Specifics of Open Access Publishing and Retrodigitization in Mathematics: An Experience from DML-CZ and EuDML Projects

Petr Sojka
<sojka@fi.muni.cz> (Faculty of Informatics, Masaryk University, Brno)

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.
Specifics of mathematics: Mathematics is the Queen of Science and Arithmetics is the Queen of Mathematics (C.F. Gauss)
Digital processing is full of challenges
Amount of math ever published is about 100,000,000 pages.
Specific publishing workflow and document engineering (tough reviewing, formulae handling)
Specific problems of digitization (OCR, digital library development)
Challenges of math formulae search and indexing
Challenges tackled during DML-CZ in Brno, CZ (pdf\TeX’s born town)
DML-CZ document engineering—best practices and tools
Towards European Virtual Library of Mathematics: bottom up DML processing towards EU (or worldwide) scale
European Digital Mathematics Library
EuDML—from local data collections to the virtual digital library

Specifics of OA Publishing and Retrodigitization in Math: DML-CZ and EuDML

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.
EuDML document engineering—scalable tools development
Yes, you can!
End of talk overview, quiz!
Open-ended story of a Vision

At the beginning there was a vision of all mathematical knowledge, *peer reviewed and verified* (100,000,000 pages) in ‘one stop shop’ (in the digital form), articulated at the beginning of millenium: reference mathematics library, grounding truths for not only mathematicians, but also other sciences.

Progress of IT, cheap disk space, new information retrieval technologies (Google).

AMS supported NSF preparation grant (in 2003) for WDML—Worldwide digital mathematics library, planned to be funded by de Moore foundation (some nine digit sum in $’s requested). Application was not successful—top down approach failed.
Vision of (European) Digital Mathematics Library

Several attempts to fund development of DML on European level (FP5, FP6) also was not successful.

Now, it starts to be realized: three year EU project EuDML (programme EU CIP-ICT-PSP, type Pilot B, EU contribution 1.6 MEur) from February 2010

The strategy:

• to master the technology, develop tools and offer them;

• concept of *moving wall* to motivate and engage commercial publishers.

• to collect (bottom up) [virtual] *digital library*, ‘one-stop shop’ and achieve critical mass in the domain → ‘me too’ effect then.
From paper to digital processing, from local to the whole
Bottom up—from building bricks of regional repositories

As building bricks current (regional or publisher’s) Digital Mathematics Library (DML) repositories as DML-CZ or NUMDAM, DML-PL, DML-PT, RusDML,…(from local repositories bottom-up to build the final thing—to realize the vision).

Example of DML-CZ: up and running digital mathematic library with nearly 30,000 papers. For more, see (who, what, browse, browse similar, how to search).

Live project—all comments to DML-CZ welcome!
DML-CZ: main facts

- Czech Academy of Sciences grant (program Information Society) 2005–2009, *full* (retro)digitization of 50,000 pages of mathematical literature per year, 8M CZK in total.

- 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. 3) The design of the work-flow aiming at mathematical knowledge stored in digital library.

- IPR part: sharing/delivery (economic models for knowledge sharing due to interests of content owners/publishers).

- Production part: dig. center Jenštejn, overestimated costs.
DML-CZ data: what is there

• 12 mathematics journals, 7 OA, 5 with moving wall.
• 8 publishers of math in CZ and SK, close cooperation.
• 27,000+ articles with rich metadata (full bibliographic record, similarities, MSC classifications, translated titles).
• heavily multilingual (7+ languages).

Links on who, what, browse, browse similar, how to search).
DML-CZ – data: scientific math published in Czech and Slovak

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

$$P(K, v) = \int_{\hat{K}} f(x) \, dx.$$  \hfill (89)  

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

$$\int_{\hat{K}} f(x) \, dx = \int_{\hat{O}} f(\varphi(y)) \cdot |T(y)| \, dy = \int_{\hat{O}} \tilde{f}(y) \, dy.$$  \hfill (90)  

From theorem 50 (and relation (86)) we see that $P(K, v) = P(\hat{K}, \hat{O})$; relations (89), (90) show therefore that $P(\hat{K}, \hat{O}) = \int_{\hat{O}} \tilde{f}(y) \, dy$, which completes the proof.

Remark. The reader may compare this paper with [6].

REFERENCES


Резюме

ПОВЕРХНОСТНЫЙ ИНТЕГРАЛ

ЯН МАРЯЦКИ (Jan Mařík), Прага.
(Поступило в редакцию 10/4 1955 г.)

Пусть $m$ — натуральное число; пусть $E_m$ — $m$-мерное евклидово пространство. Для всякого ограниченного измеримого множества $A \subset E_m$ положим $|A| = \sup \sum \frac{d\varphi_i(x)}{dx_i}$, где $v_1, \ldots, v_m$ — многоугольники такие, что $\sum v_i(x) \leq 1$ для всех $x \in A$. Пусть $\mathfrak{M}$ — система всех ограниченных измеримых множеств $A$, для которых $|A| < \infty$. Теорема 18 тогда утверждает: 

Пусть $A \in \mathfrak{M}$; пусть $D$ — граница множества $A$. Тогда на системе $\mathfrak{M}$ всех борелевских подмножеств множества $D$ существует мера $\mu$ и на
Bottom up processing—local (Brno, CZ) document engineering
The approach used in DML-CZ

A successfully built repository (e.g. set of workflows) needs a coordinated effort of librarians, IT specialists and representatives of users—content specialists: (D+M+L)=success ‘equation’.

*Design, technical* and *political decisions* behind building the *Czech Digital Mathematics Library DML-CZ* (<http://dml.cz>) in the context of other thematical community projects (PubMed Central, ADS, INSPIRE, SCOAP3 and EuDML) have been solved. *No wheel reinvention.*

Our framework integrates workflow for the articles scanned from a paper (*math OCR*), for documents from retro-born digital period (data available in some type of electronic form) and for born-digital ones. As much automated (using robots :-) as possible.
Math handling poses challenges—math OCR, math indexing,...
Document engineering—from paper to digital *workflow*
Specifics of OA Publishing and Retrodigitization in Math: DML-CZ and EuDML

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.
DML-CZ document engineering—data processing
DML-CZ now serves about 275,000 pages of math papers.

Problems of migration of existing workflows (born-digital, retro-digital) into the repository. Negotiations with Google Scholar towards better visibility, indexing and search, and problems of copyright and sustainability issues, visualization, space and processing demands,....
Document engineering—digitization, digital library development
New tools and best practices for [meta]data processing

Data heterogenity, plethora of formats, validation and conversions:

**retro-digital period:** scanning, geometrical transformations (BookRestorer), OCR (FineReader, InftyReader), two-layer PDF

**retro-born-digital period:** not complete .tex or .dvi data, bad formats, bitmap fonts of low resolution

**born-digital period:** typesetting by \( \TeX \) with export of [meta]data into digital library

**world of authors:** \( \LaTeX \), \( \TeX \) notation of mathematics

**world of applications/data exchange:** XML, MathML
Verified and proven technologies (in DML-CZ)

- scanned image processing and transformations, mathematical optical character recognition: OCR
- digital signature of PDF: pdfsign
- web-based long distance metadata editing: web application metadata editor
- optimization and recompression of PDF: downsizing of more than 50% without quality loss
- article similarity computations, demo.
- retroborn paper automated classification by MSC.
Verified and proven technologies (cont.)

- born-digital publishing system [for Archivum Mathematicum and other 4 journals] and conversions.
- retro-born-digital paper conversions and enhancements.
- data visualization, browsing: adaptation of Visual Browser
- math retrieval: math formula indexing and search
- citation linking: CiteCrawl

many open areas/challenges: multilingual retrieval?. MathML indexing,…
Bottom up processing towards EU or worldwide scale
EuDML as a virtual library portal

EuDML will be a virtual library based on data from smaller data providers, DLs and publishers:

![Diagram of EuDML's architecture]

- Web User Interface
- Accessibility Component
- Annotation Component
- Search Engine
- External Services Interface
- Metadata Repository
- Association Analyser
- Metadata Enhancer
- Local Repositories
European Digital Mathematics Library
EuDML – data: legacy scientific math

- By 2013, EuDML should integrate 12 repositories, have content from 200 integrated collections (journals, book series, conference proceedings,…), more than 160,000 digital items (papers, book chapters), 500,000 links between database objects.

- It should be ‘live’ DL, having more than 1,000 users contributing annotations, and more than 10,000 annotation by 2013.

- Concept of moving wall: legacy data even from commercial publishers.

But how to actually implement it?

Experience from project partners from current digital library development.
EuDML—from data collection to the virtual digital library

Specifics of OA Publishing and Retrodigitization in Math: DML-CZ and EuDML
EuDML service based architecture
Alternative GUIs—visualization research

Search

Search Results

- Došňová, Ondřej: A remark on power comparison theorem for half-linear differential equations
- Došňová, Ondřej: The multiplicity criteria for zero points of second order differential equations
- Došňová, Ondřej: Spectral properties of fourth order differential operators
- Došňová, Ondřej: On some problems in the oscillation theory of self-adjoint linear differential equations
- Došňová, Ondřej: On the existence of conjugate points for linear differential systems
- Došňová, Ondřej: The Pólya inequalities for a class of partial differential equations
- Došňová, Ondřej: On the Liouville-type transformation for differential systems
- Došňová, Ondřej: Sixty years of professor František Neuman
- Došňová, Ondřej: A remark on conjugacy of half-linear second order differential equations
- Došňová, Ondřej: Qualitative theory of half-linear second order differential equations
Visual Browser development (DML-CZ)

Motivation

Overview

DML-CZ

EuDML

Specifics of OA Publishing and Retrodigitization in Math: DML-CZ and EuDML

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.
EuDML document engineering—scalable tools development
EuDML service based architecture II

Based on existing YADDA (used in Driver, Driver II) and REPOX (used in EuropeanaLocal, Telplus) projects – both are verified and mature platforms (implemented in Java)

Math specifics needed to develop (TEX to MathML converter, math OCR, math in metadata support,…

MU offers: Metadata editor and other tools and expertise, mainly to be used in WP7 Metadata Enhancements
PDF Re-compression

New tools developed (Radim Hatlapatka) to re-compress [bitonal] PDF files:

<table>
<thead>
<tr>
<th></th>
<th>Original PDF</th>
<th>After using PDF re-compressor</th>
<th>After using pdfsizeopt.py</th>
<th>After both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of whole PDF</td>
<td>100%</td>
<td>74.61%</td>
<td>50.02%</td>
<td>40.23%</td>
</tr>
<tr>
<td>Size of image and other objects</td>
<td>69.46%</td>
<td>37.14%</td>
<td>45.14%</td>
<td>35.36%</td>
</tr>
</tbody>
</table>

May be used for any PDF 1.4 (since Acrobat 5 released in 2001) file—JBIG2 compression.
Metadata Editor http://editor.dml.cz

Web-based client-server tool, developed (MU) for [meta]data import, editing, validation and checking. For testing, try <http://editor.dml.cz:9129>, admin/admin
Metadata Editor localization (open source development)
Yes, you can!
Publishing and retrodigitization of mathematics poses many challenges, some yet to solve.

DML-CZ: up and running, most open access.

EuDML: work in progress, some DML-CZ experience and tools developed to be (re)used.

Invitation: EuDML meeting with content providers in Prague, October 15th: math publishers wanted to cooperate.

Further reading: DML workshop series proceedings (Openly accessed in DML-CZ).

Comments, cooperation offers welcome! We are Open and AND-minded (no black&white thinking :-).
End of the talk

Questions?

Specifics of OA Publishing and Retrodigitization in Math: DML-CZ and EuDML

COASP 2010, Prague, CZ, August 24th, 2010, 9:40 a.m.
DML-CZ team.  
*Materials about DML-CZ, project publications* [online, cit. 2010-08-24].

EuDML team.  
*EuDML project info* [online, cit. 2010-08-24].

EuDML team.  
*EuDML webpage* [online, cit. 2010-08-24].

EuDML at MU team.  
*EuDML at MU project info* [online, cit. 2010-08-24].