Jiří Bařinka (359835)

(sem. 1, roč. 1.)

3.12.2013

ResourceSync

nový standard pro synchronizaci zdrojů na webu

http://www.openarchives.org/rs/toc

Úvod

ResourceSync protokol je určen pro synchronizaci jakýchkoli zdrojů na internetu, nezávisle na zdroji i rychlosti změny dat. Je to nástupce za standart OAI-ORE (Open Archive Initiative – Object Reuse and Exchange) a využívá zkušeností s OAI-PMH (Protocol for Metadata Harvesting). OAI-PMH byl pouze pro XML metadata a složitě synchronizoval obsah mezi zdroji. Existuje mnoho projektů a služeb pro synchronizaci zdrojů, ovšem většinou všechny jsou tvořené ad-hoc, pro každý případ zvlášť a není potom jednoduché použít tyto služby pro jiný projekt. ResoursourceSync se tedy snaží o to, aby byl obecnější a měl širší možnosti využití. Protokol ResourceSync je vytvořen jako modulární, každý modul má svou určitou funkci. Tyto moduly je možné vzájemně kombinovat a není nutné použít všechny. Technicky je realizovaný pomocí XML na bázi Sitemap a Siteindex, ale rozšířený o vlastní elementy.

Historie & aktuální verze

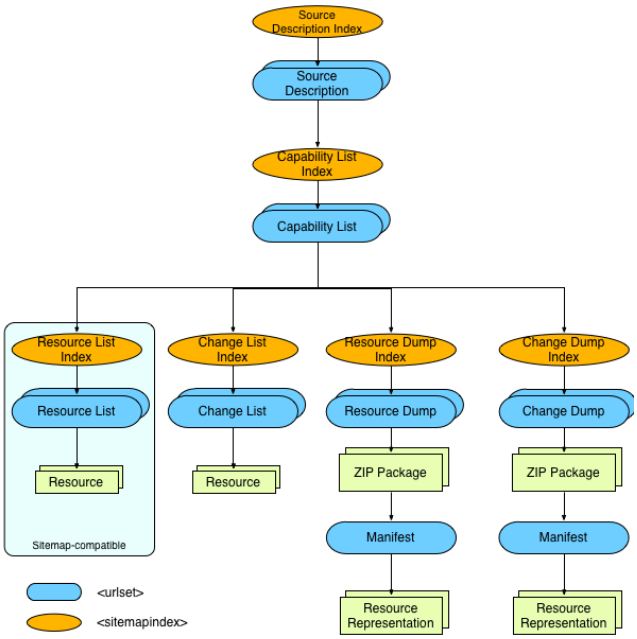
Projekt vznikl v srpnu roku 2012 pod skupinou kolem Herberta van de Sompela (Los Alamos National Laboratory). Samozřejmě se na vývoji podílí i ostatní zástupci firem jako RedHat, JISC a jiné. Prozatím poslední (5.) verze návrhu (draft) s označením 0.9.1 (Beta verze), je z letošního Srpna. Protokol by měl být hotový na přelomu letošního roku.

Cíle projektu

Cílem projektu je navrhnout postup pro synchronizaci v souladu s Webovou architekturou, která má šanci využití v různorodých situacích. Musí být lepší než dosavadní HTTP HEAD/GET.

ResourceSync protokol

Protokol pracuje s konceptem identifikace pomocí URI (Uniform Resourcer Identifier) a synchronizuje se právě URI nezávisle na jeho formátu. Mohou se takto synchronizovat data od malých webových stránek, až po velké repozitáře. Změnová frekvence může být různá od týdnů/měsíců až po sekundy. Důležité jsou v tomto případě dva pojmy, a to zdroj (Source) a cíl (destination) kam se synchronizuje. Volitelné funkce, vlastnosti a postupy, které protokol ustanovuje, musí umět zdroj i cíl, aby bylo možné celou synchronizaci provádět.



obr. – Struktura ResourceSync protokolu

Vlastnosti protokolu

- **Resource List** - Jako první krok pro snadnější synchronizaci se zdrojem je zveřejnění seznamu zdrojů (**ressource list**). Ten přenáší URI zdrojů pro synchronizaci. Je tvořen jako sitemap a jak je vidět v příkladu 1.1 zdroj má adresu URI v tagu <loc>, ten je jako větev tagu <url>. Seznam může být navíc rozšířený například o hash pod tagem <rs:md> a další.

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability="resourcelist"
         at="2013-01-03T09:00:00Z"/>
  <url>
      <loc>http://example.com/res1</loc>
  </url>
  <url>
      <loc>http://example.com/res2</loc>
  </url>
</urlset>
```

obr.: 1 – ukázka resource list

- **Change List** - Tato schopnost je implementovaná pro případ, kdy už je počet změn(případně obsah) natolik veliký, nebo se data mění tak často, že už na to vlastnosti ResourceList nestačí. Obsahuje informace například o tom, jak byl změněn v atributu „change", v jaký čas, to jsou atributy „from“ „until“. Možné operace jsou vytvoření, vylepšení nebo smazání. Příklad v obr. 1.1

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability="changelist"
         from="2013-01-02T00:00:00Z"
         until="2013-01-03T00:00:00Z"/>
  <url>
      <loc>http://example.com/res2.pdf</loc>
      <lastmod>2013-01-02T13:00:00Z</lastmod>
      <rs:md change="updated"/>
  </url>
  <url>
      <loc>http://example.com/res3.tiff</loc>
      <lastmod>2013-01-02T18:00:00Z</lastmod>
      <rs:md change="deleted"/>
  </url>
</urlset>
```

Obr.: 1.1 – ukázka change list

- **Resource Dump** – Cíl může poskytovat HTTP GET požadavky pro všechny zdroje URI obsažené v seznamu zdrojů (ResourceList). Pro velké seznamy zdrojů (ResourceList) to ovšem může být náročné. Proto je vytvořen Resource Dump, který je implementován jako sitemap a obsahuje na ukazatele na zabalený obsah do ZIPu. Každý takový ZIP obsahuje zdrojový datový tok spolu s Resource Dump Manifest, ten je také implementován jako Sitemap

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
  <rs:md capability="resourcedump"
         at="2013-01-03T09:00:00Z"/>
  <url>
      <loc>http://example.com/resourcedump.zip</loc>
      <lastmod>2013-01-03T09:00:00Z</lastmod>
  </url>
</urlset>
```

Obr.: 1.2 – ukázka Resource Dump

Vlastnosti synchronizace (PULL)

Synchronizace má 4 možnosti – **popis obsahu, zabalení obsahu, popis změn a zabalení změn.**

- **Popis obsahu** – V případě popisu dat, může zdroj obsahovat Up-to-date Resource List. Základní Resource List obsahuje minimálně URI zdrojů, které jsou dostupné pro synchronizaci. Nicméně je možné přidat informace obsahující například čas změny, kontrolní sumu, Hash nebo délku. Dále se zveřejňuje seznam zdrojů dostupných pro synchronizaci s cílem, při počátečním načtení.
- **Zabalení obsahu** – Pro případ, že potřebujeme data dostupné pro stažení, vytvoří zdroj Resource Dump, který bude obsahovat ukazatele na jeden nebo více balíčků. Každý z nich

obsahuje datový proud zdroje. K dispozici je také Resource Dump Manifest, který obsahuje metadata o datovém proudu. Reource dump vždy obsahuje alespoň URI a cestu ke zdroji ZIP souboru.

- **Popis změn** – Pokud potřebujeme synchronizovat archív s kratším intervalem synchronizace, zdroj poskytne Change List. Obsahující informace o změnách ve zdroji. Change list pokryje minimálně změny URI ve zdroji, aby zůstaly synchronní.
- **Zabalení změn** – Podobné jako u sbalení obsahu, jen tady se pracuje s Change Dump a Change dump Manifest

Vlastnosti notifikací (PUSH)

K omezení synchronizačního zpoždění a k optimalizaci samotné synchronizace slouží tyto notifikace:

- **Notifikace o změnách** - upozornění na změny jednotlivých zdrojů. (např. že byl vytvořen/upraven/smazán)
- **Framework notifikace** – upozornění na změny ve vlastnostech change list (např. že byl vytvořen/upraven/smazán)

Archivní možnosti (ARCHIVES)

Zdroje mohou uchovat historická data. Například je možné, aby cíl počkal na události, které nestihl, nebo znova navštívil předchozí stavy. Zdroj může také vydat archivní dokument, který obsahuje výpis historických vlastnosti dokumentů.

- **Resource List Archive**
- **Resource Dumb Archive**
- **Resource Change Archive**
- **Resource Change Archive**

Synchronizační vlastnosti

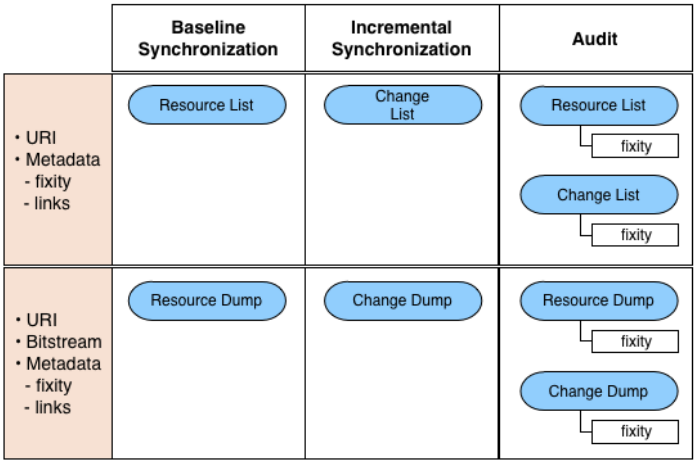
- **Prozkoumání možností** – podpora cílových stanic v prohledávání všech nabízených schopností protokolu. Aplikují se metody **PULL, PUSH, ARCHIVES**

- **Propojení odpovídajících zdrojů** – poskytuje odkazy ze zdroje subjektu k synchronizaci s odpovídajícím zdrojem. Aplikují se metody **PULL** a **PUSH**

Typy synchronizace

- **Základní Synchronizace (**initial load + catch-up**)** – Pro synchronizaci se zdrojem, musí cíl vytvořit inicializační kopii zdrojových dat. Cíl navíc může poskytnout Resource List s URI zdroje. Adresy poté postupně dereferencuje jednu po druhé.
  - Nesmí být nastavení bez spojení (Out-of-band)
- **Inkrementální Synchronizace (**průběžné up-to-date změny**)** – Cíl může využívat základní synchronizaci stále dokola. Pro zmenšení odezvy zdroj může využít ke komunikaci Change List. Dále může vytvořit Change Dump, který odkazuje na jeden nebo více balíků.
  - Subjekt má určité zpoždění, minimálně u vytvoření/update/vymazaní.
  - Umožňuje dohnat zpoždění potom, co byla cílová stanice offline

- **Audit (**kontrola synchronizovanosti**)** - Cílová stanice by měla být schopná určit jestli je synchronizovaná se zdrojem. Jestli synchronizovaná data odpovídají stávajícím datům u zdroje. Také se požaduje seznam obsahující metadata zdroje, kde jsou obsažené informace jako poslední čas změny, délka a hash.
  - Jde především o pokrytí a přesnost



Obr.: 2. – Perspektiva cíle ResourceSyc

Aktuální stav

ResourceSync protokol umožňuje synchronizaci řadou způsobů: Iniciální synchronizace všeho, inkrementální synchronizace, audit (co se povedlo/nepovedlo). Jako další selektivní synchronizace, to je zrovna jedno z témat, které není prozatím vyřešené (hlavně věci jako že cíl se nemůže volně měnit a co se bude synchronizovat). Celkově je to poměrně komplikovaný problém.

Další věc je, že samotné zdroje by měli posílat notifikace cílům (v různých definovaných intervalech), protože systém není postaven na bázi „busy waitingu“. Toto je prozatím také v experimentální fázi, ale v zásadě, dle několika příkladů z praxe by to v zásadě problém být neměl, jelikož notifikace představují jen malý objem dat při provozu v síti.

Závěr

Resource sync je poměrně nový protokol. Stále se na něm pracuje, jak jsem psal výše je teprve v Beta verzi a dokončený by měl být na konci roku. Podle mě až se dokončí, mohl by být dobrý, rychlý a měl by splňovat původní představy. Tím, že je postaven na XML a sitemap bude i rychle pochopitelný a použitelný, jelikož to nejsou složité techniky.

Literatura

ResourceSync: A Web-Based Resource Synchronization Framework, [cit. 3.12.2013]. Dostupné na: http://www.slideshare.net/OpenArchivesInitiative/resourcesync-tutorial

ResourceSync Framework Specification - Beta Draft, [cit. 3.12.2013]. Dostupné na: http://www.openarchives.org/rs/0.9.1/resourcesync#TimeAttributeReqs

Závěrečná zpráva k OAI 8, Cern, Ženeva, Švýcarsko, 2013 [cit. 3.12.2013]. Dostupné na: http://www.akvs.cz/pdf/zprava-krejcir-2013.pdf

Vlastíkova ženevská anabase (OAI8) 2013 [cit. 3.12.2013]. Dostupné na: http://www.akvs.cz/aktivity/ba-2013/ba2013-krejcir.pdf

Metadata DC

<?xml version="1.0" encoding="UTF-8"?>

<metadata xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="ResourceSync">
<dc:title>ResourceSync - novy standard pro synchronizaci zdroju na webu</dc:title>
<dc:creator>Jiří Bařinka</dc:creator>
<dc:subject>ResourceSync</dc:subject>
<dc:description>standard pro synchronizaci zdroju na webu</dc:description>
<dc:date>2013-12-3</dc:date>
<dc:format>PDF</dc:format>
<dc:language>CZE</dc:language>
</rdf:Description>
</metadata>