Sémantisation des données du site LeReacteur

Anthony Pena, Jérémy Bardon, Nicolas Brondin

Sommaire

1 Introduction

Le site internet leReacteur regroupe des informations sur des expositions de tout types : vernissage, projection ou encore performance. Il s'agit d'un site de crowdsourcing qui regroupe les événements de plusieurs villes de France ajoutés par les utilisateurs.

2 Données disponibles

Les informations visibles sur le site sont toutes enregistrées dans une base de données SQL qui répartie les données sur 3 tables : expositions, lieux et images.

Un CMS étant utilisé, certaines informations ne sont pas directement liées aux expositions mais plutôt à la gestion interne du site. C'est pourquoi nous avons dû comprendre le sens de chaque champ afin de n'extraire que les données qui sont intéressantes.

Ensemble des données		
Exposition	Lieu	Image
Nom	Nom	Nom
Description	Catégorie	
Image	Position (latitude)	
Catégorie	Position (longitude)	
Artistes	Horaires	
Lieu	Adresse (rue)	
Site web	Ville	
Festival	Site web	
Date		

Pour extraire les données de la base nous avons choisi avec l'administratrice du site d'utiliser des vues afin que l'opération soit simplifiée avec l'outil de gestion phpMyAdmin [1] qu'elle utilise quotidiennement.

3 Sémantisation

Afin de donner du sens aux données nous avons cherché du vocabulaire à l'aide du projet lov qui nous a conseillé d'utiliser en grande partie celui mis à disposition sur schema.org.

1. Outil de gestion de base de données grand public disponible à cette adresse

Ce contributeur ne faisant pas l'unanimité au sein de la communauté notre choix c'est ensuite porté sur du vocabulaire plus générique et largement utilisé afin qu'il soit facilement identifiable par ceux qui voudraient utiliser nos données.

Vocabulaire utilisé

Exposition	dc :Event
Nom	foaf :name
Description	dc :description
Catégorie	dc :type
Artistes	dc :contributor
Site web	foaf :homepage
Festival	dc :hasPart
Date	dc :date

Image	foaf :Image
Nom	foaf :name
Description	dc :description

Lieu	dc :Location
Nom	foaf :name
Catégorie	dc :type
Position (latitude)	geo :lat
Position (longitude)	geo :long
Horaires	dc :hasOpeningHoursSpecification
Adresse (rue)	vcard :street-address
Ville	vcard :locality
Site web	foaf :homepage

4 Extraction des données

Afin d'extraire les données de la base phpMyAdmin, nous avons choisi de créer une vue qui regroupe ce que nous voulons extraire pour ensuite pouvoir utiliser l'extraction en xml par défaut.

Ensuite, pour ajouter nos données dans LodPaddle nous avons dû utiliser l'openDataWrapper [2] qui demande une structuration particulière du fichier xml pour l'importation de datasets. Dans le but de palier à ce problème, nous avons développé un programme qui fait la conversion.

2. Wrapper de l'université de Nantes disponible sur GitHub à cette adresse

Listing 1 – Export phpMyAdmin

```
<pma_xml_export version="1.0">
  <database name="reacteur">
    <table name="opendata_nantes">
      <column name="e_name">Le paraphile</column>
          ...
        </table>
    ...
  </database>
</pma_xml_export>
```

Listing 2 – Format du wrapper

```
<document>
  <data>
    <element>
          <e_name>Le paraphile</e_name>
          ...
        </element>
        ...
  </data>
</document>
```

Une fois ce fichier entré dans le wrapper, nous avons obtenu les données au format turtle et RDF.

5 Liaisons possibles

5.1 Localisation

Nous avons à disposition le lieu de chaque événement avec sa position géographique précise ainsi que ces horaires d'ouverture.

On peut imaginer regrouper d'avantage d'informations à propos de ces lieux ou encore afficher d'autres types de lieux (restaurants, monuments, etc.) à proximité et même, en fonction des horaires proposer une sorte d'emploi du temps pour la journée.

5.2 Artistes

La liste des artistes est associée à chaque exposition ce qui laisse penser que l'on peut retrouver des informations plus précises sur chacun d'eux par

exemple sur DBpedia. Avec ces nouvelles données il est alors possible de montrer d'autres oeuvres ou encore de proposer des événements regroupant des artistes ayant un style proche de ceux présents à l'exposition.

6 Requêtes SPARQL

6.1 Événements du 01/01/2014

Requête qui sélectionne l'ensemble de tous les évènements qui se passent le 1^{er} Janvier 2014.

```
SELECT ?name
WHERE
{
  ?event   rdf:type    dc:Event;
           dc:date     ?date;
           foaf:name   ?name.
  filter(?date = "2014-01-01"^^xsd:date)
}
```

6.2 Expositions auxquelles participe un artiste

Requête qui sélectionne l'ensemble de toutes les expositions à Nantes auxquelles Nathalie Lamotte participe.

```
SELECT ?name , ?date
WHERE
{
  ?event     dc:contributor    ?author;
             dc:date           ?date;
             foaf:name         ?name;
             dc:Location       ?location.
  ?location  vcard:locality    ?city.
  filter(
     ?city = <http://dbpedia.org/resource/Nantes>
     and ?author = "Nathalie Lamotte"
  )
}
```

7 Avancée du projet

7.1 Fini

— Sélection des données
— Choix du vocabulaire
— Export à partir de phpMyAdmin
— Formatage en xml spécial wrapper
— Import du dataset dans le wrapper
— Remplissage du fichier de mapping
— Conversion au format turtle

7.2 Reste à faire

— Conversion au format RDF
— Linkage du dataset avec ceux des autres groupes
— Utilisation des ontologies RDFS/OWL