Recent Developments in Digital Mathematics Libraries

Jiří Rákosník[1], Peter Stanchev[2,3], Radoslav Pavlov[2]

[1] Institute of Mathematics AS CR, Czech Republic
[2] Institute of Mathematics and Informatics, BAS, Bulgaria
[3] Kettering University, Flint, USA
rakosnik@math.cas.cz, pstanche@kettering.edu, radko@cc.bas.bg

Abstract. The paper presents recent developments in the domain of digital mathematics libraries towards the envisioned 21st Century Global Library for Mathematics. The Bulgarian Digital Mathematical Library BulDML and the Czech Digital Mathematical Library DML-CZ are founding partners of the EuDML Initiative and through it contribute to the sustainable development of the European Digital Mathematics Library EuDML and to the global advancements in this area.

Keywords: Digital Mathematical Library, EuDML, BulDML, DML/CZ, Open Access, Archiving, Institutional Repositories, DSpace, Electronic publishing.

1 Introduction

For the mathematicians nowadays the web-based access to digital resources is absolutely essential. Modern publishing and disseminating methods increased the speed of preparation and circulation of documents and contributed to the ever growing volume of produced scholarly literature in mathematics. This in turn requires new efficient tools for archiving, providing access to and searching through data included in abundance of various repositories and sophisticated retrieval of the included highly structured information.

The quickly developing environment of electronic publishing offers numerous opportunities. The common tools allow users to search through the digital content using attributes such as subjects, titles, authors, dates, and keywords and chains of citations. Interlinking of documents, databases and other information resources is also becoming standard. A new and fast developing topic is the search by mathematical formulas. There have been several attempts to address these issues in the past which led to the design and implementation of a solution for mathematics retrieval [15]: MathDex, Symbolab, LeActiveMath, MathWebSearch, LaTeXSearch, EgoMath2, MIaS. Each of them employs different approaches. The most distinct of them is MathWebSearch which is not based on full-text indexing. Further features which may soon become popular include social networking from the most common Facebook and Twitter to the specialised ones like Mendeley, CiteULike, BibSonomy, and annotations systems for blogging, providing comments to the displayed texts, managing discussion threads, enabling personalization of the tools etc. Many more possibilities of devel-

opment belonging rather to the realm of visions for farther future have been discussed in the recent report initiated by the International Mathematical Union and the US National Research Committee and supported by the Alfred P. Sloan Foundation [9].

As soon as various projects aiming at digitization of mathematical literature had been emerging at the turn of millennium, the US National Science Foundation launched a project coordinated by the Cornell University toward establishment of the World Digital Mathematics Library which would "make the entirety of past mathematics scholarship available online, at reasonable cost, in the form of an authoritative and enduring digital collection, developed and curated by a network of institutions" [7].

At the same time several national and regional DML initiatives in Europe tried to coordinate efforts under auspices of the European Mathematical Society. Eventually, the European Digital Mathematics Library (EuDML) project, partly funded by the European Commission for three years from February 2010, created a network of 13 institutions from 8 European countries and established a single distributed library with about 233,000 unique items across 14 collections.

To ensure the continuation and further development of the EuDML services the international association without legal personality called EuDML Initiative has been recently established in 2014 by 12 partnering organizations including content providers and developers. The European Mathematical Society has a prominent role in the Initiative securing that the EuDML remains a sustainable public service to the worldwide scientific community.

Both the Bulgarian Digital Mathematics Library (BulDML) [4] and the Czech Digital Mathematics Library (DML-CZ) [8] have been included in the EuDML and their respective curating institutions, the Institute of Mathematics and Informatics BAS in Sofia and the Institute of Mathematics ASCR in Prague are founding members of the EuDML Initiative.

2 The Corpus of Mathematics Literature

The digital corpus of mathematics literature is rather extensive. A major part of the mathematics literature of the 20th century is now available digitally, partly due to retro digitization activities and partly as a result of production of born-digital documents. The most important mathematical research of the second half of the 19th century also has been digitized through the Electronic Research Archive for Mathematics project [26], the local DMLs (e.g., NUMDAM [22], DML-CZ, JSTOR [16]) as well as through independent efforts of publishers.

The database zbMATH (including its digitized predecessor Jahrbuch) [14] indexes more than 3,000,000 publications and approximately 3,000 journals and 170,000 books from 1868 year to the present. Similarly, the MathSciNet [1] database includes approximately 2.9 million publications, about 2,000 journals and 100,000 books from 1940 to the present.

The Digital Mathematics Library compiled by Ulf Rehmann [10] currently lists 4609 digitized books and 577 digitized journals and serials representing altogether

almost 5 million pages. This is essential part of the entire volume of mathematical literature which is estimated at total of 50 million pages [13].

The current annual increment in database zbMATH makes around 100,000 new publications. According to the American Mathematical Society there has been an average growth of about 40 new journal titles per year in mathematics since 1997. Most of them are, of course, born-digital.

Taking into account the fact that the mathematical literature never becomes obsolete, these figures show the necessity and benefits of a comprehensive digital mathematics library equipped with effective sophisticated services.

3 Challenges for (not only) the EuDML Initiative

3.1 Extending the Content

The current EuDML content resulting from the project should be substantially extended. One million of items would represent a save point of no return. Negotiations are held with further content providers such as Euclid, Math-Ru.Net and EMS Publishing House.

Unfortunately, negotiations with the big commercial publishers like Springer or Elsevier were so far not productive, partly because the EuDML Policy is to provide public control on the full-texts, which should become eventually freely accessible (possibly after some reasonable finite period, the so called moving wall) and once they are made open access due to this policy, they cannot revert to close access later on. Increasing the content and awareness of the EuDML would improve the negotiating position of the EuDML Initiative.

3.2 Metadata

Proper operation of a digital library depends on a well thought and sufficiently reach metadata schema. It is even more important for a distributed system like EuDML where the digital content is contributed by numerous partners using different metadata schemas. The most used Dublin Core metadata fields, i.e. the set title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, does not fit specific requirements for mathematics. Therefore EuDML created a unified metadata schema [3], [11], [25] based on the framework provided by the Journal Archiving and Interchange Tag Suite (JATS) [27]. Three levels of metadata have been identified.

Obligatory metadata is the required minimum of metadata in order to unambiguously identify and handle a relevant mathematical publication in the scope of EuDML: item type, authors, original title, bibliographic reference for this publication with enough structure to enable collection's browsing, unique identifier, URL of full text.

Fundamental metadata is what satisfies the functional requirements for browsing, searching and reference matching over the collections at item level. It enables basic digital library interaction with the EuDML corpus.

Supplemental metadata is whatever goes beyond fundamental metadata (e.g. relations to subject ontologies, authority lists, MR/ZM IDs, multilingual, multi-script, bibliographies/references, interlinking, math handling, etc.). They are relevance to the EuDML's corpus specificities and EuDML system functionalities.

The preferred way to contribute content to EuDML is to set-up an OAI-PMH server to export XML metadata structured according to the EuDML schema version 2.0, providing the obligatory elements and tagged according to the best practices specified on the EuDML project website [12]. Content providers who cannot export metadata prepared according to these recommendations can employ transformation from various OAI Dublin Core variants which have been developed in the project and which are performed on-the-fly during the ingestion time.

Having in mind the aim of extending the EuDML content, the EuDML Initiative will have to review and improve the metadata harvesting and serving technology developed during the project in order to ease contributing the digital content by further partners.

3.3 Searching Mathematical Formulae / Mathematical Aware Search

The peculiarity of scholarly mathematical texts lies in the abundance of mathematical symbols and formulae which require special search methods. An approach based on Presentation MathML using similarity of math sub formulae is suggested and verified by implementing it as a math-aware search engine based on the state-of-the-art system Apache Lucene [2]. The EuDML is using MIaS (Math Indexer and Searcher) designed in the Masaryk University in Brno [19] for the formula search. The MIaS system is being further developed and one of the first tasks of the EuDML Initiative is to implement its latest version. The math search is currently available only through full-texts. Experiments should be done with text and math searching also in metadata (titles, bibliographies, abstracts). The reliability of the outputs is, however, limited by the technical quality of digital documents. It is desirable to be able searching all XML metadata which is usually of much better quality than OCR full-text, especially in the case of scans of old prints.

3.4 Securing the Long Term Preservation

The complexity of the electronic publishing ecosystem emphasizes the question of permanence of documents and their long term preservation. One of the EuDML Policy principles states that the digital full text of each item contributed to EuDML must be archived physically at one of the EuDML member institutions. However, the safety of archives must be considerably increased by proper technical tools and arrangements, backlogging, duplications, mirror sites etc.

Besides archiving services provided by commercial organizations like Portico there is also an open-source library-led digital preservation system LOCKSS built on the

principle that “lots of copies keep stuff safe” [18]. The EuDML Initiative will investigate creating a private LOCKSS network. This might offer a prototype for further content providers, especially for smaller publishers.

3.5 Getting Users Involved

Although it seems that the nowadays popular social networking systems did not yet find a wide acceptance among mathematicians it is slowly changing, particularly among the young generation. Besides the most common Facebook and Twitter and more professionally oriented LinkedIn or ResearchGate there are specialised ones Mendeley, CiteULike, BibSonomy which besides networking offer further services like managing references and organizing and sharing bookmarks and lists of scholarly literature. The EuDML includes links to these systems.

Another not yet widely exploited feature to engage the mathematical community in enriching the library’s knowledge base is annotations. The annotation component developed in the EuDML enable users to personalize their use of the digital library, to compose personal lists of items and possibly share them with selected users, to create comments, discussion threads, tutorials, reviews and reading lists that can be attached to individual items in the collection. The user can also suggest corrections to metadata.

Various widgets for search, notes and lists provide a mechanism to easily integrate EuDML resources on non-EuDML sites.

3.6 Open Access

The EuDML strongly promotes the idea of eventual open access meaning that as much as possible of the digital mathematical corpus should become open access after some embargo period. As mentioned above, the EuDML Policy requires at the same time that once the EuDML items become open access, they cannot revert to close access later on. Responsibility for observing this sine qua non rule is assumed by the corresponding content providers who should negotiate the corresponding conditions with publishers.

The issue of Open Access is very complex and it is, in particular, closely connected with the questions of business models and, consequently, with quality control of publication. It should be emphasized in this connection that another important principle of the EuDML Policy requires that the texts included in the EuDML must have been scientifically validated and formally published.

3.7 Advanced Exploration of the Mathematical Content

The ever growing extent of scholarly mathematical literature, its complexity and specific ways of exploration and utilization puts new requirements for electronic archives, digital libraries, metadata schemes, display methods and sophisticated tools. This brings new challenges for developers as well as demands on producers of the documents: authors and publishers. The existing and the foreseen methods and algo-

rithms can assist humans in identifying mathematical concepts and objects (theorems, proofs, sequences, groups etc.) contained in the research literature; still, it will require a non-trivial amount of human work to enrich the source texts with necessary semantic information.

An extensive visionary survey of possible developments in this area has been presented in the recent report "Developing a 21st Century Global Library for Mathematics Research" for the IMU and NRC [9]. There already exist various specialized (non-exhaustive) resources of information, on mathematical objects like propositions ([17], [20]), proofs ([5], [20], [24]), sequences [23], functions [21], [28], [29] etc. The envisioned future global library of mathematical knowledge neatly connecting such resources will not only provide access to metadata and electronic texts, search for text and mathematical formulae and interlinking between documents. It would also offer an efficient navigation on a high level enabling the user to quickly find additional information about an object (such as other articles discussing the same object, a list of reference resources, citation graph, tracking article-to-article reading etc.) just by clicking on it.

Such visions may sound too futuristic but the development is going fast forward. The ongoing efforts in the national and regional DMLs, the EuDML Initiative activities and the first steps of the IMU following the report [9] as well as the scientific achievements in the MKM conferences and DML workshops organized yearly in frames of the Conferences on Intelligent Computer Mathematics [6], encouraged by the demands of the mathematical community, are paving the way to the success.

4 Conclusions

The dynamics in the domain of digital mathematics libraries has recently got new incentives. The international association EuDML Initiative has been established under auspices of the European Mathematical Society and the International Mathematical Union with the US National Research Council initiated a visionary report on the future global digital mathematics library. Many possibilities for new advancements to meet demands of the general mathematical community and to keep up with the technological development have been identified and organizational measures connected with a corresponding research have been launched.

BulDML and DML-CZ are regional digital libraries built to preserve and provide free access to national heritage in mathematics. Although developed independently, they have a lot in common with respect to structure and development of content, services to users, access policy of guaranteed eventual free access, and the DSpace software used for presentation. Both libraries fulfill metadata requirements for their content to be presented in the unique framework of the European Digital Mathematics Library of which they are partners. Both their parent institution are founding members of the EuDML Initiative through which they participate on further development of the EuDML and the overall DML environment.

Acknowledgment

This work was supported in part by the EU project "2nd Generation Open Access Infrastructure for Research in Europe" (OpenAIRE+) and in part by the joint research project "Content Presentation and Services Integration in Czech and Bulgarian Mathematical Digital Libraries" between the IMI BAS and IM ASCR.

References

1. American Mathematical Society, MathSciNet, http://www.ams.org/mathscinet/, accessed July 30, 2014.
2. Apache Software Foundation, Apache Lucene Core, http://lucene.apache.org/core/, accessed July 30, 2014.
3. Bouche, T., Goutorbe, C., Jorda, J.-P. and Jost, M.: The EuDML Metadata Schema: Version 1.0. 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. 45–61, http://dml.cz/bitstream/handle/10338.dmlcz/702602/DML_004-2011-1_9.pdf.
4. Bulgarian Mathematics Digital Library, http://sci-gems.math.bas.bg.
5. Category: Proof assistants, Wikipedia, http://en.wikipedia.org/wiki/Category:Proof_assistants, last modified September 21, 2011.
6. Conference on Intelligent Computer Mathematics, last modified July 6, 2014, http://cicm-conference.org/2014/.
7. Cornell University Library, Digital Mathematics Library. S.E. Thomas, principal investigator, R.K. Dennis and J. Poland, co-principal investigators, last updated December 2, 2004, http://www.library.cornell.edu/dmlib/.
8. Czech Digital Mathematics Library, http://dml.cz.
9. Developing a 21st Century Global Library for Mathematics. Report of the Committee on Planning a Global Library of the Mathematical Sciences; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council; ISBN 978-0-309-29848-3, http://www.nap.edu/catalog.php?record_id=18619.
10. DML: Digital Mathematics Library. http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_links.html, accessed July 30, 2014.