4, 16_3)\$$ (4-29)	193-218
<input type="checkbox"/> (#2) The theory of characters of finite commutative semigroups (30-58)	219-247
<input type="checkbox"/> (#3) System of congruence relations on lattices (59-93)	248-282
<input type="checkbox"/> (#4) Sur les espaces à connexion affine partiellement projectifs (94-101)	283-(290)
<input type="checkbox"/> (#5) Characters of commutative semigroups as class functions (102-103)	(291)-(292)


Delete Articles | Change Ranges | Save Contents


(193a) [2]  [edit ocr scan](#)


(193b) [3]  [edit ocr scan](#)


Move Pages | Create Article


**(#1) Über zwei neue ebene Konfigurationen  $\$(12_4, 16_3)\$$**  **193-218**  
[Obzah](#)

193 [4] 

194 [5] 

195 [6] 

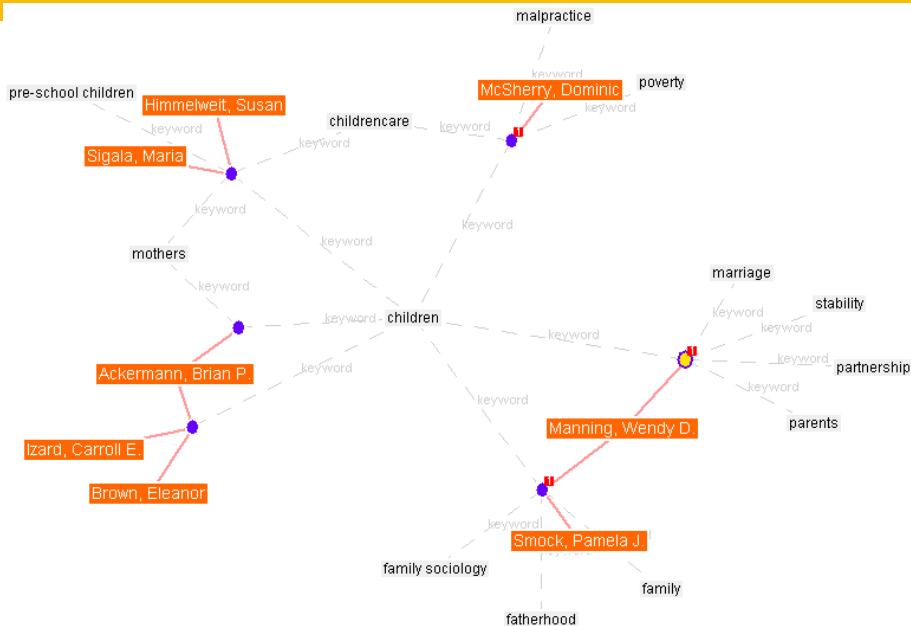
196 [7] 

197 [8] 

# 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

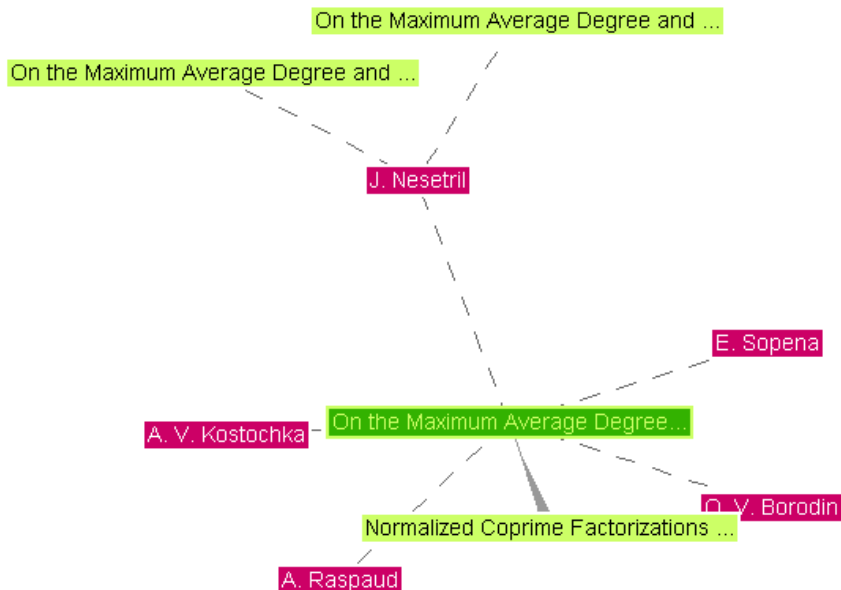


**visualization techniques** 'lost in hyperspace fear', vizualization of document clustering, Visual Browser (different user's eyes).

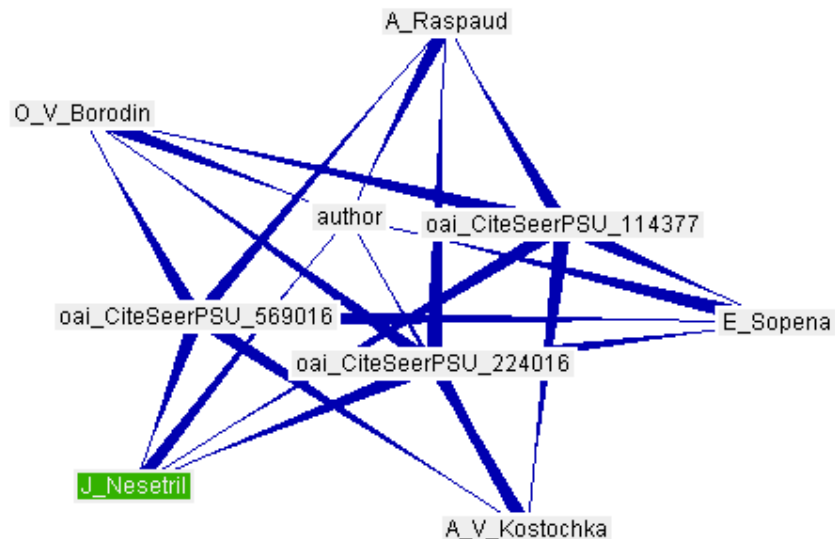
**delivery** customised digital library system DSpace (open source, created at MIT) for final articles delivery, search. Manakin interface.



# Visualization in Visual Browser



# Visualization in Visual Browser



**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))?

# 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, document 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, synopsise, 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
- ④ old papers were not classified when published or reviewed

# 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;

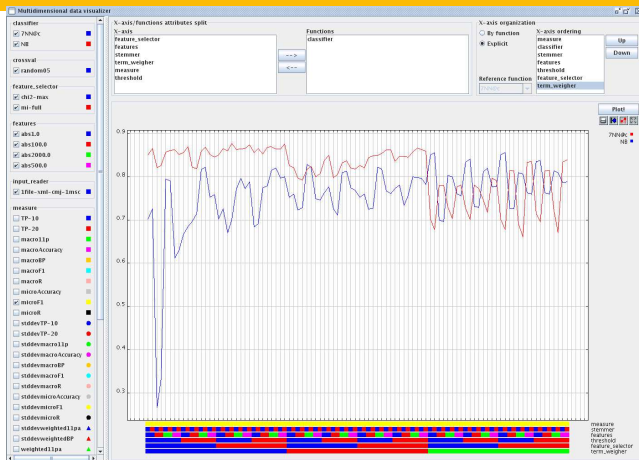
**term weighting:** how the features will be weighted (**tfidf** variants and weights normalizations (**atc** (augmented term frequency), **bnn** and **nnn**));

**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 (ROC), Normalized Cross Entropy (NCE).

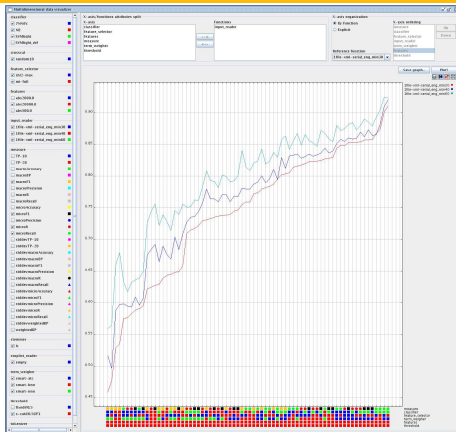
# GVP Framework for comparing learning methods



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 leads to which particular value.

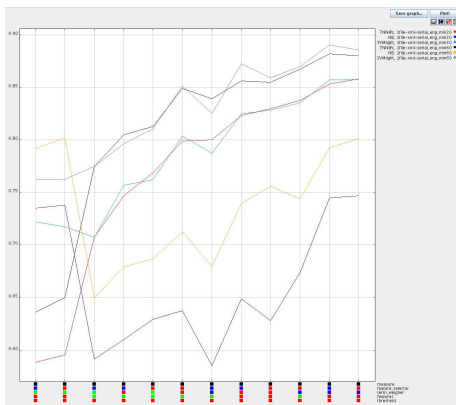


# Dependency of performance on the number of examples per class limit



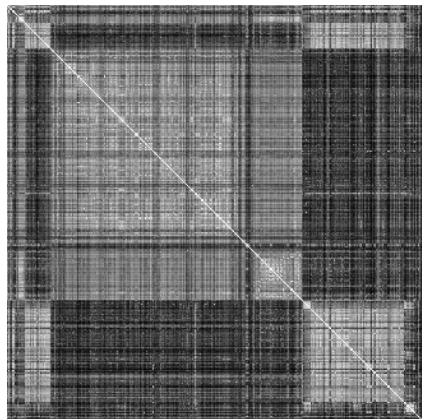
From the three curves one can see that by increasing the threshold of minimum category size one gets better results in every aspect (color square combination at the bottom).

# 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 and semantic information available is 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  $\text{\LaTeX}$  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 3 issues 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**.

- 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](http://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](http://www.inftyproject.org) the only available solution for structural math recognition.
- ⑤ No out-of-the-shelf 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

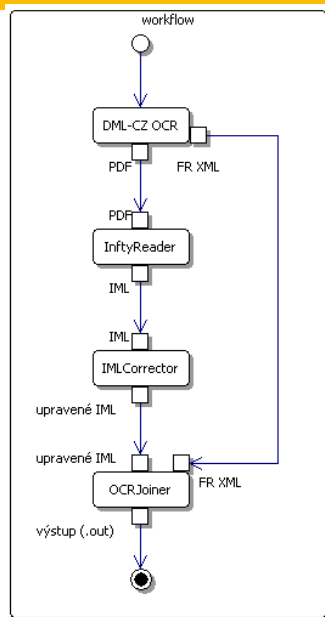
- ① combining both, using FineReader and InftyReader in a pipe to let every system to do what it is good for, then 'vote'
- ② top-level (Java) program to **automate** the process **and fix** some indeficiencies



- 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

- 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).

- ➊ Export the result as layered PDF (+FineReader XML).
- ➋ Importing this PDF by InftyReader.

- 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.

- 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.

- 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.

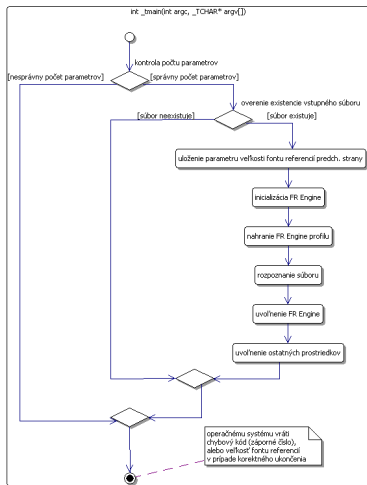
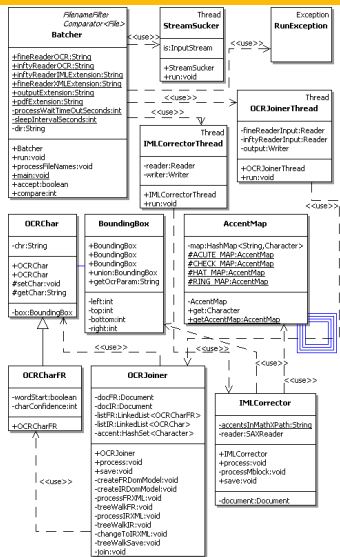


```
<mblock>
  ...
  <munit entity="1" ocrparam="685,1746,704,1758,0">
    check
    <mlink type="under">
      <munit ocrparam="684,1761,707,1794,0">s</munit>
    </mlink>
  </munit>
  ...
</mblock>
```

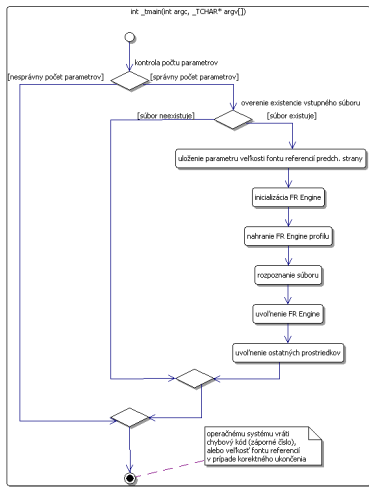
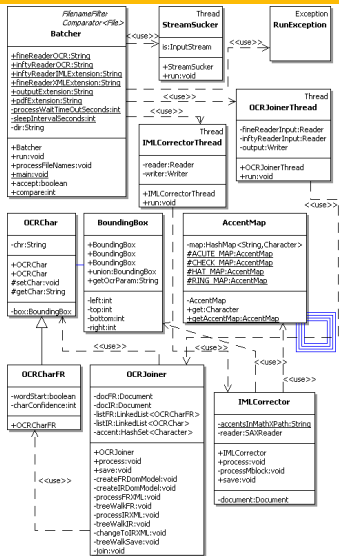
is transformed to

```
...
<char ocrparam"684,1746,707,1794" entity="1">\v{s}</char>
...
```

# DML-CZ OCR Workflow Implementation Gory Details



# DML-CZ OCR Workflow Implementation Gory Details



Contact me, no secrets, no patents!

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%

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

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

- ☞ 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 precision
- ☞ several bachelor (Vystrčil) and diploma thesis (Panák, Mudrák) using FR SDK



## Contributions (part two)

*Workflow of bulk retro-digitization. We have described and developed a new complex workflow to digitization, fine-tuned to the specifics of the mathematical community. Procedures were engineered with respect to maximal quality of results, scalability, maximal automation, efficiency and effectiveness of processing of large volumes of text and graphics.*

*The optimization of optical character recognition. We have verified and implemented the OCR technology based on an automated several phase character recognition. We have evaluated the technology to reach less than 1% character error rate, counting even errors in character font type and size.*

A new framework for retro-born-digital documents. We have designed and implemented a workflow for processing journal issues from the period, where (semi)final data are available in electronic form. Procedures for conversions of (meta)data needed for a digital library were developed and as a testbed data for **Archivum Mathematicum** from period 1992–2007 were prepared.

The foundation for born-digital document processing. We have designed and implemented workflow for processing born-digital mathematical journal issues in such a way that all metadata needed for a digital library are secured and exported during preparation of printed issue simultaneously. The workflow is applied by a production team of **Archivum Mathematicum** published by Masaryk University and respects all today's demands of search engine optimization.

## Contributions (part two, cont.)

Contributions to the *design of digital library of mathematical papers.*

*A design of procedures realized digital mathematics library*

The *solution of automated classification of mathematical documents.*

*Machine learning approach to the classification of mathematical papers according to widely accepted Mathematical Subject Classification.*

# 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)**

Technology of **competing patterns** development: methods of stratification, bootstrapping and multi-level generation shown on numerous segmentation problems, results significantly outperformed previous ones (e.g. used in everyday  $\TeX$  installation), lots of trees saved :-).

**Hyphenation** methods for several languages are in every day use, method applied with success to the problem of Thai segmentation.

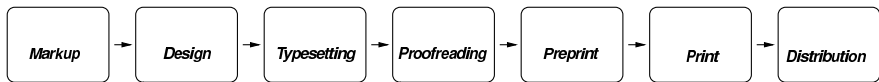
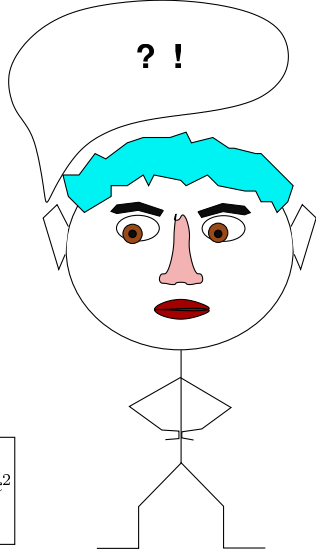
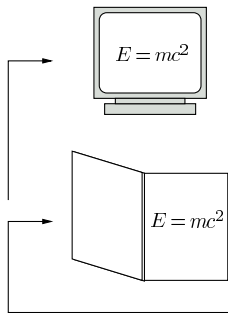
**Single-source publishing** shown very effective to deliver documents for different output devices and needs.

## Summary and Conclusion (cont.)

Methodology for **digitization** of 40.000+ pages of Ottův slovník naučný and 200.000+ pages of mathematical content in DML-CZ: <http://dml.cz/> and <http://project.dml.cz/>.

Machine learning methods for **automated classification** and **similarity** of mathematical papers.

Collection of 14 papers (out of 70+ coauthored with David Antoš, Mirek Bartošek, Han The Thanh, Jan Holeček, Martin Lhoták, Zuzana Nevěřilová, Jiří Rákosník, Radim Řehůřek, Michal Růžička, Martin Šárfy, Jiří Zlatuška). 50+ citations.





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/LNAI 5144, Springer, pp. 543–557.



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*. DML 2008, pp. 17–27, Birmingham, UK.



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.*



David Antoř, Mirek Bartořek, Han The Thanh, Jan Holeček, Martin Lhoták,  
Zuzana Nevěřilová, Jiří Rákosník, Radim Řehůřek, Michal Růžička, Martin Šarfy,  
Jiří Zlatuška