# Web Sémantique
# Des données ouvertes vers des données 5 étoiles

CARON Dylan
PINEL Félix

2017-2018

## Sommaire

# 1 Présentation des données

Notre Dataset renseigne certaines statistiques sur l'insertion professionnelle des diplômés de licences professionnelles. Elle regroupe par domaine et discipline le pourcentage de diplômés ayant obtenus un emploi, le salaire médian ainsi que son premier et troisième quartile ainsi que d'autres statistiques toujours en lien avec l'activité professionnelle. Cependant les données ne couvrent seulement que l'année 2013.

## 1.1 Lien vers GitHub et slides de présentation

Lien vers les slides de la présentation: Slides
Lien vers le github: GitHub

# 2 "Sémantification" des données

Le but est donc de transformer le CSV récupéré précédemment sur le site et de le transformer en un fichier RDF. Pour cela nous avons utilisé TARQL, grâce à ce dernier nous avons pu faire un mapping des valeurs présentes dans le csv pour en faire un RDF. Le but est donc de passé de données en 3 étoiles à des données 4 étoiles grâce à l'utilisation d'un format standardisé (RDF).

Le fichier "construct.sparql" se trouve dans le répertoire "Construct", de ce fait il est possible de voir en détail le mapping. Il est à noter que pour certains prédicats, nous n'avons pas trouvé de préfixe et nous avons donc utilisé un préfixe personnalisé (ici il se nomme "ex"). De plus, au départ nous avions "BIND" toutes les colonnes du CSV car les colonnes contenaient des symboles comme '%' et l'apostrophe ce qui nous obligeait à mapper. Malheureusement cela ne marchait pas pour toutes les colonne (notamment celle commençant par un '%') de ce fait nous avons renommé nous mêmes, à la main, les colonnes du CSV

## 2.1 Requête 1

```
SELECT  ?domaine ?discipline ?academie ?femmes
WHERE {
  ?p  ac:domaine ?domaine;
      ac:discipline ?discipline;
        ac:academie ?academie;
    ex:femmes ?femmes.
FILTER (?femmes != "ns" && ?femmes != "nd")
}
ORDER BY DESC (?femmes)
```

Voici la première des deux requêtes exigées, elle permet d'avoir le pourcentage de femmes par académie, domaine et discipline. Le FILTER est présent dans la seule optique de supprimer les lignes où le pourcentage n'est pas renseigné.

### 2.2 Requête 2

```
SELECT ?discipline ?academie ?nbFullTimeJob
WHERE{
  ?p ac:academie ?academie;
     ac:discipline ?discipline;
     stats:numberOfFullTimeJobs  ?nbFullTimeJob.
  FILTER(?nbFullTimeJob != "ns" && ?nbFullTimeJob != "nd"
    )
}
ORDER BY ASC (?academie)
```

Cette deuxième requêtes va récupérer le pourcentage de diplômés ayant obtenu un emploi à plein temps par discipline et académie. De la même façon que dans la première requête, le FILTER n'est présent que pour supprimer les lignes n'ont renseigné pour ce pourcentage.

## 3 Liaison de nos données avec celles d'autres groupes

Cette étape du projet a été pour nous la plus difficile dès le début, car aucun autre groupe n'avait de sujet en rapport avec le notre, et donc peu d'informations avec lesquelles aurions pu croiser nos données.

Nous avons constaté que le groupe composé de Clément Jehanno, Romain Duclos et Pierre Caillaud avaient l'attribut "identifiant de l'académie" en commun avec notre dataset. Le sujet de leur dataset était les effectifs d'étudiants inscrits dans les établissements et les formations de l'enseignement supérieur, recensés pour les années 2001-2002 à 2015-16. Il n'est précisé nul part sur le site de l'open data de quel "norme" provient cet identifiant d'académie, nous avons donc utilisé le terme aiiso:code, qui semble approprié à notre situation, pour décrire cet attribut.

Leur dataset étant conséquent et peu flexible à charger dans des éditeurs (Fichier CSV d'environ 60mo, JSON de plus de 240mo!), nous avons récupéré un sous ensemble de leur données grâce à l'API de data.enseignementsup-recherche.gouv.fr qui nous a permis de faire une requête et ainsi de récupérer seulement certaines colonnes et une partie des lignes concernant l'année 2013, qui correspond à la période de notre dataset. Le fichier récupéré était au format JSON, que nous avons reconverti en CSV grâce à des outils en ligne. Nous avons aussi récupéré la requête CONSTRUCT utilisée par nos collègues et l'avons modifié pour correspondre à nos besoins et lier nos datasets par l'identifiant de l'académie grâce à la propriété sameAS de OWL.

La requête fédérée que nous avons effectué consiste à afficher les effectifs par regroupement de formations ou d'établissements, et de complémenter ces informations en affichant l'académie correspondant à chaque regroupement, là où le dataset original de nos collègues ne possédait que l'identifiant.

```
SELECT DISTINCT ?effectif ?academie ?
    rgp_formations_ou_etablissements
WHERE{
  ?x          ac:academie    ?academie;
              aiiso:code     ?uri_academie.
  ?uri_academie    owl:sameAs     ?uri_academie_other.
  SERVICE <http://localhost:3030/data_ees> {
    ?z          aiiso:code     ?uri_academie_other;
                sch:EducationalOrganization ?
    rgp_formations_ou_etablissements;
              dbo:numberOfStudents ?effectif.
  }
}
```

# 4 Inférence

Pour l'inférence nous avons fait 3 fichiers. Le premier est un fichier de configuration qui renseigne les chemins aux fichiers model.ttl, myrules.rules et data.ttl. Le fichier data.ttl contient juste les triples de la base obtenu par le construct (faits à l'étape de la "Sémantification" des données). Ensuite vient le fichier model.ttl, ce dernier sert à spécifier les ontologies RDFS et OWl qui permetteront de faire les inférences. Enfin vient le fichier myrules.rules, celui-ci sert à énoncer les règles d'inférences. Par manque de temps, nous n'avons pas de règles d'inférence dans le fichier myrules.rules. Cependant un exemple peut se ressembler à ceci :

Dans le fichier model.ttl :

a:example rdf:type owl:Class;
rdfs:subClassOF c:example;
Dans le fichier myrules.rules :

[rule1: (a:example rdfs:subClassOf c:example)
(?s a:example ?o)→ (?$sc$ : $example$?$o$)]

Ceci n'étant qu'un exemple sur l'inférence subClassOf mais cela donne la façon d'écrire ses propres inférences.

# 5 Liaison des données d'ESRI au Cloud de Linked Data

Une fois le traitement de nos données effectués, l'objectif aurait été de pouvoir lier les données récupérées sur le site de l'Open Data de Enseignement supérieur Recherche et Innovation au Cloud de Linked Data. Cela correspond à la cinquième "étoile" du programme de déploiement des données dans l'Open Data. Il existe quatre "règles" qui permettent la certification linked open data des données :

1. Use URIs to name (identify) things.
2. Use HTTP URIs so that these things can be looked up (interpreted, "dereferenced").
3. Provide useful information about what a name identifies when it's looked up, using open standards such as RDF, SPARQL, etc.
4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

Nos données utilisent des URI pour définir chaque entité et chaque information que nous voulons lier avec d'autres (ici, l'identifiant de l'académie), et elles ont la volonté d'utiliser le protocole HTTP. Bien sûr, certaines de nos informations utilisent des vocabulaires et des URL fictifs (ex, stats, ac) faute de bien connaître l'étendue des vocabulaires possibles qui pourraient répondre spécifiquement à nos informations. Tous les vocabulaires "officiels" que nous avons pu trouver contenant des informations correspondant à nos besoins sont tout de même référencés par des liens HTTP réels.

Il serait aussi possible de lier nos données avec d'autres sur le Cloud comme nous l'avons fait pour les données de nos collègues, sur des ensembles de données à plus grande échelle, avec des ontologies OWL.

# 6 Utilisation du vocabulaire de description VOID

Le vocabulaire VoID (Vocabulary of Interlinked Datasets) permet d'ajouter des métadonnées décrivant des Datasets, au format RDF.

Nous avons donc écrit un fichier au format Turtle donnant quelques informations sur l'Université de Nantes, le Dataset que nous avons utilisé, le Dataset de nos collègues avec qui nous avons croisé nos données, ainsi que nous même.

Cela permettrait à des potentiels utilisateurs de notre travail d'avoir des informations utiles comme les liens des sources des données, les licences attribuées, ainsi que des informations sur les personnes ayant travaillé sur ces données, dans le cadre de remise de crédit par exemple.