Application de la fouille de texte et des ontologies pour assister la conduite des revues systématiques de la littérature

Camille Demers, Doctorante en sciences de l'information | EBSI, Université de Montréal | ✉ camille.demers@umontreal.ca

Résumé

Ce projet doctoral porte sur l'application de méthodes de fouille de textes pour l'extraction, la représentation et la synthèse des connaissances scientifiques par le biais d'ontologies. Celui-ci vise à évaluer l'applicabilité de ces méthodes pour assister la conduite des revues systématiques de la littérature.

Problématique

Revues systématiques de la littérature (RSL) [1]

- ✓ **Accès rapide aux résultats** de recherche sur un sujet
- ✓ Favorisent les **pratiques fondées sur des données probantes**
- ✓ Orientent les **investissements de recherche futurs**

Enjeux liés à la conduite des RSL

- × Deviennent **rapidement désuètes**, justifiant le besoin d'outils facilitant leur **mise à jour** [2]
- × Avec l'accroissement continu de la production savante, leur conduite s'avère **de plus en plus laborieuse** : des outils pour **automatiser certaines étapes** sont donc indispensables [3]

Automatisation des RSL

- Plusieurs travaux sur la sélection des articles et le développement de plateformes de collaboration, mais **peu sur l'extraction automatique des données** [3]

Ontologies et RSL / automatisation des RSL

- Des initiatives valorisant la **représentation des données** des RSL au moyen d'**ontologies** (ex. PICO, UMLS) ont été proposées au cours des dernières années [4]
- Les ontologies offrent un moyen concret d'appliquer les **principes de données FAIR** aux RSL [5, 6]: interopérabilité, mise à jour et réutilisation des données

Devis méthodologique : Approche quantitative en fouille de textes

Population

Population cible	Toutes les RSL publiées à ce jour
	Toutes les publications analysées dans les RSL publiées à ce jour
Population accessible	RSL indexées dans certaines bases de données (BD) bibliographiques (ex. PubMed, IEEE, JSTOR)
	Publications indexées dans certaines BD bibliographiques et accessibles en texte intégral

Plan d'échantillonnage (RSL)

300 RSL issues de trois domaines de recherche : santé, ingénierie, sciences sociales

Technique : Choix raisonné (non probabiliste)

Critères d'inclusion

- ✓ Référence au standard PRISMA [1]
- ✓ Liste complète des publications analysées
- ✓ Analyse d'articles de recherche et d'actes de communications

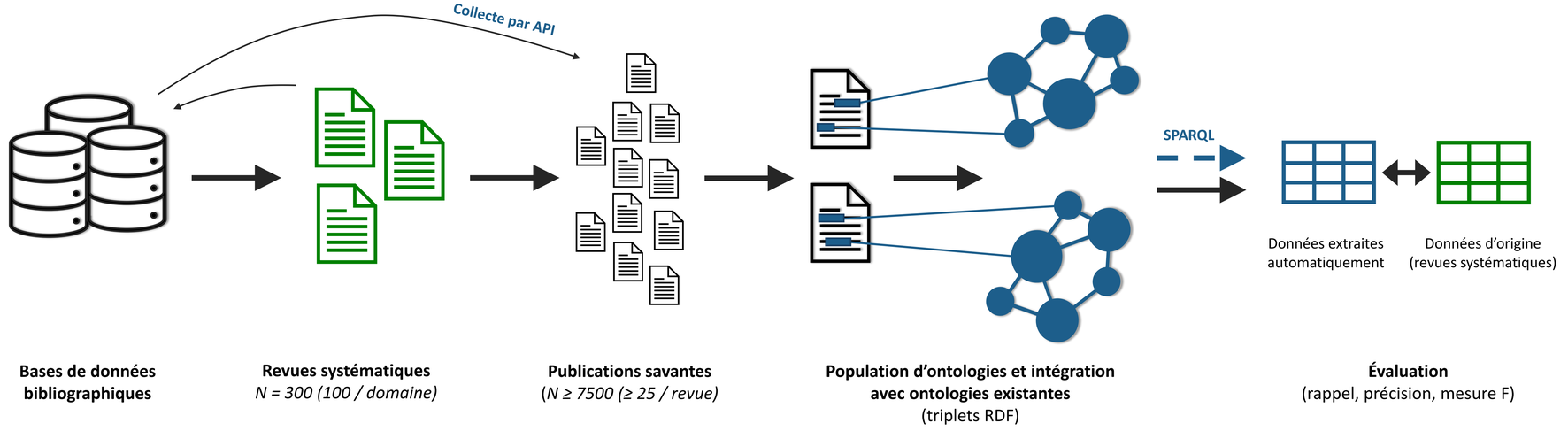
Collecte des données

- **Corpus de publications** (articles de recherche et actes de conférence) ayant été analysées dans les RSL sélectionnées
- **Collecte des documents** via les interfaces de programmation d'application (**API**) des BD bibliographiques

Analyses : Fouille de textes / Traitement automatique du langage

- Extraction terminologique
- Reconnaissance d'entités / Annotation sémantique des textes
- Population d'ontologies à partir des textes

Approche proposée



But et objectifs

But de la recherche : Comparer différentes méthodes de population d'ontologies reposant sur des techniques de fouille de textes pour assister l'étape d'extraction des données des RSL.

Objectifs spécifiques

1. **Caractériser** les enjeux relatifs à la **représentation des connaissances scientifiques** par le biais d'**ontologies.**
2. **Comparer** différentes **méthodes** permettant de **populer automatiquement des ontologies** à partir des **textes intégraux** de **publications savantes.**
3. **Identifier** les **apports et défis** soulevés par ces méthodes pour assister l'étape **d'extraction des données des RSL.**

Qualité de la recherche

Validité interne

- ✓ Revue de la littérature
- ✓ Formations techniques

Fidélité & Objectivité

- ✓ Publication du code, documentation

Validité externe

- ✓ Couverture disciplinaire large (3 grands domaines, plusieurs bases de données)
- × Limitée aux critères d'inclusion et d'exclusion de l'échantillon

Importance et retombées

- **Regard nouveau** sur la **pertinence d'utiliser des ontologies** pour la représentation des données des **RSL** ainsi que leur adéquation aux **principes de données FAIR**
- **Portrait comparatif** de méthodes permettant d'**enrichir automatiquement des ontologies** à partir de **publications savantes**
- **Recommandations** sur l'application de ces méthodes pour **faciliter la conduite des RSL**

Références citées

[1] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. et PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-analyses: the PRISMA Statement. Annals of Internal Medicine, 151(4), 264-269.

[2] Bastian, H., Glasziou, P. et Chalmers, I. (2010). Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Medicine, 7(9), e1000326.

[3] van Dinter, R., Tekinerdogan, B. et Catal, C. (2021). Automation of Systematic Literature Reviews: A Systematic Literature Review. Information and Software Technology, 136.

[4] Cochrane Linked Data Project Board. (2013). CochraneTech to 2020: The Role of Linked Data in Meeting our Strategic Goals. https://linkeddata.cochrane.org/

[5] Abu Ahmad, R., D'Souza, J., Zloch, M., Otto, W., Rehm, G., Oelen, A., Dietze, S. et Auer, S. (2024). Toward FAIR Semantic Publishing of Research Dataset Metadata in the Open Research Knowledge Graph. arXiv e-prints. https://doi.org/10.48550/arXiv.2404.08443

[6] Ali, A. et Gravino, C. (2018). An Ontology-Based Approach to Semi-Automate Systematic Literature Reviews. 12th International Conference on Open Source Systems and Technologies.

Université de Montréal

2024-05-02