Propojení IS VUT s knihovními systémy Aleph a DSpace

Tomáš Kreuzwieser, Jan Skůpa, Antonín Vaishar, Martin Fasura

Abstrakt

Příspěvek se zabývá problematikou zpřístupňování závěrečných prací (VŠKP) na VUT v Brně. Závěrečné práce jsou studenty odevzdávány do vlastního informačního systému. Zde jsou studenty vyplněny metadatové údaje a přiložen plný text práce. Po obhajobě a vložení posudků jsou metadata a plný text práce exportovány do institucionálního repozitáře Digitální knihovna VUT, který je provozovaný systémem DSpace. Data z IS jsou dále využívána pro zpřístupňování tištěných verzí VŠKP v knihovnách VUT, kdy jsou metadata importována do knihovního systému Aleph.

Klíčová slova

knihovní systémy, Aleph, DSpace, integrace systémů, VŠKP, VUT v Brně, DSpace REST API

1. Úvod

Ukládání VŠKP začala Ústřední knihovna VUT (ÚK) řešit v roce 2007, kdy vešla v platnost Směrnice rektora č.9/2007 – „Úprava, odevzdávání a zveřejňování vysokoškolských kvalifikačních prací na VUT v Brně“. Touto směrnicí VUT v Brně reagovalo na požadavky vzniklé schválením novely zákona č. 111/1998 sb. (Zákon o vysokých školách, novela zákona č. 552/2005) a zavedlo tímto povinné odevzdávání elektronických závěrečných prací. Jako systém pro ukládání plných textů byl zvolen systém DigiTool od společnosti Ex Libris, která vyvinula i knihovní systém Aleph a nabízela se zde možná jednoduchá propojení a také technická podpora. Za účelem pořízení a zprovoznění serveru a systému DigiTool byl podán projekt ve Fondu pro rozvoj vysokých škol. Tento systém byl v provozu v letech 2008-2010. V roce 2011 došlo s ohledem na dosavadní průběh a také zastavení vývoje systému ze strany Ex Libris k přehodnocení situace a výběru nového systému - DSpace. Na jeho základě je nyní provozován institucionální repozitář Digitální knihovna VUT dostupný z adresy https://dspace.vutbr.cz.

2. Vysokoškolské závěrečné práce na VUT v Brně

Objemem největší sbírkou repozitáře jsou vysokoškolské závěrečné práce (VŠKP). Do této sbírky se dostávají plné texty z devíti pracovišť - 8 fakult a

jednoho vysokoškolského ústavu. Dříve probíhalo zveřejňování plných textů ještě před obhajobou a tento proces tak splňoval zákonem dané povinnosti. V roce 2018 vstoupila v platnost směrnice, která změnila dosavadní praxi a nařizuje zveřejňovat plné texty VŠKP v elektronické podobě až po obhajobě (ať úspěšné či neúspěšné). Tato změna však způsobila, že musí být vyvinut tlak na větší konzistenci dat v IS (např. nevyplněné datum obhajoby v systému nyní zabraňuje faktickému zveřejnění).

O propojení IS a DSpace se stará vlastní synchronizační program využívající DSpace REST API.

3. Vznik záznamu v Informačním systému

Na VUT v Brně se VŠKP práce odevzdávají v rozhraní informačního systému. Metadata o VŠKP jsou uložena v centrální databázi CDB a soubory prací jsou z důvodu efektivnějšího zálohování databáze uloženy na sdíleném file systému. Po odevzdání práce studentem je práci přiřazeno unikátní ID, vloženy posudky oponentů a vedoucím práce potvrdí správnost údajů. Po obhajobě je VŠKP synchronizována do Digitální knihovny VUT.

4. Export z IS do repozitáře

4.1 Synchronizace dat

Proces synchronizace je spouštěn každých 60 minut ze serveru, na kterém je provozován systém DSpace a to tak, že synchronizace je vždy kompletní. Záznam nebo jiný objekt v Digitální knihovně je aktualizován při změně některého z kontrolovaných atributů nebo při vzniku zcela nového záznamu v centrální databázi.

Na straně CDB je zdrojem dat pro synchronizaci JSON struktura REST rozhraní IS. Tato JSON struktura je plněna SQL dotazem. O online tvorbu těchto datových struktur na straně informačního systému VUT se stará kód exportního rozhraní pro komunikaci s externími systémy. Na straně Digitální knihovny je postupným voláním metody GET REST rozhraní Digitální knihovny v paměti vytvořena stromová struktura a tyto dvě velké datové struktury jsou následně vlastním synchronizačním skriptem (napsaným v programovacím jazyku Perl) porovnány. Případné nalezené rozdíly mezi daty jsou do Digitální knihovny zapracovány opět pomocí volání REST rozhraní systému DSpace. Automaticky jsou tak prováděny operace od zakládání komunit, kolekcí, záznamů a bitstreamů po nejrůznější přesuny, aktualizace až po mazání.

Při první synchronizaci záznamu VŠKP do Digitální knihovny je pro záznam vygenerován unikátní handle a tento handle je vrácen informačnímu systému,

takže je možné se z informačního systému do Digitální knihovny prokliknout a také vzniká pevná vazba mezi interním identifikátorem CDB a unikátním handle.

Pro zjednodušení ladění případných chyb je synchronizační skript rozdělen na dvě části, kde uprostřed těchto dvou skriptů je diskový repozitář. První částí synchronizace je synchronizace mezi databází informačního systému CDB a mezi diskovým repozitářem a druhá je mezi diskovým repozitářem a mezi systémem DSpace. Diskový repozitář je kompatibilní se standardem DSpace Simple Archive Format, takže je pro účely ladění možné jednu část skriptu nahradit pomocí příkazu DSpace „dspace import", případně „dspace export". Diskový repozitář zároveň slouží jako cache dat pro DSpace. REST rozhraní DSpace totiž neumožňuje rychlé získávání dat o všech uložených záznamech a cache tak umožňuje stahovat plná data pouze při změnách některého z časů modifikace v DSpace.

4.2 Transformace metadat

V průběhu načítání dat z informačního systému jsou prováděny transformace, které modifikují data tak, aby strukturou odpovídaly aktuálnímu metadatovému schématu DSpace (systém DSpace ukládá metadatové záznamy ve formátu Dublin Core). Ve snaze oddělit transformaci dat od řízení synchronizace jsme se rozhodli řešit tvorbu metadatového záznamu v Dublin Core prostřednictvím XSL transformace.

Metadata ve formátu JSON, která vrací REST rozhraní IS, jsou pro účely transformace jednoduše převedena do XML. Šablona, kromě prostého mapování polí, provádí různé úpravy dat. Dochází například k normalizaci jmen, generování bibliografické citace a rozdělení klíčových slov, která se ve zdrojových metadatech nacházejí v jednom poli.

Kromě generování metadat ve formátu Dublin Core slouží XSLT šablona ještě k dalšímu účelu. Formát Dublin Core je jednorozměrný, nestrukturovaný, neumožnuje zachycení vztahů mezi dílčími entitami. To se ukazuje jako problém ve vztahu k posudkům práce, kterých může být několik a je vhodné u nich zachovat informaci o autorovi, aby bylo jasné, kdo je autorem jakého posudku.

Tento problém řešíme (ne zrovna elegantně) právě pomocí XSL transformace, kdy ze strukturovaných zdrojových metadat generujeme HTML soubory s posudky, ve kterých je zachován vztah autor-posudek a ty jsou pak nahrány do DSpace jako přílohy práce. Tato transformace je součástí stejné šablony, která generuje metadata. Jedna XSL transformace tak má dva výstupy: XML soubor s Dublin Core záznamem a HTML soubor s posudky. Oba soubory jsou pak v rámci Simple Archive Format balíčku nahrány do DSpace.

4.3 Omezení způsobu synchronizace

REST rozhraní DSpace je pro synchronizace využitelné, ale je zde několik úskalí, se kterými je třeba počítat. Především je velmi časově náročné vyčíst aktuální stav ve velkém objemu dat z DSpace. Tento problém je v současné době řešen využitím

právě výše zmíněné diskové cache. Dále je pomocí REST rozhraní nemožné změnit primární kolekci záznamu a zcela chybí operace pro změnu handle v DSpace, což v našem případě nastává, pokud záznam byl přidán do DSpace, následně má být smazán kvůli například změně v licenci a pak má být znovu přidán.

Z tohoto musí být změny „owning collection“ a změny handle realizovány přímo pomocí SQL v databázi DSpace.

Do budoucna počítáme s nahrazením GET metody REST rozhraní DSpace pomocí řádově rychlejšího SQL přístupu do databáze a zrušení diskové cache. Po odladění provozem si už dovolíme sloučení dvou částí skriptu do jednoho, čímž se proces synchronizace zjednoduší. Pro vytváření a modifikace existujících dat je REST rozhraní plně vyhovující.

5. Export z IS do knihovního systému

Kromě propojení IS školy se systémem DSpace realizuje VUT také poměrně analogické propojení IS s knihovním systémem Aleph. Mimo elektronických verzí VŠKP existují pochopitelně také verze tištěné, které se v některých případech dostávají do knihoven. Z tohoto důvodu není realizovaná úplná synchronizace, jsou nahrávány pouze ty práce, o jejichž nahrání si knihovny explicitně požádají. Dalším rozdílem, který výrazně snižuje náročnost implementace, je to, že se záznamy nahrávají pouze jednorázově. Je to hlavně z toho důvodu, že knihovní katalogizátoři v záznamech provádějí ručně změny a není žádoucí, aby byly tyto změny při periodické synchronizaci ztraceny.

5.1 Požadavky na nahrání záznamu

Sběr požadavků knihoven na nahrání práce probíhá také online. Knihovníci zadávají požadavky přes formulář na intranetovém portálu knihoven VUT. Požadovaným parametrem je ID práce, případně soubor s ID prací, je-li jich větší počet. Takto zadaný požadavek se propaguje do systému activeCollab (systém pro zprávu projektů a požadavků provozovaný ÚK VUT). Z požadavku vzniklý úkol pak obslouží systémový knihovník, který provede nahrání práce. Buď prostřednictvím služby v GUI klientu knihovního systému, nebo přímo z příkazové řádky na serveru (hlavně při objemnějších požadavcích).

5.2 Nahrání záznamu do knihovního systému

Vlastní nahrání provádí perlový skript, který volá příslušné rozhraní knihovního systému (tzv. X server), o tvorbu metadat se stará opět XSL transformace. Zdrojem je stejné rozhraní IS jako v případě integrace DSpace, požadovaným výsledkem jsou zde metadata ve formátu OAI MARC, což je jeden z možných XML zápisů formátu bibliografických záznamů MARC21. Šablona sdílí s tou pro DSpace část provádějící normalizaci jmen, naopak chybí rozdělení klíčových slov, věcné

zpracování je jeden úkolů, který zůstává na fakultním katalogizátorovi. Ten může nahraná klíčová slova zpracovat (tzn. rozdělit do jednotlivých polí), nebo je může zahodit a provést věcné zpracování sám. Další metadata, která musí ručně doplnit katalogizátor, souvisí s fyzickým popisem tištěné verze a v IS se nenacházejí, např. počet stran apod.

5.3 Obohacení záznamu o handle

Záznam v knihovním systému je rovněž obohacen o handle elektronické verze práce. I když je tento identifikátor záhy po vygenerování vrácen do IS a stává se součástí metadat, je pro účely obohacení bibliografického záznamu získáván jinak – přímo od zdroje. Handle se získává ze SOLR indexu, který je součástí DSpace a má vhodné API rozhraní, které na dotaz obsahující ID práce vrátí právě handle. Ve výsledku je pak pomocí handle prolinkován nejen IS a DSpace, ale také DSpace a Aleph.

6. Primo

Záznamy zveřejněné v Digitální knihovně jsou indexovány discovery systémem Primo (dostupný z adresy primo.vutbr.cz). Ten prohledává i dostupné elektronické zdroje (databáze předplácené univerzitou nebo volně dostupné) a také knihovní katalog VUT. Zde musí docházet k deduplikaci záznamů, protože jedna závěrečná práce může být jak elektronicky v Digitální knihovně, tak jako fyzický výtisk v jedné z fakultních knihoven. Deduplikace je nastavena na dva základní parametry - název a na jednoznačný identifikátor práce, který obsahují oba záznamy - Dublin Core v Digitální knihovně a MARC v knihovním katalogu.

7. Závěr

I přes dílčí komplikace se propojení IS jak s Digitální knihovnou, tak s knihovním systémem osvědčilo. V průběhu času jsme tento mechanismus, kterým probíhá synchronizace IS s DSpace v případě VŠKP, v mírně upravené podobě nasadili také na problematiku zveřejňování odborných děl (článků, příspěvků z konferencí aj.) vědeckých a akademických pracovníků VUT. V tomto případě autor přikládá plný text k vykazovanému výsledku (ze zákonné povinnosti a RIV) a sám žádá o vložení plného textu do repozitáře. Každý plný text je posuzován administrátorem z důvodu autorsko-právních aspektů.

Literatura

DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description, Dublin Core Metadata Initiative [online]. *Dublin Core Metadata Initiative*, 14.6.2012 [cit. 2018-04-24]. Dostupné z: http://dublincore.org/documents/dces/

LAGOZE, Carl, Herbert VAN DE SOMPEL, Michael NELSON a Simeon WARNER, Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: An XML Schema to represent MARC records. *Open Archives Initiative* [online]. Ithaca (New York): Open Archives Initiative [cit. 2018-04-24]. Dostupné z: https://www.openarchives.org/OAI/2.0/guidelines-oai_marc.htm

DONOHUE, Tim, Updating Items via Simple Archive Format. *Dashboard: DuraSpace Wiki* [online]. 17.3.2015 [cit. 2018-04-24]. Dostupné z: https://wiki.duraspace.org/display/DSDOC6x/Updating+Items+via+Simple+Archive+Format

Kontakt

Ing. Tomáš Kreuzwieser, VUT v Brně, Centrum výpočetních a informačních systémů, Antonínská 548/1, kreuzwieser@vutbr.cz

Jan Skůpa, VUT v Brně, Ústřední knihovna, Antonínská 548/1, skupa@lib.vutbr.cz

Mgr. Antonín Vaishar, DiS., VUT v Brně, Ústřední knihovna, Antonínská 548/1, vaishar@lib.vutbr.cz

Ing. Martin Fasura, VUT v Brně, Ústřední knihovna, Antonínská 548/1, fasura@lib.vutbr.cz