

Petr Sojka

Towards a Digital Mathematical Library

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Paris, France, July 7-8th, 2010. Masaryk University Press, Brno, Czech Republic, 2010. pp. 3--6.

Persistent URL: <http://dml.cz/dmlcz/702566>

## Terms of use:

© Masaryk University, 2010

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://project.dml.cz>

# Towards a Digital Mathematical Library

## On the Road

Petr Sojka

Masaryk University, Faculty of Informatics  
Botanická 68a, 60200 Brno, Czech Republic  
Sojka@fi.muni.cz

**Abstract.** The workshop's objectives were to formulate the strategy and goals of a global mathematical digital library and to summarize the current successes and failures of ongoing technologies and related projects.

There is already some experience with building smaller DMLs and/or building big thematical scientific digital libraries. Why there are already big fulltext digital library in some domains like PubMed in biomedical one, but none in others? We try to pose such and other questions, and try to find some answers in papers of this proceedings.

*There is no royal road to mathematics.* (Menaechmus, 380–320 BC)

## 1 The Dream

Mathematicians dream of a digital archive containing all peer-reviewed mathematical literature ever published, properly linked and validated/verified. It is estimated that the entire corpus of mathematical knowledge published over the centuries does not exceed 100,000,000 pages, an amount easily manageable by current information technologies.

“There is no royal road to mathematics”, heard Alexander the Great thousands years ago. It seems that there is no royal road to DML, either. To fulfill the dream, concerted action of Digital specialists, computer scientists, Mathematicians, topical experts, and Librarians, curators, information specialists is needed.

Our workshop's objective is to pave the road for DML, built bottom-up from smaller repositories. It is to formulate the strategy and goals of a global mathematical digital library and to summarize the current successes and failures of ongoing technologies and related projects, asking such questions as:

- \* What technologies, standards, algorithms and formats should be used and what metadata should be shared?
- \* What business models are suitable for publishers of mathematical literature, authors and funders of their projects and institutions?
- \* Is there a model of sustainable, interoperable, and extensible mathematical library that mathematicians can use in their everyday work?

- \* What is the best practice for
  - retrodigitized mathematics (from images via OCR to MathML and/or  $\text{\TeX}$ );
  - retro-born-digital mathematics (from existing electronic copy in DVI, PS or PDF to MathML and/or  $\text{\TeX}$ );
  - born-digital mathematics (how to make needed metadata and file formats available as a side effect of publishing workflow [CEDRAM model, Euclid])?

The intention was to have the workshop as a forum for presentation and discussion of the latest developments in the the field of digitization of mathematics, based on the previous bilateral discussions and successful workshops DML 2008 held as part of CICM multiconference in Birmingham, UK, and DML 2009 held in Grand Bend, Ontario, Canada last year.

*Topics of the Workshop included*

- \* search, indexing and retrieval of mathematical documents;
- \* ranking of mathematical papers, similarity of mathematical documents;
- \* math OCR with MathML/ $\text{\TeX}$  output;
- \* document conversions from/to MathML, OpenMath,  $\text{\LaTeX}$ , PostScript and [tagged] PDF;
- \* mathematical document compression;
- \* processing of scanned images;
- \* algorithms for crosslinking of bibliographical items, intext citations search;
- \* mathematical document classification, MSC 2010;
- \* mathematical text mining;
- \* mathematical documents metadata exchange via OAI-PMH and/or OAI-ORE;
- \* long term archiving, data migration:
- \* reports and experience from math digitization projects;
- \* math publishing with long term archival goal;
- \* software engineering aspects of creating, handling MathML, OMDoc, OpenMath documents, and displaying them in web browsers.

*The four branches of arithmetic – ambition, distraction, uglification and derision.*  
(Lewis Carroll: Alice in Wonderland)

## 2 On the Road

This year we concentrate on core technologies for building DML, and it is the first year of realizing EuDML, The European Digital Mathematics Library, designed as virtual library over existing smaller repositories. EuDML is presented on page 11 with a paper referring about history of European activities towards DML, about project plans and steps to reach its goals.

Invited talk by Masakazu Suzuki speaking about Infty and his tools to automate math OCR brings Japanese know-how of digitizing big volumes of mathematical papers including structural information and mathematical

formulae. This is the only verified way of getting texts with mathematics for fulltext search, math retrieval and trustworthy paper similarity computations, on larger scale today. Handling fulltexts with math together with semantic handling of math is the hard way towards semantic-aware math retrieval — for more about semantic ground truth see pages 37–42.

David Ruddy's paper on page 27 gives timely information on possible solution of mathematical metadata exchange format based on the experience with running Euclid DML.

There is a *DML Building Experience* block of papers sharing experience gained during building DML-CZ repository. pdfjblm and pdfsign tools (pages 45–55) are general-purpose tools with impressive results. Other papers refer about Metadata editor and validation tools for shaping metadata into usable shared database, and Visual Browser offers alternative graphical interface for DML (meta)data.

*Digitization Reports* block starting on page 79 brings not only information about new Italian baby in the family of DML repositories, but also gives insight about how the dream of high-energy physicists about their digital library is becoming a reality.

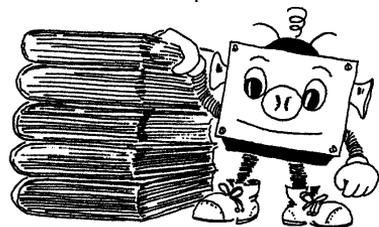
Finally, respected reader can read a case study about mathematical communication and representation in virtual learning environments on page 83. Tralics as a tool for bridging authors' world of  $\text{\TeX}$  and applications' world of XML/MathML is presented on page 105. Tough arena of mathematical formulae semantics might be entered on the page 119 with a paper about symbol declaration in mathematical writings.

*Ring the bells that still can ring  
Forget your perfect offering  
There is a crack in everything  
That's how the light gets in.*  
(Leonard Cohen: Anthem)

### 3 Summary

This volume contains the Proceedings of the Workshop *Towards a Digital Mathematics Library (DML 2010)*, organized by the Faculty of Informatics, Masaryk University with the help of CNAM, Paris and held on July 7–8<sup>th</sup>, 2010 in Paris, France, as a satellite event of CICM 2010 (Conference on Intelligent Computer Mathematics). The Proceedings is divided into five parts:

1. Towards a Digital Mathematics Library,
2. Digitization Technologies and Platform,
3. DML Building Experience,
4. Digitization Reports, and
5. Tools and Techniques.



My very special thanks go to the Program Committee members for their hard work during review periods. Most of the submitted papers were reviewed

by three members of the Program Committee, some even by four. We employed rebuttal phase, where authors were given the possibility to comment on the preliminary review reports and to answer anonymous reviewer's questions. It helped to increase the quality of final paper versions considerably.

I would also like to express my appreciation to the members of the Organizing Committee for their efforts in organizing the Workshop and ensuring its smooth running, and to CICM general chairs Renaud Rioboo and Laurence Rideau.

Last but not least, the cooperation of Masaryk University as a publisher of these Proceedings is gratefully acknowledged.

DML 2010 offered a rich program of presentations, short talks/posters, technical papers and [panel] discussions. I hope that another step on the road towards fulfilling the dream of the world-wide Digital Mathematics Library has been successfully completed.