LE BARS Yannis
FORTIN Guillaume
TRAVERS Clément

Groupe de TD : B

Rapport
Projet de web sémantique

UFR Sciences et Techniques - Nantes

M1 ALMA - 2017/2018

UNIVERSITÉ DE NANTES

Table des matières

I - Introduction **2**

II - Le projet **2**

Etape 1 2

Choisir un ensemble de données ouvertes à « sémantifier » 2

Description du sujet 2

Méthode de réalisation 3

Proposer deux requêtes SPARQL « intéressantes » 4

Première requête 4

Deuxième requête 5

Etape 2 6

Lier nos données 6

Proposer une requête 6

Etape 3 7

Etape 4 8

Etape 5 9

ANNEXE A **10**

RÉSULTAT DE LA PREMIÈRE REQUÊTE 10

RÉSULTAT DE LA SECONDE REQUÊTE 10

ANNEXE B **11**

RÉSULTAT DE LA REQUÊTE SUR DONNÉES LIÉES 11

I - Introduction

L'objectif de ce projet est de transformer des données ouvertes de l'Enseignement supérieur, de la Recherche et de l'Innovation en données sémantiques et de lier ces données sémantiques au cloud de "Linked Data : Connect Distributed Data across the Web"

Le projet se déroulera en plusieurs étapes:

Etape 1 : Choisir un ensemble de données ouvertes à « sémantifier » et proposer deux requêtes SPARQL « intéressantes » réalisables sur ces données.

Etape 2 : Lier nos données à des données des autres groupes et proposer une requête réalisable sur au moins 2 ensemble de données.

Etape 3 : Utiliser des ontologies RDFS ou OWL et faire des inférences.

Etape 4 : Proposer des propriétés et des liens pour lier les données de ESR au cloud de linked data.

Etape 5 : Utiliser le vocabulaire VOID pour décrire nôtre datasets.

II - Le projet

Lien github de notre projet: https://github.com/Astlo/ProjetTarql

Etape 1

Choisir un ensemble de données ouvertes à « sémantifier »

Description du sujet

Sujet : Appels à projets Horizon 2020 - Projets retenus et participants identifiés des appels du Programme-Cadre de Recherche et d'Innovation (PCRI)

Nombre d'enregistrements : 19.101

Nombre d'attributs : 22

Date de création : 1 juin 2017

Le dataset regroupe les projets ayant commencé entre le 1 janvier 2014 et le 1 juin 2016, et devant se terminer aux alentours de 2020. Dans les faits ces données ne sont pas “propres”, en effet, on retrouve dans le dataset des projets commencés le 2 décembre 2017 et d’autres se terminant entre 2014 et 2023.
Chaque projet a un ou plusieurs partenaires venant d’un ou plusieurs pays, une durée, un montant, un résumé, etc.

Méthode de réalisation

Pour sémantifier les données nous avons choisi d’utiliser l’outil Tarql.
Nous avons dû créer un code permettant de recevoir des données en entrée sous la forme d’un fichier csv, et avec ce code, Tarql renvoie en sortie les données sémantiques en rdf.
Voici la description de nos uris:

```
?URI rdf:type doap:Project;
       frapo:hasCode ?code_du_projet;
       doap:name ?titre;
       dbp:shortName ?acronyme;
       ?URI7 ?URI6;
       time:hasBeginning ?date_de_debut;
       time:hasEnd ?date_de_fin;
       doap:description ?resume;
       time:Duration ?duree;
       dbp:number ?montant;
       foaf:topic ?theme;
       ?URI8 ?code_rcn;
       doap:homepage ?lien_cordis;
       dcat:keyword ?mots_cles;
       org:linkedTo ?URI2.
```

Notre première URI identifie un projet. Nous nous sommes aidés d’exemples trouvés sur les internets. ?code_du_projet, ?titre, ?acronyme, … sont les propriétés du projet. Pour les cas spéciaux comme ?appel_a_projet et ?code_d_appel_a_projet, nous avons défini notre propre vocabulaires spécifiques car nous n’avons pas trouvé de vocabulaires correspondant. Pour le reste, nous avons utilisé du vocabulaire préexistants : dpb, dpo, owl, doap, foaf, rdf …

```
?URI2 rdf:type dbo:Organisation;
       ?URI5 ?identifiant_de_partenaire;
       rdfs:label ?libelle_de_partenaire;
       dbo:type ?URI3;
       dbo:locationCountry ?URI4.
```

Notre deuxième URI est de type Organisation, elle est lié aux projets avec le vocabulaire org:linkedTo, c’est pour représenter les partenaires qui sont liés aux projets.

```
?URI3 rdf:type owl:Thing;
      rdfs:label ?type_de_partenaire;
      rdfs:label ?code_du_type_de_partenaire.
```

Notre troisième URI permet l'identification des types de partenaire des projets.

```
?URI4 rdf:type dbo:PopulatedPlace;
      rdfs:label ?pays_du_partenaire;
      rdfs:label ?code_pays.
```

Notre quatrième URI permet l'identification des pays des partenaires du projet. Nous avons pas lié les pays à des URI déjà existantes, car ?pays_du_partenaire et ?code_pays sont fait de valeurs multiples.

```
?URI5 rdf:type owl:DatatypeProperty.
```

URI5 identifie les participants par leur numéro siren ou autre selon leur nature.

```
?URI6 rdf:type owl:Class;
      rdfs:label ?appel_a_projet;
      frapo:hasCode ?code_d_appel_a_projet.

?URI7 rdf:type owl:DatatypeProperty.
```

L'URI7 s'appelle "Appel_Projet", c'est une "DatatypePropertyeProperty". L'URI6 contient l'appel du projet et le code de l'appel.

```
?URI8 rdf:type owl:DatatypeProperty;
      rdfs:range xsd:nonNegativeInteger.
```

URI8 est un vocabulaire pour le code rcn et c'est un entier non négatif.

Proposer deux requêtes SPARQL « intéressantes »

Première requête

La première requête consiste à retourner les 100 projets ayant reçus les subventions les plus importantes par rapport à leur durée prévue, elle se présente de la forme suivante (*cf Annexe A*) :

```
SELECT ?titre ?duree ?montant ?rapport {
       ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://usefulinc.com/ns/doap#Project>;
          <http://usefulinc.com/ns/doap#name> ?titre;
          <https://www.w3.org/2006/time#Duration> ?duree;
          <http://purl.org/cerif/frapo/BudgetedAmount> ?montant.
       BIND (<http://www.w3.org/2001/XMLSchema#integer>(?duree) AS ?duree2)
       BIND (<http://www.w3.org/2001/XMLSchema#float>(?montant) AS ?montant2)
       BIND (<http://www.w3.org/2001/XMLSchema#decimal>(?montant2/?duree2) AS
?rapport)
}
ORDER BY DESC(?rapport) LIMIT 100
```

On renvoie, pour les 100 projets ayant reçus le montant le plus important par rapport à leur durée, leur titre, leur durée, le montant reçu et le rapport montant / durée, récupéré grâce à un BIND.

Deuxième requête

La seconde requête retourne les projets ayant la plus grande différence entre la durée prévue et la durée effective (correspondant à la période écoulée entre la date de début et la date de fin effective) (*cf Annexe A*) :

```
SELECT ?titre ?difference {
       ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://usefulinc.com/ns/doap#Project>;
          <http://usefulinc.com/ns/doap#name> ?titre;
          <https://www.w3.org/2006/time#hasBeginning> ?date_de_debut;
          <https://www.w3.org/2006/time#hasEnd> ?date_de_fin;
          <https://www.w3.org/2006/time#Duration> ?duree.
       BIND( strbefore( ?date_de_debut, \"-\" ) as ?anneeDebut)
       BIND( strbefore( ?date_de_fin, \"-\" ) as ?anneeFin)
       BIND( strafter( ?date_de_debut, \"-\" ) as ?moisjourDebut)
       BIND( strafter( ?date_de_fin, \"-\" ) as ?moisjourFin)
       BIND( strbefore( ?moisjourDebut, \"-\" ) as ?moisDebut)
       BIND( strbefore( ?moisjourFin, \"-\" ) as ?moisFin)
       BIND((<http://www.w3.org/2001/XMLSchema#integer>(?anneeFin) -
<http://www.w3.org/2001/XMLSchema#integer>(?anneeDebut))*12 +
<http://www.w3.org/2001/XMLSchema#integer>(?moisFin) -
<http://www.w3.org/2001/XMLSchema#integer>(?moisDebut) AS ?mois_effectif)
       BIND((<http://www.w3.org/2001/XMLSchema#integer>(?duree) -
?mois_effectif) AS ?difference)
}
ORDER BY desc(?difference) LIMIT 100
```

Pour les 100 projets ayant la différence la plus importante, on retourne le titre du projet et la valeur de cette différence. Beaucoup de BIND sont réalisés dans cette requête, afin de permettre de traiter les différentes informations dont on a besoin pour calculer la durée effective du projet.
La partie la plus difficile de cette requête fut de séparer le mois et l'année des chaînes de caractère date.
Le résultat des deux requêtes est disponible en annexe A.

Etape 2

Lier nos données

Nous avons choisis de lier nos données avec le groupe de Romain Brohan et Samy Gascoin-Fontaine, car leur dataset porte sur un sujet très similaire au notre, et se compose des mêmes données. Leur sujet est également une présentation d'un ensemble de projets proposés pour un appel à projet européen, dans le cadre du 7ème Programme-Cadre de Recherche et de Développement Technologique (PCRDT).

Sont recensés les projets commencés entre le 1er janvier 2007 et le 31 décembre 2013. Tout comme dans notre dataset, un projet dispose d'un ou plusieurs partenaires, une durée, un montant, etc…

Nombre d'enregistrements : 26.353

Nombre d'attributs : 23

Date de création : 6 juillet 2016

Proposer une requête

La requête proposée sur les données liées permet d'analyser les montants attribués par pays et par an (*cf Annexe B*) :

```
SELECT ?anneeDebut ?pays (SUM(?money2) AS ?somme) (AVG(?money2) AS ?moyenne) {
	{
		?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://usefulinc.com/ns/doap#Project>;
		   <http://purl.org/cerif/frapo/BudgetedAmount> ?money;
		   <https://www.w3.org/ns/org#linkedTo> ?y;
		   <https://www.w3.org/2006/time#hasBeginning> ?date_de_debut.
		?y <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Organisation>;
		   <http://dbpedia.org/ontology/locationCountry> ?pays.
		BIND( strbefore( ?date_de_debut, \"-\" ) as ?anneeDebut)
		BIND(<http://www.w3.org/2001/XMLSchema#float>(?money) AS ?money2)
	} UNION {
		GRAPH ?g {
			?x <http://www.w3.org/2000/01/rdf-schema#type>
<http://usefulinc.com/ns/doap#Project>;
			   <http://purl.org/cerif/frapo/budgetedAmount> ?money;
			   <https://www.w3.org/ns/org#linkedTo> ?y;
			   <https://www.w3.org/2006/time#hasBeginning> ?anneeDebut.
			?y <http://www.w3.org/2000/01/rdf-schema#type>
<http://usefulinc.com/ns/doap#projectParticipant>;
			   <http://dbpedia.org/ontology/locationCountry> ?pays;
			   <https://www.w3.org/2006/time#hasBeginning> ?anneeDebut.
			BIND(<http://www.w3.org/2001/XMLSchema#float>(?money) AS
?money2)
		}
	}
}
GROUP BY ?anneeDebut ?pays
ORDER BY ?anneeDebut ?pays
```

On retourne, pour chaque année et chaque pays, la somme et la moyenne des montants attribués aux projets. On réalise la jointure entre les deux datasets en utilisant une UNION, et en accédant au graphe de l'autre groupe via GRAPH.
On peut très bien penser à trier par la moyenne ou la somme pour voir quels pays en quels années a reçu le plus de budget.
Le résultat de cette requête est disponible en annexe B.

Etape 3

Dans cette étape, il nous était demandé de faire des inférences à l'aide d'ontologies RDFS ou OWL, nous avons donc choisi d'utiliser les règles d'inférences RDFS vues en cours que nous avons entrées dans un fichier rules.txt. Voici un aperçu de ce fichier avec les règles :

```
[ruleRDF1: (?u ?a ?y) -> (?a rdf:type rdf:Property)]
[ruleRDFS2: (?a rdfs:domain ?x) (?u ?a ?y) -> (?u rdf:type ?x)]
[ruleRDFS3: (?a rdfs:range ?x) (?u ?a ?v) -> (?v rdf:type ?x)]
```

Ensuite nous devions utiliser ce fichier ainsi que notre fichier de données RDF pour obtenir un nouveau fichier avec les inférences. Pour cela nous avons utilisé le reasoner de la bibliothèque apache jena. Il nous a suffit de créer un model de notre graphe RDF de base, puis de créer un reasoner pour y charger les règles créées. Ce reasoner est ensuite chargé avec le modèle de base pour créer un modèle d'inférence qui est ensuite utilisé pour créer un nouveau fichier de données RDF contenant les données inférées du modèle.

aperçu du code :

```
Model model = ModelFactory.createDefaultModel();
model.read( "donnees.ttl" );
GenericRuleReasoner reasoner = new GenericRuleReasoner( Rule.rulesFromURL(
"rules.txt" ));
InfModel infModel = ModelFactory.createInfModel( reasoner, model );
```

Etape 4

Pour pouvoir lier nos données au cloud de linked data on a besoin de respecter 4 règles :

- utiliser des URIs pour nommer les éléments du dataset,
- utiliser des HTTP URIs pour rendre les URIs précédents accessibles et consultables par autrui,
- Quand quelqu'un cherche un URI, fournissez des informations utiles, en utilisant les standards (RDF *, SPARQL),
- Utiliser des liens vers d'autres URIs, pour permettre de trouver toujours plus de choses.

Tout d'abord, nos données sont bien au format RDF et nos ressources sont définis avec des URIs qui possèdent une adresse unique http. Par exemple:

```
BIND (URI(CONCAT('http://ex.org/Projet/', ?code_du_projet)) AS ?URI)
```

Ensuite, nous utilisons des liens vers d'autres URIs, ceux de dbpedia concernant les pays :

```
 BIND(IF(bound(?pays_du_partenaire),?pays_du_partenaire,"null") AS ?test)
  ?separation apf:strSplit(?test ";")
  BIND(IF(?separation = "null",?nothing,URI(CONCAT('http://dbpedia.org/page/',
?separation))) AS ?pays)
```

Nous avons également des ressources liées à doap et dbpedia grâce à nos prefix. A première vu, on pourrait donc dire que notre graphe rdf respecte toutes les règles pour lier nos données au cloud de linked data.

Etape 5

La description utilisant le vocabulaire VOID est présente dans le fichier VoID.txt. Il contient les informations basiques à propos du dataset. Chaque littéral est décrit en utilisant le vocabulaire.

Exemples :

Titre du dataset :

```
:Horizon20 dcterms:title "Appels à projets Horizon 2020 - Projets retenus et participants identifiés"
```

Participant au dataset :

```
:Horizon20 dcterms:publisher :Le_Bars_Yannis
```

Label de ce participant :

```
:Le_Bars_Yannis rdfs:label "Yannis Le Bars"
```

ANNEXE A

RÉSULTAT DE LA PREMIÈRE REQUÊTE

titre	duree	montant	rapport
"EUROfusion"	"60"	"856961937.57"	14282699.0
"WAYTOGO FAST"	"24"	"139300194.25"	5804174.5
"SeNaTe"	"36"	"181080566.0"	5030015.5
"TAKE5"	"36"	"150301979.75"	4175055.0
"H2020"	"24"	"89619171.0"	3734132.0
"HBP SGA1"	"24"	"89000000.0"	3708333.25
"GrapheneCore1"	"24"	"89000000.1"	3708333.25
"GN4-1"	"16"	"44177389.5"	2761086.75
"PowerBase"	"36"	"87613740.0"	2433715.0
"ENABLE-S3"	"36"	"68136162.28"	1892671.125
"IoSense"	"36"	"65269715.0"	1813047.625

RÉSULTAT DE LA SECONDE REQUÊTE

titre	difference
"First InnovativeWeek"	11
"MACIVIVA"	7
"PhiSi"	7
"SmartPosition"	7
"HElAIrcOPT"	7
"CYPRESS"	7
"COLOURTEST"	7
"CELLETEST"	5
"Theseus"	5
"TRANSFORMER"	4
"ERCSC-VPRES-SUP2014"	1

ANNEXE B

RÉSULTAT DE LA REQUÊTE SUR DONNÉES LIÉES

anneeDebut	pays	somme	moyenne
"2014"	<http://dbpedia.org/page/%C3%89tats-Unis>	"2037770.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2037770.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Allemagne>	"7.6685552E7"^^<http://www.w3.org/2001/XMLSchema#float>	"5477539.5"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Australie>	"20847771.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2.0847772E7"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Autriche>	"3.2529816E7"^^<http://www.w3.org/2001/XMLSchema#float>	"8132454.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Belgique>	"6.8868528E7"^^<http://www.w3.org/2001/XMLSchema#float>	"6260775.5"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Bulgarie>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Canada>	"2.532777E7"^^<http://www.w3.org/2001/XMLSchema#float>	"1.2663885E7"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Chypre>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Croatie>	"2082673.75"^^<http://www.w3.org/2001/XMLSchema#float>	"2082673.8"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Danemark>	"2.6847772E7"^^<http://www.w3.org/2001/XMLSchema#float>	"1.3423886E7"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Espagne>	"7.4664072E7"^^<http://www.w3.org/2001/XMLSchema#float>	"5743390.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Estonie>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Finlande>	"1.2671461E7"^^<http://www.w3.org/2001/XMLSchema#float>	"4223820.5"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/France>	"8.4681304E7"^^<http://www.w3.org/2001/XMLSchema#float>	"4032443.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Gr%C3%A8ce>	"6.4722204E7"^^<http://www.w3.org/2001/XMLSchema#float>	"8090275.5"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Hongrie>	"4.4137772E7"^^<http://www.w3.org/2001/XMLSchema#float>	"2.2068886E7"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Irlande>	"1.1E7"^^<http://www.w3.org/2001/XMLSchema#float>	"5500000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Isra%C3%ABl>	"3.3410444E7"^^<http://www.w3.org/2001/XMLSchema#float>	"8352611.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Italie>	"7.0959336E7"^^<http://www.w3.org/2001/XMLSchema#float>	"5913278.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Lettonie>	"2.929E7"^^<http://www.w3.org/2001/XMLSchema#float>	"1.4645E7"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Lituanie>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Malte>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Maroc>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Norv%C3%A8ge>	"1.1E7"^^<http://www.w3.org/2001/XMLSchema#float>	"5500000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Pays-Bas>	"6.189632E7"^^<http://www.w3.org/2001/XMLSchema#float>	"6189632.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Pologne>	"3.5201276E7"^^<http://www.w3.org/2001/XMLSchema#float>	"7040255.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Portugal>	"4.0201276E7"^^<http://www.w3.org/2001/XMLSchema#float>	"6700212.5"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/R%C3%A9p.%20Tch%C3%A8que>	"2830457.5"^^<http://www.w3.org/2001/XMLSchema#float>	"1415228.8"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Roumanie>	"2.9963316E7"^^<http://www.w3.org/2001/XMLSchema#float>	"9987772.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Royaume-Uni>	"4.992736E7"^^<http://www.w3.org/2001/XMLSchema#float>	"4538851.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Russie>	"6000000.04"^^<http://www.w3.org/2001/XMLSchema#float>	"6000000.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/S%C3%A9n%C3%A9gal>	"2908622.0"^^<http://www.w3.org/2001/XMLSchema#float>	"2908622.0"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Serbie>	"1671461.25"^^<http://www.w3.org/2001/XMLSchema#float>	"1671461.2"^^<http://www.w3.org/2001/XMLSchema#float>
"2014"	<http://dbpedia.org/page/Slov%C3%A9nie>	"9673157.0"^^<http://www.w3.org/2001/XMLSchema#float>	"4836578.5"^^<http://www.w3.org/2001/XMLSchema#float>