Web des données
TP n° 6 – Web des Données & RDF

Romain LELONG – romain.lelong@gmail.com

Avant de commencer ...

Outre le présent sujet, les fichiers `guillaume_canet_simple.rdf` et `marion_cotillard_simple.rdf` sont à récupérer sur Universitice avant de débuter le TP.

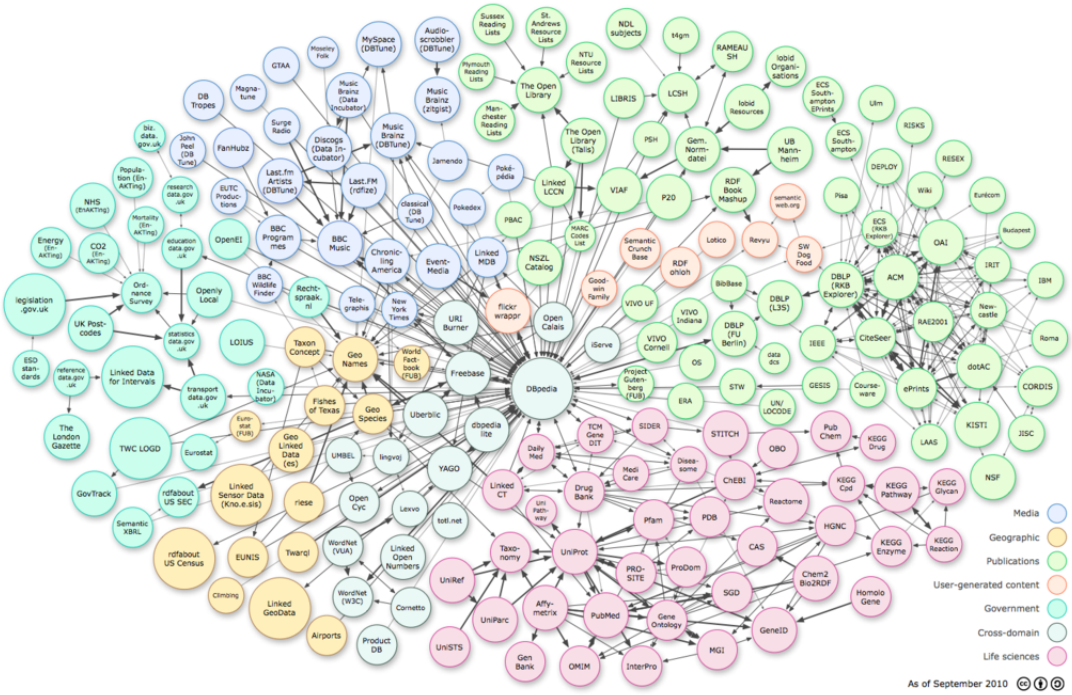
Ce TP ne sera pas noté et aucun compte-rendu n'est demandé. Vous pouvez le réaliser en binôme ou seul. Les divers éléments abordés dans celui-ci pourront nénanmoins faire l'objet d'une micro-évaluation prochaine.

1 DBpedia

DBpedia est un projet initialement mené par deux chercheurs d'universités allemandes (Berlin et Leipzig) et maintenant soutenu par une communauté internationale. C'est donc une base de connaissance en ligne dont le but est d'extraire les informations de *Wikipédia* et de les structurer au format du web sémantique, afin de les rendre accessibles sous forme de données liées et interconnectées..

DBpedia s'inscrit pleinement dans le « *linked open data* ». Il est interconnecté avec d'autres ensembles de données ouvertes provenant du Web des données.

En pratique, DBPedia met à disposition ses resources au sein de dépôt. Ces resources sont dérivés de l'encyclopédie *Wikipédia*. Ainsi, pour chaque document encyclopédique, il existe une ressource *DBPedia* contenant (en partie) les mêmes données mais sous un format propre au web sémantique.



2 À vous de jouer

Étape n° 1 : Dessiner un schéma incluant les éléments suivants :

1. Triplestore
2. Requête SPARQL
3. Graphe RDF
4. Réponse
5. protocole HTTP

Indication : *Vous possédez un cerveau ! Servez vous en ... pour l'instant, il fonctionne encore mieux que ChatGPT !*

Étape n° 2 : Il apparaît assez évident que l'accès à un *triplestore* (illustré dans le schéma de l'étape n° 1) présente des similitude à l'accès à un Système de Gestion de Base de Données Relationnel classique. Quel brique de l'architecture dépeinte à l'étape n° 1 présente cependant selon vous une plus-value majeurs par rapport à une architecture basée sur un SGBDR ?

Étape n° 3 : Dans la section précédente, qu'entend t-on par « *au format du web sémantique* » ?

Étape n° 4 : Ouvrir, dans le navigateur Web Firefox, la page Wikipedia dédiée à l'actrice *Marion Cotillard*. Choisir la langue Anglaise. (Noter l'URL de cette page)

Étape n° 5 : Pour accéder aux données *DBPedia* relatives à une ressource/page *Wikipedia*, il est simplement nécessaire de remplacer dans l'URL « `en.wikipedia.org/wiki` » par « `dbpedia.org/resource` ». Par exemple :

`https://en.wikipedia.org/wiki/Semantic_Web` ⟶ `https://dbpedia.org/resource/Semantic_Web`

Accéder à la ressource *DBpedia* correspondant à *Marion Cotillard* à l'aide de votre navigateur ?

Étape n° 6 : Sous quel format le navigateur renvois t-il la réponse (i.e. XML, JSON, HTML etc.) ? L'URL sur laquelle vous aboutissez est-elle celle renseignée initialement ? Pourquoi ce résultat n'est pas celui auquel ou pouvait s'attendre ? Quelle conjecture peut-on faire sur la cause de ce résultat ? Peut-on néanmoins parvenir au résultat escompter ?

Étape n° 7 : Dans le contexte du Web Sémantique, à quoi correspondent (i.e. quels rôles jouent) chacune des deux colonnes `Property` et `Value` du tableau présentés dans la réponse obtenu ?

Étape n° 8 : Pourquoi pourrait-on considérer qu'il manque une colonne à ce tableau ? Pourquoi, à votre avis cette colonne n'est elle pas présente ?

Étape n° 9 : Démarrer le module « *Outils de développement web* » de Firefox puis recharger la ressource `DBpedia` afin d'observer l'entête de la requête HTTP générée (Firefox » Outils de développement web » Réseau voir Figure 1). Renseigné vous sur les rôles des entêtes :

— « `Accept` » dans les requêtes HTTP ?

— « `Content-Type` » dans les réponses HTTP ?

Étape n° 10 : Le module « Outils de développement web » de Firefox permet de modifier l'entête « `Accept` » des requête HTTP. Renvoyer la requête HTTP permettant d'accéder à la ressource *DBPedia* de *Marion Cotillard* en spécifiant dans l'entête :

```
Accept: application/rdf+xml
```

Attention : *Attention à bien spécifier l'URL correspondant à la ressource DBPedia. L'URL en question doit bien débuter par* `https://dbpedia.org/resource/...` *et non par* `https://dbpedia.org/page/...`.
Attention : *Il est possible que Firefox bloque les contenus mixtes et que le contenu de la réponse HTTP ne soit par conséquent pas directement accessible depuis l'onglet « Réseau ». Dans ce cas il est conseillé de jeter un œil dans l'onglet « Console ».*

Étape n° 11 : Enregistrer la réponse obtenue dans un fichiers nommé `marion_cotillard_full.rdf`.

Étape n° 12 : Quel syntaxe de sérialisation est utilisée dans ce fichier ?

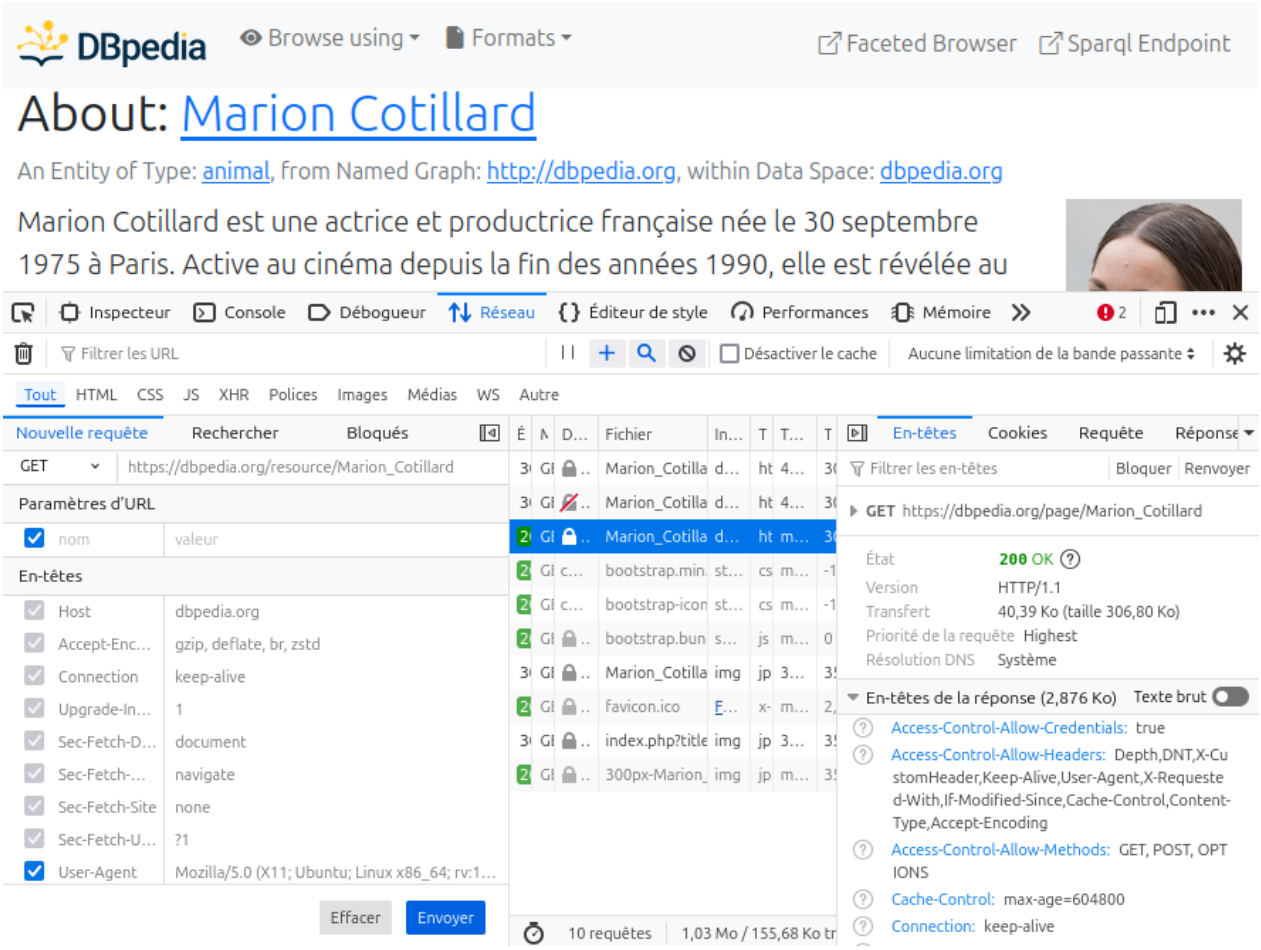


FIGURE 1 – Visualisation et/ou modification d'une requête HTTP sous Firefox

Étape n° 13 : Dans cette étape, on souhaite à nouveau récupérer la ressource *DBPedia* de *Marion Cotillard*. Cependant, on se propose maintenant d'utiliser l'outil *client URL request library* (*cURL*). Comme son nom l'indique, ce dernier est un outil en ligne de commande qui permet d'effectuer des requête HTTP. Ainsi pour effectuer une requête HTTP vers une URL *url* il suffit d'utiliser la ligne de commande :

`curl -L` *options url*

Le paramètre *options* est facultatif mais peut notamment contenir les options suivantes :

— L'option « `-H "`*header*`"` » permet de spécifier l'entête *header* dans le requête HTTP.

— L'option « `-o` *file.ext* » permet d'enregistrer le résultat dans un fichier de nom *file.ext*.

Utiliser *cURL* afin d'obtenir la ressource DBPedia correspondant à *Marion Cotillard* au format Turtle (MIME type `text/turtle`) puis enregistrer la réponse obtenue dans un fichiers nommé `marion_cotillard_full.ttl`.

Étape n° 14 : Le fichier `marion_cotillard_simple.rdf` fourni avec le TP est une version simplifiée du fichier obtenu à l'étape n° 11 dans lequel seules certaines informations ont été préservées. Extraire de ce fichier les portions qui contiennent les informations permettant de répondre aux questions suivantes puis réunir ce portions au sein d'un fichier `etape-14.rdf` autonome et bien formé :

— *Qui est le companion de Marion Cotillard ?*

— *Quelle est la date de naissance de Marion Cotillard ?*

Étape n° 15 : Remplir un tableau de même structure que celui ci-dessous à l'aide des triplets RDF apparaissant dans le document `etape-14.rdf`.

sujet	prédicat	objet	Information véhiculée par le triplet
...	...	...	...
...	...	...	...
...	...	...	...

Remarque : *On veillera à ce que les valeurs des colonnes sujet, prédicat et objet soit* ***complètes*** *et de* ***bonne nature****. On rappel qu'il s'agit ici d'un graphe RDF.*

Étape n° 16 : Le W3C fourni un service de validation RDF en ligne accessible à l'URL :

`https://www.w3.org/RDF/Validator/`

En utilisant ce service vérifier que le document `marion_cotillard_simple.rdf` est un document RDF valide. Générer le graphe correspondant à ce dernier.

Étape n° 17 : Le fichier `guillaume_canet_simple.rdf` est, à l'instar de du fichier `marion_cotillard_simple.rdf`, un extrait des données RDF disponible pour la ressource DBpedia relative à l'acteur *Guillaume Canet*. Réunir les données des deux fichiers au sein d'un seul et même fichier intitulé `merge.rdf`.

Étape n° 18 : Vérifier que le fichier constitué est un fichier RDF valide.

Étape n° 19 : Expliquer en quoi le fichier `merge.rdf` illustre le principe de *données liées* ? Par le biais de quel(s) mécanisme(s) ce principe est-il rendu possible ?
Indication : *Une réponse à la fois complète et synthétique rédigée par vos soin est attendue.*

Étape n° 20 : À l'aide d'une recherche sur internet déterminer le nom, le prénom et la date de naissance des enfants de Marion Cotillard et de Guillaume Canet. Modifer le fichier `merge.rdf` afin d'y ajouter ces informations.
Indication : *On utilisera l'espace de nommage* ***`https://www.univ-rouen.fr#`*** *pour l'ajout de ces informations.*