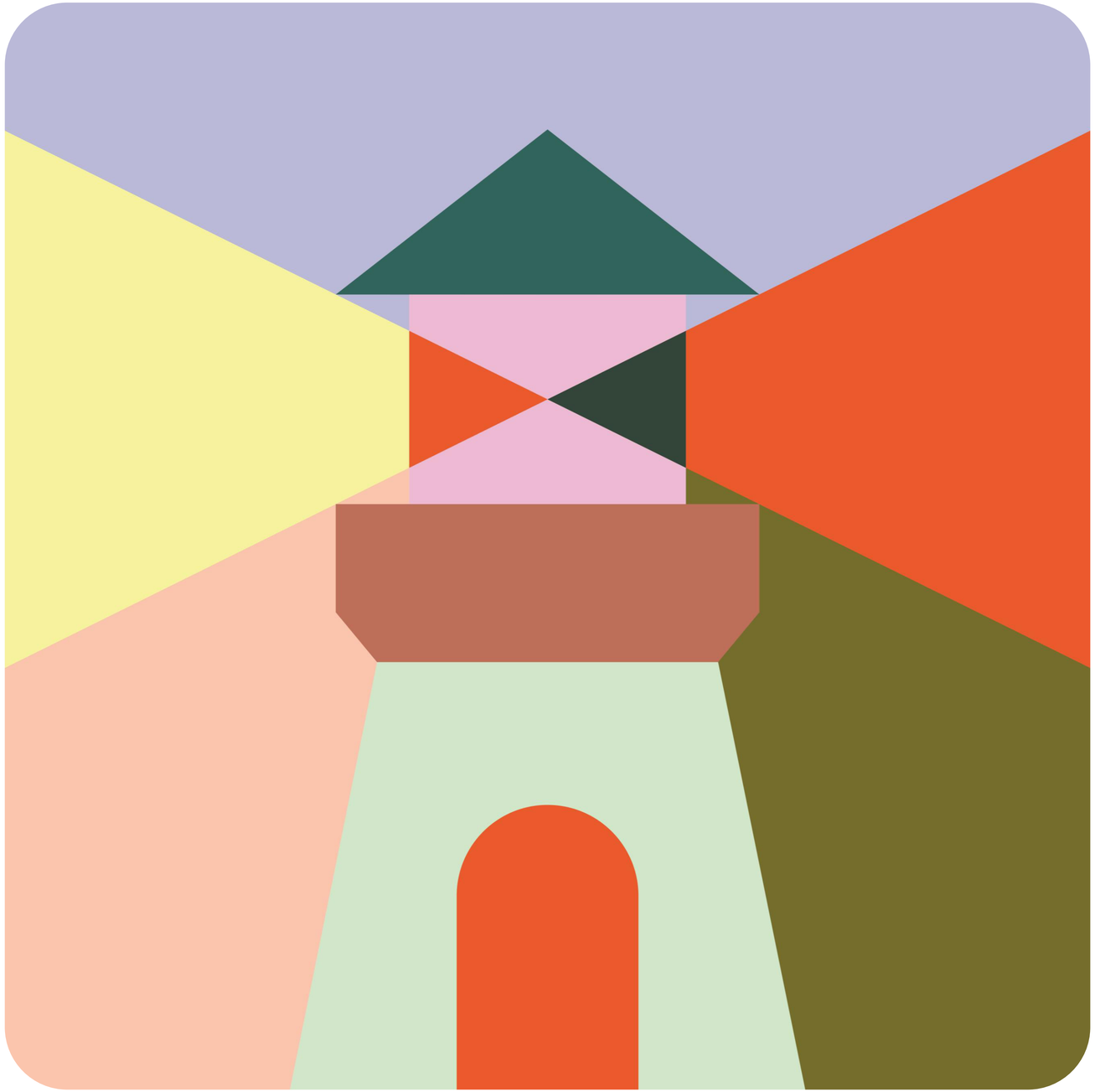
Améliorer la découvrabilité de la littérature grise en français d'un dépôt institutionnel : le cas PhareSST

Catherine Dufresne, technicienne en documentation
et Bénédicte Nauche, MBSI

Avec la collaboration de Samuel Hébert, technicien en documentation

Congrès des professionnel·le·s de l'information
Montréal, 5 novembre 2026



BY NC ND

Assisté par l'IAG

Plan

1. Invisibilité de la littérature grise en français
2. Présentation de l’IRSST et de son dépôt PhareSST
3. Actions déjà mises en œuvre
4. Spécificités des moteurs de recherche
5. Spécificités des grands modèles de langage
6. Conclusion

Littérature grise

Accessibilité difficile et découvrabilité limitée

Infrastructures Anglo centriques

Obstacle pour toute autre langue (Dony, 2024 et Garcia, 2025)

Filtre de langue

Résultats non anglophones relégués aux derniers rangs par les algorithmes (Rovira, 2021)

Biais de citation

Désavantage créé par les algorithmes qui privilégient le volume de citations (Rovira, 2021)

L'invisibilité n'est pas un accident!

Pourquoi publier en français est important

- Donner accès à la science aux communautés sans barrière de langue et de contexte culturel (Anctil, 2025, Matas, 2025)
- La pensée est modélisée en fonction de notre langue maternelle (Charton, 2025)
- La langue est associée à notre vision du monde et à notre culture (Bowker *et al.*, 2025)
- Accès aux nuances de la langue (Charton, 2025)
  - Les sciences humaines dépendent particulièrement des nuances socioculturelles

> La découvrabilité est la possibilité d'un contenu de se faire découvrir aisément, notamment par quelqu'un qui ne cherchait pas précisément le contenu en question
> (Office québécois de la langue française, 2021)



Comment rendre ces contenus **plus découvrables** au sein d'un dépôt institutionnel?

Un peu de contexte

Institut de recherche Robert-Sauvé en santé et en sécurité du travail

Fondé en 1980

Mission : contribuer à la santé et à la sécurité des travailleuses et travailleurs par la recherche, l'expertise de ses laboratoires, **ainsi que la diffusion et le transfert des connaissances**, et ce, dans une perspective de prévention et de retour durables au travail



150 employés, dont ~60 scientifiques

Centre de documentation : 2 techniciens en documentation, 2 bibliothécaires

PhareSST, le dépôt institutionnel (DI) de l'IRSST

Un choix stratégique pour le libre accès : centraliser, pérenniser et démocratiser l'accès à la production scientifique de l'IRSST
(Alayon *et al.*, 2013; Marsh, 2015)

Être accessible pour et découvrable par :



Milieux de travail :
Travailleurs, préventeurs, ergonomes…



Communauté scientifique :
Chercheurs, professeurs, étudiants



Décideurs publics :
Impact sur les politiques gouvernementales



International :
Rayonnement global

PhareSST

Plateforme choisie : Digital Commons

Forces

- Solution clé en main
- Interopérabilité (protocole OAI-PMH)
- Standardisation des métadonnées
- Indexation par les moteurs de recherche
- Capacité de produire des rapports

Limites

- Interface native en anglais uniquement (travail de traduction interne requis)
- Intégration partielle et rigide des balises Dublin Core
- Certaines fonctionnalités système non traduisibles

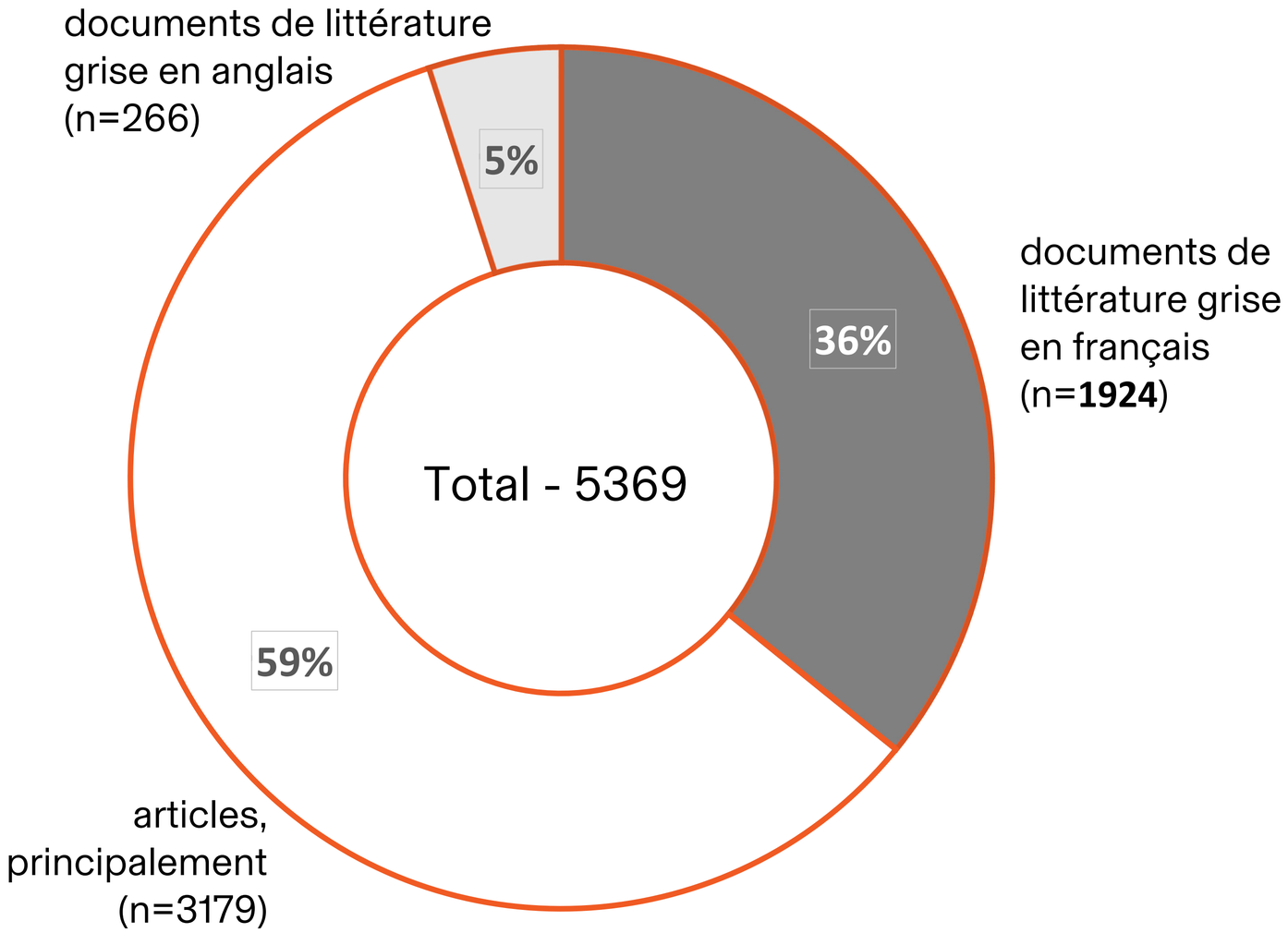
Naissance et état actuel

Été 2024 : ajout de DOI sur nos nouvelles publications

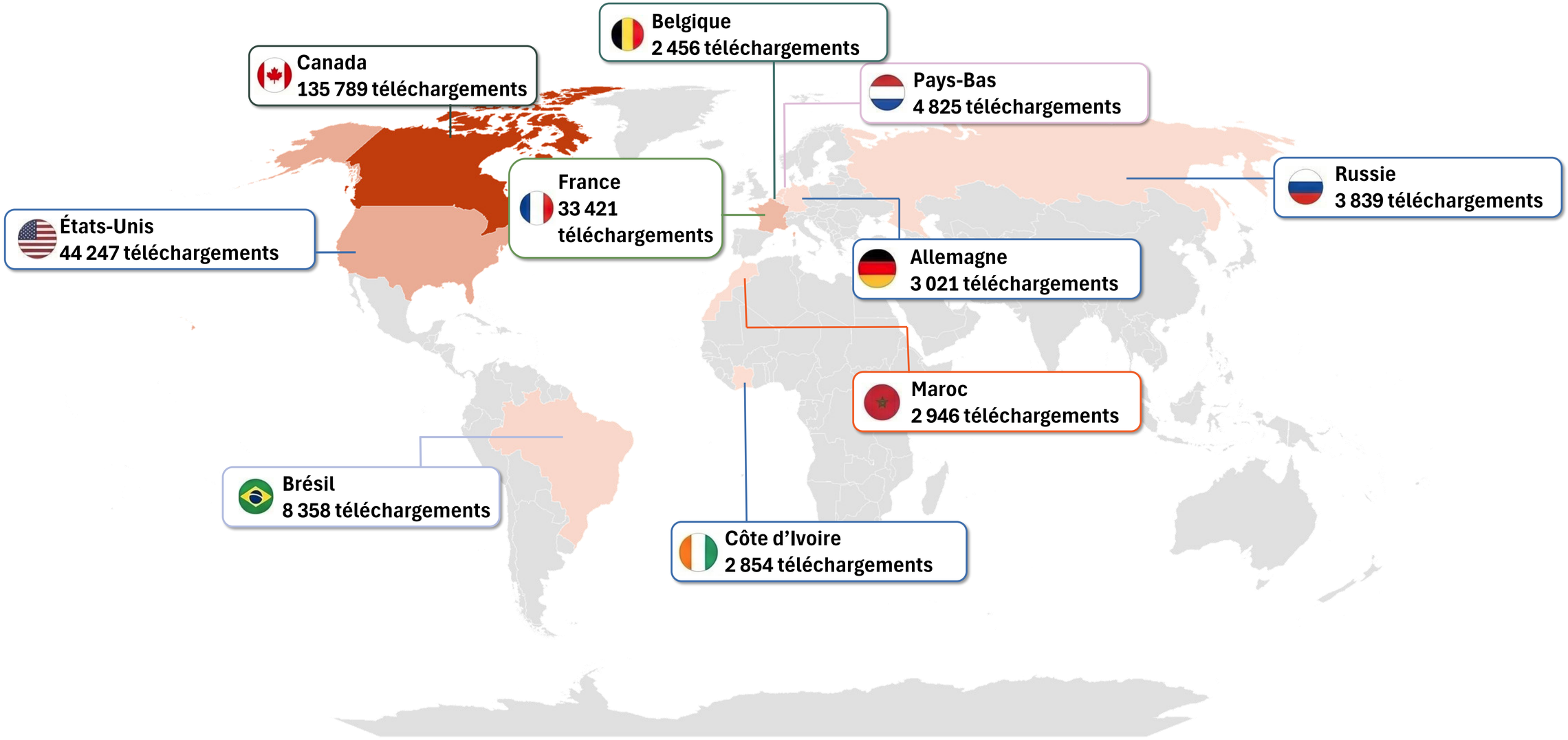
Septembre 2024 : lancement
(auparavant, solution de mise en ligne maison)

Mars 2025 : indexé par Google/Google Scholar, mais aussi dans OpenAlex et Semantic Scholar

RÉPARTITION GLOBALE DU CORPUS AU 31 DÉCEMBRE 2025



Répartition géographique des principaux pays qui téléchargent sur PhareSST



© Australian Bureau of Statistics, GeoNames, Geospatial Data Edit, Microsoft, Navinfo, Open Places, OpenStreetMap, Overture Maps Fundation, TomTom, Wikipedia, Zenrin
Avec Bing

Actions déjà mises en œuvre pour augmenter la découvrabilité du contenu francophone de PhareSST

Au niveau documentaire

Traduction des résumés

Saisie en anglais et français.
Utilisation de DeepL.
Relecture systématique.
(Bowker et al., 2025; Fiorini, 2025; McElroy et Bridges, 2018; Meurs, 2025)

Bémol : un seul champ pour le résumé en français et en anglais; travail chronophage et exigeant.

Indexation

Utilisation d'un vocabulaire contrôlé; tous les mots-clés sont inscrits en français et en anglais.
(Chipangila et al., 2024; COAR Task Force on Supporting Multilingualism and non-English Content in Repositories, 2023)

Bémol : Un seul champ pour les mots clés en français et en anglais; pas de vraie intégration du thésaurus.

Métadonnées ajoutées aux PDF

Ajout manuel de données dans les propriétés du fichier et modification de la configuration pour que les métadonnées ne soient pas chiffrées afin que les moteurs de recherche puissent les lire.

Bémol : les moteurs académiques ignorent souvent le PDF au profit du HTML. La page de garde reste « reine ».

OCRisation

Rendre les textes des anciennes publications lisibles par les machines.

Bémol : travail chronophage.

Au niveau « système »

Enregistrements périphériques

Google Scholar, ROAR, OpenDOAR, Open Alex, Overton (Arlitsch et O'Brien, 2012)

Bémol : Absence de contrôle sur ce que font ces moteurs de recherche avec l'indexation de nos documents.

Attribution de DOI

Enregistrement rigoureux des métadonnées, incluant les bibliographies, auprès de Crossref (Dépôt fédéré de données de recherche, s.d.; Go FAIR, s.d.; Stoll et de Lamotte, 2025)

Utilisation des ORCiD

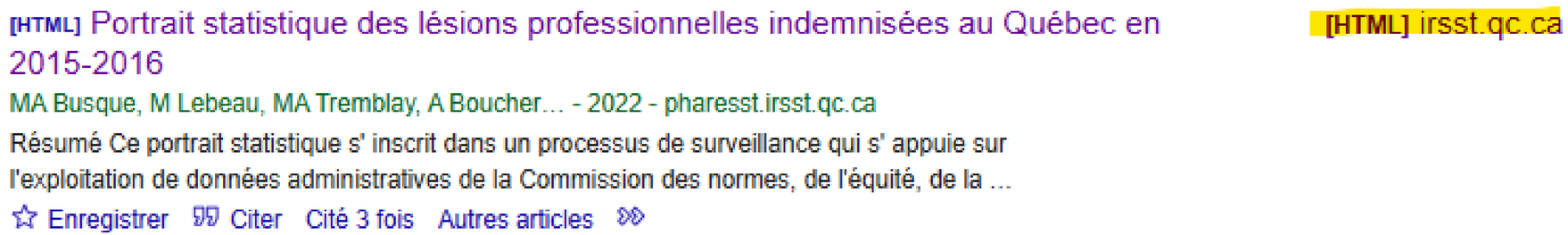
Sous forme d'hyperlien vers la page ORCiD (COAR Task Force on Supporting Multilingualism and non-English Content in Repositories, 2023; Dépôt fédéré de données de recherche, s.d.; Go FAIR, s.d.; Stoll et de Lamotte, 2025)

Bémol : pas de réelle intégration avec l'auteur et l'organisation ORCiD.
Solution : intégration Auteur-ORCiD faite dans Crossref.

Pour les moteurs de recherche : une "découvrabilité des liens"

Enjeux

- Indexation



- Indexation inégale dans OpenAlex et mauvais classement de l’accès (fermé au lieu d’ouvert)
- Balise HTML sur la langue dans les pages du dépôt : « en »

Leviers au sein de la plateforme

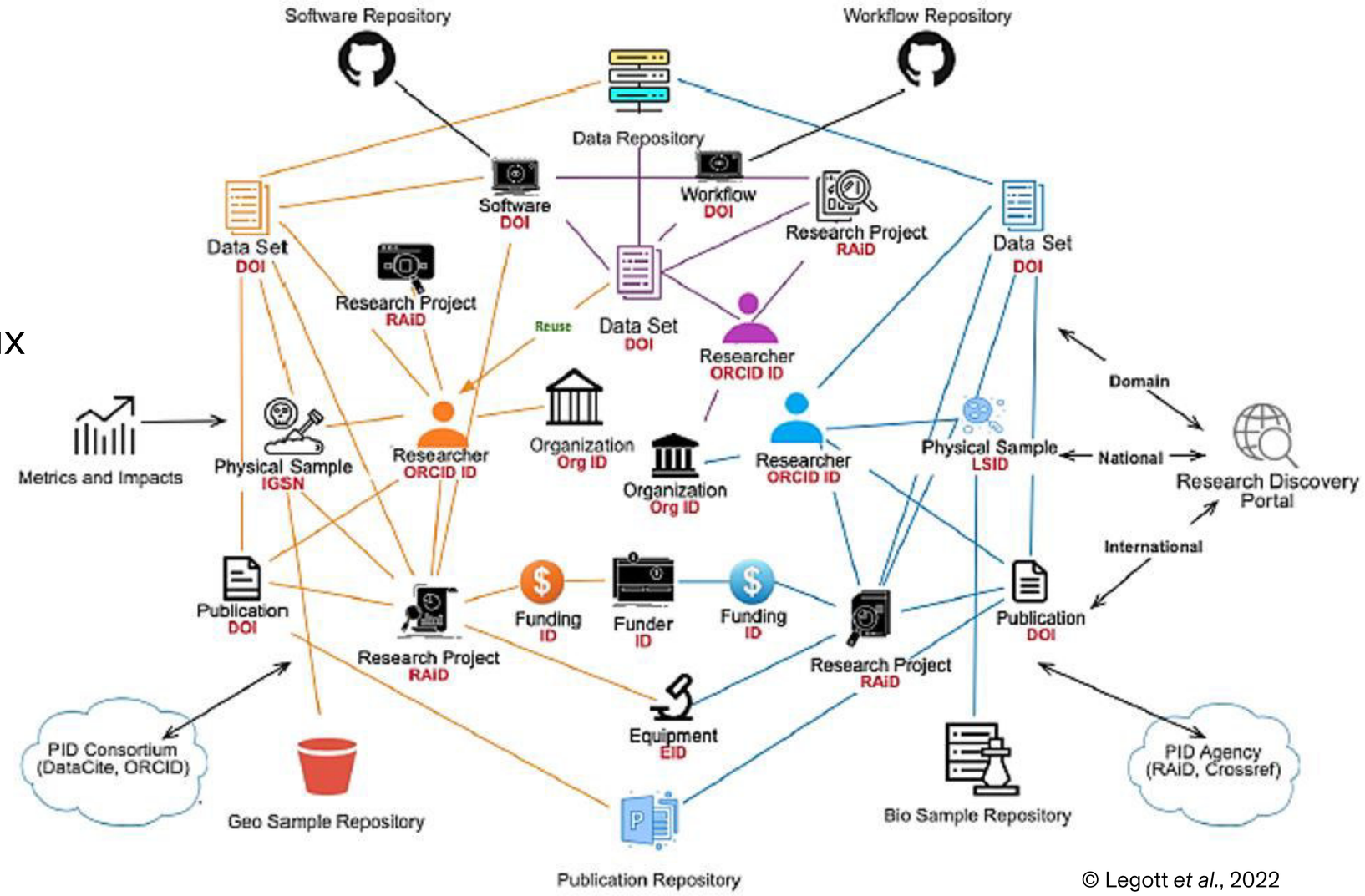


En attente de la refonte en cours

- Plus grande intégration des identifiants pérennes comme les ORCID par exemple
- Meilleure application des standards internationaux

Leviers externes à la plateforme : Crossref et les identifiants pérennes (PIDs)

- DOI comme conteneur riche
- Ajout rétrospectif des DOI, ORCiD, ROR et licences CC aux publications de l'IRSST
- Explorer les autres PIDs



© Legott *et al.*, 2022

Pour les grands modèles de langage (GML) ?
Une "découvrabilité de la réponse"

La lecture d’un DI par les GML se fait à deux niveaux

1. Pré-entraînement (mémoire de l’IA)
2. RAG (Retrieval-Augmented Generation)
   - RAG spécialisés (Elicit/Consensus)
   - RAG généralistes (Perplexity/ChatGPT)

Pour s’assurer d’être référencé

- Avoir un résumé sur la page HTML
- Incorporer schema.org
- Utiliser tous les identifiants pérennes possibles
- Métadonnées complètes
- Déclarer le libre accès via les licences CC

Conclusion

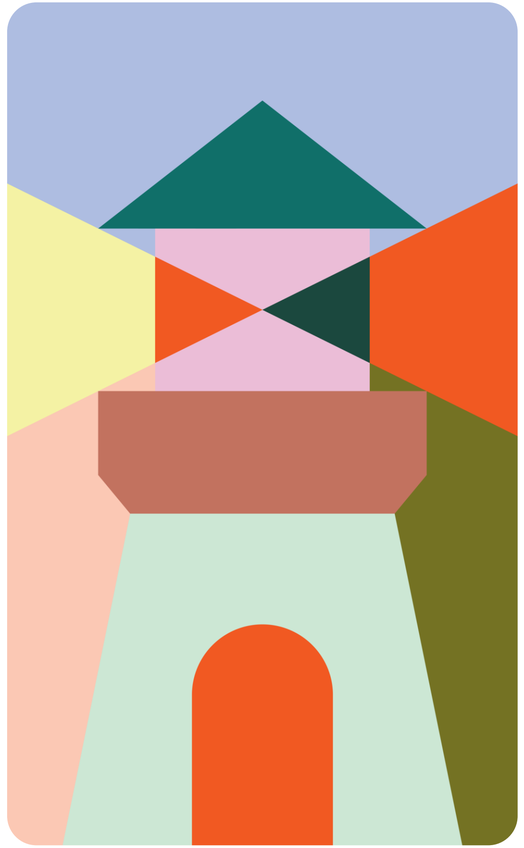
La plateforme ne fait pas tout. L’interopérabilité multilingue est hors de notre contrôle.

Certaines actions permettent quand même d'optimiser la découvrabilité et elles doivent être adaptées pour les engins de recherche ET pour les GML.

Les développements technologiques devraient amener des nouvelles opportunités pour améliorer la découvrabilité de la littérature savante dans d’autres langues que l’anglais (Matas, 2025).

Améliorer la découvrabilité n’est pas un projet ponctuel. Pour l’IRSST, c’est même un objectif du Plan stratégique 2026-2028 (à paraître).

Si la littérature grise n’est pas visible pour les machines aujourd’hui, elle risque tout simplement de ne pas exister dans les connaissances de demain.



Bibliographie

https://tinyurl.com/ykbwcus7

Merci ! Questions ?