An Automated Processing of Journal Articles for a Digital Library[1]

Petr Sojka, Michal Růžicka et al.

DML-CZ
Faculty of Informatics, Masaryk University, Brno

June 11th, 2008

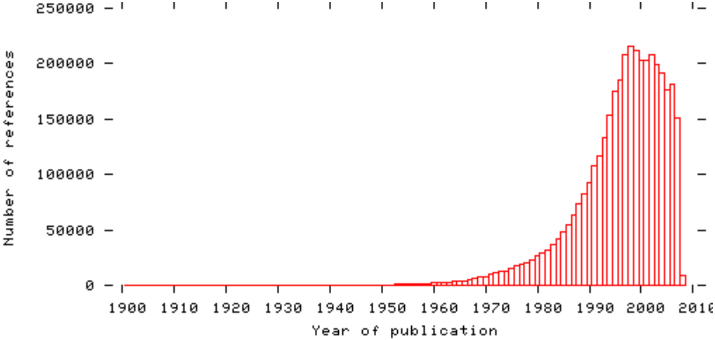
[1] Supported by the Academy of Sciences of Czech Republic grant #1ET200190513

Digital Libraries and Electronic Journal Subscriptions

- Trends to put everything to Digital Libraries (Europeana, Springer Link, JStor) and search (Zbl, miniDML, EuDML or Google Scholar).
- Digital journal editions needed for electronic journal subscriptions anyway, authors prefer publishing in electronic journals as it increases impact (citations).
- Journal production have to be adopted to produce content optimized for the new media and possibilities, crosslinking, etc.
- May cause problem for small publishers (costs), the need for automated solutions.
- Success in biomedical domain (PubMed Central).

The need for automated and unified journal production

- 60,000 new reviews, 80,000 new papers in MR **per year** and the number increases (but 600,000+ in PubMed Central).
- Number of citations 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*

- problems with reviewing

Existing approaches

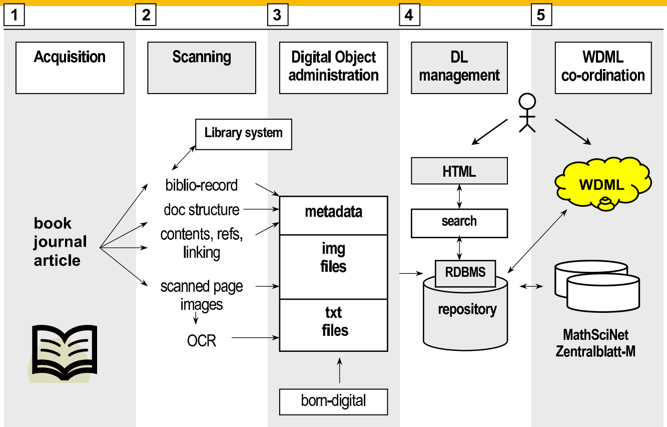
- Big publishers' journal workflow usually employ the use of SGML/XML, going LaTeX⟶XML ⟶ LaTeX⟶ DVI ⟶ PDF route.
- CEDRAM project (cedram.org) at MathDoc Grenoble offers journal production support for French publishers of mathematics.
- We decided is DML-CZ to follow this model, to save costs.
- Pilot project of **Archivum Mathematicum** journal published in Brno by Masaryk University since 1965.

Journal Publishing Periods

There are three main periods of time that must be addressed within a digital library project.

1. A retro-digitization period – The documents are available only in paper format and must be digitized for the needs of the digital library.
2. A retro-born-digital period – The documents are already born-digital but they have been made without awareness of the digital library. Therefore the format of the documents is often not suitable for the needs of the digital library.
3. A born-digital period – The documents are born-digital and they are made in such a way as to meet the needs of both the publisher and the digital library.

DML-CZ workflow steps



1 Acquisition

book
journal
article

2 Scanning

Library system

biblio-record

doc structure

contents, refs, linking

scanned page images

OCR

3 Digital Object administration

metadata

img files

txt files

born-digital

4 DL management

HTML

search

RDBMS

repository

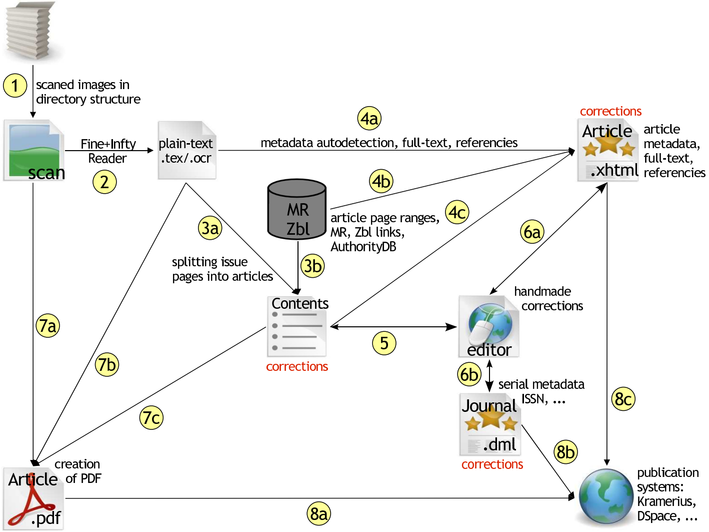
5 WDML co-ordination

WDML

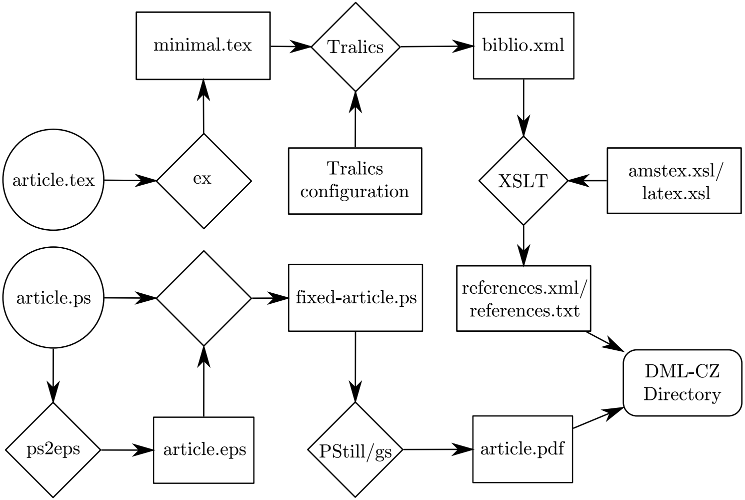
MathSciNet
Zentralblatt-M

physical document	digitization	digital document	digital library	global DL
- doc selection - IPR clearing - scan preparation	- workflow - scanning - img enhancement	- quality assessm. - conversions - struct. content	- DL system - access - archiving	- standards - interoperability - co-operation

Top-level DML-CZ workflow overview (simplified)



Schema of retro-born-digital period workflow



Archivum Mathematicum, 1992–2007

1. Attempt 1: recompilation: *even incomplete sources*
2. Attempt 2: partial recompilation: references only
3. AMSTeX 50%, LaTeX 50%
4. use of Tralics to convert AMSTeX or LaTeX to XML (references.xml).

References of Archivum Mathematicum, 1992–2007, TeX

```
\documentclass{archivum}
\begin{document}
  \Refs
  \ref\key1\by Gancarzewicz, J., Michor P. W.
      \paper Natural...
  \endref
  \ref\key2\by Zajtz, A.\paper On the order of natural...
  \endref
  ...
  \endRefs
\end{document}
```

References of Arch. Math., 1992–2007, Tralics defs

```
...
\gdef\SETENDELEM{\gdef\ENDELEM{\end{xmlelement}}}
\gdef\DELENDELEM{\gdef\ENDELEM{\gdef\ENDELEM{%
  \end{xmlelement}}}\gdef\POTENTIALENDELEMENT{\end{xmlelement}}}
\gdef\POTENTIALENDELEMENT{}
\SETENDELEM
def\Refs{\begin{xmlelement}{Refs}}
\def\ref{\begin{xmlelement}{ref}}
\def\key{\begin{xmlelement}{key}}
\def\by{\ENDELEM\begin{xmlelement}{by}}
\def\paper{\ENDELEM\begin{xmlelement}{paper}}
...
\def\endref{\POTENTIALENDELEMENT%
  \gdef\POTENTIALENDELEMENT{}\ENDELEM\ENDELEM}
\def\endRefs{\ENDELEM}
...
```

References of Archivum Mathematicum, 1992–2007, XML

```
<?xml version="1.0" encoding="UTF-8"?>
<references>
  <reference id="1">
    <prefix>[1]</prefix>
    <title>Natural...</title>
    <authors>Gancarzewicz, J., Michor P. W.</authors>
    ...
  </reference>
  <reference id="2">
    <prefix>[2]</prefix>
    <title>On the order of natural...</title>
    <authors>Zajtz, A.</authors>
    ...
  </reference>
```

Conversions, bitmap enhancements

① Only PostScripts with 300 DPI available

② dvips output changes in time (font identification)

③ attempt 1: *exchange by FixFont program*

④ attempt 2: *fonts exchange during PostScript conversion to PDF*

⑤ PStill – *exchange bitmap fonts by outline ones during distilling*

⑥ FontRep Adobe Acrobat plugin

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.

⑤ CEDRAM *cedram.org* project.

⑥ *economy of scale/ unification of workflow.*

Born-digital phase: pilot project of Archivum Math.

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,...

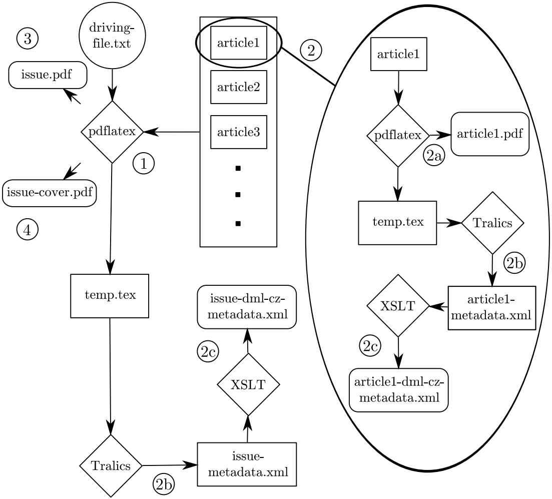
⑤ use of configurable Tralics converter to XML

⑥ high automation by program make

⑦ automated import to DML-CZ

⑧ first issue already available

Schema of born-digital period workflow



Top level TeX source of the Archivum Mathematicum born-digital (2008⟶)

```
\documentclass[AM,english,RedoBibTeX,Volume,Couverture,XML]
% volume number, issue number, month, year
\IssueInfo{44}{1}{}{2008}
\SetFirstPage{1}
\begin{document}
\makefront
\articles
  \includearticle{article1}
  \includearticle{article2}
  \includearticle{article3}
  ...
\makeback
\end{document}
```

Born-digital journal processing issues

1. using `\write18` call from within LaTeX
2. Jabref (Java), GUI application for references, conversion into `references.xml`
3. final generation under Linux with program `make`.
4. `make` goals for printed (mirrored) version, EMIS web page, for Zbl review forms, for export into DML-CZ

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.

② No single OCR system with acceptable results: high error rate, working only for specific purposes (plain English text), direct use was not possible.

③ Fine Reader by ABBYY gave good results for (even multilingual) text, and allows for typeface learning.

④ InftyReader by **`www.inftyproject.org`** the only available solution for structural math recognition.

⑤ No out-of-the-shelf solution.

Our OCR Solution

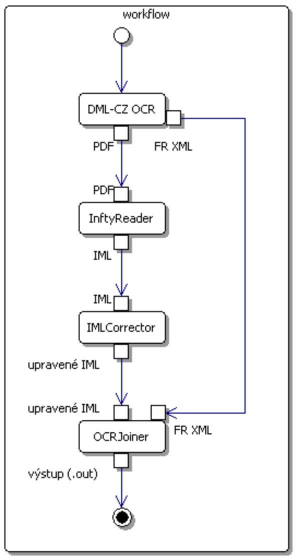
① combining both, using FineReader and InftyReader in a pipe to let every system to do what it is good for, then 'vote'

② top-level (Java) program to **automate** the process **and fix** some indeficiencies

③ instant setup unusable: **fine-tuning** and **gradually enhancing** the OCR procedure and program parameters so that OCR results would be acceptable for DML-CZ purposes

④ trying to improve the results further by close cooperation with the team of prof. Suzuki (Infty Project leader, Kyushu University, Japan, 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

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.

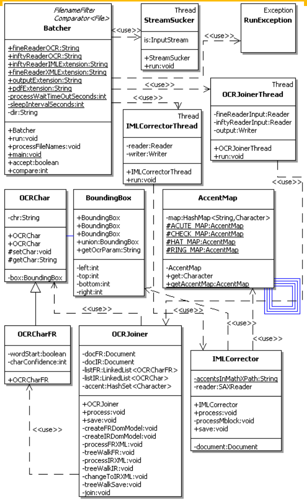
OCR XML Postprocessing

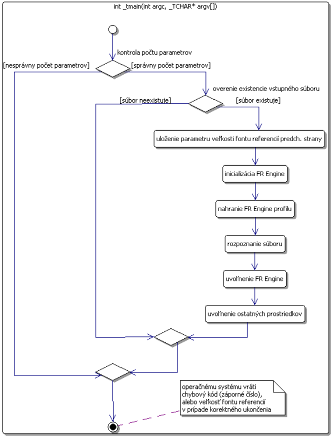
```
<mblock>
...
<munit entity="1" ocrparam="685,1746,704,1758,0">
check
<mlink type="under">
<munit ocrparam="684,1761,707,1794,0">s</munit>
</mlink>
</munit>
...
<mblock>
```

is transformed to

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

DML-CZ OCR Workflow Implementation Gory Details




int _tmain(int argc, _TCHAR* argv[])
kontrola počtu parametrov
[nesprávny počet parametrov]
[správny počet parametrov]
overenie existencie vstupného súboru
[súbor neexistuje]
[súbor existuje]
uloženie parametru veľkosti fontu referencií predch. strany
inicializácia FR Engine
nahranie FR Engine profilu
rozpoznanie súboru
uvoľnenie FR Engine
uvoľnenie ostatných prostriedkov
operačnému systému vráti chybový kód (záporné číslo), alebo veľkosť fontu referencií v prípade korektného ukončenia

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).
- ☞ 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://dml.cz/ and http://project.dml.cz/.

Archivum Mathematicum ready with data from both retro-born and born-digital period ready.

DML 2008 workshop invitation

Towards Digital Mathematics Library: DML 2008 workshop:
http://www.google.com/search?q=DML+2008

Submissions at
http://www.easychair.org/conferences/?conf=dml2008 by the end of May!

- 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.