Web sémantique

De plus en plus de données étant disponibles sur internet, il y a une réelle nécessité de traiter et d'interpréter ces données. Or ces données ne sont pas toujours structurées et donc difficilement lisibles par la machine. On s'oriente alors vers une transition d'un web dit « sémantique ».

Le Web Sémantique c'est quoi ?

Le web sémantique, ou ***Web 3.0*** est souvent appelé le web des données.
Le but est de pouvoir utiliser le web comme une immense base de données, que l'on peut enrichir (les projets gouvernementaux fournissent des jeux de données utilisables par tous), ou utiliser (recherche d'information pour enrichir sa propre base de données).
L'idée est de parvenir à un Web intelligent, où les pages des sites seraient gérées par une base de données intelligente ce qui permettrait de mettre les objets au service des personnes.

Pour cela il faut que les données soient interopérables et soient accompagnées de leur sémantique (***ontologie***). L'ontologie constitue un modèle de données qui représente les concepts d'un domaine ainsi que les relations entre ces concepts. Les ontologies permettent de représenter la donnée pour qu'elle soit compréhensible par un ordinateur. Il s'agit d'une sorte d'étiquette : Usain Bolt est une personne et un athlète, il s'agit là de deux ontologies.

En plus de données interopérables, il est nécessaire d'utiliser un dispositif langagier normalisé afin qu'il soit utilisable par tous de la même manière, ce qui permettra l'accès universel et l'utilisation intelligente de la donnée par les ordinateurs.

Normaliser le format des données

Les données doivent être formalisées sous un format standard.

Pour cela on utilise un graphe appelé ***RDF*** (= Resource Description Framework) qui est un modèle de données composé du triplet suivant :
- Le ***sujet*** (l'élément à décrire)
- Le ***prédicat*** (une propriété de cet élément)
- L'***objet*** (la valeur de cette propriété, qui est un autre élément)

On peut les voir comme une phrase reliant 2 éléments, comme par exemple "U2 a écrit l'album 'Songs of Innocence' ".
U2 est le sujet, 'a écrit l'album' est le prédicat et 'Songs of Innocence' est l'objet.

Le but du web sémantique étant de relier des données du monde entier, ces 3 éléments doivent être uniques. RDF est donc un triplet d'***URI*** (Uniform Resource Identifier). Le but principal d'une URI est justement de fournir un nom universellement unique à une ressource afin qu'il soit possible de lier des données de différentes sources partout dans le monde. Cela permet également d'ajouter une sémantique à l'information, c'est-à-dire que l'information est décrite dans l'URI.

RDF est un langage qui permet de stocker de l'information sous une certaine forme. Il existe plusieurs formats afin de sauvegarder les triplets RDF en un flux d'octets (***sérialisation***), les trois principaux étant ***RDF/XML*** (.rdf),qui est la syntaxe originale mais plus verbeuse que les autres, ***N3*** (.n3) et ***Turtle***(.tll), ce dernier étant un standard W3C. La plupart des outils savent jongler entre ces trois formats, notamment en reconnaissant l'extension.

Interroger une base de données

Lorsqu'une base de données est créée, il faut aussi savoir l'interroger pour récupérer des informations. Pour cela il faut interroger le serveur à l'aide d'un langage de requête. Chaque requête permet de chercher des informations dans les ontologies. Le langage le plus connu et le plus utilisé est SPARQL.
SPARQL permet principalement de chercher de l'information dans des bases de données structurées, d'où l'intérêt de structurer l'information.

Dans le cas de corpus de texte non structurés, il est également possible d'extraire de l'information, en utilisant le machine learning. On appelle cela le ***text mining*** ou text analytics. La traduction française moins communément utilisée est l'extraction de connaissances ou fouille de textes.
Le text mining permet de transformer un texte non structuré en texte structuré par l'analyse d'une collection de ressources écrites. Il permet d'identifier les faits, les relations entre ces faits ainsi que les assertions.
L'immense avantage de ce concept par rapport à la recherche par mots clefs (keyword search) est qu'il permet de reconnaître des concepts similaires même quand ils sont exprimés différemment.

Vocabulaire standard

On a dit que pour construire une base de données, celle-ci doit respecter le schéma RDF recommandé par W3C afin de construire un Web normalisé, compréhensible de tous. Cependant, il est également nécessaire d'avoir un langage standard, construit sur le modèle de données RDF. Ce langage, appelé ***Web Ontology Language*** (**OWL**) permet de définir des ontologies web structurées.

Toujours dans la perspective d'une utilisation universelle des données, des ontologies web écrites avec le standard OWL ont été créées pour être utilisées par tous. Cela permet d'éviter à chacun de créer tout un ensemble d'ontologies lors de la création d'une base de données, mais également cela permet à tout le monde d'avoir les mêmes définitions de concepts.

Voici une liste non exhaustive des principales ontologies, accompagnées du préfixe communément utilisé. A noter qu'il est possible de choisir n'importe quel préfixe étant donné que c'est une abréviation. Utiliser les préfixes standards permet simplement une compréhension plus universelle de son code.

- Le ***FOAF***, littéralement Friend Of A Friend, permet de décrire des entités. Elle fournit des informations diverses sur des personnes ou choses (quelles qu'elles soient) ainsi que leurs relations.
  foaf: http://xmlns.com/foaf/0.1/

- Le ***Dublin Core*** permet de décrire des ressources numériques ou physiques (titre, créateur, éditeur, sujet, langue, description, format, date ..)
  dc: http://purl.org/dc/elements/1.1/
  dct : http://purl.org/dc/terms/
  Par exemple http://purl.org/dc/elements/1.1/title est le titre d'un document, d'un livre.
- Le vCard de W3 qui concerne le monde des affaires :
  v :http://www.w3.org/2006/vcard/
  Par exemple http://www.w3.org/2006/vcard/title est l'intitulé du poste d'une personne.
- Les relations entre objets liés au schéma RDF ainsi que la syntaxe des éléments.
  rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
  rdfs: http://www.w3.org/2000/01/rdf-schema#
- Pour le format d'un element
  xsd: http://www.w3.org/2001/XMLSchema#
  Par exemple "4"^^xsd:integer

Où la donnée est-elle stockée ?

En théorie, on pourrait accéder à toutes les données du Web avec le langage standard SparQL. Cependant ce n'est pas le cas pour le moment puisque seule une partie des fichiers du Web est standardisée au format RDF.
Néanmoins certaines personnes ou institutions, ont créé des fichiers ayant pour vocation d'être utilisés universellement. Il s'agit donc en quelques sortes de bases de données de ressources et de propriétés libres. Ces fichiers respectent le schéma RDF ainsi que les standards OWL.
Ces bases de données (SGBD) relationnelles spécifiques au format des triplets sont appelées RDF RDBMS en anglais (SGBD=RDMS) Triplestore ou Quadstore.
Quelques exemples : OpenLink Virtuoso, AllegroGraph, Stardog, Neo4J, MarkLogic

Ces RDF RDBMS sont stockées sur internet pour être accessibles, à l'aide de ***SPARQL endpoint***, ou point de terminaison. Cela fonctionne comme pour les moteurs de recherche classiques : on a une URL racine (base URL) et on lui ajoute les paramètres de la requête.

Si elles ne sont pas mises à jour, les bases de données deviennent obsolètes. Or mettre à jour une base de données prend du temps et ne peut pas être fait toutes les minutes. Sur chaque point de terminaison est stockée une base de données à une version v, à un instant t. Il est donc important de choisir un point de terminaison récent lorsqu'on souhaite interroger une base de données.

DBpedia possède différents SPARQL endpoint dont l'un est mis à jour régulièrement: http://live.dbpedia.org/sparql . DBpedia live est le seul point de terminaison qui mette à jour régulièrement la base en ne mettant à jour que les données modifiées.

Voici d'autres SPARQL endpoints mais dont la mise à jour est en général bien antérieure à celle de DBpedia Live :
http://dbpedia.org/snorql/
http://demo.openlinksw.com/sparql
http://librdf.org/query/

Il est normalement impossible d'interroger une base de données d'un autre point de terminaison que celui que vous utilisez sauf si vous incluez des informations à ce propos par exemple à l'aide de « SparQL federation ». Ce dernier permet d'interroger plusieurs points de terminaison SparQL en même temps pour obtenir un résultat combiné.

Aide pour construire des requêtes

Les points ci-dessous sont quelques clés pour écrire une requête SparQL.

Si l'on souhaite récupérer des informations de Wikipédia, il faut chercher l'élément sur DBpédia de la manière suivante :

- Récupérer l'orthographe exacte sur Wikipédia : https://en.wikipedia.org/wiki/U2 on récupère U2
- L'inclure dans l'URL DBpédia : live.dbpedia.org/page/U2

Pour avoir des informations sur les ontologies :
http://mappings.dbpedia.org/server/ontology/classes/

Pour connaitre le lien possible entre 2 objets, regarder la catégorie Rdfs :domain
Pour ce qui est de l'objet du lien Rdfs :range

Pour connaitre les propriétés d'une classe :

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select distinct ?property where {
?property rdfs:domain <http://dbpedia.org/ontology/Person> . }
```

Pour connaître la classe mère ou classe fille d'un élément : rdfs:subClassOf

Ex :

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT * WHERE {
?game a dbo:Game .
?subclass rdfs:subClassOf dbo:Game .
}
```

Sources :

- bases du web sémantique :
http://www.zeblogsante.com/web-3-0-definition/
http://infomesh.net/2001/swintro/
https://www.youtube.com/watch?v=r795n3AffgA&list=PL2gqTPNkhkw9sjvDDK4joBZpOc690DELl&index=2

-Langages de requêtes (SPARQL):
https://www.dropbox.com/s/zgvxvna132yhj1h/DuCharme_Learning%20SPARQL.pdf?dl=0
https://en.wikipedia.org/wiki/RDF_query_language

https://www.youtube.com/watch?v=FvGndkpa4K0&index=3&list=PL2gqTPNkhkw9sjvDDK4joBZpOc690DELl

-DBpedia:
https://www.youtube.com/watch?v=BmHKb0kLGtA&index=4&list=PL2gqTPNkhkw9sjvDDK4joBZpOc690DELl

-Base de données :
https://en.wikipedia.org/wiki/Triplestore

-Ontologies et vocabulaire
https://fr.wikipedia.org/wiki/Dublin_Core
https://fr.wikipedia.org/wiki/Web_Ontology_Language

-Text mining :
https://archinfo41.hypotheses.org/571
https://www.youtube.com/watch?v=f7XN3RuDzmc&list=PL2gqTPNkhkw9sjvDDK4joBZpOc690DELl&index=1

-Endpoint :
https://stackoverflow.com/questions/48229394/why-different-endpoints-do-not-query-same-datasets

https://bibliotheques.wordpress.com/2012/11/19/sparql-endpoint/
https://bibliotheques.wordpress.com/2012/06/14/sparql-premier-pas-installation-de-loutil-2/