European Digital Mathematics Library

Jiří Rákosník[1], Radoslav Pavlov[2]

[1]Institute of Mathematics AS CR, Czech Republic
[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
rakosnik@math.cas.cz, radko@cc.bas.bg

"Making mathematics literature published in Europe available online"
www.eudml.org

Abstract. The aim of this paper is to survey the European Digital Mathematics Library project goals and achievements as well as an outlook for sustainable development.

Keywords: Digital Mathematical Library, Open Access, Institutional Repositories, DSpace, Publishing

1 Introduction

The idea of a freely accessed Digital Mathematics Library (DML) has been endorsed by the International Union of Mathematicians [1], [2], [3] for more than two decades. In between, numerous local initiatives have emerged in Europe ([4], [5], [6], [7], [8], [9], [10], [11], [12], [13]) creating a huge volume of digital material as a base for a global DML and the European Mathematical Society (EMS) promoted several attempts to get a public support for a project of the European Digital Mathematics Library.

A major step forward has been achieved during the course of the pilot project of the European Digital Mathematics Library (EuDML) partially funded by the European Commission from February 2010 to January 2013. The project has been accomplished by a consortium of 17 partners representing a wide spectrum of stakeholders including local digital libraries, technology developers, universities, research institutes, private companies and organizations representing community of mathematicians:

- Instituto Superior Técnico, Computer Science Department, Lisabon (overall management & technical coordination)
- Cellule MathDoc at Université Joseph Fourier, Grenoble (scientific coordinattion)
- University of Birmingham, School of Computer Science
- Fachinformationszentrum Karlsruhe / Zentralblatt MATH
- Masaryk University Brno, Faculty of Informatics

- Univerzytet Warszawski, Interdiscyplinarne centrum modelowania matematycznego i komputerowego
- Instituto de Estudios Documentales sobre Ciencia y Tecnología – IEDCYT, Madrid (left the project for technical reasons)
- Édition Diffusion Presse Sciences, Paris
- University of Santiago de Compostela, Institute of Mathematics
- Institute of Mathematics and Informatics BAS, Sofia
- Institute of Mathematics AS CR, Praha
- Ionian University, Department of Informatics, Corfu
- Made Media Ltd, Birmingham
- Centre National de la Recherche Scientifique /Cellule MathDoc, Grenoble
- European Mathematical Society
- Niedersächsische Staats- und Universitätsbibliothek Göttingen
- Biblioteca Digitale Italiana di Matematica (entered in the course of the project).

In the rest of the paper the project and its outcome are highlighted.

2 The Project and its Outcomes

EuDML is designing and building a collaborative digital library service that is collating currently distributed content from the diversity of its providers. This is achieved by implementing a single-access platform for heterogeneous and multilingual collections (www.eudml.org).

The project was organized in eleven mutually intertwined work packages. Besides the standard ones (Project Management; Assessment and Evaluation), the remaining nine were devoted to particular building blocks and features. We will analyze each of them.

2.1 Policies, Exploitation and Dissemination

The work package in which all partners took part comprised a variety of activities for promotion and dissemination of project results, bringing together stakeholders, academic community, potential partners and digitization initiatives to agree on crucial issues, setting-up a scientific advisory board under the auspices of the European Mathematical Society and developing a sustainable business model.

The dissemination included creation of the EuDML brand identity (logo, document templates, web sites for the project and for the DML portal), promotional flyers, electronic newsletter sent to subscribers, and more than 200 press releases, articles, interviews, advertisements, lectures, presentations and posters, a series of DML workshops in frames of the Conferences on Intelligent Computer Mathematics, several workshops for different stakeholders and a round table in the 6th European Congress of Mathematicians.

The European Mathematical Society established the Scientific Advisory Board of 10 distinguished representatives of stakeholders worldwide. The Board acknowledged

that EuDML gained some kind of worldwide leadership in the international effort to build a global DML, supported its sustainability plan and provided advices on the EuDML policies.

The EuDML policies can be summed up to the following three items:

1. The texts in EuDML must have been scientifically validated and formally published.
2. EuDML items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on.
3. The digital full text of each item contributed to EuDML must be archived physically at one of the EuDML member institutions.

The business model for EuDML sustainable development is based on creation of an association without legal personality called EuDML Initiative, in which 11 partners providing digital content, technical and political support and human resources will maintain and develop the EuDML under umbrella of the European Mathematical Society for the initial period of at least three years, during which the possibility/necessity of transforming the EuDML Initiative to another model involving legal personality and financial issues will be investigated.

2.2 Content Aggregation

This work package had the following objectives: identify all the metadata schemas used by content providers, define the common EuDML metadata schema, export each metadata set from each provider to the EuDML format, recursively run these tasks as long as the other work packages feed new metadata, and function as an entry point for possible new content providers.

The work started with a content analysis bringing detailed information on available digital collections and metadata. This was used to set up the first version of EuDML metadata schema. The subsequent metadata harvest provided a feedback for individual content providers to improve their metadata and cleared the way for specification of the final version of metadata schema.

After an extensive study of the existing metadata schemas and their actual use by its content providers many different strategies and existing schemas were evaluated. This led to the decision to investigate further the framework provided by the Journal Archiving and Interchange Tag Suite, which has been created by The National Center for Biotechnology Information of the National Library of Medicine in the USA. The tag suite defines a set of XML schema modules for storing and exchanging content of scholarly publications and provides readily usable schemas for journal articles and books which represent 97% item types in EuDML. In between, the NLM Journal Archiving and Interchange Tag Suite used as main reference for the EuDML project has been passed over to the NISO standard body. In fact, the new Journal Archiving and Interchange Tag Set schema with NISO version 1.0 [14] took into account all the changes that had to be incorporated in the EuDML version.

The EuDML metadata schema specification is described on the project site [15]. For content providers to check the validity of their data in EuDML format before publishing it the EuDML XML Data Validation Tool has been developed [16]. It takes as input an XML file and validates it, first against the chosen EuDML schema, verifying that the overall XML structure and namespaces are correct, and second against the schematron rules [17], to make sure that the data is presented correctly according to the EuDML best practices.

By the end of project the EuDML has been aggregating approx. 225 000 items after deduplication (altogether 2 600 000 pages) from 22 datasets of 12 content providers (Tables 1 and 2).

Table 1. Number of items collected in the EuDML

Item type	Number of items
Journal article	221 293
Proceedings contribution	2 962
Book chapter	42 520
Book: monograph	1 724
Book: conference	66
Book: volume	1 179
Multiple volume work	296
Total	270 040

Table 2. Items contributed to the EuDML by individual content providers

Country	Projects	Contributed
Germany	GDZ Mathematica, ELibM	100 000
France	Gallica-Math, NUMDAM, CEDRAM	57 000
Czech	DML-CZ	28 000
Russia	RusDML	17 000
Poland	DML-PL	14 000
Spain	DML-E	6 400
Greece	HDML	3 000
Italy	BDIM	2,000
Portugal	SPM/BNP	1 300
Bulgaria	BulDML	600

2.3 System Architecture and Design

The primary objective of this work package was to design the architecture and to create technical specifications of the system and its interfaces upon which the components were implemented within other work packages. Its role was to coordinate technical activities of the project, analyse the requirements, design the architecture, plan and supervise software development cycle, supervise and orchestrate integration efforts (Figure 1).

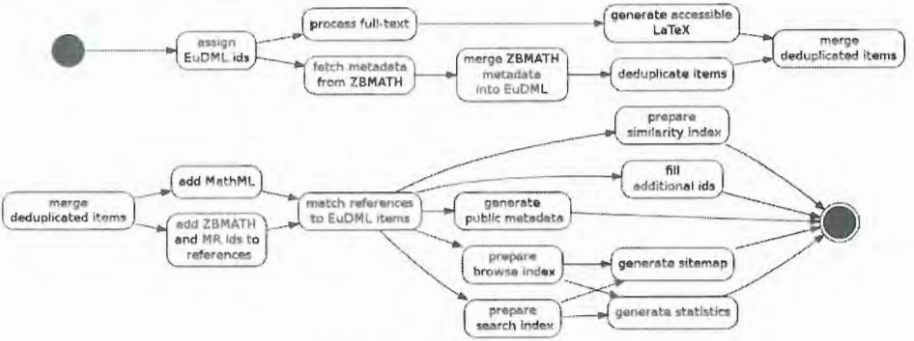

Fig. 1. Processing workflow

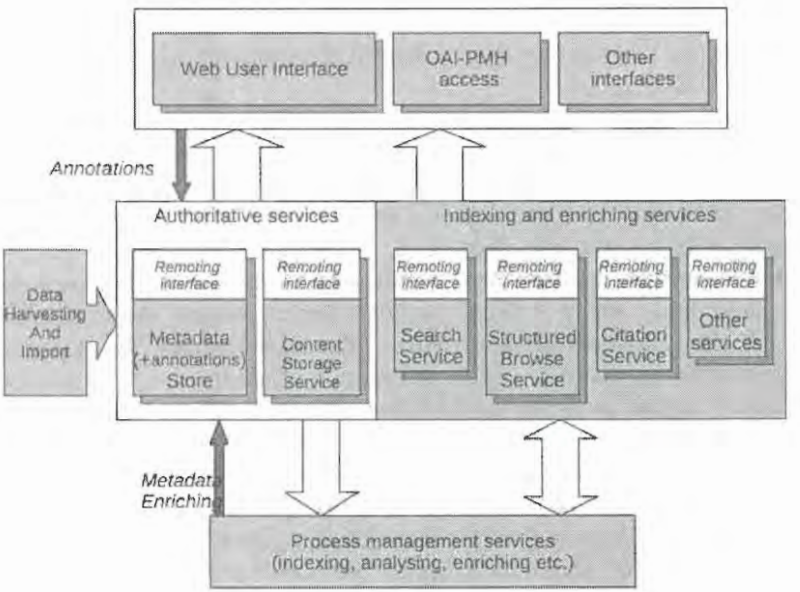

Fig. 2. EuDML core and extension services

The EuDML system is based on Service Oriented Architecture. It is designed to embrace a set of core services (sufficient for the basic system operation) and a number of enriching services (e.g. publication metadata store, the indexing and the search services, the content storage system and structured publication browsing services) (Figure 2).

The extensibility and the multidimensional scalability of the EuDML platform are its key features: allowing easy addition of new services (and content), additional volume, new content's structure, concurrent users, etc., without performance or reliability degradation. A natural solution towards these concepts is a modular, distributed architecture [18].

2.4 Metadata Repository and Search Engine Implementation

The main tasks were to provide core services (metadata processing, index and browse services) and a framework for other services to integrate with. Other goals included user management, storage service, workflow engine and formulae search.

The EuDML metadata repository is formed by REPOX [19], a framework to manage metadata spaces. It comprises several channels to import metadata from data providers, services to transform metadata between schemas according to user's specified rules, and services to expose the results to the exterior. REPOX works as an OAI-PMH client and server. It allows to schedule metadata ingests and monitoring OAI-PMH servers – the system automatically sends a notification to the administrator in case of unavailability. REPOX is a core component of the EuDML metadata ingestion workflow from the Content Providers to the YADDA Storage System (EuDML main data repository) [20], [21].

The EuDML search service supports basic search for the term in fulltext index and an advanced search, where user may combine number of terms in complex boolean queries. Each search result set may be refined by facets, which include journal, document type, publication year and author. List of mathematical terms and topics for autocompletion helps users with dyslexia.

Search also supports mathematical formula search, where user can type in a mathematical formula in either LaTeX (popular format among mathematicians) or MathML. It allows finding different forms of the same formula. The quality of the results depends on the metadata and content given by providers.

The search engine is based on YADDA Search with the indexes stored in Lucene/Solr-based system.

The EuDML browse service provides two main usages within application. First is browsing of the journals. The journals can be listed, and filtered by title. Then on the journal page user may browse journal contents, using hierarchical tree. This is implemented using dedicated hierarchical browse service. Second browsing method is 'browse by subject' based on MSC2010 categorisation of the articles. User may browse articles for specific topic, and navigate down in hierarchical category tree. Only articles are visible in this browsing view.

2.5 Web and Service Interface Implementation

This work package covers the human and machine interfaces of the system. Its goals were to achieve a high quality set of interfaces to EuDML that have a consistent look and feel promoting the EuDML brand, are internationalised across a range of European languages and provide accessibility options to visually-impaired and dyslexic users, support integration on external web sites via configurable widgets and provide interoperability support with other relevant systems.

The EuDML service interface allows the EuDML system to be used by external applications and third party systems. It consists of several RESTful interfaces: OAI-PMH server, OpenSearch, Batch Ref, Reverse Ref, Similar Items, All Pointers, Batch Ids, Handled Ids, Metadata. Demonstration can be found under [22].

2.6 Metadata Enhancer Toolset Implementation

Objectives of this work package were to provide tools and workflows to extract textual and mathematical metadata, validate and merge the discovered metadata with that already registered for the target items and design and implement metadata enhancement workflow.

Large number of tools that have been adapted or developed for the EuDML include text extraction (MaxTract for recreating latex sources from PDFs, PDF Text Extractor, INFTY OCR for mathematical formula recognition, PdfToTextViaOCR based on Tesseract), mathematical formulae processing (Tex2NLM for converting TeX code to MathML and EnhanceNLMTeXwMML), metadata enrichment and deduplication and merging tools. For the detailed reference see the EuDML Enhancer toolset demos [23].

2.7 Association Analyser Implementation

The goal was to provide tools to identify various types of referential and semantic connections between different items in the content repositories and also between such items and external resources, to integrate them into workflow and integrate the results with the UI.

The toolset includes two key functionalities. The citation interlinking and matching aims to create a network of documents within the collection by automatic parsing and linking of citations. It is based on the UJF Citation Matcher which provides a robust method for resolving (incomplete or possibly incorrect) citations to a particular document or identifier. The service allows matching citations within EuDML by REST and Web User Interface. External database item matching identifies items from external databases (Zentralblatt Math and Mathematical Reviews) and identifies references to documents in these databases. By the end of the project about 100 000 internal citations and about 1 100 000 links to external databases have been resolved. In addition, an interactive lookup allows the user input bibliographic citation (as a string) and get back near matches when they are found, and the BatchRef tool can be used to

process a batch of reference strings and get back EuDML identifiers when they are found.

The document similarity service is finding ordered set of documents semantically similar to the given one using MathML formulas. It is using GENSIM similarity engine [24] which implements unsupervised clustering algorithm which uses various machine learning methods to measure semantic similarity based on word co-occurrences. The tool is available through REST service.

2.8 Annotation Component Implementation

A modern digital library should do more than just obtain items to read for the user - it should provide richer interactions between the user and the library and support collaboration between users. The objectives of the work package included provision of replicated, synchronised management of user annotations to items in the content repositories. Annotations can be comments, discussion threads, tutorials, reviews, reading lists, or other user contributed elements that can be attached to individual items in the collection.

The service provides a range of different item annotation types. The user can add comments to items for other users to see and provide links to important related items that the uninitiated might not be aware of. Replies can be added to previous notes to build a conversation. The user can create lists of items relevant to research projects or interests for oneself, sharing them among a team of collaborators or publishing them to the world. Improving the library for everyone is allowed by community sourced feedback and correction to the library metadata. The users can share items with other users via email, social networks (Tweet, Facebook "Like", Google "+1") and bibliography and reference managers (Mendeley, CiteULike, BibSonomy). Widgets allow easy integration of these resources on non-EuDML sites. The connection to EuDML from widgets is implemented on top of a remote service interface which accepts URI queries for simple queries, SPARQL queries for more complex queries and returns JSON records with the query results.

2.9 Accessibility Component Toolset Implementation

Motivation and objectives of this work package include providing support not normally available for mathematical documents for users with special needs (visually impaired or dyslexic) and towards automatic language translation, by generating formats for mathematical documents that allow for access like LaTeX, MathML, Braille, supporting existing tools for speech synthesis of text and math, implementing search term correction for mathematical vernacular and creating a multilingual thesaurus to enable knowledge discovery within multilingual resources.

Full accessibility has been provided for users with special needs for some documents. The user can choose accessible file formats. Dyslexia support based lexicon of terms extracted from the Encyclopedia of Mathematics comprises search term completion suggested as soon as three letters are entered. Multilingual support via cross-indexing of terms in a thesaurus enriches indexed documents in various languages

with English translation of the occurring terms. The index can be used to find documents in other languages referring to the particular English mathematical term.

All these functionality is considered experimental and can have significant limitations. However, it shows how a digital library can provide dedicated tools tailored to mathematics to allow for support in very advanced areas like accessibility and multilingual search.

3 Conclusions

During the three years project the consortium of 17 partners across Europe produced a fully functional European Digital Mathematics Library EuDML with a critical mass in content. The EuDML mathematics oriented features and services include MathML metadata, math mining, MSC, links to/from math databases. The EuDML web site offers unique navigation tools adapted to the user community (internal and external deep interlinking, MSC browsing, reference lookup). The project outputs include a number of productivity and interoperability devices enabling the main service (some production ready, some more experimental). The consortium formed a growing cooperation network and established an external cooperation model [25] and formulated a business plan [26] for EuDML sustainability and further development.

References

1. Jackson, A.: The digital mathematics library. Notices Amer. Math. Soc. 50 (2003), no. 8, 918–923. Dostupné z: http://www.ams.org/notices/200308/comm-jackson.pdf
2. Ewing, J.: Twenty centuries of mathematics: Digitizing and disseminating the past mathematical literature. Notices Amer. Math. Soc. 49 (2002), no. 7, 771–777.
3. The Future World Heritage Digital Mathematics Library: Plans and Prospects. http://ada00.math.uni-bielefeld.de/mediawiki-1.18.1/index.php/Main_Page/
4. NUMDAM: Numérisation de documents anciens mathématiques, http://numdam.org/
5. Cedram. http://www.cedram.org/
6. bdim: Biblioteca Digitale Italiana di Matematica. http://www.bdim.eu/
7. eLibrary of Mathematical Institute of the Serbian Academy of Sciences and Arts. http://elib.mi.sanu.ac.rs/pages/main.php/
8. Bulgarian Digital Mathematics Library. http://sci-gems.math.bas.bg/
9. Biblioteca Digital Española de Matemáticas. http://dmle.cindoc.csic.es/
10. Polska Biblioteka Wirtualna Nauki. Kolekcja Matematyczna. http://pldml.icm.edu.pl/
11. Hellenic Digital Mathematics Library. http://www.hdml.gr/
12. Göttinger Digitalisierungszentrum. http://gdz.sub.uni-goettingen.de/
13. Czech Digital Mathematics Library. http://dml.cz/
14. U.S. National Library of Medicine National Center for Biotechnology Information. Journal archiving and interchange tag library, niso jats version 1.0, August 2012. Full online documentation at http://jats.nlm.nih.gov/1.0/
15. EuDML Metadata Schema Specification. https://project.eudml.org/eudml-metadata-schema-specification-v20-final/
16. EuDML XML Data Validation Tool. http://eudml.mathdoc.fr/eudml-validation-demo/
17. Schematron. http://www.schematron.com/

18. Pavlov, R.: No Royal Road but at least a Gateway to the Mathematical Knowledge: The EuDML Project, In: Proceeding of the Fortieth Jubilee Spring Conference of the Union of the Bulgarian Mathematicians "Mathematics and Education in Mathematics", 70–79 (2011).
19. REPOX. http://bd2.inesc-id.pt:8081/repox/
20. YADDA. http://ceon.pl/en/software/yadda-main-en
21. The EuDML Metadata Registry and Repository – Final. Deliverable D5.4 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library. https://project.eudml.org/sites/default/files/D5_4_v1.0.pdf
22. EuDML API-Tester. http://project.eudml.org/api-tester/
23. EuDML Enhancer Toolset. https://project.eudml.org/tools-technical-specifications/
24. gensim. http://nlp.fi.muni.cz/projekty/gensim/
25. Bouche, T.: Reviving the Free Public Scientific Library in the Digital Age? The EuDML project. In Klaus Kaiser, Steven Krantz, and Bernd Wegner, editors, Contemporary Issues in Mathematical Publishing, 20 pp., 2013. To appear.
26. Bouche, T. and Rákosník, J.: Report on the EuDML external cooperation model. In Klaus Kaiser, Steven Krantz, and Bernd Wegner, editors, Contemporary Issues in Mathematical Publishing, 8 pp., 2013. To appear.