Přenos VŠKP pomocí protokolu OAI-PMH

Jan Mach

machj@vse.cz

Abstrakt

Příspěvek prezentuje zkušenosti s implementací přenosu metadat EVSKP-MS verze 1.1 z VŠE do NR VŠKP pomocí standardu OAI-PMH. Na příkladech je ukázán princip harvestování vybraných metadat z lokálního registru (VŠE) do registru centrálního (NR VŠKP), následuje volitelné stažení plných textů, jejichž URL adresy jsou uvedeny v metadatech. Mechanismus přenosu metadat je možné, díky použitému mezinárodnímu standardu, využít i pro potřeby Národního úložiště šedé literatury a jiných registrů, např. na bázi DSpace. Potřebné skripty byly na VŠE naprogramovány v jazyce PHP.

Klíčová slova

OAI-PMH, Elektronické vysokoškolské kvalifikační práce, Metadata, Harvestování, Vysoká škola ekonomická v Praze, NR VŠKP

1. OAI-PMH – The Open Archives Initiative Protocol for Metadata Harvesting

Petr Žabička [1] definuje protocol OAI-PMH jako „jednoduchý protokol, umožňující „poskytovatelům služeb“ automatické získávání (sklízení) metadat nabízených „poskytovateli dat“. Díky těmto vlastnostem a díky mnoha volně dostupným softwarovým komponentám je protokol možné snadno implementovat do stávajících systémů digitálních knihoven.“

Vědomi si výhod protokolu OAI-PMH, v rámci své práce v Odborné komisi pro otázky elektronického zpřístupňování vysokoškolských kvalifikačních prací AKVŠ (http://www.evskp.cz/, dále jen Komise eVŠKP) doporučují členové komise tento formát pro výměnu metadat mezi na straně jedné „poskytovatelem dat“ – lokálním registrem vysokoškolských kvalifikačních prací (VŠKP) na jednotlivých školách, a na straně druhé „poskytovatelem služeb“ – Národním registrem VŠKP (http://www.theses.cz; dále též NR VŠKP).

Při implementaci protokolu se vycházelo z oficiální definice protokolu OAI-PMH verze 2.0 [4]. Tento dokument definuje tyto pro nás významné pojmy:

- **Harvester**
  - Národní registr VŠKP je systémem sbírajícím metadata, na jejich základě plné texty a poskytuje s těmito daty další služby
- **Repozitář**
  - Repozitář je definován jako primární zdroj metadat, v našem příkladě se jedná o Databázi VŠKP VŠE.
- **Jednotka**
  - Jednotkou přenosu je kvalifikační práce obhájená na VŠE, která je popsána metadaty. Jednotka může být popsána různými formáty metadat.
- **Unikátní identifikátor**
  - Každý jednotka je jednoznačně identifikován svým unikátním identifikátorem. Tento identifikátor je odlišný od identifikátoru dc:identifier, který je obsahem

metadat. Příkladem unikátního repozitáře v OAI-PMH na VŠE je řetězec *oai:vse.cz:vskp/4367*.

- **Záznam**
  - Metadatový záznam VŠKP popisuje jednotku přenosu – kvalifikační práci obhájenou na VŠE. Na VŠE se používá formát EVSKP-MS ve verzi 1.1 a formát kvalifikovaný Dublin Core, který je v rámci OAI-PMH povinnost implementovat. Protokol OAI-PMH neřeší přenos plných textů. Řešení přenosu plných textů, na základě poskytnutých metadat, je popsán v kapitole xxxxx níže.
- **Sada**
  - Repositář může být rozdělen do jednotlivých sad, každá sada může být harvestována samostatně. Slouží tedy jako určitý filtr nad repositářem. Repositář může obsahovat členění dat podle fakult či, jako je tomu na VŠE, podle typu prací. Na VŠE je definována sada *100* pro *Vysokoškolské kvalifikační práce*.
- **ResumptionToken**
  - V případě velmi dlouhých seznamů je možné použít tzv. ResumptionToken, unikátní řetězec, který repozitář zašle na konci částečného seznamu metadat a harvester může, za použití tohoto řetězce, požádat v následujícím dotazu o pokračování požadovaného seznamu metadat. Použití ResumptionTokenu je volitelné, na VŠE nebylo využito a veškeré seznamy se proto zasílají v celku jako jeden soubor.

2. Výchozí situace

Vysoká škola ekonomická měla na začátku roku 2008 již několik let fungující Databázi VŠKP (http://www.vse.cz/vskp) na bázi PHP a MySQL s několika tisíci odevzdaných kvalifikačních prací v elektronické podobě. Databáze evidovala metadata VŠKP podle vývojové verze 0.1 formátu EVSKP-MS [2]. Komise eVŠKP dne 15. 7. 2008 publikovala oficiální verzi 1.1 EVSKP-MS [3], která byla přijata jako importní formát pro NR VŠKP a připravované Národní úložiště šedé literatury (http://nusl.stk.cz). Databáze VŠKP na VŠE byla v průběhu roku napojena na Integrovaný studijní informační systém VŠE (http://isis.vse.cz)
Vzhledem k tomu, že VŠE je členem Centralizovaného projektu MŠMT C1/2008 „*Národní registr VŠKP a systém odhalování plagiátů*“ a projektu "*Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace*" MK ČR, rozhodl se řešitel VŠE pro rozšíření aplikace Databáze VŠKP o export metadat ve formátu EVSKP-MS verze 1.1 pomocí protokolu OAI-PMH verze 2.0. Toto řešení díky své standardizaci lze použít nejen pro NR VŠKP, ale i pro registry/archivy další.

3. Zadání implementace OAI-PMH

Pro implementaci OAI-PMH na Vysoké škole ekonomické v Praze bylo zvoleno prostředí PHP pro generování WWW stránek s XML záznamy repozitáře VŠE a databáze MySQL, která obsahuje data o VŠKP VŠE.
Požadavky a odpovědi OAI-PMH je možné rozdělit na statické, kde odpovědí je převážně neměnný soubor XML, a dotazy dynamické, kde obsahem odpovědí jsou metadata generovaná převážně z databáze MySQL. Z povahy věci je zřejmé, že implementace neměnných odpovědí na statické dotazy je jednodušší než generování metadat EVSKP-MS.
Základní URL adresa sloužící pro dotazování repozitáře VŠE je http://www.vse.cz/oai, jednotlivé typy dotazů a parametry jsou podle protokolu OAI-PMH uváděny v parametrech

této URL adresy. Konkrétní příkaz pro OAI-PMH server je specifikován v parametru *verb* této základní URL adresy.

Odpovědi statické

První kategorii dotazů tvoří identifikace serveru, dotaz na seznam podporovaných formátů metadat a dotaz na seznam datových sad záznamů.

Identify

Příklad: http://www.vse.cz/oai?verb=Identify
Odpověď ve formátu XML obsahuje m.j. jméno repozitáře, e-mail administrátora, indikaci podpory smazaných záznamů v repozitáři (VŠE nepodporuje), formát identifikátoru aj.

List Metadata Formats

Příklad: http://www.vse.cz/oai?verb=ListMetadataFormats
XML záznam obsahuje seznam metadatových formátů – prefix pro označení formátu, použité schéma XML a jmenný prostor formátu. Každý repozitář musí povinně poskytovat metadata ve formátu **nekvalifikovaný Dublin Core** (prefix *oai_dc*), v případě VŠKP se použije **formát EVSKP-MS** (prefix *oai_evskpms*)

List Sets

Příklad: http://www.vse.cz/oai?verb=ListSets
Každý repozitář může interně členit záznamy do datových sad, které jsou označeny číslem. Při požadavcích na metadata je možné specifikovat, o jakou sadu máme konkrétně zájem. Na VŠE se používá sada číslo 100 pro kvalifikační práce. V dalších sadách repozitářů by mohly být např. výroční zprávy, kvalifikační práce rozdělené dle fakult apod.

Odpovědi dynamické

Konkrétní identifikátory poskytuje dotaz List Identifiers, konkrétní metadata potom dotazy List Records a Get Record.

List Identifiers

Příklad: http://www.vse.cz/oai?verb=ListIdentifiers&from=2008-01-01&
until=2008-01-03&metadataPrefix=oai_evskpms
URL adresa obsahuje kromě označení příkazu v parametru verb povinně formát, v jakém budeme chtít metadata získat (v našem příkladu prefix *oai_evskpms* pro formát EVSKP-MS), nepovinně datovou sadu a hlavně časový rozsah, kdy mělo dojít k modifikaci metadat (ve formátu EVSKP-MS uloženo v prvku *evskp:modified*).
Každý záznam v repozitáři v daném standardu změněný v požadovaném období je v XML označen identifikátorem specifickým pro daný OAI-PMH server, např. na VŠE prvek *header*, opakující se pro každý záznam, vypadá např. takto:

```
<header>
   <identifier>oai:vse.cz:vskp/4823</identifier>
   <datestamp>2008-01-02T11:04:55Z</datestamp>
</header>
```

List Records

Příklad: http://www.vse.cz/oai?verb=ListRecords&from=2008-01-01&
until=2008-01-02&metadataPrefix=oai_evskpms

Na rozdíl od požadavku na identifikátory List Identifiers v této variantě získáváme přímo metadatové záznamy v požadovaném standardu, zde EVSKP-MS. Metadatový záznam je uveden jako vnořený prvek v rámci prvku *metadata*, který se opakuje pro každý záznam splňující omezující podmínky v URL.

Get Record

Příklad: http://www.vse.cz/oai/?verb=GetRecord&metadataPrefix=oai_evskpms&identifier=oai:vse.cz:vskp/4840
Pokud máme seznam identifikátorů, zaslaný jako XML v požadavku List Identifiers, můžeme záznamy stahovat po jednom – v URL adrese specifikujeme požadovaný prefix formátu metadat a identifikátor. Námi požadovaný záznam je vložen opět v rámci prvku *metadata*, tentokráte se již na rozdíl od List Records tento prvek neopakuje.
Příklad odpovědi OAI-PMH serveru VŠE na výše uvedený požadavek je uveden v příloze 1.

4. Realizace exportu na VŠE

Základní naprogramování OAI-PMH serveru na VŠE trvalo přibližně 40 člověkohodin včetně studia dokumentace, převedení EVSKP-MS záznamů z verze 0.1 na verzi 1.1 a odladění validity XML. Pro aplikaci bylo využito stávající databáze VŠKP postavené na bázi PHP a MySQL.
Základní adresa OAI/PMH http://www.vse.cz/oai je obsluhována jedním PHP skriptem, který dle typu požadavku využívá šest dalších stránek. Stahování plných textů bylo řešeno pomocí prvku *evskp:transfer*, URL adresa plného textu má povolen přístup z IP adres serveru theses.cz, který stahuje XML záznamy a následně i plné texty.
Funkčnost byla testována pomocí online aplikace Repository Explorer http://re.cs.uct.ac.za/. Získané XML záznamy, konkrétně prvek *evskp:metadata*, byly kontrolovány validátorem na serveru http://validator.nu oproti Relax NG schéma, které je připraveno pro poslední verzi EVSKP-MS standardu.
V současné době je z VŠE naimportováno do NR VŠKP přes protokol OAI-PMH téměř 7000 záznamů ve formátu EVSKP-MS verze 1.1, připravuje se na VŠE import nových záznamů z nové aplikace pro evidenci VŠKP do databáze výše zmiňované, zajišťující popsaným způsobem OAI-PMH export dat do NR VŠKP.

Použité zdroje:

1. Žabička, Petr. *OAI-PMH: Protokol pro metadatovou imperoperabilitu* [online]. 2003 [cit. 2008-10-18] Dostupný na WWW: <http://knihovny.cvut.cz/akp2003/sbornik/05_zabicka.pdf >
2. *EVSKP-MS : Metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR : Návrh* [online]. Zpracovatelé E. Bratková, E. Bulínová, J. Mach. Verze 0.1. 2005-11-07 [cit. 2008-10-19]. Dostupný na WWW: <http://www.evskp.cz/standardy/metadata/navrh-2005-11-07.html>.
3. *EVSKP-MS : Metadatový soubor pro elektronické vysokoškolské kvalifikační práce v ČR* [online]. Zpracovatelé Eva Bratková, Jan Mach. Verze 1.1. Praha : Odborná komise pro otázky elektronického zpřístupňování VŠKP AKVŠ ČR, 2008-07-15 [cit. 2008-10-19]. Dostupný z WWW: <http://www.evskp.cz/standardy/evskp/1.1/>.
4. *The Open Archives Initiative Protocol for Metadata Harvesting* [online], Protocol Version 2.0 of 2002-06-14 [cit. 2008-10-05]. Dostupný na WWW: <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Příloha 1 – ukázka metadatového záznamu pro požadavek Get Record

Příklad URL: http://www.vse.cz/oai/?verb=GetRecord&metadataPrefix=oai_evskpms&identifier=oai:vse.cz:vskp/4840

Odpověď OAI-PMH:


```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2008-12-09T17:57:12Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_evskpms"
identifier="oai:vse.cz:vskp/4840">http://www.vse.cz/oai</request>
- <GetRecord>
- <record>
- <header>
  <identifier>oai:vse.cz:vskp/4840</identifier>
  <datestamp>2008-01-02T11:10:35Z</datestamp>
  </header>
- <metadata>
- <evskp:metadata version="1.1" xmlns:ccz="http://www.evskp.cz/standardy/corpcz/"
xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dctype="http://purl.org/dc/dcmitype/" xmlns:evskp="http://www.evskp.cz/standardy/evskp/"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:pcz="http://www.evskp.cz/standardy/perscz/"
xmlns:thesis="http://www.ndltd.org/standards/metadata/etdms/1.0/">
  <dc:title xml:lang="cs">Finanční deriváty v zobrazení IFRS</dc:title>
- <dc:creator>
- <pcz:person>
- <pcz:name>
  <pcz:foreName>Jaroslava</pcz:foreName>
  <pcz:surName>Remková</pcz:surName>
  </pcz:name>
  <pcz:dateOfBirth>1983-03-28</pcz:dateOfBirth>
  </pcz:person>
  </dc:creator>
  <dcterms:abstract xml:lang="cs">Cílem této diplomové práce je přiblížit stávající účetní zobrazení finančních derivátů, způsoby jejich oceňování a požadavky kladené na jejich zveřejňování a prezentaci stanovené v rámci Mezinárodních standardů finančního výkaznictví platných k 1. 1. 2007.</dcterms:abstract>
- <dc:publisher>
- <ccz:universityOrInstitution>
  <ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
  <ccz:email>webmaster@vse.cz</ccz:email>
  <ccz:homepage>http://www.vse.cz/</ccz:homepage>
  </ccz:universityOrInstitution>
  </dc:publisher>
- <dc:contributor thesis:role="advisor">
- <pcz:person>
- <pcz:name>
  <pcz:foreName>Libor</pcz:foreName>
  <pcz:surName>Vašek</pcz:surName>
  </pcz:name>
  </pcz:person>
  </dc:contributor>
- <dc:contributor thesis:role="referee">
- <pcz:person>
```

```
- <pcz:name>
  <pcz:foreName>Petr</pcz:foreName>
  <pcz:surName>Ryneš</pcz:surName>
  </pcz:name>
  </pcz:person>
  </dc:contributor>
  <dcterms:created>2007-12-20</dcterms:created>
  <dcterms:dateSubmitted>2008-01-02</dcterms:dateSubmitted>
  <dcterms:dateAccepted>2008-02-04</dcterms:dateAccepted>
  <dcterms:modified>2008-02-04T13:57:48Z</dcterms:modified>
  <dc:type xml:lang="cs" evskp:typeType="TypVSKP">Diplomová práce</dc:type>
  <dcterms:medium>application/pdf</dcterms:medium>
  <dc:identifier>http://www.vse.cz/vskp/eid/4840</dc:identifier>
  <dc:language>cs</dc:language>
- <thesis:degree>
  <thesis:name>Ing.</thesis:name>
  <thesis:level xml:lang="cs">Magisterský studijní program</thesis:level>
  <thesis:discipline xml:lang="cs">Hospodářská politika a správa/Účetnictví a finanční řízení
podniku</thesis:discipline>
- <thesis:grantor>
- <ccz:universityOrInstitution>
  <ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
  </ccz:universityOrInstitution>
  </thesis:grantor>
  </thesis:degree>
  <evskp:contact contactID="3190" />
  <evskp:fileNumber>3</evskp:fileNumber>
  <evskp:fileProperties fileID="14998" fileType="thesis" fileName="14998_xremj02.pdf"
fileSize="657189" format="application/pdf">Hlavní práce</evskp:fileProperties>
  <evskp:fileProperties fileID="16549" fileType="advisorReview" fileName="16549_kubova.pdf"
fileSize="674174" format="application/pdf">Oponentura</evskp:fileProperties>
  <evskp:fileProperties fileID="17458" fileType="refereeReview" fileName="17458_kubova.pdf"
fileSize="85398" format="application/pdf">Hodnocení vedoucího</evskp:fileProperties>
  <evskp:transfer accessRights="domain"
fileID="14998">http://www.vse.cz/vskp/id/14998</evskp:transfer>
  <evskp:transfer accessRights="domain"
fileID="16549">http://www.vse.cz/vskp/id/16549</evskp:transfer>
  <evskp:transfer accessRights="domain"
fileID="17458">http://www.vse.cz/vskp/id/17458</evskp:transfer>
- <evskp:server>
- <ccz:universityOrInstitution>
  <ccz:name xml:lang="cs">Vysoká škola ekonomická v Praze</ccz:name>
  <ccz:place>Praha</ccz:place>
- <ccz:department>
  <ccz:name xml:lang="cs">Centrum informačních a knihovnických služeb</ccz:name>
  <ccz:email>webmaster@vse.cz</ccz:email>
  <ccz:homepage>http://ciks.vse.cz/</ccz:homepage>
  </ccz:department>
  </ccz:universityOrInstitution>
  </evskp:server>
  <evskp:modified>2008-02-04T13:57:48Z</evskp:modified>
  </evskp:metadata>
  </metadata>
  </record>
  </GetRecord>
  </OAI-PMH>
```