Konverze metadat z formátu Marc21 do Dublin Core v prostředí univerzitního repozitáře

Irena Baranayová

Abstrakt
Na Univerzitě Karlově v Praze se buduje Digitální repozitář eVŠKP v systému DigiTool od roku 2006. Vysokoškolské kvalifikační práce jsou v současné chvíli popisovány pomocí bibliografického formátu Marc21. Bibliografické záznamy jsou do Digitálního repozitáře importovány z Centrálního katalogu Univerzity Karlovy. Formát Marc21 ovšem není vyhovující pro potřeby projektu univerzitního repozitáře a proto je nutná konverze záznamů do metadatového formátu Dublin Core. V příspěvku budou prezentovány první zkušenosti s řešením konverze mezi bibliografickým formátem Marc21 a metadatovým popisným formátem Dublin Core v prostředí univerzitního repozitáře.

Projekt Digitálního univerzitního repozitáře na Univerzitě Karlově v Praze byl spuštěn v roce 2006 v systému DigiTool izraelské firmy ExLibris. Repozitář byl zprvu naplňován pouze elektronickými verzemi kvalifikačních prací, časem byl rozšířen o historické listiny Archívu Univerzity Karlovy, historické mapy Přírodovědecké fakulty, preprinty a příspěvky ze sborníků zaměstnanců Matematicko-fyzikální fakulty a 2. lékařské fakulty, studijní a výukové materiály, technické a výzumné zprávy, výroční zprávy a grantové zprávy. Plánuje se rozšíření i o multimediální materiály. Technicky a metodicky chod Univerzitního repozitáře zajišťuje Ústav výpočetní techniky Univerzity Karlovy, jednotlivé fakultní knihovny zajišťují sběr dat a jejich vkládání. Přístup do Univerzitního repozitáře je omezen na IP adresy Univerzity Karlovy, vzdálený přístup je zajištěn pro studenty a zaměstnance Univerzity Karlovy pomocí služby EZProxy. Obecné informace o projektu jsou dostupné na http://digitool.cuni.cz, samotný Univerzitní repozitář je na http://repozitar.cuni.cz
Elektronické verze vysokoškolských kvalifikačních prací v současné době dodává 11 fakult Univerzity Karlovy. V současné době je v Repozitáři skoro 2000 kvalifikačních prací. Pro autory prací je připravená šablona ve formátu DOC a TeX (na http://evskp.cuni.cz/EVSKP-18.html), pro potřeby autorů i knihovníků je na http://digitool.cuni.cz/DIGITOOL-44.html vypracován seznam volně dostupných PDF konvertorů, které převádí práce do PDF formátu kompatibilního se systémem DigiTool.
Sběr prací probíhá buď na základě na základě nařízení děkana (Fakulta sociálních věd, 2. lékařská fakulta) nebo na bázi dobrovolnosti (Přírodovědecká fakulta). Studenti odevzdávají dvě tištěné kopie a jednu elektronickou. Jedna tištěná práce jde do Archívu Univerzity Karlovy, druhá spolu s elektronickou verzí (obvykle na CD-ROM) jsou odevzdány do příslušné fakultní knihovny, kde proběhne zpracování do Centrálního katalogu Univerzity Karlovy (systém Aleph 500, v. 18) a do Univerzitního repozitáře.
Projekt konverze bibliografických záznamů z Centrálního katalogu do Univerzitního repozitáře vznikl jako nouzové řešení. Plánované propojení Informačního systému s Univerzitním repozitářem (http://evskp.cuni.cz/EVSKP-6.html), kdy by práce byly vkládány a popisovány samotnými autory a do Univerzitního repozitáře by se dostaly po obhajobě práce a zkontrolování správnosti bibliografických dat knihovníkem, se z finančních a z důvodů změn priorit na straně Informačnáho systému nepovedlo zrealizovat. Knihovníci se tak dostali do nepříjemné situace, kdy zpracovávali tištěné práce do Centrálního katalogu a zároveň do Univerzitního repozitáře. Logickým řešením situace bylo propojení katalogu a repozitáře, aby knihovníkům odpadla zbytečná duplicitní práce. Workflow zpracování kvalifikačních prací je následující – do knihovny jsou z jednotlivých

katederních pracovišť dodány kvalifikační práce v tištěné a elektronické podobě. Nejprve je zpracována tištěná verze do Centrálního katalogu, poté je elektronická verze vložena do Univerzitního repozitáře, kde se k ní vyrobí technická metadata, plnotextový index a náhled titulní strany a z Centrálního katalogu je pomocí Z39.50 k objektu připojen bibliografický záznam. Tím odpadla duplicitní výroba popisných metadat.
Bibliografické záznamy v Centrálním katalogu jsou vytvářeny ve formátu Marc21, v Univerzitním repozitáři je pro bibliografický popis užit metadatový formát Dublin Core, resp. formát EVSKP-MS verze 1.0. Pomocí Z39.50 jsou bibliografické záznamy připojeny k objektům v Marcu21. To vedlo k určitým problémům při zařazování prací do jednotlivých sbírek. Navíc je část prací popsána ve formátu EVSKP-MS a část v Marcu21. Je nutné podobu záznamů sjednotit, a to do formátu EVSKP-MS doporučeném Odbornou komisí pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací.
Prvním předpokladem pro vytvoření konverze bylo vypracování mapování mezi Marcem21 a Dublin Core resp. formátem EVSKP-MS. Národní knihovna má vypracované mapování pro konverze mezi Marcem21 a Dublin Corem, které se také stalo základem pro konverze dat mezi Centrálním katalogem a Univerzitním repozitářem. Nicméně bylo potřeba mapování přizpůsobit dle formátu EVSKP-MS a dle interních požadavků systému DigiTool. Po srovnání vyplynula nutnost přidat do bibliografických záznamů v Marcu 21 dvě nová pole - pole pro Formát eVŠKP (v Dublin Core se jedná o pole dc:format) a pole Kód logické sbírky (v Dublin Core se jedná o pole dc:collection). Dále bylo nutné obohatit některá marcová pole jako např. pole 520 pro abstrakty – formát EVSKP-MS vyžaduje atribut jazyka, který se v Marcu neuvádí. Také se ukázalo, že některá pole nelze pro potřeby konverze využít. Typickým příkladem je marcové pole 710, kdy kvůli propojení na Národní autority obvykle chybí údaj o fakultě. Proto se údaj pro dcterms:grantor přebírá z pole 502.
Velkým úskalím celého projektu je velice rozdílná kvalita záznamů kvalifikačních prací v Centrálním katalogu. Záznamy monografií a seriálů jsou kontrolovány také v souvislosti s přispíváním do Souborného katalogu ČR, na záznamy kvalifikačních prací se kontroly dosud uplatňovány nebyly. V souvislosti s konverzí je plánováno i pravidelné kontrolování těchto typů záznamů.
Knihovníci byli upozorněni na nová pole, která je nutné vyplňovat, aby konverze proběhla bezproblémově. Pečlivé vyplnění těchto údajů ve výsledku zajistí konrektní výsledek konverze. Pro knihovníky byl taktéž vypracován manuál s povinnými poli a příklady vyplňování. Ten je i součástí nápověd v systému Aleph.
Nejnovější marcovské záznamy, připojené do Digitálního univerzitního repozitáře přes Z39.50, budou v pravidelných časových intervalech exportovány do speciálního NFS adresáře. Obsah tohoto adresáře (před noční zálohou systému) projde konverzní nástrojem (s mapováním popisných polí Marc21 do Dublin Core) a duplicitní marcovské záznamy budou odstraněny.
Výhledově bude nutné odladit konverzní nástroj pro vyšší efektivitu výsledků. Dalším krokem je synchronizace záznamů mezi Centrálním katalogem a Univerzitním repozitářem pomocí technologie OAI tak, aby se záznamy z Centrálního katalogu automaticky promítly do Univerzitního repozitáře. Pokud bude práce pouze v elektronické podobě, bude probíhat konverze z Univerzitního repozitáře do Centrálního katalogu.
Testování konverzí bude probíhat na pracích odevzdaných v zimním semestru. Cílem je, aby bibliografické záznamy prací odevzdané v letním semestru 2009 již byly do Univerzitního repozitáře importovány automaticky bez nutnosti zásahů ze stran knihovníků.