Universal-endpoint.com : une plateforme d'accès simple au Web des Données

Thomas Raimbault, Abdellah Sabry, Sonia Djebali

Léonard de Vinci Pôle Universitaire, De Vinci Research Center, ESILV, Paris La Défense
{thomas.raimbault,abdellah.sabry,sonia.djebali}@devinci.fr

Résumé. Universal-endpoint.com est une plateforme web permettant un accès simple au Web des Données par trois aspects : (i) une plateforme de correspondance, pour l'accès aux bases du Web des Données depuis un seul point d'accès centralisé, (ii) le langage SimplePARQL, pour une écriture intuitive de requêtes sous forme de triplets à la manière de SPARQL mais ne nécessitant pas une connaissance préalable des bases du Web des Données, et (iii) une aide à la rédaction de requêtes SPARQL.

1 Introduction

Le Web des Données – encore appelé Linked Data ou Web Sémantique (Berners-Lee et al., 2001) – est constitué de centaines de bases RDF (Klyne et al., 2014) inter-liées formant un vaste réseau de milliards de triplets RDF. Une base RDF est composée d'un ensemble de *triplets*, où chaque triplet s'exprime sous la forme (sujet, prédicat, objet). Les triplets peuvent être vus comme des phrases élémentaires sujet–verbe–complément, c'est à dire « Le 'sujet' a pour 'prédicat' la valeur 'objet' ». Chaque ressource est identifiée de manière unique au sein de la base RDF où elle est stockée. Les identifiants sont généralement des IRI [1] pour un accès à travers le Web [2]. Pour récolter et manipuler les données d'une base RDF, SPARQL (Prudhommeaux et Seaborne, 2008) est le langage de requêtes recommandé par le W3C. L'écriture d'une requête SPARQL reste cependant difficile pour la plupart des utilisateurs potentiels du Web des Données. En effet, une des raisons principales est qu'il est souvent nécessaire de connaître les IRI des ressources et propriétés manipulées pour pouvoir interroger les bases.

Notre contribution, avec la plateforme universal-endpoint.com, pour l'interrogation du Web des Données est triple. Premièrement, l'utilisateur peut rédiger des requêtes en SimplePARQL, à la manière de SPARQL où des *ressources imprécises* peuvent être utilisées au sein de triplets – en sujet, en prédicat et/ou en objet. Deuxièmement, la plateforme agit comme une plateforme de correspondance depuis laquelle l'utilisateur peut accéder à différentes bases du Web des Données à la fois. Troisièmement, l'utilisation de SimplePARQL peut-être vu comme une aide à la rédaction de requêtes SPARQL. La figure FIG. 1 présente le schéma de fonctionnement de la plateforme universal-endpoint.com.

Cet article est organisé comme suit. La Section 2 présente les différents services proposés par la plateforme universal-endpoint.com. La Section 3 conclue cet article.

1. *Internationalized Resource Identifier*, norme RFC 3987 (2005) généralisant les adresses URI.
2. Seuls les « nœuds blancs » (*blank nodes*) sont des ressources uniquement accessibles localement à la base.

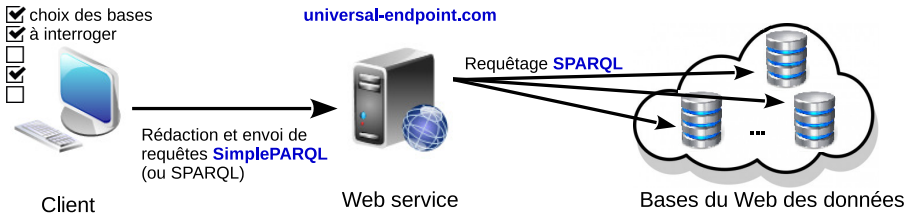

FIG. 1 – *Fonctionnement de la plateforme* `universal-endpoint.com`

2 Les services de la plateforme universal-endpoint.com

2.1 Requêtes SimplePARQL

Dans les bases RDF du Web des Données, les ressources sont décrites par leurs liens aux autres ressources et leurs liens à des valeurs littérales. La sémantique des bases RDF est donc contenue dans ces relations. Cependant, il y a un écart entre la représentation structurée que perçoit l'utilisateur et celle physiquement présente dans une base RDF. Par exemple, l'information (Einstein, lieu-naissance, Ulm) pour l'utilisateur est concrètement stockée sous la forme (ressource1,propriété2,ressource3), (ressource1,foaf:name,"Einstein"), (propriété2,rdfs:label,"birth place"), (ressource3,rdfs:label,"Ulm"). Le rôle de l'approche SimplePARQL est de réduire cet écart. Pour cela, SimplePARQL est une généralisation du langage SPARQL, utilisant quasiment la même syntaxe et la même grammaire mais où les différentes composantes d'un triplet – le sujet, le prédicat et/ou l'objet – peuvent être des *ressources imprécises* (en plus de pouvoir être des variables, des IRI ou éventuellement des blank nodes).

Le premier travail théorique sur SimplePARQL date de 2015 (Djebali et Raimbault, 2015), et à cette même date un premier moteur de requêtes SimplePARQL, peu fonctionnel, était disponible. La présentation faite ici par l'exemple repose sur les fonctionnalités de la nouvelle version du moteur SimplePARQL disponible sur universal-endpoint.com depuis octobre 2017.

Supposons que nous cherchions le lieu de naissance d'Albert Einstein. Dans *DBpedia*, on utiliserait la requête SPARQL REQ. 1, tandis que dans *Wikidata* la requête REQ. 2. Dans les deux cas, la connaissance de (l'identifiant de) la ressource Einstein et de (l'identifiant de) la propriété birth place est nécessaire. Si ce n'est pas le cas, l'utilisateur – expérimenté – sera amené à écrire une requête plus complexe ou à écrire des requêtes préliminaires pour l'obtention de ces identifiants. Mais alors la vision simple et suffisante sous la forme d'une phrase élémentaire exprimée par l'unique triplet « Albert Einstein a pour lieu de naissance la valeur inconnue recherchée » n'est plus perçue par l'utilisateur.

REQ. 1 – requête SPARQL pour *DBpedia*

```
SELECT ?p  WHERE {
<http://dbpedia.org/resource/Albert_Einstein> <http://dbpedia.org/property/birthPlace> ?p
}
```

REQ. 2 – requête SPARQL pour *Wikidata*

```
SELECT ?p  WHERE {
   <http://www.wikidata.org/entity/Q937> <http://www.wikidata.org/entity/P19> ?p
}
```

Avec SimplePARQL, l'utilisateur a la possibilité d'utiliser en plus de ce qui est possible avec SPARQL des *ressources imprécises* au sein des triplets. Cela peut être :

1. Un mot unique (*e.g.* Einstein) : toutes les ressources dans la base interrogée qui sont liées – via un certain prédicat – à un littéral contenant ce mot doivent être retournées.
2. Plusieurs mots entre barres obliques (*e.g.* /birth place/) : toutes les ressources liées à un littéral possédant ces mots (ordre et casse insensibles) doivent être retournées.
3. Un ou plusieurs mots entre double quotes (*e.g.* "Albert Einstein") : toutes les ressources liées à cette expression exacte doivent être retournées [3].

✎ A chaque fois, le filtrage sur la langue peut être précisé (*e.g.* /Einstein Albert/@de).

Concrètement, une requête SimplePARQL est réécrite en un ensemble de requêtes SPARQL offrant la possibilité de *matcher* une ressource imprécise avec des ressources RDF dans la/les base/s interrogée/s. En fonction de la position de la ressource imprécise dans le triplet, les requêtes SPARQL sont générées selon des règles de réécriture et selon certaines priorités que nous avons définies (non présentées ici, mais disponibles en ligne sur la plateforme). Les différents résultats et les différentes pages de résultats sont obtenus selon ces stratégies de réécriture.

La requête REQ. 3 est un exemple de requête SimplePARQL contenant deux ressources imprécises : /Einstein Albert/ et /birth place/. Les résultats de cette requête, fournis par la plateforme universal-endpoint.com, sont présentés en figure FIG. 2 où en plus des valeurs trouvées pour la variable ?p (comme en SPARQL) la plateforme apporte la précision sur les correspondances qui ont été trouvées pour chaque ressource imprécise.

REQ. 3 – requête SimplePARQL (multi-bases).

```
SELECT ?p  WHERE { /Einstein Albert/  /birth place/ ?p }
```

p	Albert Einstein	birth place
http://dbpedia.org/resource/German_Empire	http://dbpedia.org/resource/Albert_Einstein has for label Albert Einstein@en	http://dbpedia.org/ontology/birthPlace has for label birth place@en
http://dbpedia.org/resource/German_Empire	http://dbpedia.org/resource/Albert_Einstein has for label Albert Einstein@de	http://dbpedia.org/ontology/birthPlace has for label birth place@en
http://dbpedia.org/resource/Ulm	http://dbpedia.org/resource/Albert_Einstein has for label Albert Einstein@en	http://dbpedia.org/ontology/birthPlace has for label birth place@en
http://dbpedia.org/resource/Kingdom_of_Württemberg	http://dbpedia.org/resource/Albert_Einstein has for the proprety http://xmlns.com/foaf/0.1/name the value Albert Einstein@en	http://dbpedia.org/ontology/birthPlace has for label birth place@en
http://dbpedia.org/resource/German_Empire	http://dbpedia.org/resource/Albert_Einstein has for the proprety http://dbpedia.org/property/name the value Albert Einstein	http://dbpedia.org/ontology/birthPlace

FIG. 2 – *Extrait des réponses en interrogeant DBpedia par la requête SimplePARQL* REQ. *3*

Il est intéressant de noter que SimplePARQL cherchant à faire coïncider une ressource imprécise en explorant les voisinages des ressources dans la base RDF questionnée, on peut écrire sa requête par exemple en utilisant le français pour décrire une ressource imprécise. Ainsi la requête SimplePARQL "SELECT ?p WHERE{/Einstein Albert/ /lieu naissance/ ?p}" a des chances d'aboutir si un label en français est associé à la propriété correspondante dans la base.

2.2 Plateforme de correspondance

La plateforme universal-endpoint.com agit comme une plateforme de correspondance (un point d'accès central, un *hub*) depuis laquelle l'utilisateur peut accéder à différentes bases du

3. Si l'expression est en position d'objet, il s'agit d'un littéral en SPARQL (et non d'une ressource imprécise).

Web des Données. La plateforme se charge d'interroger en SPARQL les bases sélectionnées via leurs *endpoints* SPARQL publics, puis de centraliser les réponses. L'utilisation de SimplePARQL trouve toute sa place ici, puisque les requêtes peuvent être écrites sans nécessité l'usage d'IRI qui sont propres à une base, et donc une même requête SimplePARQL peut être utilisée pour interroger plusieurs bases différentes.

2.3 Aide à la rédaction de requêtes SPARQL

L'utilisation de SimplePARQL peut-être vu comme une aide à la rédaction de requêtes SPARQL où après chaque requêtage en SimplePARQL, l'utilisateur peut choisir la ressource qu'il souhaite manipuler en lieu et place d'une ressource imprécise. Pour réaliser ceci, sur la figure FIG. 2 l'utilisateur clique simplement sur la double flèche correspondant à la ressource souhaitée pour remplacer la ressource imprécise dans la requête SimplePARQL. Ainsi, à partir d'une requête SimplePARQL l'utilisateur obtient au final (éventuellement par itérations successives) une requête SPARQL – sans ressource imprécise, puisque toutes désambiguïsées.

3 Conclusion

La plateforme universal-endpoint.com propose un ensemble de services pour un accès simplifié aux données de Web des Données. Les travaux futurs sont dans le groupement de résultats intra et inter-bases lorsque les ressources résultats correspondantes aux ressources imprécises sont similaires (soit même IRI soit ressources liées par une propriété owl:sameAs).

Références

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The Semantic Web. *Scientific American 279*(5), 34–43.

Djebali, S. et T. Raimbault (2015). SimplePARQL : A New Approach Using Keywords over SPARQL to Query the Web of Data. In *Proc. of SEMANTICS'15*, pp. 188–191. ACM.

Klyne, G., J. J. Carroll, et B. McBride (2014). RDF 1.1 Concepts and Abstract Syntax. `http://www.w3.org/TR/rdf-concepts/`.

Prudhommeaux, E. et A. Seaborne (2008). SPARQL Query Language for RDF. `www.w3.org/TR/rdf-sparql-query/`.

Summary

Universal-endpoint.com is a web platform allowing easy access to the Web of Data for three reasons: (i) it acts as a hub platform, to access to semantic web databases from a single point centralized access, (ii) the SimplePARQL is a SPARQL-like language that allows more intuitive writing of queries, always in the form of triplets but without requiring prior knowledge of the databases, and (iii) help writing SPARQL queries.