Des données ouvertes vers des données en 5 étoiles

L'objectif du projet est de transformer les données ouvertes de l'Enseignement supérieur, de la Recherche et de l'Innovation (https://data.esr.gouv.fr/FR/) en données sémantiques dans et de lier ses données sémantiques au cloud de "Linked Data : Connect Distributed Data across the Web" (http://linkeddata.org). Un livre sur le Linked Data est disponible à l'adresse : (http://linkeddatabook.com).

Le projet est réalisé en groupe. Chaque groupe est composé de 2-3 étudiants. Le projet se déroule en 4 étapes.
Etape 1 : chaque groupe choisit un ensemble de données ouvertes à « sémantifier ».
Il existe des outils (lod2Stack, DataLift, les wrappers de lodpaddle, tarql, etc) qui permettent de transformer des données structurées en RDF. Utiliser, si possible, des vocabulaires partagées (foaf, doap, etc). En examinant, par exemple, des ensembles de données similaires à vos données, vous pouvez utiliser les vocabulaires les mieux adaptés à vos données. Proposez deux requêtes SPARQL « intéressantes » réalisables sur vos données.
Etape 2 : lier vos données à des données des autres groupes. Chaque groupe doit proposer une requête réalisable sur au moins 2 ensemble de données. Le groupe qui propose la requête la plus intéressante aura des « bonus ».
Etape 3 : utiliser des ontologies RDFS ou OWL et faire des inférences. Exécuter vos requêtes sur vos données individuelles et sur les données liées. Que constatez vous ?
Etape 4 : proposer des propriétés et des liens pour lier les données de ESR au cloud de linked data.
Etape 5 : Utiliser les vocabulaires VOID pour décrire vos datasets.

Agenda du projet :

- Le **07/11**, présentation des étapes 1 et 2.
- Le **1/12** rendu final du projet.

A rendre :

- 4-5 slides de présentation par groupe sur l'étape 1, et 2-3 slides de présentation de l'étape 2 pour le 07/11. Vous devez ajouter un lien vers votre présentation.
- Les slides contiennent une description des données ouvertes choisies et la méthode utilisée pour réaliser l'étape 1 du projet, et les deux requêtes.
- Rapport final du projet (4-5 pages) qui englobe les résultats obtenus par les 4 étapes du projet avec vos remarques et conclusions. A déposer le 1/12 sur madoc.

Hints: Il existe plusieurs outils pour sémantifier les données, par exemple:

- Tarql un outil simple et facile d'utilisation : Documentation: https://tarql.github.io/, github: https://github.com/tarql/tarql
- DataLift: http://datalift.org/

Pensez à utiliser l'ontologie de dbpedia, par exemple:

http://dbpedia.org/ontology/EducationalInstitution

Sinon, il existe un site qui regroupe les vocabularies:

http://lov.okfn.org/dataset/lov/about

Groupe	lien vers data set	Description de data set	Lien vers la présentation de l'étape 07/11
Samy GASCOIN-FONTAINE Romain BROHAN	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-aap-fp7-projets-retenus-participants-identifies/table/?disjunctive.identifiant_de_partenaire	Appels à projets 7ème Programme-Cadre de Recherche et de Développement Technologique (PCRDT). - Projets retenus et participants identifiés	slides de présentation
Loïc BOUTIN Ivan DROMIGNY--CHEVREUIL Baptiste AUFFRAY	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-rd-moyens-entreprises-intensite/table/?sort=valeur	Les moyens consacrés à la R&D : Les entreprises par intensité technologique	slides
Nathan SALAUN Antoine MAGNIN Martin LAVILLE	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/principales-institutions-executant-ou-finacant-la-recherche-hors-etablissements/table/	Principales institutions exécutant ou finançant la recherche (hors établissements d'enseignement supérieur)	https://docs.google.com/presentation/d/1ymg8n7fAtPMIrpX-vLdHpieYI0M4X6eeCU-K7gtiF10/edit?usp=sharing
Fortin Guillaume Le Bars Yannis	https://data.enseignementsup-recher		https://docs.google.com/presentation/d/1x

Travers Clément	che.gouv.fr/explore/dataset/appels-a-projets-horizon-2020-projets-retenus-et-participants-identifies/information/		8Ejp1bEC7LryAf3h_wGiKno5P06oElris_hysR8FeM/edit?usp=sharing
Florent ALAPETITE Roxane BELLOT Alexandre BOUDINE	fr-esr-etoile-de-l-europe	Résultats du concours des étoiles de l'Europe	Slides
JEHANNO Clément DUCLOS Romain CAILLAUD Pierre	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-atlas_regional-effectifs-d-etudiants-inscrits/export/?sort=-rentree	Effectif des étudiants sup en Pays de Loire	https://docs.google.com/presentation/d/1CscPCMNfQfRWbI06LCL3Si4Ie-eseh1ImG1O0SY2Q0s/edit?usp=sharing
MENARD Mica WIBAUX Robin LE GOUVELLO Quentin	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-initiatives-pour-la-lutte-contre-les-violences-sexistes-et-sexuelles/table/?disjunctive.zone_geographique&sort=academie	Initiatives pour la lutte contre les violences sexistes et sexuelles	https://docs.google.com/presentation/d/1p0_tIQJrS9i-gyZ4KfjNwOkqST7I-od47Kcc5ws1sEo/edit#slide=id.gc6f9e470d_0_0
GAILLARD Florent JEAN Marvin DUBOIS Hippolyte	dataset	Les moyens consacrés à la R&D : Les administrations par type d'organisme	Accès aux slides
AHMED Daniel BAH Thierno	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-finalistes-et-laureats-du-concours-ma-these-en-180-secondes-france/	C'est un concours qui permet à chaque étudiant doctorant doit présenter sa thèse devant un auditoire en seulement 3 minutes (Ma thèse en 180 secondes).	https://docs.google.com/presentation/d/1uLPDMwRWjXX51wl8mxouNW93cKVY_91Bz_L4W-zRocM/edit?usp=sharing

BELHADJ KACEM SOUFYANE YAMEN ALNAJM KONE SADA OUMAR	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-implantations_etablissements_d_enseignement_superieur_publics/table/?disjunctive.bcnag_n_nature_uai_libelle_editi&disjunctive.services&disjunctive.type_uai&disjunctive.nature_uai&sort=siege_lib	Implantations des établissements d'enseignement supérieur publics	https://docs.google.com/presentation/d/1vIdhi1WkRjKeCGHkxXqUL20GucfqPTJpqLFj5tLwv8E/edit?usp=sharing
MAHIER Loïc PHALAVANDISHVILI Demetre COUILLEROT Carol	budgets-de-recherche-et-de-transfert-de-technologie-rt-des-collectivites	Budget de recherche et de transfert de technologie (R&T) des collectivités territoriales.	Transparent
MOUSTAFA Ossama	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr_crous_logement_france_entiere/	Ensemble des logements proposés aux étudiants par le réseau des CROUS	

BRASSIER Maëlle MARINIER Ophélie LARDY Florian	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-iuf-les-membres/table/	Les membres de l'Institut universitaire de France	Slides
MAZOUA Quentin	Data set	Établissements publics et privés impliqués dans la recherche et développement	https://docs.google.com/presentation/d/1m_3GwvpfPmbcvHaslXAasyFYXKiP3soOkeHqFsI6KrE/edit?usp=drivesdk
CARON Dylan PINEL Félix	https://data.enseignementsup-recherche.gouv.fr/explor	Insertion professionnelle des diplômé.e.s de	slides

	e/dataset/fr-esr-insertion_professionnelle-lp/?disjunctive.academie	Licence professionnelle en universités et établissements assimilés	
MARIONNEAU Corentin BARZILAI Merlin	Dataset	Les bénéficiaires de la prime d'excellence scientifique	Slides
JAMET Félix LE FALHUN Mattis	https://data.enseignementsup-recherche.gouv.fr/explore/dataset/fr-esr-enseignants-titulaires-esr-public-national/?disjunctive.groupe_cnu&disjunctive.section_cnu&disjunctive.classe_age5	Les enseignants titulaires de l'enseignement supérieur public (national)	

Exemple de reqête:

```
# Fichier de mapping du dataset "fa-par-region" pour le projet de Web Sémantique
# Auteurs : Alexis Giraudet & Thomas Minier, M1 ALMA 2015
PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia: <http://fr.dbpedia.org/resource/>
PREFIX erdf: <https://erdf.opendatasoft.com/>
PREFIX schema: <http://schema.org/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX semanco:
<http://semanco-tools.eu/ontology-releases/eu/semanco/ontology/SEMANCO/SEMANCO.owl#>

CONSTRUCT {

        # mapping
   ?URI_REGION rdf:type dbpedia:Region;
                owl:sameAs ?URI_REGION_DBPEDIA;
                owl:sameAs ?URI_REGION_PARC;
                owl:sameAs ?URI_BILAN_POWER;
                schema:GeoShape ?GEO_SHAPE;
                schema:longitude ?LONGITUDE;
                schema:latitude ?LATITUDE.
```

```
        ?URI_PRODUCTION rdf:type ?TYPE_PRODUCTION;
                                        dbpedia:region ?URI_REGION;
                                        erdf:NbInstallation ?NOMBRE_INSTALLATIONS;
                                        dbpedia:Megawatt ?PUISSANCE_CUMULEE.

 }
 FROM <file:fa-par-region.csv#delimiter=%3B;encoding=utf-8>
 WHERE {
        # on construit le type pour chaque champ
        BIND( STRDT(?Geo_Shape, xsd:string) AS ?GEO_SHAPE)

        BIND( STRDT(?Nombre_d_installations, xsd:integer) AS
?NOMBRE_INSTALLATIONS)

        BIND( STRDT(?Puissance_cumulée, xsd:float) AS ?PUISSANCE_CUMULEE)

        # on extrait la latitude et on y associe le bon type
        BIND( STRDT(STRBEFORE(?Geo_Point, ","), xsd:float) AS ?LATITUDE).

        # on extrait la longitutde et on y associe le bon type
        BIND( STRDT(SUBSTR(STRAFTER(?Geo_Point, ","), 2), xsd:float) AS
?LONGITUDE).

        # on construit l'URI de la ressource par région
        BIND(
                URI(
                        CONCAT("https://erdf.opendatasoft.com/fa-par-region/",
                                REPLACE(?Région_administrative, " ", "_")
                        )
                ) AS ?URI_REGION).

        # on construit l'URI de la ressource par type de production
        BIND(
                URI(
                        CONCAT("https://erdf.opendatasoft.com/fa-par-region/",
                                CONCAT(
                                        REPLACE(?Région_administrative, " ",
"_"),
                                        CONCAT("/", ?Type_de_production)
                                )
                        )
                ) AS ?URI_PRODUCTION).

        # on construit l'URI de la région sur DBpedia (avec le cas spécial de la Région
Centre)
```

```
	BIND(
		URI(
				CONCAT("http://fr.dbpedia.org/resource/",
					REPLACE(REPLACE(?Région_administrative, " ", "_"),
"Centre", "Région_Centre")
				)
		) AS ?URI_REGION_DBPEDIA).

	# on construit l'URI du type de production
	BIND(
		URI(

CONCAT("http://semanco-tools.eu/ontology-releases/eu/semanco/ontology/SEMANCO/SEM
ANCO.owl#",
				REPLACE(
					REPLACE(
						REPLACE(

REPLACE(?Type_de_production,
								"Bio énergie", "Biomass"
							),
							"Eolien", "Wind_Energy"
						)
					,"Photovoltaïque", "Solar_Energy"
					)
					, "Hydraulique", "Hydro_Energy")
				)
		) AS ?TYPE_PRODUCTION).

	# Liens avec les autres graphes
	# on construit l'URI de la ressource par région pour le graphe "parc raccorde par
region"
	BIND(
		URI(

CONCAT("https://erdf.opendatasoft.com/parc-raccorde-par-region/",
				REPLACE(?Région_administrative, " ", "_")
			)
		) AS ?URI_REGION_PARC).

	# même chose pour le graphe "bilan électirque par puissance installée"
	BIND(
		URI(

CONCAT("https://erdf.opendatasoft.com/bilan-electrique-puissance-installee/",
				REPLACE(?Région_administrative, " ", "_")
```

```
                                    )
                        ) AS ?URI_BILAN_POWER).
    }
:
```