****

|  |
| --- |
| TP Web Sémantique  **Web des données** |

Colin FRAPPER

Paul-Alexandre TESSIER

Quentin LEXERT

Laure DUCHEMIN

## I ) Introduction

Dans le cadre du module Web des données, nous devons réaliser un TP par groupes de 4. L’objectif du TP est de manipuler des données et des ontologies déjà existante ou non pour donner une forme sémantique de ces données. Nous avons choisi le domaine économique parmi les 4 thèmes proposés. Il contient des données sur les représentants des entreprises et des données SIRENE qui permettent une identification de celles-ci. Les données proviennent d’Open Data Soft. Nous devions donc trouver, améliorer ou créer les ontologies correspondant à ces données. Une fois, cela fait, nous devions rédiger un fichier RML pour ensuite exécuter l’outil RML Mapper pour avoir les données au format sémantique. C’est avec ce fichier RDF qu'on peut par la suite exécuter des requêtes SPARQL par exemple.

## II ) Outils

L’outil principal, nouveau pour nous, est RML Mapper. Il permet de générer les données au format RDF à partir des ontologies, du fichier RML et des données. Nous avons eu quelques difficultés pour l’installer sur nos machines. Nous avions 2 ordinateurs avec Windows et 2 avec un MacOS. Nous avons réussi à le faire tourner uniquement sur les Mac. Il aurait fallu une machine avec Linux pour éviter ces soucis techniques.

## III ) Installation

Pour la réalisation de ce TP, nous avons eu besoin d’installer l'environnement de développement RML Mapper. Comme il a été énoncé plus tôt nous avons des difficultés pour installer celui-ci sur tous les postes des membres du groupe et seule l’installation sur les poste MacOS fut un succès. Pour l’installation, nous avons procédé de la façon suivante :

* Nous avons installé sur nos postes l’outil RML Mapper par l'intermédiaire de la plateforme GitHub
* Nous avons lancé la construction de l’outil à l’aide de la commande “bin/RML-Mapper”
* Cependant une fois arrivés à ce niveau, nous avons commencé à rencontrer des problèmes d’installations. L’installation nous demandez en pré-requis d’avoir Maven d’installer et son installation fut compliqué
* Pour installer Maven nous avons fini utiliser la commande “brew install maven” avec en pré-requis Homebrew d’installé
* Une fois ces opérations effectuées, Homebrew et Maven installés la construction de RML Mapper peut se faire sans encombre.

## IV) Démarche

Nous avions 2 sources de données, les représentants des entreprises et les données SIRENE. Nous avons défini ensemble les différents champs et attributs provenants des données d’OpenDataSoft. Les données de SIRENE contiennent beaucoup de données par rapport aux chefs d’entreprise. Certaines étaient inutile, il était donc important de faire un tri préliminaire.

Ensuite, pour les ontologies, nous avons cherché le plus possible existant. Ce sont les ontologies qui vont nous permettre de lier les données entre elles et avec des données extérieures à notre schéma. En nous appuyant avec le cours, nous savions déjà comment en concevoir une, mais pas de cette taille. Il était donc important de récolter le plus d’ontologies pour nous éviter d’aller dans la mauvaise direction. Nous avions comme sources DBpedia, qui contient beaucoup d’informations sur les données SIRENE. Nous avions aussi le site LOV (Linked Open Vocabularies) qui contient quelques ontologies. Il était aussi proposé un diagramme, mais il était beaucoup trop lourd à utiliser et n’était pas pratique pour récupérer les informations.

Pour la rédaction des fichiers RML nécessaire pour le fonctionnement d’RML Mapper, nous avons choisi de nous partager le travail. Nous nous sommes séparé en binômes pour que chacun réalise un RML correspondant à une source de donnée. Paul et Colin ont réalisé pour les représentants d’entreprise, Quentin et Laure les données SIRENE. Ces binômes se sont créé naturellement car habitués à travailler ensemble tout au long de l’année sur d’autres projets ou TP.

i) Base des Représentants des entreprises

Pour représenter cette base, nous avons donc recherché les différentes ontologies que l’on pouvait recenser. Nous en avons trouvé quelques-unes et avons commencé à les implémenter. Afin de trouver une ontologie, il s’agissait pour nous de tout d’abord, définir l’objectif de cette ontologie et surtout de savoir comment l’utilisation de cette ontologie nous permettra d’atteindre nos objectifs. Par rapport à notre base, il nous était demandé de représenter une base de dirigeants d’entreprise. Dans un premier temps, nous avons recherché à travers les différents sites, les ontologies qui pouvaient se rapprocher des différents thèmes que l’on aurait pu aborder, nous en avons trouvé plusieurs, les plus importantes selon nous et c’est le plus important pour nous. En effet il ne s’agit pas créer un document RDF avec deux trois valeurs et le document en lui-même ne veut rien dire, il y a besoin de relations entre les entités de ce document. De plus lors de notre première implémentation, nous avons choisi de définir l’URI de notre sujet (rr:template “http://ods.dataset.com/Representant/{$.recordid}” ), car le recordid était un identifiant unique et qu’il nous semblait logique d’utiliser cet identifiant. Néanmoins, nous avons discuté avec Benjamin peu de temps après et il nous éclaircit sur le choix de cet identifiant, en effet, il était plus judiciable de choisir un identifiant que l’on puisse réutiliser après dans d’autres ontologies. C’est à ce moment que l’on a fait le rapprochement entre le numéro SIREN qui est un identifiant unique et donc nous avons choisi de changer pour ce numéro. Car nous pensons avec du recul qu’il faut comprendre à quoi et comment notre travail pourrait servir pour les “générations” futures dans le sens où si nous avions gardé le numéro recordid, d'une il n’était pas très parlant et en plus, personne n’aurait compris de quoi il s’agissait. Alors qu’avec notre numéro de SIREN, notre travail peut potentiellement être réutilisé et c’est tout simplement comme ça que le web sémantique est né, les différentes relations, l’assemblage d’ontologies pour en créer de nouvelles et ça ne fait que grandir. C’est pourquoi il est toujours plus intéressant de choisir une URI qui est parlante pour tout le monde afin qu’elle puisse être utilisée un jour... qui sait ?

ii) Base SIRENE

Pour la représentation de la base SIRENE, nous avons dans un premier temps étudié les informations qui sont disponibles dans la base et les avons classifiés dans un tableau à l’aide d’un code couleur. Ce code couleur nous a permis de trier les informations selon leur représentation, données géographiques, personnes, entreprises… Cette classification permet de cibler plus facilement les ontologies, et nous avons pu commencer nos recherches. Nous nous sommes principalement concentrés sur les ontologies disponibles sur “DBpedia” pour cette base, elle disposait des ontologies ciblés sur les entreprises, mais aussi sur les données géographiques et celles liées aux personnes. Nous avons aussi voulu intégrer dans notre représentation l’utilisation d’autres sources pour nos ontologies et avons décidé d’intégrer celle de “Geonames” pour les ontologies géographiques non-disponibles sur “DBpedia”. Les informations classifiées dans la base SIRENE étant assez conséquentes, il nous a fallu effectuer un tri dans les données modélisées et il nous a été impossible de trouver toutes les ontologies correspondantes. Nous avons alors créé les ontologies qui n’étaient pas encore existantes pour représenter les données qui nous semblaient intéressantes de conserver.

## 

## V ) Conclusion

Cela nous a permis d’y voir plus clair sur la démarche à suivre pour construire le web des données. La différence entre les ontologies, les données, RDF, RML est beaucoup plus clair après les avoir utilisés. De plus, ce TP nous a permis de voir les différentes données déjà présentes sur le web, et c’est plutôt impressionnant la quantité de données présente, quand on sait que c’est en constante évolution, on ne peut être qu’impatient. A travers ce travail, nous avons donc pu découvrir les différents objectifs d’une ontologie, comme le fait de pouvoir partager la signification de termes (comme le numéro SIREN par exemple) dans un domaine donné, notre domaine pour ce TP était plus centré sur les données économiques, de plus les ontologies permettent la standardisation du vocabulaire par exemple certains vocabulaire comme la Greffe si on reprend nos bases des représentants, sont un frein à la bonne compréhension de tous les parties. C’est pourquoi les ontologies sont également ici pour améliorer la compréhension entre les parties. Fondamentalement, le rôle des ontologies est d’améliorer la communication entre humains, mais aussi entre humains et ordinateurs et finalement entre ordinateurs.

Pour conclure, et en étant honnête le TP n’était pas facile, car c’était un travail que l’on a jamais réalisé auparavant et l’approche était différente ce qui permet d’en découvrir de nouvelles tout simplement, ce qui ne peut être que bénéfique pour nous. Comme dit précédemment, nous avons pu voir à travers notre exemple l’étendue d’une ontologie et plus largement d’un modèle de graphe RDF.

## VI ) Annexe

# Ontologie “Base Sirene”

<http://blog.sparna.fr/2013/12/07/ontologie-thesaurus-taxonomie-web-de-donnees/>

<http://www.geonames.org/>

<http://xmlns.com/foaf/spec/>

<http://mappings.dbpedia.org/server/ontology/classes/>

<http://rdf.insee.fr/def/index.html>

http://dbpedia.org/ontology/siren

A construire

|  |  |  |
| --- | --- | --- |
| Colonne | Lien |  |
| Identifiant de l’entreprise |  |  |
| Numéro interne de classement |  |  |
| SIRET |  |  |
| Première ligne de l’adressage de l’établissement |  |  |
| Deuxième ligne de l’adressage de l’établissement |  |  |
| Troisième ligne de l’adressage de l’établissement |  |  |
| Quatrième ligne de l’adressage de l’établissement |  |  |
| Cinquième ligne de l’adressage de l’établissement |  |  |
| Sixième ligne de l’adressage de l’établissement |  |  |
| Septième ligne de l’adressage de l’établissement | <http://mappings.dbpedia.org/server/ontology/classes/Settlement> |  |
| Numéro dans la voie |  |  |
| Indice de répétition |  |  |
| Type de voie de localisation de l’établissement |  |  |
| Libellé de voie de localisation |  |  |
| Code postal |  |  |
| Code CEDEX |  |  |
| Région de localisation de l’établissement |  |  |
| Libellé de la région |  |  |
| Département de localisation |  |  |
| Libellé du département de l’établissement |  |  |
| Arrondissement de location |  |  |
| Canton de localisation |  |  |
| Commune de localisation |  |  |
| Libellé de la commune de localisation |  |  |
| Département et commune de localisation |  |  |
| Libellé de la taille de l’unité urbaine |  |  |
| Numéro de l’unité urbaine |  |  |
| Localisation de l’établissement |  |  |
| Tranche de commune détaillée |  |  |
| Zone d’emploi |  |  |
| Qualité de siège ou non de l’établissement |  |  |
| Enseigne ou nom de l’exploitation |  |  |
| Indicateur du champ du publipostage |  |  |
| Libellé de la nature de l’établissement |  |  |
| Libellé de l'activité principale |  |  |
| Année de validité de l’activité principale |  |  |
| Libellé de la tranche d’effectif |  |  |
| Effectif salarié de l’établissement |  |  |
| Année de validité de l’effectif salarié |  |  |
| Libellé de l’origine de la création |  |  |
| Année et mois de création de l'établissement |  |  |
| Date de début d’activité |  |  |
| Libellé de la nature de l’activité |  |  |
| Lieu de l’activité de l’établissement |  |  |
| Libellé du type de magasin |  |  |
| Caractère saisonnier ou non de l’activité |  |  |
| Libellé de la modalité de l’activité |  |  |
| Caractère productif de l’établissement |  |  |
| Libellé de la participation particulière |  |  |
| Caractère auxiliaire de l’activité |  |  |
| Nom ou raison sociale de l’entreprise |  |  |
| Sigle de l’entreprise |  |  |
| Nom de naissance |  |  |
| Prénom |  |  |
| Civilité des entrepreneurs individuels |  |  |
| Numéro d’identification au répertoire |  |  |
| Numéro interne de classement de l’établissement |  |  |
| Région de localisation du siège |  |  |
| Département et commune de localisation |  |  |
| Adresse mail |  |  |
| Libellé de la nature juridique |  |  |
| Activité principale de l’entreprise |  |  |
| Section |  |  |
| Classe |  |  |
| Sous-classe |  |  |
| Groupe |  |  |
| Division |  |  |
| Activité principale au registre |  |  |
| Libellé de l’appartenance au champ de l'économie sociale et solidaire |  |  |
| Date ESS |  |  |
| Libellé de la tranche d’effectif salariés de l’entreprise |  |  |
| Effectif salarié de l’entreprise |  |  |
| Année de validité de l’effectif salarié |  |  |
| Catégorie de l’entreprise |  |  |
| Année et mois d'introduction de l’entreprise dans la base |  |  |
| Libellé de la modalité de l’activité |  |  |
| Caractère productif de l’entreprise |  |  |
| Année de validité des entreprises |  |  |
| Libellé de la tranche de CA |  |  |
| Première activité secondaire déclarée |  |  |
| Deuxième activité secondaire déclarée |  |  |
| Troisième activité secondaire déclarée |  |  |
| Quatrième activité secondaire déclarée |  |  |
| Date de traitement de la MAJ |  |  |
| Coordonnées |  |  |
| Téléphone |  |  |

# Ontologie “Représentants des entreprises”

|  |  |  |
| --- | --- | --- |
| Colonne | Lien | Commentaires |
| Siren | <http://dbpedia.org/ontology/siren>  rdfs:label **siren** number @en  localName **siren** |  |
| Qualité | rdfs:label |  |
| Nom Patronymique | http://dbpedia.org/ontology/name |  |
| Prénoms | http://dbpedia.org/ontology/Surname |  |
| Numéro de Gestion | http://dbpedia.org/ontology/number |  |
| Type | http://dbpedia.org/ontology/type |  |
| Greffe | http://dbpedia.org/ontology/city |  |
| date\_greffe | http://dbpedia.org/ontology/date |  |
| Nom d'usage | http://dbpedia.org/ontology/name |  |
| Pseudonyme | http://dbpedia.org/ontology/name |  |
| Dénomination | http://dbpedia.org/ontology/denomination |  |

VII) Sources

<http://www-igm.univ-mlv.fr/~dr/XPOSE2009/Le%20Web%203.0/concepts.html>

Nous nous sommes également appuyés sur le cours dispensé par Mr Desmontils - Web des données et des connaissances.