Projet Analyse des données, Web des données, Web sémantique.

Le but du projet est de créer un graphe de connaissances et de l'exploiter à l'aide de requêtes SPARQL. Pour mener ce projet de web sémantique, il est important de suivre une approche qui balaie toutes les étapes ci-dessous et de commencer avec un jeu de données de taille restreinte. Lorsque vous aurez franchi toutes ces étapes vous pourrez augmenter la taille et la complexité de votre jeu de données (ajouter relations, types de données …).

PRENEZ UN DOMAINE QUE VOUS AUREZ PLAISIR A TRAITER.

1. Choisir un domaine d'application

- **Définir le domaine** du projet qui vous intéresse, qui possède suffisamment de sources de données, qui vous permettra de créer une ontologie et qui peut avoir des ressources intéressantes sur le web pour faire du linked-data.
- **Vérifiez que vous avez les compétences** suffisantes pour comprendre le domaine afin de développer un modèle d'ontologie et exploiter ces données.
- **Fixer les objectifs de votre projet en fonction des données disponibles que vous avez trouvé**
    - Inventaire des données : passez en revue les données que vous avez. Par exemple, si vous choisissez le domaine des produits alimentaires, vous pourriez avoir des données sur : les ingrédients, les valeurs nutritionnelles, et les effets sur la santé.
- Pour structurer votre ontologie posez-vous des questions comme :
    - Quelles informations je recherche ?
    - Comment je pourrai bénéficier d'une structure sémantique des données ?
    - Quels besoins de relations je peux représenter entre différents concepts ou jeux de données ?
    - Que puis-je faire avec les données que j'ai choisies ?
    - Quelles informations sont les plus importantes ou la plus utiles ?
    - Y a-t-il des requêtes qui pourraient valoriser mon ontologie et les jeux de données que j'ai choisis ?

2. Sélectionner les Données

- **Choisir des sources de données** ouvertes (qui ne vous demanderons pas un nettoyage important), accessibles dans différents formats (CSV, BD relationnelles, API, ou graphe RDF).
- Exemples de sources disponibles :
    - Données gouvernementales : https://data.europa.eu/en , https://data.gov/ , https://www.opendatasoft.com/en/ , https://www.data.gouv.fr/fr/ , https://www.data.gouv.fr/fr/ , https://geodatamine.fr/ , ONU https://data.un.org/ , https://worldpopulationreview.com/
    - Les données géographiques sur geoportal.org, earthdata.nasa.gov, ign sont des sites intéressants mais difficile à exploiter.
    - Données économiques et statistiques : https://data.worldbank.org/ , https://catalogue-donnees.insee.fr/fr/catalogue/recherche ,

https://ec.europa.eu/eurostat/web/main/data/database , https://transport.data.gouv.fr/infos_reutilisateurs
    - Données culturelles : https://data.culture.gouv.fr , https://pro.europeana.eu/discover-the-data/apis#edm ,
    - Transition écologique (ADEME) : https://data.ademe.fr/ , https://www.statistiques.developpement-durable.gouv.fr/environnement
    - Data Education : https://data.education.gouv.fr
    - Produits alimentaires : https://world.openfoodfacts.org/data
    - Bibliothèques : librairie du congrès américain https://www.loc.gov/search/?in=&q=LOC+Datasets&new=true , Bnf https://api.bnf.fr/fr/dumps-de-databnffr
    - Musique : https://musicbrainz.org/ , Spotify API Documentation , Kaggle Music Datasets , Last.fm API , Discogs API , http://millionsongdataset.com/ , https://acousticbrainz.org/ ,
    - Sport : https://www.thesportsdb.com/ , https://www.football-data.org/ , http://nbasense.com/nba-api/ , https://nextgenstats.nfl.com/ , https://odf.olympictech.org/2024-Paris/paris_2024_OG.htm , https://sportradar.com/betting-gaming/products/betting/sports-data/ , https://france.sport/sport-data-hub/138-D%C3%A9ploiement-op%C3%A9rationnel-du-Sport-Data-Hub/90-Data-Warehouse-(DWH) ,
    - Cinema : https://developer.themoviedb.org/docs/getting-started , https://www.omdbapi.com/ , https://developer.imdb.com/non-commercial-datasets/
- **Web Scraping** : Si les données ne sont pas directement accessibles, vous pouvez envisager de faire du scraping (en respectant les droits d'accès et la législation) pour extraire des informations des sites web.

3. Comprendre et préparer les Données

- **Identifier dans** vos données les concepts et les variables qui les définissent et les relations entre concepts.
- **Nettoyage et préparation des données**
    - **Éliminer les données dupliquées** passez en revue vos jeux de données et nettoyez-les.
    - **Structuration uniforme** : Assurez-vous que les formats de données (comme les dates, les unités de mesure) sont uniformisés.

4. Concevoir une Ontologie

- **Créer ou réutiliser une ontologie** adaptée à votre domaine. Une ontologie définit les concepts et les relations dans un domaine donné, ce qui permet de donner du sens à vos données.
- **Utiliser des vocabulaires existants** lorsque possible (comme FOAF, Dublin Core, ou Schema.org) pour garantir l'interopérabilité.
- **Création d'une ontologie** : Si aucune ontologie ne convient, vous devrez en créer une, par exemple en utilisant Protégé, un outil de conception d'ontologies.
- Outils utiles pour créer des ontologies : **Protégé,**

5. Peupler votre ontologie

- **Avec les données** que vous avez nettoyées, créer un graphe RDF en utilisant le vocabulaire défini par votre ontologie.

6. Enrichir les Données

- Vous pouvez associer vos données à des liens vers des entités externes (par exemple, associer des noms de lieux à des URI de DBpedia ou Wikidata). https://dumps.wikimedia.org/wikidatawiki/entities/ . Ces liens permettent de relier vos données avec d'autres jeux de données du web sémantique, facilitant ainsi l'interopérabilité.

7. Stocker

- **Utiliser un triplestore (une base de données RDF**) pour stocker les données et interroger les données.
- Charger vos jeux de données RDF dans le triplestore que vous avez choisi.

8. Interroger les Données

- **Créer des requêtes SPARQL** (SPARQL Protocol and RDF Query Language) pour interroger les données.
- Par exemple, vous pourriez écrire une requête pour trouver toutes les entités d'une classe spécifique ou pour rechercher des relations spécifiques entre les entités.
- **Sauvegarder et exporter les résultats** : Une fois la requête exécutée, vous pouvez exporter les résultats sous différents formats comme CSV, JSON, ou RDF, directement depuis l'interface.

9. Interroger des données liées (Linked Data)

- Proposer des exemples de requêtes fédérées exploitant les liens avec des ressources externes que vous avez faits dans le point 6

10. Outils Utiles

- **Pour ce point vous pouvez vous reporter à la note https://lig-membres.imag.fr/genoud/teaching/coursSW/tps/OUTILS_SW/index.html#section01** que vous avez déjà utilisé en TP. Vous pouvez adapter ces outils à un framework qui vous convient (Python ou autres).
- Protégé : pour la création d'ontologies.
- Apache Jena Fuseki, GrapheDB ou autres : pour le stockage RDF et pour les requêtes SPARQL
- DBpedia, Wikidata : sources de données liées utiles pour l'enrichissement de votre base de connaissances.
- A voir éventuellement des outils de transformation de données comme : OpenRefine avec l'extension RDF et Tarql (CSV vers RDF).
- Télécharger des Jeux de Données Complets Wikidata (Dumps) met régulièrement à jour des dumps complets de sa base de données, que vous pouvez télécharger pour

effectuer des analyses hors ligne. Page de téléchargement des dumps : Wikidata Dumps