

Aleš Nevřela, UČO: 208069, Fakulta informatiky Masarykovy univerzity
obor: Informační systémy, 2. ročník

Projekt: **EuDML**, nositel: CIP, URL: <http://www.eudml.eu/>

Rok: 2011

Abstrakt: Studentská práce o projektu EuDML (European Digital Mathematics Library – Evropská digitální matematická knihovna). Po základních informacích rozděluje projekt do 11 pracovních balíčků (workpackages), ke kterým podává podrobnější informace. Na závěr mapuje aktuální situaci a autorovu zkušenost s implementací webového portálu EuDML.

1. Úvod

Matematické důkazy, články, recenze, knihy, diplomové práce, přednášky z konferencí a další odborné texty. Většina z těchto prací je publikována online a přístupná prostřednictvím celosvětové sítě Internet. K loňskému roku byl rozsah matematické literatury odhadován na méně než 100 miliónů stran textu[2]. Ten je uložen v repozitářích univerzit, na kterých autoři pracují, na jejich osobních stránkách, v národních digitálních knihovnách, případně na dalších místech. Problémem pro případného zájemce o přístup k vybranému svazku se stává jeho vyhledání mezi množstvím těchto úložišť. Právě s tímto chce uživatelům pomoci EuDML[1].

2. Cíle projektu

EuDML (European Digital Mathematics Library – Evropská digitální matematická knihovna) je projekt podporován a spolufinancován Rámcovým programem pro konkurenceschopnost a inovace Evropské komise (Competitiveness and Innovation Framework Programme – CIP). Svůj vývoj započal 1. února 2010 a je plánován po dobu tří let do 31. ledna 2013. Jeho cílem je vybudovat novou vícejazyčnou službu pro vyhledávání a prohlížení si obsahů existujících matematických portálů. Jejím základem by se mělo stát rozsáhlé úložiště metadat a jejich shromažďování a textů kolekcí digitálních a digitalizovaných děl (článků, knih a dalších). Služba prohledá jednotlivá díla a vyextrahuje základní informace jako je autor nebo autoři, bibliografické odkazy a matematické myšlenky. Ty budou nadále odkazovat na odpovídající dokument. Podobný mechanismus bude nabídnut také jako webová služba přístupná dalším aplikacím, stránkám a uživatelům sítě Internet[3].

3. Rozdělení projektu

Vypracování celé digitální knihovny bylo rozděleno celkově do 11 pracovních balíčků (Workpackages) [4]:

3.1. Project management (Správa projektu)

Zahrnuje obvyklou správu projektu jako jsou finanční a administrativní záležitosti a celkový dohled nad vývojem[4].

3.2. Policy, exploitation and dissemination (Strategie, užítkování zdrojů a rozšíření projektu)

Hlavním záměrem je propagace a šíření výsledků EuDML mezi veřejnost, koordinace spolupráce mezi partnery a získávání dalších. Bylo vytvořeno logo, webové stránky a informační leták, který byl přeložen i do dalších světových jazyků, a dále sestavena Vědecká rada (Scientific advisory board), která bude hodnotit první dosažené úspěchy a napomáhat v pro-

jektování dalšího rozvoje EuDML. Jsou pořádány konference a prezentace pro odbornou i laickou veřejnost[5].

3.3. Content Aggregation (Shromáždění obsahu knihovny)

Zajišťuje shromažďování dokumentů a interoperabilitu mezi jejich poskytovateli a centrálním systémem EuDML. To zahrnuje identifikaci kvality a formátů existujících materiálů a používaných metadat za účelem definování vlastního schématu, který se bude využívat v projektu. V rámci vývoje bylo nashromážděno 174 000 záznamů od 7 partnerů v nejrůznějších formátech zahrnující články v časopisech, disertační práce, monografie i vícesvazkové publikace.

Návrh schématu metadat vychází z frameworku Journal Archiving and Interchange Tag Suite (JATS), který je *de facto* standardem používaným v mnoha institucích po celém světě. Ten byl pro potřeby EuDML rozšířen. Základními XML elementy pro popis dokumenty jsou:

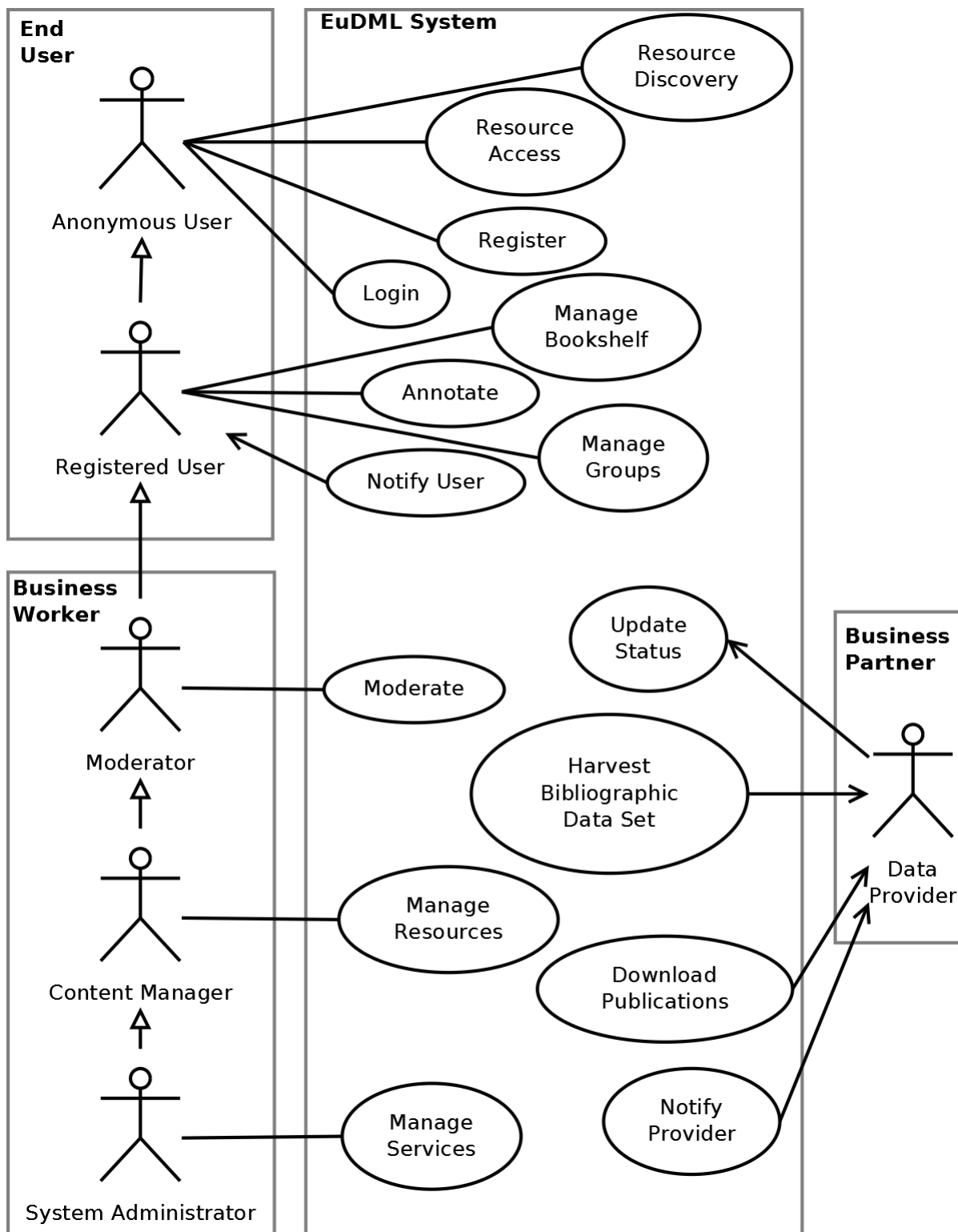
- `<article>` pro popis článků v časopisech s minimální změnou oproti JATS
- `<book>` pro popis jednotlivých knih. V EuDML specifikaci metadat sem patří také popis pouhé části knihy (kapitoly atd.)
- `<mbook>` pro popis prací vícesvazkových děl

Pro zvýšení interoperability s ostatními nepartnerskými systémy byla vyvinuta jednoduchá transformace do OAI-DC (Open Archives Initiative – Dublin Core), takže jsou metadata projektu přístupná jejich sběračům (harvester)[6].

3.4. System architecture and design (Systémová architektura a design)

Cílem tohoto balíčku je navrhnout celkovou architekturu systému. Ta je rozdělena do dvou základních oblastí: architekturu procesů a dat.

Základem celého systému se stávají jednotlivé služby (pro vyhledávání, zpracovávání dat, správu uživatelů, jejich registraci a další) a komunikace mezi nimi – servisně orientovaná architektura. Vše je vyjádřeno v následujícím diagramu případů užití[7]:



Využívat základních funkcí pro vyhledávání a nahlížení do nalezených dokumentů budou moci všichni uživatelé. Registrací získají možnost přidávat si svazky do své osobní knihovničky, vytvářet skupiny a další výhody. Bude zavedena podpora pro přihlašování se prostřednictvím veřejných webových služeb pro autentizaci uživatelů. Zvažovány jsou OpenID a Google Accounts[7].

Projekt EuDML je zaměřen na práci s metadaty publikací, a tudíž je snaha strukturovat informace tak, aby bylo v co nejvyšší míře podporováno jejich zpracování. Vzhledem k navrhovanému systému metadat jsou veškerá data ukládána a je s nimi pracováno ve for-

mátu XML. Jako systém řízení báze dat byla zvolena Postgres SQL databáze pro metadata a menší soubory a souborový systém pro ty velké.

Mezi základní požadavky patří nezávislost na operačním systému nebo webovém prohlížeči. Programovacím jazykem, ve kterém bude projekt implementován, byla zvolena JAVA právě pro svoji multiplatformovost[7].

3.5. Metadata repository and search engine implementation (Repozitář metadat a implementace vyhledávače)

Hlavní náplní je implementace repozitáře metadat a vyhledávání podle specifikací stanovených v pracovním balíčku 4. K tomu jsou využity softwarové nástroje vyvinuté nebo používané partnery projektu, a to YADDA [8] framework a systém REPOX [9]. Ten sbírá metadata od poskytovatelů dat a transformuje je do požadovaného formátu.

Vyhledávač je založen na technologii Solr a dále na systému MIaS (Math Indexer and Searcher) pro indexaci a vyhledávání v matematických digitálních knihovnách vyvinutým na Masarykově univerzitě[10, 11].

3.6. Web and service interface implementation (Implementace rozhraní webu a webové služby)

Cílem je navrhnout a vytvořit webový portál pro uživatele EuDML, prostřednictvím kterého jim budou zpřístupněny jednotlivé služby, a také rozhraní pro webovou službu umožňující dalším systémům komunikovat s tímto. V rámci vývoje byla provedena studie podobně zaměřených služeb: NumDam, Project Euclid, Google Scholar a Zentralblatt Math [12].

3.7. Metadata enhancer implementation (Implementace optimalizátoru metadat)

Nejdůležitějším požadavkem na projekt typu EuDML je efektivní vyhledávání mezi metadaty. Ty by tudíž měly být co nejobsáhlejší a nejpřesnější. Záměrem pracovního balíčku je identifikovat a poskytnout soubor nástrojů pro získávání a zpracování metadat. Zahrnuje nástroje pro skenování a analýzu dokumentů, převod mezi formáty, extrahování metadat a jejich optimalizaci. Úplný seznam s odkazy k jejich použití je k nalezení na domovských stránkách projektu. V úvahu byly brány zkušenosti partnerů, např. Masarykovy univerzity s vývojem České digitální matematické knihovny [13].

3.8. Association analyser implementation (Implementace analyzátoru vztahů mezi dokumenty)

Práce v matematice jsou vždy založeny na předcházejících a ty zase na dalších. Právě na tuto skutečnost je zde upřena pozornost. Cílem je zasadit dílo do kontextu souvisejících dokumentů:

- práce, které jsou citovány v této
- práce, které citují tuto
- recenze

K jeho dosažení jsou využity dvě technologie: indexování citací a seskupování (clustering) dokumentů.

Indexování citací zahrnuje automatické parsování a propojování citací. Bývá standardně zavedena ve větších digitálních knihovnách [14].

3.9. Annotation component implementation (Implementace komponenty pro správu anotací)

Hlavní náplní je navrhnout a implementovat službu pro správu anotací. Těmi zde rozumíme komentáře, recenze, diskuzní vlákna a další příspěvky uživatelů [15].

3.10. Accessibility component implementation (Implementace komponenty pro správu přístupu)

Stará se o vytvoření souboru nástrojů, které umožní přístup k matematickým dokumentům skupinám uživatelů se specifickými požadavky, např. uživatelům s poškozením zraku nebo poruchou učení, dyslexií a dalšími. V této funkcionalitě je projekt EuDML výjimečný v rámci digitálních knihoven a repozitářů s vědeckou literaturou. Zahrnuje porovnání existujících nástrojů a jejich implementaci do systému. Hlavně je zaměřeno na čtečky obrazovky, syntetizátory řeči a automatické překladače textu [16]. Bylo přihlédnuto ke zkušenostem Masarykovy univerzity získaných s vedením centra Teiresiás [17].

3.11. Assessment and evaluation (Vyhodnocení)

Cílem je vyhodnocení projektu, prací na něm, funkcionality[18].

4. Partneri

Do projektu EuDML je zapojena celá řada partnerských organizací[1]:

Instituto Superior Técnico: Computer Science Department (Lisabon, Portugalsko)

Université Joseph-Fourier (Grenoble, Francie)

University of Birmingham (Velká Británie)

FIZ: Fachinformationszentrum (Karlsruhe, Německo)

Masarykova univerzita (Brno, Česká Republika)

Uniwersytet Warszawski, ICM (Varšava, Polsko)

5. Zkušenosti autora

5.1. EuDML Beta

V červenci 2011 byla spuštěna beta verze webového portálu EuDML[19] s možností využívat základní funkce systému. Implicitně, nehladě na nastavení prohlížeče, jsou stránky v angličtině, ale je možné si dále vybrat ze španělštiny, němčiny, polštiny, češtiny a francouzštiny. Návštěvníkovi umožňuje vyhledávat díla podle zadaných klíčových slov nebo využít pokročilého vyhledávání, při kterém lze aplikovat množství filtrů (jméno autora, jazyk, matematickou formuli atd.). Funkce procházení dokumentů podle předmětu nebo jednotlivých sesbíraných kolekcí přístupné neregistrovaným uživatelům ještě nejsou implementovány.

Záznam vybrané nalezené práce je podobný knihovnímu štítku, obsahuje informace o autorovi (případně autorech), publikaci, roku vydání. Dále abstrakt, generovanou citaci, seznam citované literatury v práci a komentáře uživatelů. Neobsahuje náhled díla, ale odkaz do repozitáře, kde je uloženo a odkud byla jeho metadata sesbírána. Po registraci vám bude vytvořen uživatelský účet, ke kterému se přihlásíte pod zadaným e-mailem a heslem. Další možností je využít služby OpenID[20]. Registrovaný uživatel získá vlastní knihovničku, do které si může ukládat jednotlivé svazky. Tato funkcionalita však ještě není implementována. Jedinou jeho výhodou prozatím zůstává zasílání komentářů.

5.2. Zhodnocení

Projekt EuDML si za svůj cíl klade zmapování a zpřístupnění matematických prací z evropských repozitářů. To považují za poměrně vysoký cíl, ale určitě ne nereálný. Sjednocení různých schémat záznamů o publikacích, vypořádání se s podmínkami specifickými pro matematickou literaturu (zápis a formát matematických vzorců a rovnic). To jsou nejdny z problému, které se muselo překonat. Avšak beta verze webového portálu ukazuje, že se vydali správným směrem a po dvou letech vývoje projektu vypadá lze očekávat rok třetí s nadějí.

Projekt evropské matematické digitální knihovny mi přijde jako rozšíření snahy jednotlivých institucí shromažďovat práce do svých repozitářů a nemusí zůstat jen u toho. Myšlenka dalšího rozšíření z evropské na celosvětovou nemusí být daleko.

6. Seznam použité literatury

- [1] EuDML [online]. 2011 [cit. 2011-12-11]. Dostupné z WWW: <<http://www.eudml.eu/>>
- [2] M. Bartošek. Česká digitální matematická knihovna. Zpravodaj ÚVT MU. ISSN 1212-0901, 2010, roč. XX, č. 3, s. 11-13
- [3] BORBINHA, José; BOUCHE, Thierry; NOWIŃSKI, Aleksander; SOJKA, Petr. Intelligent Computer Mathematics. Heidelberg : Springer, 2011. Project EuDML – A First Year Demonstration, s. 281 – 284. Dostupné z WWW: <http://dx.doi.org/10.1007/978-3-642-22673-1_21>. ISBN 978-3-642-22672-4, Doi: 10.1007/978-3-642-22673-1_21
- [4] BORBINHA, José. First Year Report[online]. <https://www.eudml.eu/sites/default/files/D1.3.1-1.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D1.3.1-1.pdf>>
- [5] BORBINHA, José; MACÍAS, Enrique; BOUCHE, Thierry; CAGO, Felipe; MARTÍNS, Bruno. Dissemination Plan[online]. <https://www.eudml.eu/sites/default/files/D2.1.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D2.1.pdf>>
- [6] JOST, Michael; BOUCHE, Thierry; GOUTORBE, Claude; JORDA, Jean-Paul. The EuDML metadata schema[online]. <https://www.eudml.eu/sites/default/files/D3.2-v1.6-1.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D3.2-v1.6-1.pdf>>
- [7] BORBINHA, José; NOWIŃSKI, Aleksander; SYLWESTRZAK, Wojtek; PEDROSA, Gilberto. EuDML global system functional specification[online]. <https://www.eudml.eu/sites/default/files/D4.1-v1-1.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D4.1-v1-1.pdf>>
- [8] YADDA [online]. 2011 [cit. 2011-12-11]. Dostupné z WWW: <<http://yaddainfo.icm.edu.pl/>>
- [9] REPOX [online]. 2011 [cit. 2011-12-11]. Dostupné z WWW: <<http://reporx.ist.utl.pt/>>
- [10] PEDROSA, Gilberto; BORBINHA, José. The EuDML Metadata Registry and Repository[online]. <https://www.eudml.eu/sites/default/files/D5.1-v1.0.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D5.1-v1.0.pdf>>
- [11] NOWIŃSKI, Aleksander; ROSIEK, Tomasz; POLITOWSKI, Michał. The EuDML Search and Browsing Service[online]. https://www.eudml.eu/sites/default/files/D5.2_The_EuDML_Search_and_Browsing_Service-v1.1.pdf [cit. 2011-12-11]. Dostupný z WWW: <https://www.eudml.eu/sites/default/files/D5.2_The_EuDML_Search_and_Browsing_Service-v1.1.pdf>
- [12] JAMES, Mark; GORMAN, Vanessa. User Interface Design[online]. <https://www.eudml.eu/sites/default/files/D6.2.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D6.2.pdf>>
- [13] SOJKA, Petr; BAKER, Josef; SEXTON, Alan; SORGE, Volker. State of the Art of Augmenting Metadata Techniques and Technology[online]. <https://www.eudml.eu/sites/default/files/D7.1-v1.2.pdf> [cit. 2011-12-11]. Dostupný z WWW: <<https://www.eudml.eu/sites/default/files/D7.1-v1.2.pdf>>
- [14] LEE, Mark; SOJKA, Petr; SORGE, Volker; BAKER, Josef. Association Analyzer Implementation: State of the Art[online]. https://www.eudml.eu/sites/default/files/D8.1_0.pdf [cit. 2011-12-11]. Dostupný z WWW: <https://www.eudml.eu/sites/default/files/D8.1_0.pdf>

- [15] JAMES, Mark; GORMAN, Vanessa. Annotation Component Design[online]. https://www.eudml.eu/sites/default/files/D9.1_0.pdf [cit. 2011-12-11]. Dostupný z WWW: <https://www.eudml.eu/sites/default/files/D9.1_0.pdf>
- [16] SORGE, Volker; LEE, Mark; SOJKA, Petr; SEXTON, Alan. State of the Art of Accessibility Tools[online]. https://www.eudml.eu/sites/default/files/D10.1_0.pdf [cit. 2011-12-11]. Dostupný z WWW: <https://www.eudml.eu/sites/default/files/D10.1_0.pdf>
- [17] Středisko Teiresiás [online]. 2011 [cit. 2011-12-11]. Dostupné z WWW: <<http://teiresias.muni.cz/>>
- [18] PAVLOV, Radoslav; PELIKAN, Peter. Report on evaluation of existing services and content providers[online]. https://www.eudml.eu/sites/default/files/D11.1_0.pdf [cit. 2011-12-11]. Dostupný z WWW: <https://www.eudml.eu/sites/default/files/D11.1_0.pdf>
- [19] EuDML [online]. 2011 [cit. 2011-12-11]. Dostupné z WWW: <<http://demo.eudml.eu/beta/home.action>>
- [20] OpenId Foundation website [online]. 2006 [cit. 2011-12-11]. Dostupné z WWW: <<http://openid.net/>>

7. Metadata v Dublin Core

TITLE=EuDML

CREATOR=Aleš Nevřela

DESCRIPTION= Studentská práce o projektu EuDML (European Digital Mathematics Library – Evropská digitální matematická knihovna). Po základních informacích rozděluje projekt do 11 pracovních balíčků (workpackages), ke kterým podává podrobnější informace. Na závěr mapuje aktuální situaci a autorovu zkušenost s implementací webového portálu EuDML.

SUBJECT.Keywords=EuDML, DML, CIP, workpackage

DATE.Created=2011-12-11

LANGUAGE=cz

FORMAT.Medium=application/pdf