Masarykova univerzita
Fakulta informatiky

OAI-ORE

PV070 Digitální knihovny – Esej

Jiří Kremser

[N-IN Informační systémy sem. 4, roč. 2]

5. 12. 2009

Úvod

Motivací pro vznik standardu *Object Exchange and Reuse*, dále jen OAI-ORE, byla potřeba jednoznačně identifikovat množinu nebo kolekci zdrojů jako jeden složený zdroj. Jako typický případ užití standardu OAI-ORE bývá často uváděno označení souvisejících výstupů jednoho vědeckého experimentu jako jeden samostatný zdroj. Různé přílohy, protokoly, statistiky měření apod. tak budou svázány se samostatnou textovou částí, která může být dostupná samozřejmě ve více formátech. Podobně různé verze téhož obrázku, lišící se pouze svým rozlišením, například na Flickru. Vše navenek se tváří jako jeden složený zdroj. Takovýto zdroj může být součástí jiného složeného zdroje a může být strojově zpracován lépe, než jeho jednotlivé části bez vazeb na celek. Záměrem *Open Archives Initiative* (OAI) zde bylo využít tento koncept v sémantickém webu.

V eseji popíši stručně historii a vznik projektu, následně popíšu vztah k architektuře webu, protože autoři se snažili co nejlépe vystihnout podstatu zdrojové orientace OAI-ORE. Popíšu koncept abstraktního datového modelu a jeho jednotlivé komponenty a náležitosti, zejména mapu zdroje, což je klíčová část modelu. Abstraktní datový model, což je v podstatě graf s předem a jasně definovanou sémantikou, je potřeba nějak zachytit v určitém formátu, tento proces je realizován prostřednictvím serializace a popsán ve stejnojmenné části. Zmíním také současný stav projektu amožné vývoje do budoucna.

Historie

Standard OAI-ORE je méně známým bratříčkem protokolu OAI-PMH a kromě stejné skupiny tvůrců s ním nemá příliš společného. V současnosti se pro zachycení struktury metadat využívá nejčastěji formát METS podporovaný americkou Library of Congress. Naproti tomu OAI-ORE sází na RDF a ATOM jako nosiče struktury.

Potřeba po standardu tohoto typu vyplynula ze setkání nazvaného „*Augmenting interoperability across scholarly repositories*“ *[1]*, tedy zlepšení interoperability mezi vědeckými repositáři. Setkání uspořádala nadace Andrew William Mellona a uskutečnilo se 20. a 21. dubna 2006. Na tomto setkání ještě nepadl název tohoto standardu, ale Herbert Van de Sompel, zde měl prezentaci na téma složených zdrojů na příkladu zdroje v archivu *arXiv.org*.

V New Yorku 11. a 12. ledna 2007 se sešli *[2]* Carl Lagoze, Herbert Van de Sompel a technická komise iniciativy *Open Archives* a společně dali vzniknout novému standardu. Spíše než technické detaily probírali obecné témata, krom jiného abstraktní datový model, architekturu webu a hranice složeného objektu. Ale vznikly už i první představy o implementacích.

V tomtéž městě 29. a 30. května 2007 stejná skupina uspořádala další setkání *[3]*, kde doladila technické detaily a pro zachycení struktury složeného objektu zvolila tzv. pojmenované grafy, tedy obecný koncept, který může být realizován například prostřednictvím RDF tripletů (vrchol, hrana, vrchol) či protokolu určeného původně pro syndikaci – ATOM.

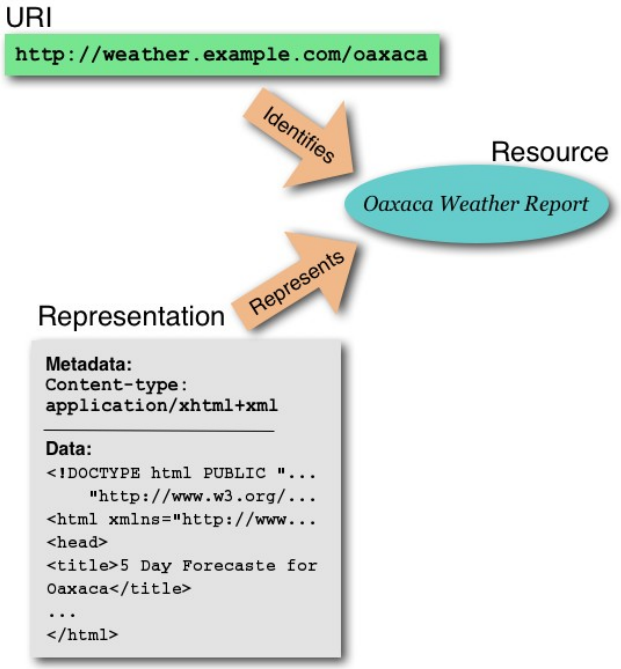
Ve verzi 1.0 bylo OAI-ORE zveřejněno 17. 10. 2008. Tato verze je současná (ke dni 5. 12 2009).

Vztah k architektuře webu

V prostředí webu rozlišujeme identifikátory – URI, zdroje a reprezentace. Zdrojem je myšlený jakýkoli předmět našeho zájmu na webu. Identifikátor identifikuje zdroj pomocí URI *[4]* a reprezentace reprezentuje (vizualizuje) zdroj, situace je patrná z *obrázku 1*. Reprezentace vzniká jako aplikace služby na identifikátor zdroje. Jeden zdroj tak může mít více reprezentací. Zdroj se tak stává tzv. občanem první kategorie a reprezentace občanem druhé kategorie. Toto členění se také nazývá ROA (*Resource Oriented Architecture*) a souvisí s myšlenkami sémantického webu. V popředí zájmu jsou v této architektuře vrcholy grafu a jejich sémantika[1]. Sémantika zdrojů v současném webu buď zcela chybí (tagy *div*), nebo je dodávána ad-hoc nesystematicky. Situace kolem digitálních objektů a digitálních knihoven je lepší, protože jsou zde pověření lidé starající se o připojování metadat k danému objektu v daném standardizovaném formátu.

„Syntax is not equivalent to nor sufficient for semantics.“ Searle, 1995 [5]

Každá reprezentace zdroje podle OAI-ORE musí mít vlastní URI. To neplatí v běžné architektuře webu, protože reprezentace je jednoznačně určena reprezentovaným zdrojem a protokolem či službou, kterou zdroj zpřístupňujeme. Digitální dokument (zdroj) je úzce spjat se svými reprezentacemi a svými dílčími částmi. Toto je potřeba formálně zachytit, aby robot, *crawler*, či jiný systém, dále jen agent, dokázal rozpoznat související celky. Podstatným rysem OAI-ORE je, že zachycuje relaci *býti částí* a umožňuje tak konstruovat stromové hierarchie celek-část. Tyto a jiné vlastnosti popisuje abstraktní datový model složeného zdroje o kterém bude následující část.



Ilustrace 1: Obrázek 1: Vztah mezi URI, Reprezentací a Zdrojem [4]

1 Oproti tomu konekcionalistický *[6]* přístup upřednostňuje propojení vrcholů a topologii celého grafu (neuronové sítě).

Abstraktní datový model

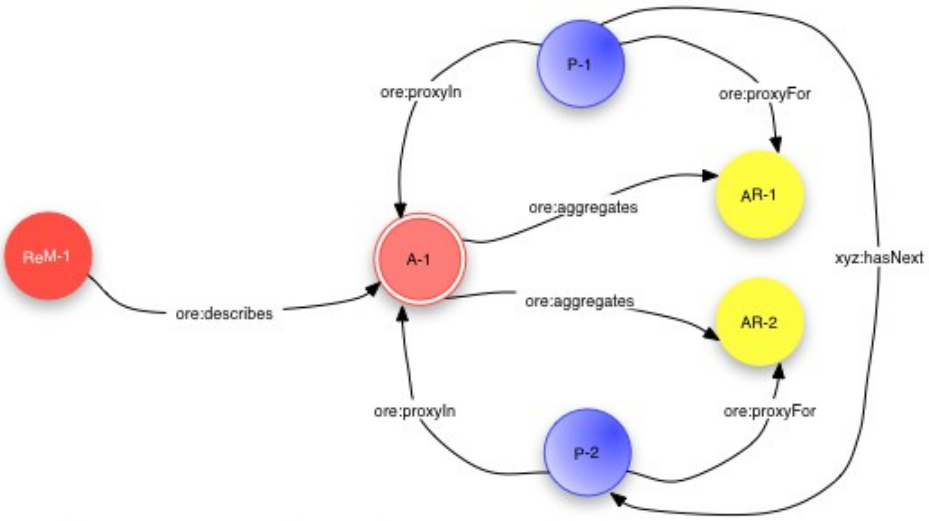
Slouží k popisu jak složeného objektu, tak jeho metadat, utváří tak pomyslnou hranici mezi tím co tvoří složený objekt a zbytkem světa. Lze jej vizualizovat formou grafu nebo zaznamenat jako seznam uspořádaných trojic – tripletů. Entitami a zároveň zdroji tohoto modelu mohou být agregace (*Aggregation, A)*, agregovaný zdroj (*Aggregated Resource, AR)*, proxy nebo mapa zdroje (*Resource Map, ReM)*.

Agregace a agregované zdroje

Agregace určuje množinu agregovaných zdrojů. Popisuje tedy strukturu složeného zdroje vztahem „*celek ore:aggregates část"*. Agregovaný zdroj může být najednou součástí více agregací.

Proxy

Zdroje typu proxy jsou jako jediný typ nepovinné a slouží k zanesení informace, která je kontextově závislá na dané agregaci. Model by totiž měl obsahovat pouze takové výroky (RDF triplety), které platí obecně. Někdy může být výhodné zachytit agregaci ne jen jako neuspořádanou množinu svých agregovaných zdrojů, ale také zachytit jejich pořadí pro určitý účel. Tato informace je ale kontextově závislá na dané agregaci, protože jiná agregace může nahlížet na tyto agregované zdroje v jiném pořadí. Proto musíme použít zástupné zdroje *Proxy* (zastupují agregované zdroje). Situace je znázorněna na *obrázku 2*.



<A-1>	<ore:aggregates>	<AR-1>
<A-1>	<ore:aggregates>	<AR-2>
<P-1>	<ore:proxyFor>	<AR-1>
<P-1>	<ore:proxyIn>	<A-1>
<P-2>	<ore:proxyFor>	<AR-2>
<P-2>	<ore:proxyIn>	<A-1>
<P-1>	<xyz:hasNext>	<P-2>

A-1	http://www.dlib.org/dlib/february06/smith/aggregation
AR-1	http://www.dlib.org/dlib/february06/smith/02smith.html
AR-2	http://www.dlib.org/dlib/february06/smith/pg1-13.html
ore:proxyFor	http://www.openarchives.org/ore/terms/proxyFor
ore:proxyIn	http://www.openarchives.org/ore/terms/proxyIn
xyz:hasNext	http://www.example.org/terms/hasNext

Obrázek 2: Použití proxy zdrojů [7]

Mapa zdroje

Mapa zdroje je zdroj, který popisuje (relace *ore:describes* popřípadě inverzní *ore:isDescribed-By*) právě jeden zdroj typu agregace. Jeden zdroj typu agregace ale může být popisován více mapami zdrojů, z nichž pouze jedna je takzvaná autoritativní mapa zdroje a ostatní jsou neautoritativní mapy zdroje. Autoritativní mapa zdroje je taková která je obdržena aplikací přístupového protokolu na URI agregace. Zpravidla se používá protokol HTTP a metody *HTTP 303 redirection*. V případě, že je pro každou agregaci použita právě jedna mapa zdroje, nebo server nepodporuje HTTP 303, je doporučováno použít k přístupu k mapě zdroje z URI agregace také metodu „*hash URIs*“,.tedy adresy se znakem #. Jako adresa agregace se tak zvolí například *http://server.org/foo.atom#aggregation*. Část za znakem # je pak podle specifikace protokolu HTTP ignorována serverem.

Důsledkem tohoto přístupu je mimo jiné také možnost vygenerovat *human-readable splash page* rovnou z mapy zdroje, která je zase *machine-readable*, a to je výsledek, který si OAI-ORE kladl na začátku. Mapa zdroje obsahuje celý abstraktní datový model daného složeného objektu, není to v rozporu s tím, že jednou z entit abstraktního datového modelu je sama mapa zdroje. Je zde totiž použito self-reference. Mapa zdroje tedy kromě výše zmíněného musí obsahovat vazbu na agregaci, alespoň jednu relaci mezi agregací a agregovaným objektem a základní metadata o sobě sama, jako autora a čas poslední modifikace. Mapa zdroje dále může obsahovat dodatečné metadata o sobě a o agregaci nebo relace mezi agregací a podobnými zdroji (například citace apod.). Takováto mapa zdroje s abstraktním datovým modelem je nejčastěji serializována do formátů ATOM a RDF/XML.

Serializace mapy zdroje

Abych nezabíhal zbytečně do detailů, tak uvedu pouze stručné charakteristiky dvou nejčastějších formátů určených k serializaci mapy zdroje, i když OAI-ORE nevylučuje jiné.

ATOM

ATOM je formát určen primárně pro syndikaci ve webovém prostředí, byl vytvořen jako nástupce formátu RSS a jeho dnešní použití je daleko širší než tvůrci původně zamýšleli. Lze jej například používat v architektuře REST, která má mimochodem hodně blízko ke zdrojově zaměřené architektuře. ATOM je dnes nedílnou součástí webu 2.0. ATOM je jednoduchý XML formát který popisuje webový zdroj jako tzv. f*eed*. Relace mezi agregací a agregovaným zdrojem je v ATOMU vyjádřena elementem *link* následujícím způsobem.

```
 ...
<link rel="http://www.openarchives.org/ore/terms/aggregates"
href="http://arxiv.org/ps/astro-ph/0601007"
title="Parametrization of K-essence and Its Kinetic Term"
type="application/postscript" hreflang="en"/>
 ...
```

Atribut `rel` zde určuje typ relace a atribut `href` objekt na druhé straně relace. Pro vyjádření relace, kde „subjekt“ není agregace je zde zabudována podpora pro RDF. Stačí vnořit příslušné RDF elementy do elementu `oreatom:triples`.

RDF/XML

RDF *(Resource Description Framework)* je model popisu metadat. Využívá zmiňované pojmenovaná grafy, které jsou ekvivalentní sérii výroků subjekt, vlastnost, objekt – RDF tripletů. RDF je často spojováno s technologií sémantického webu. Za subjekt a objekt je obvykle dosazován webový zdroj, respektive jeho URI a za vlastnost nějaká binární relace ze slovníku. Slovník je v RDF/XML představován jmenným prostorem. Takto vytvořené metadata je snadné strojově zpracovávat a budovat nad nimi ontologie, popřípadě provozovat nějaké složitější logické operace metodami strojového učení či sémantickými rozhodovacími nástroji (např. jazyk OWL DL).

Zde je příklad serializace mapy zdroje do RDF. Hodnota atribuju `rdf:about` je subjekt, `rdf:resource` je objekt a název elementu uvnitř `rdf:Description` je název relace, tedy `ore:aggregates.`

```
  ...
<rdf:Description rdf:about="http://arxiv.org/aggregation/astro-ph/0601007">
   <ore:aggregates rdf:resource="http://arxiv.org/ps/astro-ph/0601007">
    ...
</rdf:Description>
 ...
```

Současný stav

Ačkoli rodina specifikací OAI-ORE vypadá hodně pokrokově a rozumně mění podstatně koncept uchovávání metadat v institucionálních repositářích. Herbert Van de Sompel na 25. května 2009 v Praze v rámci akce Inforum tvrdil, že v tehdejší současnosti probíhá implementace OAI-ORE do všech větších repositářů, tedy Fedora Commons, DSpace a ePrints. Také se zmínil že je vyvíjen plugin do MS Word a že LoC publikuje digitalizované noviny využívaje OAI-ORE.

U gigantů jako je Fedora, implementace OAI-ORE samozřejmě není otázka pár dní. Současná současnost nasvědčuje tomu, že se standard uchytil. Aktivita na mailing listech OAI sice není vysoká a skupina na *google groups* je zaplavena spamem, ale komunita vyvíjí nové pluginy a repositáře si začínají OAI-ORE všímat[2]. *The Texas Digital Library* rozchodila OAI-ORE v prostředí Dspace[3] *[8]*. Pro repositář Fedora Commons existuje projekt *[9]* integrující OAI-ORE, bohužel nemůžu otestovat nakolik je funkční. Pro ePrints také vzniká implementace *[10]*.

Závěr

Dle mého názoru je OAI-ORE krok správným směrem. Usnadní práci archivářům, crawlerům a jiným agentům. Jako jeho největší výhodu vidím možnost generovat *splash page*. Je to další krok k oddělení formátování od vlastních dat a jejich smyslu. Nevýhody spatřuji v nejednoznačnosti hranic složeného objektu. Nějaká fuzzy klasifikace by lépe reflektovala realitu, ale zase by byla více složitá pro manipulaci. Myslím, že OAI-ORE má své místo v *Resource Oriented Architecture* a ve spojení s technologiemi ATOM, REST a sémantický web o něm ještě uslyšíme.

2 Archív JSTOR už implementoval (http://www.jstor.org)

3 Oficiálně ale ještě DSpace nepodporuje OAI-ORE

Zdroje

[1] http://msc.mellon.org/Meetings/Interop/
[2] http://www.openarchives.org/ore/documents/OAI-ORE-TC-Meeting-200701.pdf
[3] http://www.openarchives.org/ore/documents/OAI-ORE TC Meeting 200705_public.pdf
[4] http://www.w3.org/TR/webarch
[5] http://www.springerlink.com/content/9305464855t71341/fulltext.pdf
[6] http://en.wikipedia.org/wiki/Connectionism
[7] http://www.openarchives.org/ore/1.0/datamodel.html
[8] http://txspace.tamu.edu/handle/1969.1/86479
[9] http://oreprovider.sourceforge.net
[10] http://wiki.eprints.org/w/New_Features_Proposed_for_EPrints_3.2
[11] http://www.ikaros.cz/prehled-ramce-pro-vymenu-a-opetovne-vyuziti-digitalnich-objektu-v-otevrenych-archivech-oai-ore-herbe

Metadata

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="OAI-ORE" />
<meta name="DC.Creator" content="Jiří Kremser" />
<meta name="DC.Subject" content="OAI-ORE" />
<meta name="DC.Subject" content="složený objekt" />
<meta name="DC.Subject" content="sémantický web" />
<meta name="DC.Subject" content="OAI" />
<meta name="DC.Description.abstract" content="OAI-ORE je standard pro popis složených zdrojů" />
<meta name="DC.Publisher" content="Jiří Kremser" />
<meta name="DC.Date.created" scheme="W3C-DTF" content="2009-05-12" />
<meta name="DC.Type" scheme="DCMIType" content="Text" />
<meta name="DC.Format" scheme="IMT" content="application/pdf" />
<meta name="DC.Format.medium" content="computerFile" />
<meta name="DC.Format.extent" content="7 stran" />
<meta name="DC.Identifier" content="https://dspace.muni.cz/retrieve/4787/oai-ore.pdf" />
<meta name="DC.Identifier" content="http://www.mzk.cz/~kremser/oai-ore.pdf" />
<meta name="DC.Identifier" scheme="URN" content="URN:NBN:cz-nk20099278" />
<meta name="DC.Source" scheme="URL" content="http://www.openarchives.org/ore/"/>
<meta name="DC.Language" scheme="RFC3066" content="cze" />