Introduction Aux *Linked Open Data*

D'après Blaney Jonathan, *Introduction to the Principles of Linked Open Data*, ProgrammingHistorian.org, eds. Adam Crymble

Changement de Slide

Que sont les Linked open data

> Les LOD sont des informations structurées écrites dans un format pensé pour les machines, et pour cette raison ne sont pas nécessairement très agréables pour l'oeil humain.

Changement de Slide

Quand on visite un site web, on retrouve en effet beaucoup d'informations. Sil l'on allait sur la page de Sénèque sur Wikipedia, on verrait rapidement les informations suivantes : c'est un auteur, il est né vers 4 av. JC, il meurt en 65 et c'est un stoïcien. L'information est claire, lisible. Maintenant, je vais sur le site de la BNF pour lire une édition des Lettres à Lucilius. Puis je visite le site de Perseus qui contient une édition de ce texte. Pour moi, humain, la connexion entre ces différentes ressources est claire.

Changement de Slide

Malheureusement, ce n'est pas le cas pour la machine qui doit s'efforcer d'extraire ce sens à grand coup d'apprentissage automatique. Et l'apprentissage automatique n'étant pas parfait, les développeurs ont prévu un autre système.

Changement de Slide

Les projets mettent dorénavant des données à disposition : les données sont structurées suivant les mêmes standards et sont disponibles à la lecture. On mets alors ces données en groupe : on appelle cela plus communément des *datasets* et les projets hébergeurs des silos d'informations.

Pour créer ces silos et sets de données, il faut respecter **trois principes** :

1. **Utiliser un format standard** de *linked open data*: pour que les machines puissent communiquer entre elles, les données doivent être structurées. Vous avez vu jusqu'ici la structuration du texte via par exemple le XML et une grammaire. Il s'agit ici de faire la même chose, mais pour la description d'objet.
2. **Utiliser des référentiels communs** : le principe des LOD est de fournir un ensemble d'informations sur les mêmes objets. Tout cela est rendu plus facile lorsque ces objets portent le même nom.
3. **Publier ses données de manière ouverte** : pour que les données soient liables, il ne faut pas avoir à se connecter ou à payer un abonnement pour pouvoir les lire.

Changement de slide

Video Europeana

Europeana est un projet européen de bibliothèque numérique en ligne. Cette bibliothèque n'héberge aucune ressource, mais sert de portail vers les bibliothèques nationales et locales grâce à l'usage de Linked Open Data et de standards d'échange de catalogue. On retrouve par exemple les ouvrages de Gallica.

https://youtu.be/oEuDaJjEFos

Changement de slide

Vous avez dit RDF ?

RDF : Resource Data Framework

Avant d'expliquer ce qu'est le RDF, pouvez-vous me dire quelle est la structure de phrase verbale complète la plus commune ?

Changement de slide

Sujet-Verbe-COD. C'est un peu de cette structure qu'est partie le concept de RDF où les choses sont structurées en sujet prédicat objet.

Changement de Slide

Et donc si on applique au travail précédent, on aura : - Sénèque a écrit Les Lettres à Lucilius - Les lettres à Lucilius sont appelées Ad Lucilium

Faire un exemple ensemble

Le problème de ces exemples c'est qu'ils signifient quelque chose pour nous, personnes parlant le français. À votre avis, que peut-on faire pour aider la machine à comprendre cela ?

Changement de Slide

La première chose, c'est évidemment d'utiliser un référentiel. Là pour le moment, on a 2 objets distincts : Sénèque et Les Lettres à Lucilius. Étant donné que l'on a à faire à des objets traditionnels des catalogues, il s'avère que des identifiants ont déjà été fournis : il s'agit des identifiants du VIAF (Virtual International Authority File).

En l'occurrence, si on cherche Sénèque sur le VIAF on tombe sur cette page https://viaf.org/viaf/90637919/#Seneca,_Lucius_Annaeus,_approximately_4_B.C.-65_A.D . Que pouvez-vous dire sur cette page ?

- Il y a des traductions du nom de l'auteur
- Il y a des liens vers les oeuvres

Du coup, si on cherche Ad Lucilium, on trouve aussi la page https://viaf.org/viaf/184199909/#Seneca,_Lucius_A
65_A.D.|_Epistulae_morales_ad_Lucilium .

Changement de Slide

On a en fait avec ces deux pages deux identifiants : https://viaf.org/viaf/90637919 et https://viaf.org/viaf/184199909 . Maintenant, on peut écrire :

- https://viaf.org/viaf/90637919 a écrit https://viaf.org/viaf/184199909
- https://viaf.org/viaf/184199909 s'appelle Ad Lucilium

Changement de Slide

C'est bien, mais on n'est pas encore très clair pour une machine. Typiquement, une machine est habituée à quelque chose comme `x=w`. Et c'est ce qu'on va lui fournir. Car "=", ce n'est qu'un prédicat parmi tant d'autres.

C'est dans le cadre de la traduction des prédicats que s'insèrent les ontologies. Une ontologie, un vocabulaire, c'est un ensemble de verbes et d'objets qui permettent de compléter, de manière commune, ces phrases qui définissent nos objets. Par exemple : http://purl.org/dc/elements/1.1/creator

Changement de Slide

Ainsi on aura pour la phrase :

https://viaf.org/viaf/184199909 http://purl.org/dc/elements/1.1/creator https://viaf.org/viaf/90637919

Mais vu qu'on décrit un nom pour "Ad Lucilium", on va simplement utiliser une chaîne de caractère :

https://viaf.org/viaf/184199909 http://purl.org/dc/elements/1.1/title "Ad Lucilium"

Ces éléments sont des triplets, et c'est ce qui correspond à ce que l'on entend en général par RDF.

Quelques ontologies et référentiels

DC

Une ontologie produite pour les catalogues par la Dublin Core Initiative.

http://dublincore.org/documents/dcmi-namespace/

FOAF

Une ontologie produite pour les relations humaines et la description de personnes vivantes :

http://xmlns.com/foaf/0.1/

SAWS

Une ontologie produite pour décrire des sources anciennes

http://purl.org/saws/ontology

LAWD

Une ontologie spécialisée pour les mondes antiques

http://lawd.info/

SNAP

Une ontologie pour la prosopographie ancienne

http://snap.dighum.kcl.ac.uk/img/OwlVizImage.png et http://data.snapdrgn.net/ontology/snap

DBPedia : la ressource pour les objets

http://dbpedia.org/class/Book

Pleiades

Un référentiel de lieux : http://pleiades.stoa.org

Geonames

Référentiel de lieux modernes : http://www.geonames.org/

Exercice

Prendre un auteur parmi ceux proposés Dans un fichier excel, ordonner ses données sur les personnes avec le sujet à gauche. Exemple

Sujet	Verbe	Objet	Langue (si nécessaire)
https://viaf.org/viaf/96237916/	http://xmlns.com/foaf/0.1/name	"Caesar, Julius"	
https://viaf.org/viaf/96237916/	http://xmlns.com/foaf/0.1/gender	"male"	

Changement de slide

Prefixes

RDF prévoit dans sa rédaction la simplification des identifiants et des URL à travers l'utilisation de préfixes. Cela permet d'éviter une trop grande répétition, mais aussi de rendre plus lisible pour le lecteur et la lectrice les informations données. On définit alors une chaîne de caractère qui représentera ensuite l'ensemble du texte.

Changement de slide

Turtle

Turtle est un acronyme sur *Terse RDF Triple Language*. *Terse* signifie lapidaire.

Turtle est une des manières d'écrire le RDF. Si on se remémore nos 3 principes, il s'agit du principe n°1 : utiliser un format standard. Turtle en est un.

Turtle s'écrit de manière assez simple :

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix viaf: <http://viaf.org/viaf/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>

<viaf:184199909>
    dc:creator viaf:90637919.
```

Changement de slide

Mais on peut bien sûr aller plus loin et écrire plusieurs informations :

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix viaf: <http://viaf.org/viaf/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>

<viaf:184199909>
    dc:creator viaf:90637919;
    dc:title "Ad Lucilium".

<viaf:90637919>
    foaf:name "Sénèque".
```

Pour valider, on trouve en ligne un grand nombre d'applications. Certaines proposent même de convertir dans d'autres formats. http://rdfvalidator.mybluemix.net/