HAI824 - Traitement sémantique des données

Construction d'un outil d'intégration de données
— TP noté 2 —

L'objectif de ce travail est de créer un outil d'intégration de données structurées sous la forme de graphes de connaissances.

Les fonctionnalités de l'outil d'intégration de données :

- prendre en entrée 2 fichiers rdf ou, autrement dit, 2 graphes de connaissances
- proposer à l'utilisateur à configurer l'outil, par exemple
  - choix de propriétés à comparer
  - choix de mesure(s) de similarité entre les valeurs de ces propriétés
  - combinaison de plusieurs mesures de similarité (afin de prendre, par exemple, leur moyenne pondérée comme mesure finale)
  - choix du seuil de similarité
- fournir en résultat un ensemble de liens owl:sameAs sous la forme d'une liste de triplets de type (e1, owl:sameAs, s2), où e1 est l'url d'une entité venant du dataset source et e2 est l'url d'une entité venant du dataset cible.

Validation des résultats :
Pour mesurer la qualité de la performance de votre outil, utilisez un fichier d'alignement de référence (fourni avec les données) qui vous permettra de calculer la précision, le rappel et la f-*measure* de vos résultats en fonction du seuil de similarité choisi. Pour montrer l'effet du seuil de similarité sur vos résultats et sur la performance de votre outil, pensez à une représentation graphique qui montre **la courbe de la variation de la f-measure en fonction du seuil.**

Les données :
Pour le but de la construction de cet outil, nous allons nous intéresser aux données suivantes :

Graphe source: https://github.com/DOREMUS-ANR/doremus-playground/blob/master/DHT/source.ttl
Graphe cible: https://github.com/DOREMUS-ANR/doremus-playground/blob/master/DHT/target.ttl

Alignement de référence (vérité de terrain) entre les deux graphes : https://github.com/DOREMUS-ANR/doremus-playground/blob/master/DHT/refDHT.rdf

Ce sont des jeux de données réels, constitués d'un couple de graphes provenant de la Bibliothèque Nationale de France (BnF) et de la Philharmonie de Paris contenant chacun les références de 238 œuvres musicales hétérogènes dans leur description. Ils manifestent notamment des différences de description, comme le multilinguisme, des différences de catalogues, des différences d'orthographe, etc. Le fichier refDHT.rdf contient l'alignement de référence (vérité terrain) entre les œuvres que vous pouvez utiliser pour évaluer vos systèmes et valider leur performance.

Les données et leur modèle sont décrit en détail dans l'article suivant (facilement trouvable sur internet) :

- Achichi, M., Lisena, P., Todorov, K., Troncy, R., & Delahousse, J. (2018). DOREMUS: A graph of linked musical works. In *International Semantic Web Conference* (pp. 3-19). Springer.

Pour plus d'information : https://data.doremus.org/

Modalités :

- Travail à effectuer en groupes de 2 à 4 personnes
- Un unique rendu (.zip) contenant les noms des membres du groupe dans son nom avec:
  - un mini rapport (**6 pages max.**) contenant l'architecture de votre système, l'évaluation effectuée (contenant les courbes de f-measure en fonction du seuil) et un scénario de cas d'usage
  - l'intégralité des traitements automatiques (code, données externes utilisées, etc.) accessibles **dans un lien git.**