LodPaddle et OpenDataWrapper - HOW TO

Master 1 Informatique- University of Nantes

17 octobre 2014

Table des matières

1 LodPaddle et le web sémantique

1.1 Présentation de LodPaddle

LodPaddle est un projet porté par Hala Skaf-Molli du LINA [1] et fait écho au projet Européen Lod2 [2] qui a pour but de faciliter la production et l'exploitation de données au format spécifique du web sémantique. Pour ce faire, Lod2 nous propose une suite d'outils qui permettent l'extraction et la conversion de donnée, leur mise en ligne, les liens entre les différentes entités, leur enrichissement ...

Depuis quelques temps, le conseil général de la Loire-Atlantique, la région Pays de la Loire et la ville de Nantes on ouvert un pole Open Data dans leur département. Actuellement composé de 422 jeux de données sur des sujets variés comme les écoles, les loisirs, les subventions, les transports ..., ces données gagneraient énormément à être sémantifiées.

Le projet LodPaddle s'est donc associé aux acteurs du site data.paysdelaloire.fr afin de sémantifier leurs données, de permettre aux utilisateurs et développeurs de récupérer les données au format RDF [3] ou d'interroger directement un *SPARQL Endpoint* [4] et de développer une application pilote montrant que les données sémantifiées gagnent en valeur ajoutée.

1.2 Qu'est ce que le web sémantique

Aussi appelé web 3.0 ou web des données, le web sémantique est une évolution du web que nous connaissons. Actuellement, énormément d'information est dispersé sur la toile, comme sur Wikipédia par exemple, mais il est difficile de récupérer la bonne information rapidement. Par exemple, vous cherchez la liste des capitales mondiale, vous allez faire une requête sur votre moteur de recherche préféré, chercher une page qui pourrait éventuellement proposer la solution puis chercher dans la page choisie la probable information. Et celàaest vrai uniquement si une personne a crée cette liste au préalable.

Dans le domaine du web sémantique, la réponse serait instantanément une liste des capitales mondiales, au format texte, exploitable par un humain et une machine.

Listing 1– Requête SPARQL récupérant la liste des capitales mondiales

```
1 select distinct ?nom where {
2   ?pays prop-fr:capitale ?capitale.
3   ?pays rdf:type dbpedia-owl:Country.
4   ?capitale prop-fr:nom ?nom .
5 }
```

Le résultat de cette requête donnerait une liste comme celle présentée en Figure-1 .

1.3 Le format Ressource Description Framework (RDF)

Pour arriver au résultat précédent, nous avons besoin d'une syntaxe particulière permettant de décrire les données sous forme d'objet manipulables : Le

1. Laboratoire d'Informatique de Nantes Atlantique
2. LOD2 : http://lod2.eu/WikiArticle/Project.html
3. Ressource Description Framework
4. Serveur web comprenant le langage de requête SPARQL

nom
"Vienne"@fr
"Kaboul"@fr
"Buenos Aires"@fr
"Le Cap"@fr
"Bloemfontein"@fr
"Pretoria"@fr
"Riyad"@fr
"Erevan"@fr
"Luanda"@fr
"Saint John's"@fr
"Andorre-la-Vieille"@fr
"Bruxelles"@fr
"Dacca"@fr
"Brasilia"@fr
"Nassau"@fr
"Porto-Novo"@fr
"Sofia"@fr
"Belmopan"@fr
"Manama"@fr
"Thimphou"@fr
"Sarajevo"@fr

FIGURE 1 – Liste des capitales mondiales

RDF. Ce dernier fonctionne par triple de la forme Sujet - Predicat - Objet. Par exemple, le sujet « France » , le prédicat « a pour capitale » et l'objet « Paris ». Avec ce format, on se rends compte qu'on peut simplement demander une partie de ce triple en résultat ; par exemple, donne moi tous les triples qui ont pour sujet « Paris ».

1.4 Le Linked Open Data

La force du web sémantique par rapport au stockage classique des donnée par base de donnée, tableur et autres, vient du fait qu le sujet est identifié de façon unique sur le web. Ainsi, si une autre personne que vous a crée ce même sujet, vos données respectives voint pouvoir s'additionner : on appelle cela lier les données. Lorsque l'on effectue une recherche sur le sujet, vos informations, ainsi que toutes les informations produites par d'autres personnes concernant ce sujet, sont donc disponible.

2 Open Data Wrapper

L'Open Data Wrapper est un outils développé en JAVA permettant la conversion automatique des données de data.paysdelaloire.fr en données sémantifiées. Les données sont récupérées automatiquement du site à travers leur API dans

la mesure du possible.

2.1 Utilisation

Lancez Open Data Wrapper. Après un léger temps de chargement, l'application vous demande ce que vous souhaitez faire. Les propositions sont :

— lister les sources de donnée
— convertir une source de donnée au format N3
— convertir toutes les sources de donnée au format N3
— convertir une source de donnée au format XML/RDF
— convertir toutes les sources de donnée au format XML/RDF
— faire des requêtes sur vos données.
— quitter

Après selection d'une option, le traitement s'effectue.

2.1.1 Les proxies

Dans beaucoup d'organisation, le réseau est protégé par un proxy. Open Data Wrapper est configurable pour utiliser des proxies. Pour ce faire, crée un fichier proxy.pwd à la racine de votre compte (/home/utilisateur pour Linux, C :/documents and setting/utilisateur pour Windows). Ce fichier comportera les lignes suivantes :

Dans beaucoup d'organisation, le réseau est protégé par un proxy. Open Data Wrapper est configurable pour utiliser des proxies. Pour ce faire, crée un fichier proxy.pwd à la racine de votre compte (/home/utilisateur pour Linux, C :/documents and setting/utilisateur pour Windows). Ce fichier comportera les lignes suivantes :

Listing 2– Fichier de configuration des proxys proxy.pwd

```
1  proxyHost = url.du.proxy
2  proxyPort = 3128 (port du proxy)
3  authUser = nom d'utilisateur (vide si pas d'indentification)
4  authPassword = mot de passe
```

L'authentification retenu par Open Data Wrapper est de type BASIC. Si vous avez besoin de changer cela (NTLM par exemple), il vous faudra modifier le code.

2.2 Développement

Au cours du temps, les jeux de données vont s'enrichir et augmenter en nombre. Voici la procédure pour ajouter une nouvelle source de donnée

2.2.1 dataSources.xml

Ce fichier contient toutes les informations necessaires à la conversion d'une source de donnée.

Listing 3– Extrait de dataSources.xml

```
1 <source>
2   <nom>Hotel</nom>
```

```
3   <api>true</api>
4   <apiurl>
5     http://data.paysdelaloire.fr/api/publication/22440002800011
          _CG44_TOU_04815/hotels_STBL/content?format=xml
6   </apiurl>
7   <file>false</file>
8   <filepath>null</filepath>
9   <mappingFile>null</mappingFile>
10  <xsltFile>src/main/resources/xsl/hotel.xsl</xsltFile>
11  <format>XML</format>
12  <outputTtlFile>src/main/resources/output/ttl/Hotel.n3</
        outputTtlFile>
13  <outputXmlFile>src/main/resources/output/rdf-xml/Hotel.rdf</
        outputXmlFile>
14 </source>
```

Pour ajouter une nouvelles source, il vous faut créer un nouveau nœud <source></source> puis ajouter les balises suivantes :

- <nom> : le nom de la nouvelle source.
- <api> : true si la source existe sur l'API, false sinon.
- <apiurl : l'URL API de la source. Ne sera lu uniquement si api vaut true. null si aucune url.
- <file> : true si la source existe localement, false sinon. Si API vaut true, cette valeur n'est pas prise en compte.
- <filepath> : le chemin du fichier ou null sinon.
- <mappingFile> : le chemin du fichier de mapping dans le cas de fichier source spécifique. Pas utilisé encore.
- <xsltFile> : chemin vers le fichier de transformation XSL ou XSLT.
- <format> : format du fichier. XML uniquement pour le moment (CSV rapidement).
- <outputTtlFile> : le chemin du fichier turtle en sortie.
- <outputXmlFile> : le chemin du fichier XML/RDF en sortie.

Votre source est ainsi ajoutée à Open Data Wrapper.

2.2.2 ajout alternatif de source

Si la modification du XML est fastidieuse ou que vous avez une grande liste de dataset à ajouter, vous pouvez utiliser la fonction d'ajout de source. Pour ce faire, créez un fichier nommé **import.odw** à la racine de votre compte personel ($home). Dans ce fichier la syntaxe d'ajout est la suivante :

Listing 4– syntaxe import.odw

```
1 nom_du_dataset = http://url_de_l_api.com;http://desc_de_l_api.com;
      Titre_de_l_api;Publisher_de_l_api
```

Les restrictions liées à cette fonction sont les suivantes :

- impossible d'importer autre chose que des fichiers XML par API
- le nom doit être composer uniquement de caractères alphanumériques et le caractère _
- l'url de l'api ne comprend pas d'information de filtre ou de format

2.2.3 mapping.properties

Pour les données contactée par API, le format du fichier récupéré est de l'XML. Sa transformation en RDF se fait à l'aide d'une feuille de transformation XSL. Ce fichier est généré automatiquement par open Data Wrapper, en fonction du nom des balises XML. Attention, il faut que chaque nom de balise correspondent à une seule et uniquement une seule signification. De plus, le XML doit absolument avoir une balise dont le vocabulaire correspondant est traduit en **rdf :foaf**. Open Data Wrapper va prendre ses informations dans le fichier mapping.properties, lire le nom de la balise, trouver le correspondant RDF et génère un bloc de XSL. Si le nom de la balise n'est pas trouvé, un message s'affiche et le fichier mapping.properties est mis à jour avec une valeur par défaut. Vous devrez modifier cette valeur par la suite, pour lui attribuer un réel sens.

Ignored propreties during transformation exist in *src/main/resources/specific_ mapping*

2.2.4 Linking Datasets

Look at *src/main/config/linkmapping/properties* to learn about datasets mappings ..