Petr Sojka
Towards a Digital Mathematics Library. On the Crossroad

In: Petr Sojka and Thierry Bouche (eds.): Towards a Digital Mathematics Library. Bertinoro, Italy, July 20-21st, 2011. Masaryk University Press, Brno, Czech Republic, 2011. pp. 1--5.

Persistent URL: http://dml.cz/dmlcz/702596

Towards a Digital Mathematics Library
On the Crossroad

Petr Sojka

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

Abstract. The DML 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 experience with building regional DMLs or building big thematic scientific digital libraries. EuDML project reached it half-life period. While there are already big fulltext digital libraries in some domains like PubMed Central in the biomedical domain, Inspire in high-energy physics, why did not these emerge in other scientific areas? Will EuDML project in mathematics follow their success? Which crucial decisions has to be taken so that the heritage of mathematics would be sustainably at fingertips of scientists? We pose such and other questions, and try to find some answers in papers of this proceedings.

You are now at a crossroads. This is your opportunity to make the most important decision you will ever make. Forget your past. Who are you now? Who have you decided you really are now? Don't think about who you have been. Who are you now? Who have you decided to become? Make this decision consciously. Make it carefully. Make it powerfully. (Anthony Robbins)

1 The Dream

Mathematicians dream of a digital archive containing all peer-reviewed mathematical literature ever published, properly linked, with validated and verified content and form. 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" was reportedly said to Alexander the Great by fellow mathematicians as an answer when Alexander asked for a shortcut to understanding mathematics. There is no royal road to a general Digital Mathematics Library either. To make the dream a reality, concerted action of **D**igital specialists (computer scientists), **M**athematicians (topical experts), and **L**ibrarians (curators, information specialists) is needed. Mathematicians should have their say on what should constitute a DML of their choice, librarians of digital age should tell whether digital realization of the classical library metaphor of collecting, cataloguing and providing classified documents is enough for library to sustain today, and computer scientists should say what is possible technically, technologically and how. The

Petr Sojka, Thierry Bouche (editors): DML 2011, Towards a Digital Mathematics Library, pp. 1–5.
 ISBN 978-80-210-5542-1

shape of the target dream the three groups might end up with may be quite different. It may be 'just' virtual digital library mimicking the 'good old ages' of working with printed catalogue cards or with searching in basic metadata fields mathematicians are used to work with in referative mathematical databases. However, the technological progress speeds up so quickly that today there exist production systems starting to cope with semantically disambiguated texts in different languages on the fulltext level. To realize several different types of DML visions and wait for users' decision is simply beyond the socioeconomic possibilities of heterogenious mathematical community.

There are scientific domains, where the concerted action of DML preparation already happened — e.g. high-energy physics or [bio]medical scientists now work with DLs as Inspire or PubMed Central in ways that were never possible before. DLs allow for different search and linking strategies, providing new level of exploatation of scientific heritage. These communities have managed to agree on what is the common dream and how to realize it. Mathematics community is still on its road to reach the consensus and realize their DML. To help to pave the road for future DL in the domain of mathematics was the objective for setting up the DML workshop series.

2 DML Workshop Series

DML workshop series objective 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 or TEX);
  - retro-born-digital mathematics (from existing electronic copy in DVI, PS or PDF to MathML or 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 workshops have been held as satellite event of CICM multi-conferences in previous years: DML 2008 in Birmingham, UK, DML 2009 in Grand Bend, Ontario, Canada, and DML 2010 in Paris, France.

Topics of the DML workshops included

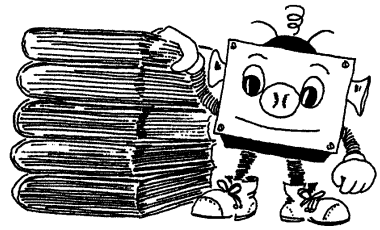
* search, indexing and retrieval of mathematical documents;

* ranking of mathematical papers, similarity of mathematical documents;
* math OCR with MathML and TEX output;
* document conversions from and to MathML, OpenMath, 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 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.

3 On the Crossroad

There are already ongoing projects as the European Digital Mathematics Library (EuDML) project, where DML is aimed to be built in a bottom-up way from smaller repositories of contributing partners. This year, six out of nine Proceedings contributions do acknowledge EuDML. The Proceedings volume is divided into three parts:

1. Digital Mathematics Library Reports,
2. Digitization Workflows and Standards, and
3. DML Building Technologies.



EuDML is designed as a virtual library over existing regional repositories and publisher archives. One of participating EuDML data providers is the project DML-CZ. On page 9, Jiří Rákosník overviews current status quo of DML-CZ content and technological achievements and experience gained in the last two years.

EuDML invites joining any content providers that curate mathematical scientific content and bdim repository is an example of a smaller DML designed for smooth joining the club. Current state of this repository development is presented in a short paper by Vittorio Coti Zelati on page 19.

Talk by Takao Namiki speaking about tools to time stamp preprints brings Japanese know-how of running electronic journal server environment — for more see pages 19–23.

Digitization Workflows and Standards are covered in the second block of papers, all motivated and prepared as parts of the EuDML project. Researchers from Warsaw University's Interdisciplinary Centre for Mathematical and Computational Modelling are working on important problems of author name

disambiguation and metadata extraction from retro-born digital documents. They present their solutions in two papers on pages 27–37 and pages 39–44, respectively.

Standard EuDML metadata schema aimed to be used as rich data container for data exchange between EuDML and its partners is described in a paper by Thierry Bouche, Claude Goutorbe, Jean-Paul Jorda and Michael Jost. It is the important standard to follow for all publishers and other entities offering data for EuDML. Only slight modifications done to the widely used NLM Journal Archiving and Interchange Tag Suite warrants wide acceptance among publishers, as most of them already archive their holdings with Portico in this format.

There is a third part of proceedings named *DML Building Technologies*. As for most DML papers only metadata and PDF will be available, for tools working with full texts and math one needs a technology to get the texts and formulae out of the PDF. Progress report by Josef Baker et al. about the development of a tool to reverse engineering the PDF documents is presented on pages 65–75. At the end, this tool should possibly allow mathematics retrieval from PDF-only sources.

Tough arena of mathematical formulae indexing and search might be entered on the page 77 with a paper by Masaryk University group about WebMIaS system that allows math-aware structure respecting scalable retrieval of mathematical documents.

Finally, the respected reader can enjoy a case study about using discourse context to interpret object-denoting mathematical expression. Although the goal of semantic disambiguation with respect to math is far on the horizon of current natural language processing technologies, Magdalena Wolska et al. took the courage to tackle the problem of interpretation of mathematical expressions given the context.

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)

4 Summary

This volume contains the Proceedings of the Workshop *Towards a Digital Mathematics Library (DML 2011)*, organized by the Faculty of Informatics, Masaryk University and held on July 20–21, 2011 in Bertinoro, Italy, as a satellite event of CICM 2011 (Conference on Intelligent Computer Mathematics). DML 2011 offered nine presentations, EuDML session and [panel] discussion. We hope that it has helped to choose the right direction on the crossroad towards fulfilling the common dream of the Digital Mathematics Library.

Acknowledgements. My very special thanks go to the workshop PC co-chair Thierry Bouche, to Programme Committee members and additional referees for their hard work during review period. Most of the submitted papers were reviewed by three members of the Programme Committee, some even by four.

I would also like to express my appreciation to the CICM local chair Andrea Asperti for assuring conference facilities support on the workshop site. Last but not least, the cooperation of Masaryk University as a publisher of these Proceedings is gratefully acknowledged.