МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2011
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2011
Proceedings of the Fortieth Jubilee Spring Conference
of the Union of Bulgarian Mathematicians
Borovetz, April 5–9, 2011

NO ROYAL ROAD BUT AT LEAST A GATEWAY TO THE MATHEMATICAL KNOWLEDGE: THE EuDML PROJECT

Radoslav Pavlov

We will deliver a truly open,
sustainable and innovative framework
for access and exploitation of
Europe's rich heritage of mathematics.
EuDML Team

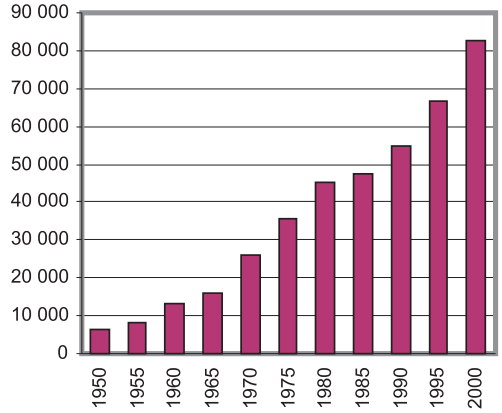
The paper presents the EuDML project – the European Digital Mathematics Library (`http://www.eudml.eu`) which main goals are:

- to create a common infrastructure for seamless navigation, searching and interacting within the deeply interlinked network of distributed validated multilingual digital mathematical content available throughout Europe, which will make mathematics readily available for all users of this resource;
- to provide a safe archival back-end so that publishers do not have to maintain their back catalogues indefinitely, and thus agree to transfer their content and to licence eventual open access to it according to their moving wall policy;
- to satisfy the demand for reliable and long-term availability of mathematical research output.

The Bulgarian contribution to the project is also presented – BulDML "The Bulgarian Digital Mathematics Library" (`http://sci-gems.math.bas.bg`).

1. The Mathematics library. It is a myth that mathematicians need nothing for their work. Research in mathematics may not require expensive lab equipment or involve costly field trips, but every mathematician needs fast and free access to the huge and growing body of mathematical knowledge accumulated since antiquity. The traditional form of storing this knowledge is the library of paper editions. A complete coverage of the classics, prompt acquisition of new items, ease of access, user-friendliness are all goals that librarians have been used to pursue. But in order to be fully useful, the mathematical library must be more than a collection of texts; it must be an infrastructure, a network with links and references which allow one to track the relations between the items. The peculiarity of mathematics among sciences is that its results never age: a valid theoretical development remains valid for ever, with new results supplementing rather than supplanting the classics.

In our times the volume of new scientific results is growing exponentially. Mathematical literature today amounts to more than 3 million publications, of a total of 100 million pages. The diagram on the right shows the number of new publications per year (by now it has reached 100 000). In 2010 Zentralblatt MATH has refereed 66 180 new items. Half of all citations aim more than 10 years back; one quarter aim more than 20 years back. This rate of growth is hard for even the largest and wealthiest traditional libraries to keep up with.



2. The Digital Mathematics Library. Together with this problem, the modern age has brought forth its solution, based on the progress of digital and communication technologies. The solution is the digital library of mathematical knowledge. Such a library can grow indefinitely, with no physical restrictions whatsoever, and offer ubiquitous access to all, from anywhere and at any time. Moreover, it allows the construction of a comprehensive system of links and references between the items in a way that paper editions do not.

The creation of digital archives is assisted by the fact that nowadays it is standard practice for new publications to be prepared in electronic form, and old texts are being converted into it by many projects worldwide. However, these efforts are not well coordinated and do not systematically cover the literature. Digital/digitised documents are often duplicated, serials are split among various providers, and permanent reliable access is not guaranteed. Good access requires numerous subscriptions and complicated search. These challenges led to the concept of a large-scale virtual mathematical library.

After pioneering generalist digitisation projects launched in the 1990s in France (Gallica), in USA (JSTOR), in EU (DIEPER), and the seminal JFM/ERAM project (funded by DFG over 1998–2002), the International Mathematical Union (IMU) published a white paper on the concept of "a virtual library containing much of the past literature—a library that could eventually grow into a World Mathematics Library". This led in 2001 to: "Twenty centuries of mathematics: Digitizing and disseminating the past mathematical literature" [1].

While new maths oriented projects were successfully launched (project Euclid in USA, NUMDAM in France), the NSF funded a one year (2002–2003) planning project in order to study further the feasibility of the Digital Mathematics Library (DML).

IMU proposed to assume further coordination, under the flag "World Digital Mathematics Library" (WDML), with supervision by the Committee on Electronic Information and Communication (CEIC). This worldwide activity culminated in 2006 when the General Assembly of IMU adopted the communication "Digital Mathematical Library: a Vision for the Future" prepared by the CEIC. However, no actual coordination emerged among interested parties, whose number was steadily growing. The ultimate WDML goal

being virtually endorsed by all mathematicians, new initiatives appeared worldwide, from a wide range of organisations with possibly different views on content selection, cooperation, and business model.

The European Mathematical Society (EMS), which had been involved in this process from the beginning, organised brainstorming meetings in Berlingen (Switzerland) in 2002 and 2003, yielding an "expression of interest" letter sent to the European Commission in 2003, and a proposal to FP5 for an infrastructure called DML-EU. Despite the lack of a formal success of this European initiative, a number of meetings, conferences and local DML activities have been organized since 2003, thus demonstrating even stronger the urgent needs of the global scientific community. This continuous work allowed participants to keep track of: the ongoing activities, emerging best practices, as well as the technical challenges related to the development of a "mathematical semantic web". Finally the idea for a common DML infrastructure for access to the mathematical knowledge in Europe was successfully materialised within the frames of the project *European Digital Mathematics Library* – EuDML.

3. The European Digital Mathematics Library – EuDML. The EuDML project obtained support from the European Commission in the framework of the Information and Communications Technologies Policy Support Programme (CIP ICT PSP, "Open access to scientific information", project no. 250503). The official start of the project was on February 1, 2010 for a duration of 36 months [2].

The project consortium consists of 14 European partners (Table 1), among which universities, public institutions and leading technology providers, dedicated to producing quality scientific information. The project partner teams include mathematicians, librarians, digital library specialists, publishers, professional information service and document engineering specialists, and computer scientists. The consortium is strengthened by its 2 associated partners: the European Mathematical Society is associated to the project as the relevant organization for setting the goals and assessing the usefulness of the project's outcome, it will chair the Scientific advisory board; the University Library of Göttingen will contribute journals digitised in the ERAM and RusDML projects, as well as the largest collection of digitised mathematical books.

The main goal of the EuDML project is to establish a pilot implementation of an integrated mathematics digital library system allowing for seamless access for otherwise dispersed digital material of the partners and associates.

The main outcome of the envisioned library service is expected to set-up a network of institutions where the digital items would be physically archived. Each *local* (partner) institution is assigned with the following tasks: selecting, acquiring, developing, maintaining, cataloguing and indexing, as well as preserving its own collections according to clear policies. Thus they will play the role of a reference memory institution for a well-defined part of the mathematical corpus.

The network of institutions as a whole would make it possible to assemble a global, virtual library providing a *one-stop gateway* to the distributed content through user-friendly retrieval interfaces. Moreover, published standards of interoperability would allow this virtual library to serve as an infrastructure layer for any component of the scientist's environment where reference to a published mathematical result is necessary, turning a mere intellectual reference into an actual link to the result.

The EuDML system (Fig. 1) is based on Service Oriented Architecture. It is de-

Table 1. The EuDML Consortium

Partner organisation name	**Technical expertise**	**Contributed content**	**Country**	**Current estimation of the existing DML content**
Instituto Superior Técnico: Computer Science Department	digital libraries	*Portugaliae Mathematica*	Portugal	2,000 items
Université Joseph Fourier: Cellule MathDoc	math metadata, journal publishing	NUMDAM, CEDRAM	France	50,000 items
EDP Sciences	journal publishing & interoperability	SMAI/EDPS journals		
University of Birmingham: School of Computer Science	math accessibility		United Kingdom	
Made Media Limited	user interface			
Fachinformationszentrum Karlsruhe: Zentralblatt MATH	math metadata	Zentralblatt, EMIS	Germany	85,000 items
Masaryk University Brno: Faculty of Informatics	math metadata, electronic publishing	Czech math journals	Czech Republic	26,000 items
Institute of Mathematics – Academy of Sciences of the Czech Republic	math, journal publishing	DML-CZ		
Warsaw University: Interdisciplinary Centre for Mathematical and Computational Modelling	access platform & interoperability	Bibliotheka Wirtualna Matematyki	Poland	13,000 items
Consejo Superior de Investigaciones Científicas: IEDCYT	electronic scientific literature	DML-E	Spain	5,000 items
University of Santiago de Compostela: Institute of Mathematics	math			
Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences	digitisation and digital libraries	BulDML	Bulgaria	3,500 items
Ionian University: Department of Informatics	math and education metadata	HDML	Greece	15,000 items

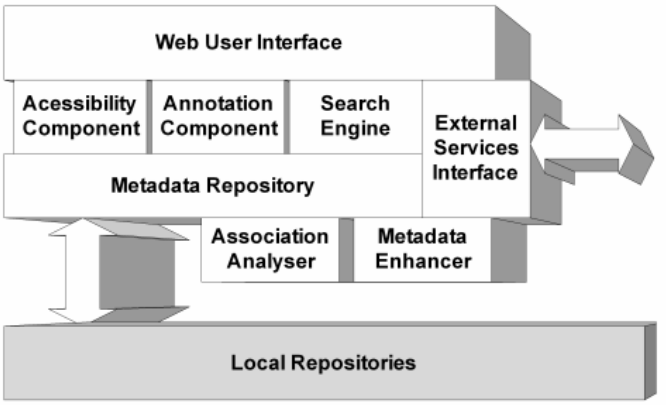


Fig. 1. EuDML functional vision

signed 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) (Fig. 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 degra-

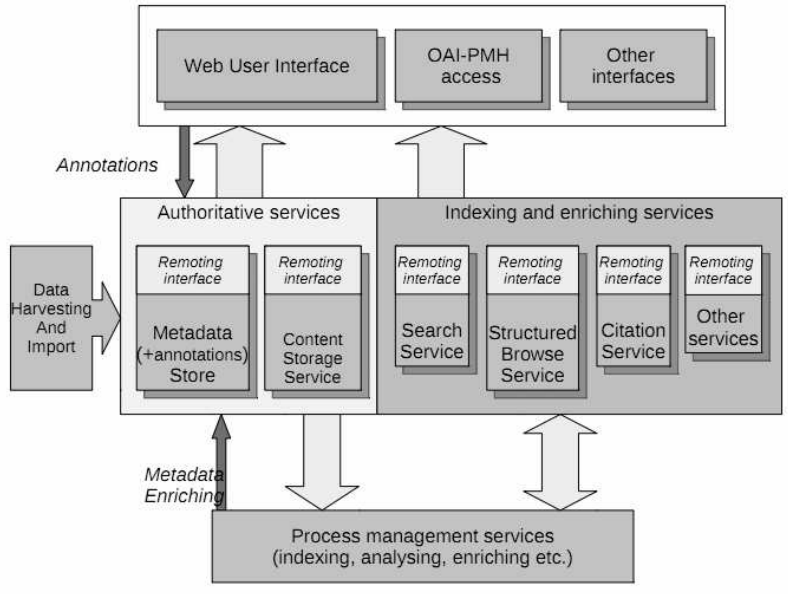


Fig. 2. EuDML core and extension services

dation. A natural solution towards these concepts is a modular, distributed architecture.

The *Metadata Repository*, which manages a replica of all the metadata for all the items in the different content repositories is at the core. It does not contain actual object items, which remain in the local repositories (this way we avoid to have to deal with IPR issues, leaving the control of that to the local repositories).

The *Metadata Repository* manages the metadata updates from the local repositories. It will support adapting the different forms of metadata that each provider has about their items to the common format required by the EuDML. To support those processes, the *Metadata Repository* also will comprise in its internal architecture a *Metadata Registry*. Some items in the content repositories may not yet be freely accessible, but may need enhancement of their metadata. Hence, a facility is planned for content providers to invoke the *Metadata Enhancer* directly on their privileged access items *via* the External Services Interface, to enhance the metadata before uploading. At the bottom is the collection of Local Repositories. Each content provider maintains its own collection of items, which can be supplied on demand. Such requests are *via* identifiers, i.e. the content repositories are not required to provide search or browse facilities, but merely to manage their collections of items, provide access *via* identifiers and manage right of access issues corresponding the to moving wall open access policy adopted by the consortium.

The *Search Engine* provides search and item identifier resolution facilities. The Metadata Repository mediates with the content repositories to fetch actual items for the higher levels of the system.

The *Metadata Enhancer* is a component that can identify items for which it might be able to enhance the existing metadata, including fetching the item and analyse it (e.g. using current OCR, math formula recognition, keyword extraction, signature file construction, bibliography analysis technologies). This component will actually consist of a range of different tools, that can be improved and extended over time, and that can handle different aspects of individual item metadata enhancement.

The *Association Analyser* works on metadata for sets of different items, rather than, as in the case of the Metadata Enhancer, on single items. It identifies related items from the metadata records and updates the metadata accordingly so that various links between documents, or links to relevant external resources, can be recorded in the Metadata Repository.

The *Annotation Component* provides mechanisms to attach new material to individual items in the repositories and maintain this new material. It is envisaged that this will support community interactivity with the collection by allowing users to add, view and update their own reviews, tutorials, comments, or recommendations to each item.

The *Accessibility Component* provides support for enhancing accessibility of items, if required, before presentation to end users. For visually impaired or dyslexic users, this will involve speech synthesis for text or for mathematical formulae (*via* MathML annotation), large print re-formatting or OCR facilities to make scanned image items accessible to Braille readers. Automatic translation facilities will be provided for texts in a language not understood by the user.

The *Web User Interface* and the *External Services Interface* provide public access to the system both *via* web browsers and by web services. This comprises bidirectional interoperability with external resources such as the Math Genealogy, MacTutor History of Mathematics, Mathworld, Wikipedia, EqWorld and other mathematical and science

related web resources. Interoperability with Europeana will be also considered at this level.

The EuDML system's architecture will be developed taking into account the detailed functional requirements specifications. The extensibility, scalability and reliability requirements make it natural to look for a fully distributed, platform independent solution. While the general architecture design will follow the Service Oriented Architecture (SOA) paradigm, the communication layer will remain abstract. This will enable the individual services to communicate through different appropriate means, including possible direct connection when deployed in a single location (host). Furthermore, universal SOAP communication will be maintained as the default for flexibility and compatibility reasons, and more lightweight remote communication will be provided when necessary.

Each of the functional modules presented above will be accomplished by means of a separate service or a group of services acting together.

The *Metadata Store* will be composed of several separate services acting together: a *Metadata Repository Manager*; a *Storage Manager*; and a *Metadata Registry*. An important requirement is also to provide the assignment of a persistent identifier to each entity, and the ability of the related resolution service to point to a local copy and back to the original resource's locations.

Despite the fact that EuDML will be used its internal common metadata schema, it is anticipated that multiple different metadata patterns will be used by different content providers and data sources, and these will have to be reversibly mapped onto the internal data structure. That implies an additional requirement for the *Metadata Store*, viz. to support the synchronization of the different metadata forms with the format required by the EuDML. For that purpose, the *Metadata Repository* will also include a *Metadata Registry* in its internal architecture [3].

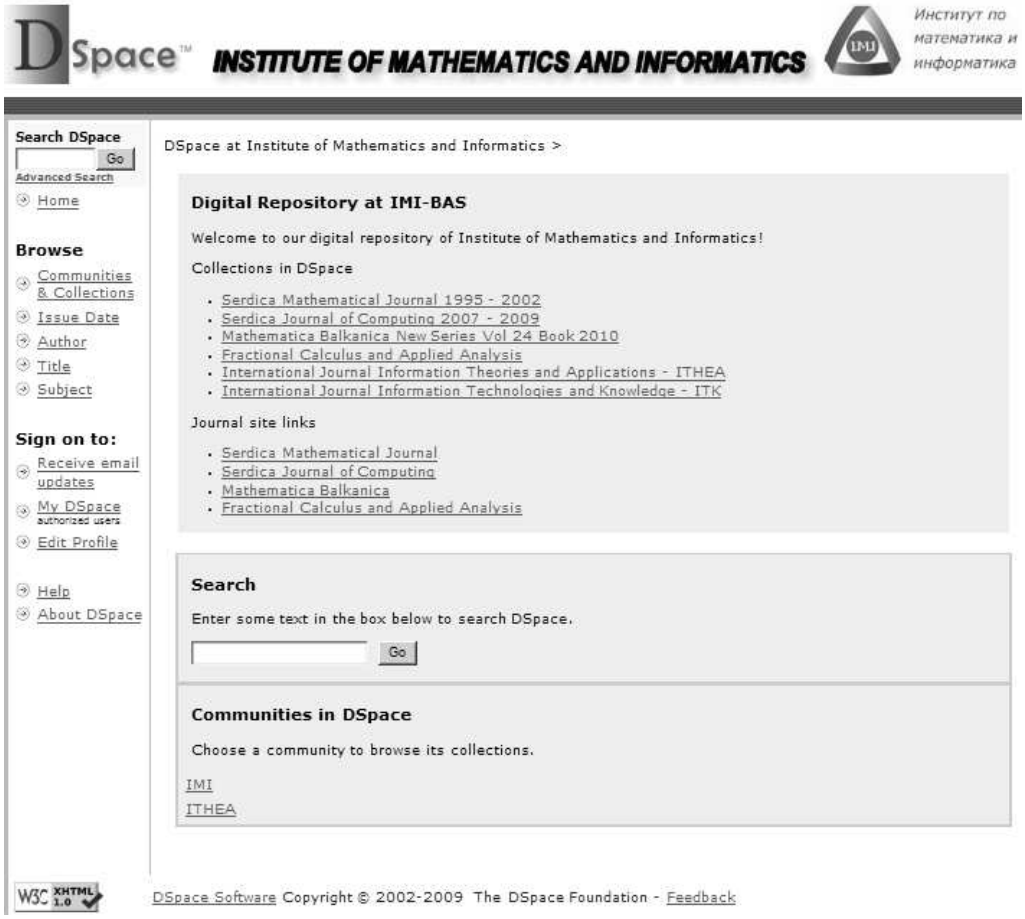
The other expected functions, including annotation, accessibility, metadata enhancer or association analyser will be realised as separate enriching services, that while following the same service design principles, will be considered extension services. Examples of such services include the Citation Service, responsible for citation resolving and indexing or Similarity Service [4], which would be able to return similar objects based on a predefined metrics and criteria. Similarly, additional extension services are hoped to be developed in the future by third parties or by the involved partners.

One or more *Web User Interface Services* will be developed, based on user requirements criteria. Functional interfaces and widgets is also envisaged to enable the inclusion of an *EuDML Search Box* in other local systems and portals. All the system's operations will be managed *via* a process management service, responsible for operation scheduling, synchronisation and timely execution, and for the overall system level integrity of the services.

4. The Bulgarian Digital Mathematics Library – BulDML. The IMI-BAS repository developed under the *EuDML* project is based on the DSpace software with an active and configured OAI-PMH interface for metadata harvesting. A permanent unique identifier from Handle.NET system is assigned to each article (not only for citation needs). These identifiers together with OAI-PMH make it possible to build a network of repositories, search portals, and centralised web services. The BulDML provides an open access to all its content.

When performing a search the user can view the individual search result (a record

representing an item or a digital object in the repository and eventually the item itself). Two options are offered – simple and full record. The simple record provides bibliographic data related to the item with labels in natural language (e.g., Title, Author or Publisher) whereas the full record provides more detailed information with Dublin Core element labels (such as *dc.title*, *dc.contributor.author* or *dc.publisher*). Both record types provide a URI (Uniform Resource Identifier), in this case represented by handle, which serves as a persistent identifier of the item.



`http://sci-gems.math.bas.bg`

The BulDML content is organized around Communities which can correspond to administrative entities such as schools, departments, labs and research centers. Within each community there can be an unlimited number subcommunities and an unlimited number of collections. Each collection may contain an unlimited number of items.

BulDML is configured to use the Dublin Core metadata schema by default and is in a full accordance with the metadata scheme developed under the EuDML project. Dublin core is made up of elements, and qualifiers. There are 15 base elements: *Title;*

Creator; Subject; Description; Publisher; Contributor; Date; Type; Format; Identifier; Source; Language; Relation; Coverage; Rights.

Currently the BulDML covers more than 3500 mathematical items – publications from 6 journals:

- Serdica Mathematical Journal
- Serdica Journal of Computing
- Mathematica Balkanica
- Fractional Calculus and Applied Analysis
- International Journal Information Theories and Applications – ITHEA
- International Journal Information Technologies and Knowledge – ITK

and conference proceedings.

BulDML captures, distributes and preserves digital research products. Here you can find articles, working papers, preprints, technical reports, conference papers and data sets in various digital formats. Content grows daily as new communities and collections are added to BulDML.

5. Conclusions. The EuDML project is a result of the solid commitment and serious efforts of the mathematical community to produce a common digital library of the mathematical knowledge. This could be the ultimate answer to the difficulties mathematicians face when in search of a mathematical result they need. The foundations of the EuDML semantic mathematical Web set would enable a user to find a previous work by using a mathematical structure or a formula rather than a keyword. Furthermore, any advance in this area achieved in the narrow context of mathematical research papers would enhance the accessibility to the whole technical and scientific documentation.

We all remember Euclid's legendary words to King Ptolemy: *There is no royal road to geometry!* Still our strong belief is that the EuDML project will build a gateway to the world's mathematical knowledge.

Acknowledgements. The author is thankful to Dr. Ivan Derzhanski, Georgy Simeonov and Maria Sendova for their support in the preparation of this paper.

REFERENCES

[1] T. Bouche. Introducing EuDML – The European Digital Mathematics Library. *EMS newsletter*, 76:11–16, June 2010. `http://www.ems-ph.org/journals/journal.php?jrn=news`.

[2] `http://www.eudml.eu`

[3] W. Sylwestrzak et al. EuDML—Towards the European Digital Mathematics Library Architecture and Design. In: In: Proceedings of DML 2010 Conference. Towards a digital Mathematics Library. Paris, France, 2010, p. 19.

[4] R. Rehurek, P. Sojka. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta, 2010, 45–50; software available at `http://nlp.fi.muni.cz/projekty/gensim`

Radoslav Pavlov
Institute of mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: radko@cc.bas.bg

АКО НЕ ЦАРСКИ ПЪТ, ТО ПОНЕ МАГИСТРАЛА КЪМ МАТЕМАТИЧЕСКОТО ЗНАНИЕ: ПРОЕКТЪТ EuDML

Радослав Павлов

Представен е проектът EuDML – Европейската цифрова библиотека по математика (http://www.eudml.eu), който цели:

- да създаде обща инфраструктура за безпроблемна навигация, търсене и взаимодействие в рамките на плътна мрежа от разпределено валидирано многоезично математическо съдържание в цифрова форма, което да е достъпно в цяла Европа, и така да направи математиката лесно достъпна за всички потребители;
- да задоволи изискването за надежден и дългосрочен достъп до математическите изследвания.

Представен е и българският принос в проекта – BulDML – цифрово хранилище за математическа литература на Института по математика и информатика на БАН (http://sci-gems.math.bas.bg).