

# Un dictionnaire de données médico-administratives pour accélérer la recherche

91e Congrès de l'Acfas

Vincent Martin-Schreiber MEng, BScN, PhD candidate

Institut du Savoir Montfort

2024-05-15

# Introduction



Besoin = un supermarché des données



# Les Métadonnées : une porte d'entrée sur le Big Data

- ▶ “Des données sur les données”.
- ▶ Big Data et IA

# Rendre les données découvrables

- ▶ Description du jeu de données
- ▶ Documentation
- ▶ recherchables
- ▶ métadonnées au niveau variable

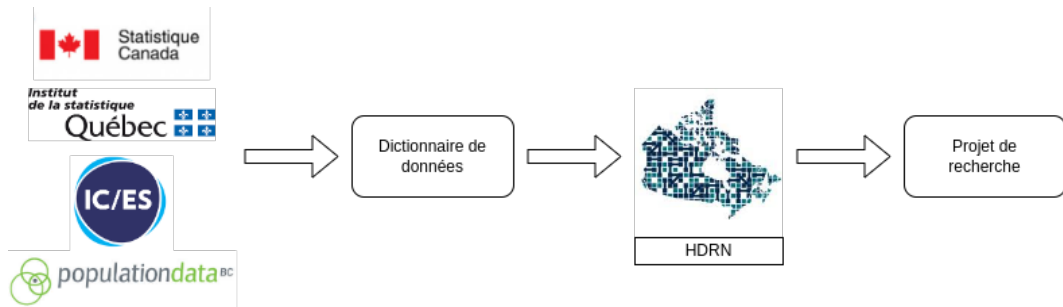
## Des défis de découvrabilité au Canada

“Les sources de données des sciences de la santé restreintes ont reçu de mauvaises notes de découverte de données en raison **d’un manque de métadonnées** (38/48, 79%), de l’**impossibilité de rechercher/parcourir les ensembles de données** (32/46, 70%) et du **manque de documentation des données** pour soutenir l’interprétabilité et la réutilisation (27/48, 56%).” (Read et al., 2022)

# Démonstration du Dictionnaire

<https://www.dictionnairededonneesensante.ca/>

# Scénario possible



- ▶ algorithms de labellisation des données : “langue parlée à la maison”, “langue fournisseur de soins”, etc. - par exemple Fredriksson et al. (2021)
- ▶ projet collaboratif pour la labellisation des données
- ▶ conversion en plateforme plus aboutie, par exemple Data Hub Project<sup>1</sup>

<sup>1</sup><https://demo.datahubproject.io/>



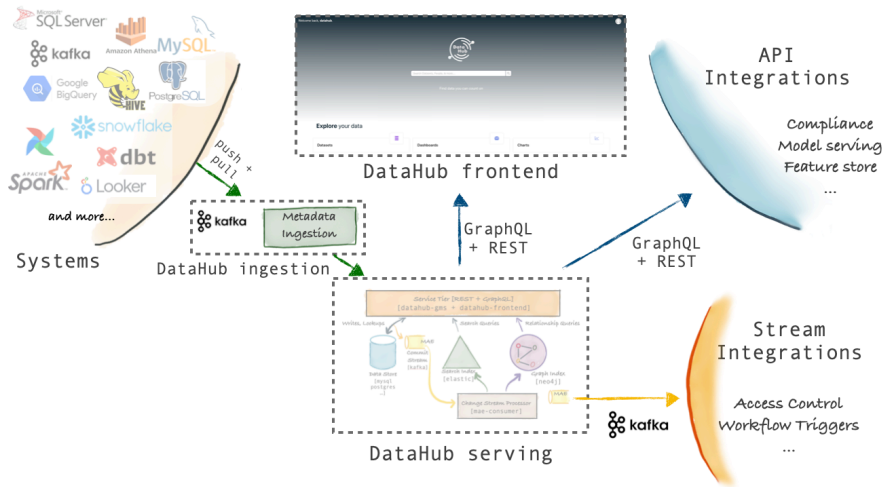
# Conclusion et Perspectives

Impact et bénéfices pour la recherche:

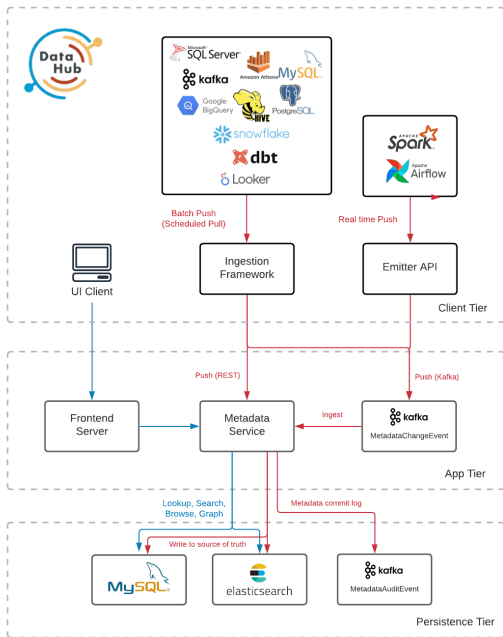
- ▶ analyse de la fréquence des variables
- ▶ économies sur les coûts d'analyse
- ▶ accélération des “time to results” : possibilité de coder l'analyse directement (R, Python, etc.)

Mais aussi:

- ▶ des défis (par ex. standardisation)
- ▶ un besoin d'élargissement du cercle classique des acteurs (notamment vers les TI)
- ▶ des opportunités méthodologiques (par ex. statistiques bayésiennes)



<https://datahubproject.io/>



# Références

- Fredriksson, T., Mattos, D. I., Bosch, J., & Olsson, H. H. (2021). An Empirical Evaluation of Algorithms for Data Labeling. *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 201–209. <https://doi.org/10.1109/COMPSAC51774.2021.00038>
- Posit team. (2023). *RStudio: Integrated Development Environment for R* (Version 2023.06.0+421) [x86\_64, linux-gnu]. Posit Software, PBC. <http://www.posit.co/>
- Read, K. B., Gibson, G. A., Leahey, A., Peterson, L., Rutley, S., Shi, J., Smith, V., & Stathis, K. (2022). *Understanding the Challenges Associated with Finding and Accessing Restricted Data in Canada: A Mixed Methods Study* [Preprint]. Open Science Framework. <https://doi.org/10.31219/osf.io/pa5fx>
- Zotero (Version 6.0.30). (2023). [Computer software]. Corporation for Digital Scholarship. <https://www.zotero.org>