Název: **Standardizace otevřených archivů: popis a výměna agregovaných webových zdrojů prostřednictvím OAI-ORE**

Autor: Petr Novák, Ústav informačních studií a knihovnictví FF UK a Ústav výpočetní techniky UK

Příspěvek na Semináři ke zpřístupňování šedé literatury, Centrum VUT Brno, 8.října 2008

---

1. Úvod

Nová aktivita Iniciativy pro otevřené archivy (Open Archive Initiative – OAI) s názvem Znovupoužití a výměna digitálních objektů (Object Reuse and Exchange) si klade za cíl nalézt způsob, jak deklarovat vazby mezi jednotlivými komponentami a prvky, ze kterých se skládají digitální objekty a jak jednotlivé komponenty a prvky komunikovat skrze počítačové sítě s cílem zachování maximální informační hodnoty možných existujících a souvisejících vazeb. Nalezený způsob je následně symbolicky vyjádřen prostřednictvím Abstraktního datového modelu (Abstract Data Model). K modelu je nabídnuto několik způsobů, jak prostřednictvím stávajících webových standardů a technologií realizovat celý proces v reálném provozu.

Aktivita je dokumentována online na webovém sídle OpenArchives.org a komunikována prostřednictvím návrhů specifikace, které k dnešnímu dni (3.10.2008) se nachází ve verzi 0.9. Vývoj specifikace je realizován prací Technického výboru OAI-ORE, Styčné skupiny OAI-ORE a Poradního výboru OAI-ORE[1]. Vývoj je financován z prostředků Mellonovy nadace, společnosti Microsoft a National Science Foundation. Veškeré aktivity jsou koordinovány Carlem Lagozem (Cornell University) a Herbertem van de Sompelem (Los Alamos National Laboratory).

1.1. Vztah k OAI-PMH

Nová aktivita nenavazuje na předchozí OAI-PMH, ani jej nerozšiřuje či na něj nenavazuje. ORE definuje datový model pro mapy zdrojů popisující agregace webových zdrojů a navrhuje formáty pro serializaci těchto map. ORE využívá webové architektury, kdy každý informační objekt je označen a zpřístupněn pomocí identifikátoru URI. Není definován nový protokol. Výměna map zdrojů je umožněna nezávisle buď přímým zpřístupněním na webu (HTTP / HTML zpřístupnění), nebo pomocí dávkových mechanismů. OAI-PMH je jeden z protokolů, který může být využit pro implementaci dávkového zpřístupnění[2].

1.2. Terminologie

URI (Uniform Resource Identifier) - identifikátor zdroje dle RFC3986. Velice zjednodušeně lze za URI považovat jak identifikaci pomocí URL (lokační funkce), tak pomocí persistentních identifikátorů URN, Handle apod.

ATOM - standard pro syndikaci webových zdrojů

agregace či agregát (aggregation) - zdroj tvořený množinou (agregovaných) webových zdrojů

---

[1] http://www.openarchives.org/ore/ORE_Community.php

[2] http://www.openarchives.org/ore/0.9/primer.html#RelationToPMH

mapa zdroje (resource map) – popis agregace opatřený dalšími metadaty a splňující požadavky datového modelu ORE

agregovaný zdroj (aggregated resource) – webový zdroj, který je součástí agregace (je v ní obsažen)

serializace (serialization) – vyjádření mapy zdroje pomocí standardizovaného formátu

výrok (assertion) – tvrzení, v kontextu ORE vyjádření vazby pomocí trojice

graf - grafické znázornění prvků (uzlů) a vazeb (hran)

1.3. RDF

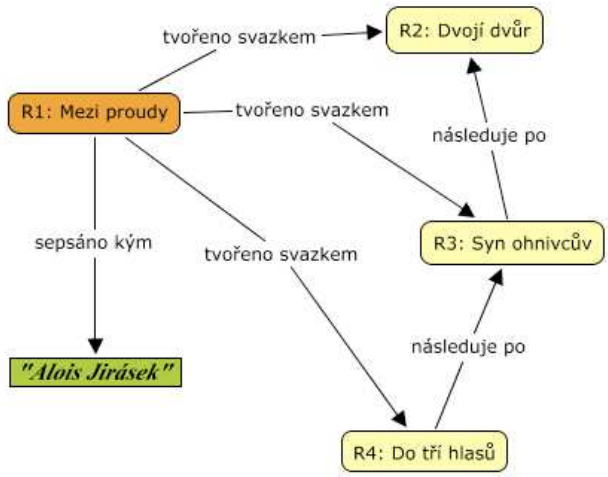
RDF (Resource Description Framework) je datový model pro popis metadat navržený W3C. Principem je připojování sémantických metadat formátovaných v XML k libovolnému dokumentu či informačnímu zdroji.

RDF graf je tvořen tvrzeními v podobě trojic (triplets), skládající se z uzlů (nodes). Prvky v trojici jsou ekvivalentní k anglickým označením větným členům:

angl.označení	**český ekvivalent**	**význam v kontextu RDF**
subject (S)	podmět	zdroj (subjekt))
predicate (P), možné též property	přísudek	vlastnost (predikát)
object (O)	předmět	objekt

1.4. Orientované (pojmenované) grafy v RDF

Mocnou zbraní RDF je vizualizace trojic, která je znázorňována jako orientovaný (pojmenovaný) graf. Subject a Object jsou uzly grafu, Predicate je hrana grafu.

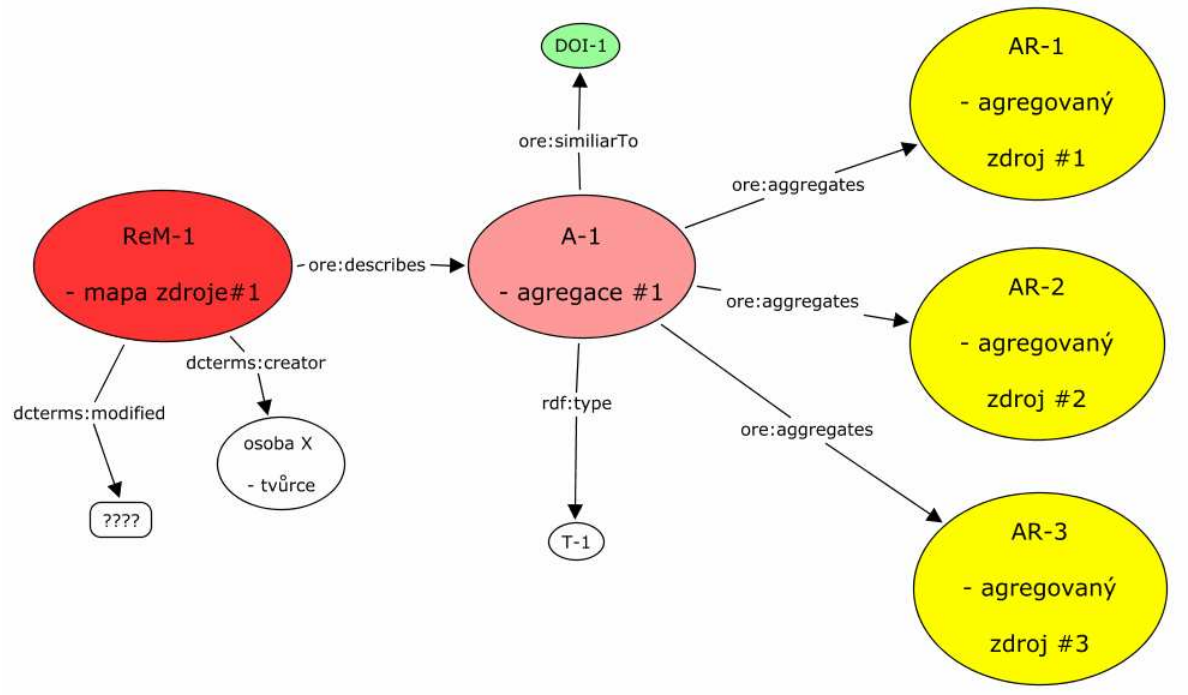


zdroj	**vlastnost**	**objekt**
R1	tvořeno svazkem	R2
R3	následuje po	R2
R1	sepsáno kým	***"Alois Jirásek"***
…	…	…

2. Abstraktní datový model (ADM) - základ ORE

Agregace je množina zdrojů. Zdroje jsou identifikovány pomocí URI. Pokud jsou tyto zdroje sdruženy v agregaci, jsou označeny jako agregované zdroje. Agregace je seskupení zdrojů, které je jako celek také identifikovatelné pomocí URI. Toto URI je jedinečné a nelze jej použít pro jiný účel.

Mapa zdroje popisuje obsah agregace, deklaruje vazby mezi agregovanými zdroji a doplňuje jej o další metadata (např. vazby na další agregace apod.). Mapa zdroje je identifikována pomocí URI.



Vazba mezi agregací a mapou zdroje je charakterizována takto:

- každá zdrojová mapa definuje (určuje) právě jednu agregaci
- každá agregace MŮŽE být vyjádřena a popsána vícero mapami zdrojů
- každý zdroj musí mít jednu serializaci (reprezentaci)

Uvedené principy zajišťují ověřitelnost a důvěryhodnost, ucelenost každé mapy zdroje.

Možnosti vazeb na mapu zdroje:

ore:describes	mapa zdroje popisuje agregaci
dcterms:modified	deklarace statusu mapy zdroje
dc:creator	tvůrce mapy zdroje či agregace (odpovědnost). Může jít i o proces zpracování, např. ingest či importní filtr
ore:similiarTo	agregace je podobná zdroji identifikovanému např. identifikátorem DOI
ore:aggregates	agregace tvoří vazbu na agregovaný zdroj (agreguje…)
rdf:type	komponenta sémantického webu, charakterizuje podstatu zdroje, který je vytvářen agregací a popsán mapou zdroje. Může definovat např. vazby na hesláře, tezaury, terminologické slovníky aj.

3. Serializace: cesta k praktickému využití teoretického konceptu

3.1. Cíle serializace

Praktické využití Abstraktního datového modelu závisí na výběru technologie, která teoretický koncept umožní zařadit do kontextu (současných) webových technologií. Mapy zdrojů mohou být serializovány pomocí vícero různých standardů a technologií, přičemž jejich výběr závisí na kompatibilitě se zdroji dat a nástroji, kterých je již užíváno.

3.2. Serializace v Atom

Jako prostředek pro reálnou implementaci se nabízí syndikace obsahu – technologie umožňující strojem čitelné předávání často aktualizovaných informací ve webovém prostoru[3]. Kromě formátů RSS existuje formát Atom, standardizovaný sdružením W3C[4].

Dle existujících srovnání Atom[5] nabízí oproti RSS 2.0 některé výhody:

- na Internetu jsou používány různé verze RSS (0.9*, 1.*, 2.*, plánována verze 3), Atom je v současné verzi 1.0 a stabilní.
- Kódování Atomu je na současné úrovni XML standardů (dodržování jmenných prostorů, schéma validace).
- Atom je vhodný k šíření širokého spektra obsahu, neboť podporuje zabalení binárních dat pomocí BASE-64 (např. obrázky, PDF dokumenty apod.).
- Atom je dobře rozšiřitelný (atribut rel pro prvek link apod).

Serializace prostřednictvím Atomu se nabízí díky možnostem tohoto formátu, zejména při využití nepovinných atributů a prvků.

3.3. Mapování ORE <-> Atom

Serializace vyžaduje proces mapování, definující vazby mezi jednotlivými prvky konceptu a mapovaného standardu:

	ORE		**Atom[6]**	
Význam	**Agregace**	<—>	**Feed (zdroj)**	Význam
URI identifikující agregaci	**URI-A**	<—>	**Feed <id>**	Persistentní identifikace zdroje
URI identifikující mapu zdroje	**URI-R**	<—>	**<link href="URI" rel="self"**	Způsob formátování URI pro daný účel
Vyjádření podobnosti	**ore:SimiliarTo**	<—>	**<link href="URI" rel="related"**	Způsob formátování URI pro daný účel
	Vlastnosti agregace a další metadata	<—>	**Metadata zdroje – např. <category>, <logo>, <rights>, <subtitle>, <contributor>**	Nepovinné elementy pro zdroj, mohou být využity v konkrétní aplikaci
	Agregovaný zdroj	<—>	**Entry (záznam)**	Základní komponenta zdroje – opakuje se
URI identifikující agregovaný zdroj	**URI-AR**	<—>	**<link href="URI" rel="alternate"**	Způsob formátování URI pro daný účel
	Vlastnosti agregovaného zdroje a související metadata	<—>	**Metadata záznamu – např. <category>, <contributor>, <published>, <source>, <rights>**	Nepovinné elementy pro záznam, mohou být využity v konkrétní aplikaci

[3] Syndikace je nejčastěji ztotožňovaná se standardy RSS (Really Simple Syndication či jiné výklady zkratky). S nejčastějším použitím RSS se lze setkat na zpravodajských serverech a portálech, kde jsou takto šířeny informace o vydání nového či o aktualizaci stávajícího článku, knihovníci nabízí pomocí syndikace novinky v katalozích knihoven či přírůstky v elektronických informačních zdrojích pomocí předdefinovaných klíčových slov (obdoba SDI profilů).

[4] http://interval.cz/clanky/atom-1-0/

[5] http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared

[6] feed v kontextu syndikace lze přeložit jako zdroj (obecné) nebo jako kanál – užíváno v případě RSS. Pro účely ORE se však označení „kanál" příliš nehodí.

3.4. Příklad serializace ORE ve formátu Atom[7]

Zdroj (feed):

```
<?xml version="1.0" encoding="utf-8"?>
<feed xmlns="http://www.w3.org/2005/Atom">

  <id>http://arxiv.org/rem/astro-ph/0601007#aggregation</id>
  <link href="http://arxiv.org/rem/astro-ph/0601007" rel="self" type="application/atom+xml"/>
  <generator uri="http://arxiv.org/">arXiv.org e-Print Repository</generator>
  <updated>2007-10-10T18:30:02Z</updated>
  <category scheme="http://www.openarchives.org/ore/terms/"
 term="http://www.openarchives.org/ore/terms/Aggregation" label="Aggregation" />

</feed>
```

Agregované zdroje – záznamy:

```
<entry>
   <id>http://oreproxy.org/r?what=http://arxiv.org/ps/astro-ph/0601007&amp;
       where=http://arxiv.org/rem/astro-ph/0601007%23aggregation</id>
   <link href="http://arxiv.org/ps/astro-ph/0601007" rel="alternate"
         type="application/postscript"/>
 </entry>

 <entry>
   <id>http://oreproxy.org/r?what=http://arxiv.org/pdf/astro-ph/0601007&amp;
       where=http://arxiv.org/rem/astro-ph/0601007%23aggregation</id>
   <link href="http://arxiv.org/pdf/astro-ph/0601007" rel="alternate"
         type="application/pdf"/>
 </entry>

 <entry>
   <id>http://oreproxy.org/r?what=http://arxiv.org/e-print/astro-ph/0601007&amp;
       where=http://arxiv.org/rem/astro-ph/0601007%23aggregation</id>
   <link href="http://arxiv.org/e-print/astro-ph/0601007" rel="alternate"/>
 </entry>
```

3.5. Další cesty k serializaci

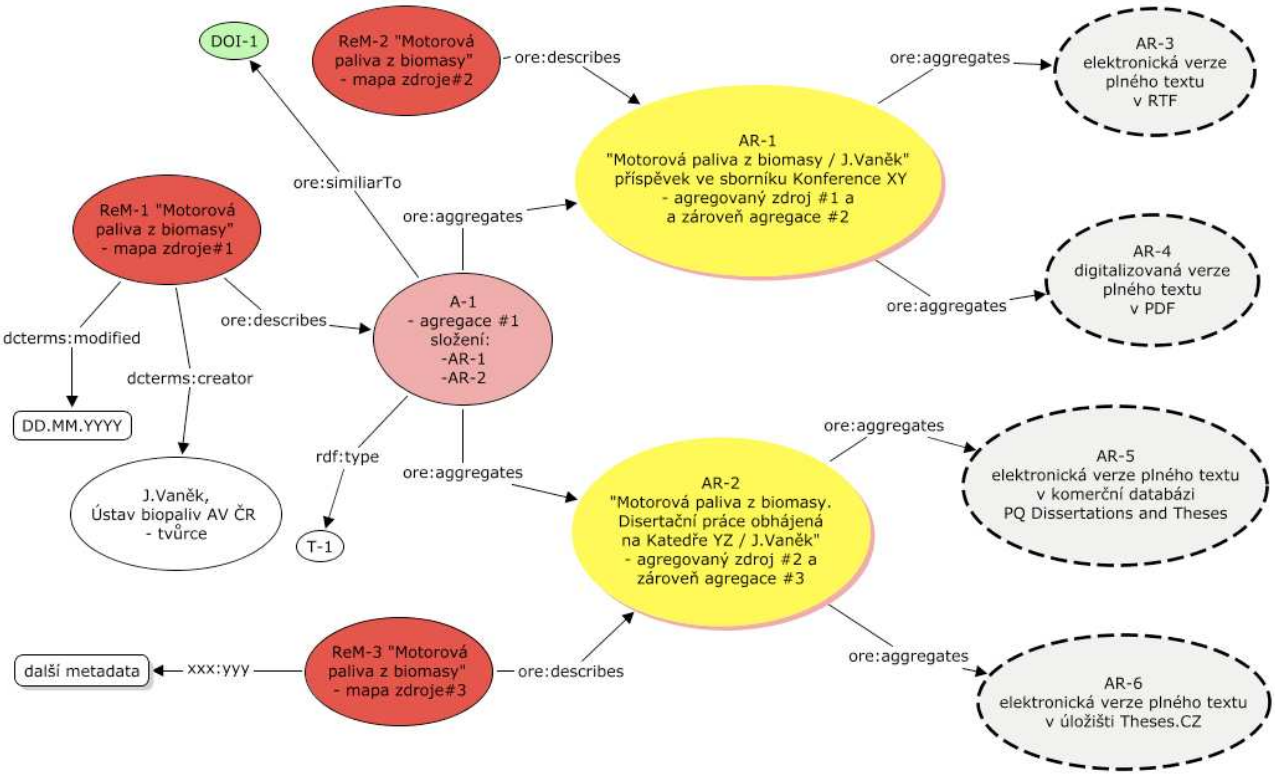
Serializace v Atom není jedinou možností. Doporučeným formátem pro serializaci je RDF/XML díky komplexním schopnostem umožňujícím přesné vyjádření bez omezení nativními elementy používanými v Atom. Jinou cestou je vložení RDF dat do XHTML dokumentu, který je zobrazován jako rozcestník po agregovaných zdrojích.

4. Možné praktické využití ORE

4.1. Provázání různých dokumentů šedé literatury na stejné téma

Navrhované řešení umožňuje pomocí mapy zdroje zachytit vazby mezi metadatovými záznamy 4 verzí 2 dokumentů. Všechny 4 verze primárních dokumentů – agregované zdroje – se nachází na různých místech, jsou zpřístupněny různými kanály a různým skupinám uživatelů dle zákonných omezení. Hlavní „mateřská“ agregace a ji definující mapa může být obsažena např. v souborném úložišti dokumentů šedé literatury.

[7] zdroj: http://www.openarchives.org/ore/0.9/atom-implementation



4.2. Provázání struktury obsahu online vědeckého časopisu

ORE lze využít pro zachycení a komunikování struktury časopis – číslo – článek – hierarchie stran. Alternativní způsob propojení pomocí map zdrojů umožní vizualizaci na úrovni preprint – publikovaný článek – postprint (plus přílohy – obrázky, kompletní datové řady, mapy aj.) nezávisle na místě uložení primárních textů jak v prostředí digitálních knihoven vydavatelů, tak v prostředí digitálních knihoven agregátorů či preprintových repozitářů. Možnosti vizualizace nastiňuje např. OREsome, Javovská knihovna vytvořená Rossem McFarlanem z University of Liverpool[8].

4.3. ORE Plugin využitelný při ingest procesu (ukládání) do repozitáře založeného na DSpace

Existující pluginy (založené např. na protokolu SWORD[9]) umožňují snadnou realizaci procesního řetězce pro ukládání obsahu do repozitářů. V případě uložení pouze primárního plného textu a popisných metadat dojde k ochuzení o vazby různého typu (struktura, souvislosti, podobnosti, složky), což lze vyřešit vkládáním ucelené mapy zdroje dle specifikace ORE. Problém řeší projekt realizovaný týmem Australian National University[10].

5. Závěr

OAI-ORE je v současné fázi ve verzi připomínkování finální verze. Po jejím uvolnění lze předpokládat

- implementaci do dostupných repozitářových systémů (Dspace, Fedora, EPrints) s cílem nabídnout uživatelům efektivní přístup ke zdrojům zachyceným v rámci

[8] http://www.openarchives.org/ore/RepoCamp2008/#OREsome
[9] http://www.ukoln.ac.uk/repositories/digirep/index/SWORD
[10] http://www.apsr.edu.au/ore/

agregace; podmínkou je úprava workflow pro vkládání metadat a primárních zdrojů do repozitářů, kdy bude nutné zajistit
    - generování vazeb mezi agregacemi a agregovanými zdroji,
    - vyplňování korektních metadat definujících mapy zdrojů,
    - zajistit vkládání časových otisků ke každému prvku, ze kterých se mapa skládá;
- návrh mechanismu pro automatizované generování map zdrojů a vytváření agregací nad stávajícími daty, které jsou již v repozitářích uloženy;
- rozšíření metadatových schémat komunikovaných pomocí infrastruktury OAI-PMH kompatibilních repozitářů o datový typ mapa zdroje a následné sklízení kompletních map.

Zda se koncepce ORE v praxi komunikování vědeckých poznatků osvědčí a ujme, závisí dle mého názoru jak na podpoře infrastruktury ze strany producentů (vydavatelé časopisů, provozovatelé repozitářů), tak na potřebě koncových uživatelů na vyhledávání informačních zdrojů „vcelku“ a v kontextu.