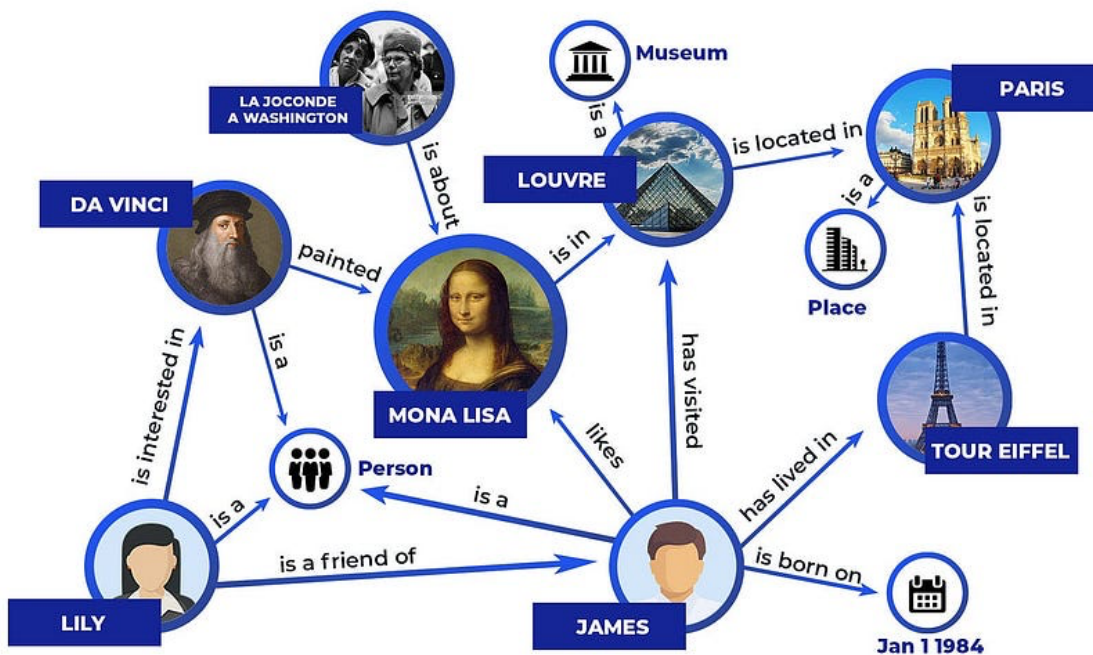


Transformer les articles scientifiques en graphe de connaissances

L'extraction d'informations pertinentes à partir de textes est une tâche cruciale pour mener des analyses approfondies dans les sciences économiques. Les liens sémantiques présents dans les textes contiennent des informations précieuses qui permettent de mieux comprendre les relations entre concepts et d'enrichir l'analyse économique. Ce projet propose d'analyser le contenu des articles scientifiques en économie afin de représenter les informations les plus pertinentes sous forme de graphe de connaissances, en s'appuyant sur les technologies du web sémantique. Il s'agira de convertir le texte non structuré en triplets RDF (Resource Description Framework) < sujet-prédicat-objet >, conformément aux recommandations du W3C.



En 2023, un groupe d'étudiants a développé une première version fonctionnelle d'un graphe de connaissances (disponible ici : <https://github.com/daJster/KG-generation/>). Cependant, pour rendre ce graphe pleinement exploitable, il est nécessaire d'améliorer et d'optimiser le processus de traitement, notamment pour gérer des volumes importants de données.

Les critères à prendre en compte incluent :

- La capacité à traiter des articles dans différentes langues.
- La possibilité de traiter plusieurs millions d'articles de manière efficace.
- L'adaptation à l'infrastructure universitaire, sans recours à des solutions basées sur des services d'abonnement comme Amazon, Azure, ou autres.

Langage de programmation : python.

Quelques articles pour commencer :

<https://medium.com/@jack16900/generative-knowledge-graph-construction-a-review-947521309ec5>

<https://app.ingemmet.gob.pe/biblioteca/pdf/SP-558-02.pdf>

<https://github.com/nicolas-hbt/pygraft>

<https://github.com/Fcabla/DBpedia-abstracts-to-RDF>