Project EuDML — A First Year Demonstration

José Borbinha[1], Thierry Bouche[2], Aleksander Nowiński[3], and Petr Sojka[4]

[1] INESC-ID, Portugal
jlb@ist.utl.pt
[2] Institut Fourier (UMR 5582) & Cellule Mathdoc (UMS 5638), Université Joseph-Fourier, (Grenoble 1), B.P. 74, 38402 Saint-Martin d'Hères, France
thierry.bouche@ujf-grenoble.fr
[3] Interdisciplinary Center for Mathematical and Computational Modelling, University of Warsaw, ul. Pawińskiego 5A, 02-106 Warsaw, Poland
A.Nowinski@icm.edu.pl
[4] Masaryk University, Faculty of Informatics, Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

Abstract. This demonstration describes the results of the first year of the EuDML project, an initiative building a new multilingual service for searching and browsing the content of existing European portals of mathematical content. We demonstrate the first versions and proofs of concept of the EuDML portal, its contents' aggregator, and a toolset for added value.

About EuDML. — EuDML, the European Digital Mathematics Library (WWW.EUDML.EU), is a project that will build a new multilingual service for searching and browsing the content of existing European mathematical portals [5,1]. It will be based on a rich metadata repository, aggregating metadata and full text of heterogeneous and multilingual collections of digitised and born digital content (articles, books, theses, etc.). The service will merge and augment the information about each document from each collection, and also will match documents and references across the entire combined library. Entities such as authors, bibliographic references and mathematical concepts will be singled out and linked to matching items in the collections; similar mechanisms will be provided as public web-services so that end-users or other external services will be able to discover and link to EuDML items. This way, EuDML will be a new major international player in the emerging landscape of scientific information discovery services, enabled for reuse in new added value chains. EuDML is partially funded by the Competitiveness and Innovation Framework Programme of the European Commission (CIP ICT PSP Digital Libraries), grant agreement no. 250.503.

The EuDML Service Architecture. — The EuDML system can be summarised by the use cases represented in Figure 1. The ultimate purpose will be to serve End Users, who can search and browse anonymously, or can register for personalised services. A set of Business Workers are intended to maintain the services and content, while external business partners contribute their content (bibliographic data and full texts for indexing and added value services).

J.H. Davenport et al. (Eds.): Calculemus/MKM 2011, LNAI 6824, pp. 281–284, 2011.

The EuDML Portal. — The first version of the EuDML portal can be accessed from the EuDML website.[1] So far, there are no access restrictions, as all the services are available for anonymous users. This demonstration contains approximately 55,000 documents, provided by a group of partners (CEDRAM, DML-CZ, DML-E, ELibM, GDZ, NUMDAM, Portugaliæ Mathematica, and RusDML).

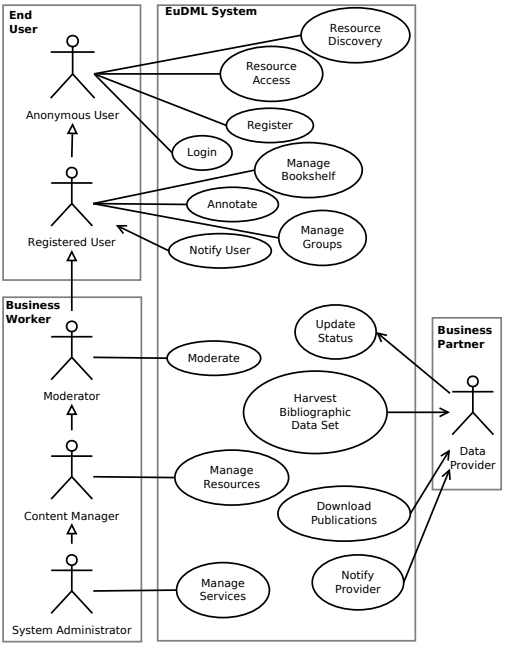

Fig. 1. The EuDML Use Cases

EuDML Content Aggregation. — One of the first project's result was a detailed analysis of the existing content formats and metadata schemas used throughout partnering projects and content providers. Informed by this study, a specification for a EuDML schema, heavily based on NLM JATS[2], was written down.

REPOX is a framework to manage data sets. It comprises multiple channels to import data from providers, services to transform data between schemas according to specified rules, and services to expose the results to the exterior. REPOX allows to monitor OAI-PMH[3] servers and schedule data ingests.

Instances of REPOX for EuDML are currently running at Instituto Superior Técnico (Lisbon) and Cellule MathDoc (Grenoble). These instances aggregate the bibliographic metadata collected up to now, converted to EuDML format, and exploited partially within the portal.

REPOX is complemented by the EuDML Profile Report, a service to inspect and create statistics and metrics on data quality, including whether the data conforms to particular standards or patterns. All these can be accessed from the page set-up in the EuDML website.[4]

The EuDML Enhancer and Association Toolsets. — This demonstration also comprises tools gathered or produced by EuDML partners as building bricks of enhancer tools, whose functionality should check, normalize and enhance metadata collected from partners, including Zentralblatt MATH, or extracted from the analysis of the full text of

[1] Go to HTTP://WWW.EUDML.EU/FIRST-YEAR-DEMOS#SYSTEM

[2] Journal Archiving and Interchange Tag Suite from the US National Library of Medicine, cf. HTTP://DTD.NLM.NIH.GOV/

[3] The Open Archives Initiative Protocol for Metadata Harvesting, cf. HTTP://WWW.OPENARCHIVES.ORG/

[4] Go to HTTP://WWW.EUDML.EU/FIRST-YEAR-DEMOS#AGGREGATION

items in the EuDML collections. Demonstration web pages allow testing and evaluation of prototypes of thirteen tools.[5]

This toolset consists of solutions for OCR, information extraction, content analysis, data conversion and document refinement. At this stage, more tools are being developed and tested mostly at the technology providers' sites, with well defined interfaces allowing further integration into subsystems of the EuDML core system site. As a next step, these tools will be merged into bigger components and installed in the central EuDML system, together with recently developed search of mathematical formulae [4].

Another set of tools targets tasks as interlinking scientifically related items in EuDML: turning citations into links [2], computing semantically similar papers or plagiarism candidate paper pairs [3]. Similar tools for linking to items in external services such as reviewing databases will be developed.

The EuDML User Interface Design and Tools. — The success of the project depends not only on the amount of aggregated data, but on the user experience as well. During the interface design process, a usability study has been performed. The study identified typical usage patterns, so it was possible to design an interface oriented toward scholar's productivity. The initial version of the portal covers the basic functionality of the digital library: searching and browsing collections, downloading the content, etc. Next step of the project is to add Web 2.0 functionalities, allowing to annotate and comment documents, and to share them with others, using both internal mechanisms and external services. The user interface design is also focused on providing efficient support for the mathematical content, both in presentation and user input (search or annotations).

The user interface is developed in Java, and is based on Spring framework[6]. Both EuDML repository services and web interface are based on the Yadda toolkit developed in ICM[7], customised and extended for required functionality.

Final Note. — The services described here are expected to evolve rapidly during the remaining project's lifetime. Up-to-date services and documentation will be linked from the resources page on our web site.[8]

References

1. Bouche, T.: Introducing EuDML—The European Digital Mathematics Library. EMS Newsletter 76, 11–16 (2010)
2. Goutorbe, C.: Document Interlinking in a Digital Math Library. In: Sojka, P. (ed.) Proceedings of DML 2009, Grand Bend, Ontario, CA, pp. 85–94. Masaryk University (July 2009), http://dml.cz/dmlcz/702560
3. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, Valletta, Malta, pp. 45–50. ELRA (May 2010), http://is.muni.cz/publication/884893/en, http://nlp.fi.muni.cz/projekty/gensim

[5] See http://www.eudml.eu/first-year-demos#toolset

[6] See http://www.springsource.org/

[7] See http://yaddainfo.icm.edu.pl/

[8] See http://www.eudml.eu/resources

4. Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries (May 2011), accepted for CICM 2011 track MKM in this LNAI issue
5. Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka, P. (ed.) Proceedings of DML 2010, Paris, France, pp. 11–24. Masaryk University (July 2010), http://dml.cz/dmlcz/702569