

2024-10-07

Rapport: Comprendre et évaluer la repérabilité et l'accessibilité des données à accès restreint au Canada : Évaluation des métadonnées de sources de données de santé

Groupe d'experts sur la découverte et les métadonnées



Alliance de recherche
numérique du Canada

Digital Research
Alliance of Canada



Autrices et auteurs

Kevin Read*, Université de la Saskatchewan, kevin.read@usask.ca

Grant Gibson, Réseau canadien des Centres de données de recherche

Amber Leahey, Scholars Portal/Borealis, Université de Toronto

Lynn Peterson, Conseil national de recherches

Sarah Rutley, Université de la Saskatchewan

Julie Shi, Scholars Portal, Université de Toronto

Victoria Smith, Alliance de recherche numérique du Canada

Kelly Stathis, DataCite

* Auteur-ressource

À propos des autrices et auteurs

Le groupe d'experts sur la découverte et les métadonnées (GEDM) de l'Alliance a créé le groupe de travail sur la découverte des données à accès limité en 2021. Ce dernier a mené une évaluation des conditions de découverte et d'accès pour les données de recherche à accès restreint ou limité actuellement disponibles au Canada.



Table des matières

Résumé	4
Contexte	4
Présentation générale du projet	4
Partie 1 : Recensement et évaluation des sources de données de santé du Canada	5
Partie 2 : Identification des métadonnées communes aux différentes sources de données	6
Comparaison du couplage des métadonnées	6
Sources de données à accès limité et prochaines étapes	7
Recommandations clés.....	7
Pour les parties prenantes de l'écosystème de GDR et de découverte des données :.....	7
Pour l'Alliance de recherche numérique du Canada :.....	8
Pour l'Alliance de recherche numérique du Canada, les trois organismes et les autres bailleurs de fonds et entités gouvernementales :.....	9
Pour les bibliothèques et les dépôts nationaux :.....	9
Conclusion	10
Ressources connexes.....	11



Résumé

Dans le cadre de ce projet, nous avons recensé les sources de données à accès limité ou restreint (n = 137) et avons évalué un échantillon (n = 48) de sources de données de santé afin de mesurer la repérabilité et l'accessibilité des jeux de données pour les chercheuses potentielles et chercheurs potentiels. Pour repérer les éléments communs aux sources de données, nous avons répertorié les éléments et les avons liés aux normes actuelles sur la découverte des métadonnées et l'accès à celles-ci. Globalement, de nombreuses sources de données n'offraient que des renseignements incomplets et des métadonnées de base sur les jeux de données. Aucune source de données de l'échantillon n'avait instauré de norme des métadonnées relative au partage des données à accès restreint, ce qui limite considérablement la découverte et la réutilisation des données de santé à accès restreint au Canada.

À la lumière de ces constatations, les parties prenantes de l'écosystème des données de recherche devraient envisager de mettre en œuvre des recommandations clés et d'établir des priorités pour que les organismes collaborateurs améliorent les politiques et les systèmes de découverte des données à accès restreint et d'accès à ces données au Canada. Le présent rapport contient ces recommandations clés ainsi qu'une description du projet.

Contexte

Ces dix dernières années, il y a eu d'importantes avancées au Canada pour favoriser l'ouverture de la communauté de recherche scientifique, dont la promotion de pratiques et de politiques relatives au gouvernement ouvert et à la science ouverte ainsi que le soutien du financement des services et de l'infrastructure nationale de recherche numérique¹. L'amélioration de la découverte et de la consultation des données de recherche au Canada met en lumière la nécessité d'une collaboration accrue dans l'écosystème élargi de recherche et de découverte des données ouvertes, de sorte que toutes les données de recherche seraient repérables, particulièrement si elles sont accessibles sous des conditions d'accès contrôlé ou limité.

La découverte des données à accès restreint pose de nombreuses difficultés, dont la présence de métadonnées décrivant les données et les renseignements d'accès, comme les procédures de demande d'accès et les licences et ententes d'utilisation. Les métadonnées ouvertes pour la découverte soutiennent la repérabilité des données, les descriptions et les normes relatives aux données, le couplage de données et les citations pour la réutilisation et la collaboration entre bailleurs de fonds, établissements et communautés de recherche ainsi qu'en recherche scientifique, puisqu'elles facilitent la réutilisation ouverte des données et la collaboration interdisciplinaire dans toutes les communautés de recherche.

Présentation générale du projet

Conscient de ces défis, le groupe d'experts sur la découverte et les métadonnées (GEDM) de l'Alliance de recherche numérique du Canada a créé le groupe de travail sur la découverte des

¹ Alliance de recherche numérique du Canada (2022), *La stratégie de GDR de l'Alliance*. Consulté le 20 juin 2024. https://alliancecan.ca/sites/default/files/2024-02/rdm_strategy_2022-09_fr.pdf



données à accès limité en 2021 pour faire le point sur la découverte des données à accès restreint et sur l'accès à ces dernières au Canada². Nous avons entrepris ce projet pour recenser et évaluer les caractéristiques des sources de données à accès limité canadiennes en matière de découverte et d'accès. Par « données à accès restreint », nous entendons les données qui ne sont pas immédiatement accessibles pour des raisons commerciales, éthiques ou juridiques, ou qui ne sont disponibles que sur demande et sous certains contrôles d'accès.

Ce projet a donné lieu à deux études distinctes, qui ont été publiées en 2024^{3 4}. Des jeux de données et des documents supplémentaires sont disponibles sur l'espace de projet Open Science Framework du groupe de travail sur la découverte des données à accès limité⁵. Nous recommandons de consulter ces ressources pour connaître notre méthodologie et nos résultats.

Dans le présent rapport, nous donnons un résumé des conclusions du projet et proposons des pistes d'analyse et des recommandations à l'intention de la communauté de la science ouverte et de la gestion des données de recherche (GDR) au Canada.

Partie 1 : Recensement et évaluation des sources de données de santé du Canada

Pendant la première partie du projet, qui s'est déroulée de janvier à novembre 2021, nous avons recensé autant de sources canadiennes de données à accès restreint que possible (n = 137). Nous avons ensuite évalué les caractéristiques relatives à la découverte et à l'accès pour un sous-ensemble de source de données de santé (n = 48/137) au moyen d'un [barème](#) conçu par le groupe de travail sur la découverte des données à accès limité. Inspiré des principes selon lesquels les données doivent être faciles à trouver, accessibles, interopérables et réutilisables (principes FAIR), ce barème tient compte d'un ensemble de conditions et d'exigences pour la découverte et l'accès et attribue une note allant de A à C pour chacune d'elles⁶. Nos résultats montrent que pour les données de santé à accès restreint au Canada, il y reste du chemin à faire dans trois grands domaines :

- **Découverte ouverte** : De nombreuses sources de données n'utilisaient pas les pratiques exemplaires ni les métadonnées ouvertes pour rendre leurs données repérables. La vaste majorité des sources de données ont eu de mauvaises notes en découverte des données en raison d'une absence de norme des métadonnées et de disponibilité des métadonnées (10/48, **soit 21 %, ont reçu une bonne note**).

² Alliance de recherche numérique du Canada (s.d.), *Réseau d'experts*. Consulté le 25 mars 2024. <https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts>

³ K. B. Read, G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (2024), « Understanding the challenges associated with finding and accessing restricted data in Canada: a mixed methods study », *FACETS*, vol. 9, p. 1-9. <https://doi.org/10.1139/facets-2023-0102>

⁴ K. B. Read, G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (16 août 2024), « Identifying metadata commonalities across restricted health data sources: A mixed methods study exploring how to improve the discovery of and access to restricted datasets », *Journal of eScience Librarianship*, 13(2):e907. <https://doi.org/10.7191/jeslib.907>

⁵ K. B. Read, G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (18 août 2022), « Identifying and evaluating restricted data source in Canada: Recommendations for improving data discovery and access ». osf.io/5nh2s

⁶ K. B. Read, G. Gibson, A. Leahey, L. Peterson, S. Rutley, V. Smith et K. Stathis (13 juillet 2022), « Access-Limited data source Grading Rubric ». <https://doi.org/10.17605/OSF.IO/KC4U9>



- ▶ **Description et documentation des données** : La plupart des sources fournissaient au moins quelques renseignements décrivant le jeu de données, comme l'étendue, l'objectif et la nature de la collecte (30/48, **soit 62,5 %, ont reçu la note A ou B**). Cependant, peu de sources donnaient accès à de la documentation sur le jeu de données, p. ex. renseignements variables ou livres de codes (21/48, **soit 43,8 %, ont reçu la note A ou B**).
- ▶ **Accès aux données** : La plupart des sources donnaient au moins quelques renseignements sur les procédures d'accès aux données restreintes (33/48, **soit 68,8 %, ont reçu la note A ou B**), mais la majorité a reçu une mauvaise note globale pour l'accessibilité des données en raison des renseignements manquants (p. ex. transparence du coût d'accès et critères d'admissibilité).

Partie 2 : Identification des métadonnées communes aux différentes sources de données

À partir de notre échantillon de sources de données de santé de la partie 1 (n = 48), de décembre 2021 à mars 2023, nous avons examiné et colligé les métadonnées de chaque source. Nous avons ensuite identifié celles qui leur étaient communes et les avons catégorisées et comparées à des normes des métadonnées existantes pour évaluer leur conformité par rapport à celles-ci.

Après avoir catégorisé les métadonnées communes, nous avons recensé 35 éléments de métadonnées communs et 27 éléments d'accès communs. Pour connaître la liste complète des éléments communs, voir la publication⁷.

Comparaison du couplage des métadonnées

Les éléments communs d'accès et de métadonnées ont ensuite été couplés à cinq schémas de métadonnées spécialisés pour déterminer dans quelle mesure les normes actuelles facilitent la découverte des jeux de données restreintes et l'accès à ces données : [DataCite \(version 4.4\)](#), [DDI-Lifecycle \(version 3.1\)](#), [DDI-Codebook \(version 2.5\)](#), [DCAT Vocabulary \(version 3\)](#) et [DATS \(version 2.2\)](#).

Les éléments communs ont ensuite été évalués selon leur correspondance (parfaite, partielle ou nulle) avec les critères de chaque schéma pour déterminer leur niveau de conformité.

Conformité des métadonnées de découverte

Le schéma DDI-Lifecycle est celui qui correspond le mieux aux éléments de métadonnées des sources de données, mais il n'est pas aussi fréquemment utilisé que d'autres schémas, comme DataCite, qui a cependant produit moins de correspondances parfaites ou partielles. En outre, aucune des sources ne prenait en charge ou ne mentionnait ces normes des métadonnées.

⁷ K. B. Read, G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (16 août 2024), « Identifying metadata commonalities across restricted health data sources: A mixed methods study exploring how to improve the discovery of and access to restricted datasets », *Journal of eScience Librarianship*, 13(2):e907. <https://doi.org/10.7191/jeslib.907>



Conformité des métadonnées d'accès

Les métadonnées d'accès des schémas ne semblent pas correspondre aux éléments communs que nous avons relevés. Bien que les cinq schémas traitent tous des métadonnées d'accès, leur portée est trop vaste pour tenir compte de la granularité et de la complexité requises pour décrire les restrictions et les processus de demande d'accès aux données restreintes. Par ailleurs, le manque d'uniformité des procédures de demandes d'accès des diverses sources de données vient compliquer l'établissement de normes.

Sources de données à accès limité et prochaines étapes

Les résultats de ce projet indiquent clairement qu'il faut normaliser les métadonnées et l'infrastructure des sources de données à accès limité pour rendre ces données plus repérables et accessibles et se plier aux normes des bailleurs de fonds et des revues scientifiques, comme la Politique des trois organismes sur la gestion des données de recherche⁸. Il faut en outre améliorer la compréhension et la mise en application des pratiques entourant les données partagées et les normes de métadonnées, et soutenir davantage la découverte et l'accès ouverts aux données restreintes ou contrôlées. Des activités en ce sens peuvent être menées par diverses parties prenantes de l'écosystème de recherche numérique, dont les bailleurs de fonds, les éditeurs de travaux savants, les établissements et les fournisseurs d'infrastructure ainsi que les chercheuses et chercheurs.

Recommandations clés

Pour les parties prenantes de l'écosystème de GDR et de découverte des données :

- ▶ Adopter ou adapter des normes des métadonnées existantes, comme le schéma universel [DataCite](#) ou des schémas propres à certaines disciplines, comme [DDI-Codebook](#) et [DDI-Lifecycle](#), lorsque possible et à propos, pour améliorer la repérabilité des données de recherche, y compris celles à accès limité.
- ▶ Investir dans l'établissement d'un consensus et le développement de normes des métadonnées nouvelles ou améliorées pour faciliter la découverte des données et le partage des renseignements relatifs à l'accès aux données restreintes ou contrôlées, notamment en indiquant où trouver les métadonnées et la documentation relative aux variables des jeux de données.
- ▶ Établir une terminologie commune pour les modalités d'utilisation, les licences, les ententes, etc., et rédiger des lignes directrices pour les programmes et services d'accès aux données restreintes afin de faciliter la tâche aux chercheuses et chercheurs qui demandent l'accès et aux personnes responsables de l'octroyer.

⁸ Gouvernement du Canada (14 mars 2021), *Politique des trois organismes sur la gestion des données de recherche*. <https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche>



- ◆ Cela pourrait comprendre des procédures et politiques de demande d'accès, des ententes de partage des données, des modèles de licence et des mécanismes d'accès aux données.
- ▶ Étudier les identifiants pérennes et encourager leur adoption pour les données à accès restreint (p. ex. DOI de DataCite pour assigner des métadonnées aux données restreintes et les rendre publiquement interrogeables et utilisables au moyen d'agrégateurs et d'outils comme Lunarix).
- ▶ Dans certains cas, l'octroi d'accès peut être simplifié par des cadres ou des services existants. On peut aussi évaluer les occasions d'investissement pour développer de telles initiatives, par exemple, pour donner accès aux chercheuses et chercheurs à des sources de données de santé des administrations, des hôpitaux ou des gouvernements des différentes provinces (p. ex., initiative DASH) ou au dossier maître de microdonnées de Statistique Canada administré par le Réseau canadien des Centres de données de recherche (RCCDR), et au fichier de microdonnées à grande diffusion par l'intermédiaire de l'Initiative de démocratisation des données (IDD).
- ▶ Élaborer des lignes directrices sur les enjeux d'intendance à long terme en matière d'accès aux données restreintes, l'objectif étant de rendre la reproductibilité possible ou d'éviter les pertes de données, étant donné les limites de temps et de ressources pour la plupart des projets de recherche et des cycles de vie des données.
- ▶ Établir un dialogue et s'engager à améliorer la capacité, l'infrastructure, la formation et les services pour renforcer la gouvernance des données à accès limité et des données de recherche au Canada, notamment par la communication et la collaboration intersectorielles au privé, au public, dans le monde universitaire et dans les organismes à but non lucratif.

Pour l'Alliance de recherche numérique du Canada :

- ▶ Approcher les établissements, les gestionnaires et les fournisseuses et fournisseurs de données à accès restreint pour mieux comprendre ce qui limite la découverte et l'accès et pour étudier des pistes de solution.
- ▶ Commencer à intégrer les sources de données nommées dans cette étude aux plateformes nationales de découverte des données, comme [Lunarix](#), pour améliorer la découverte et, éventuellement, la réutilisation des données de recherche à accès restreint (lorsqu'il est approprié de le faire).
- ▶ Appliquer des approches similaires à celles abordées dans l'étude pour évaluer les besoins relatifs aux métadonnées des données de recherche sensibles et à accès restreint au Canada dans le cadre du développement continu des services et des nouvelles initiatives, comme le [Projet de gestion de l'accès contrôlé aux données de recherche](#), une collaboration ayant pour objectif de créer une feuille de route pour établir des solutions techniques et politiques et sécuriser le stockage, le partage et la réutilisation des données à accès restreint au Canada.



Pour l'Alliance de recherche numérique du Canada, les trois organismes et les autres bailleurs de fonds et entités gouvernementales :

En particulier, Innovation, Sciences et Développement économique Canada et le Bureau de la Conseillère scientifique en chef :

- ▶ Soutenir l'élaboration de processus standardisés pour améliorer la repérabilité et l'accessibilité des données de recherche restreintes. Il faudrait proposer une gamme d'options d'infrastructure adaptable aux chercheuses et chercheurs, aux établissements et aux intendantes et intendants des données pour gérer le partage et la réutilisation éthiques et légaux des données restreintes tout en respectant les compétences, les mécanismes, les processus de travail et les politiques de partage des différents territoires.
- ▶ Proposer du financement afin de soutenir l'élaboration et l'adoption de normes relatives à la découverte des données de recherche et à l'accès à celles-ci pour l'infrastructure et les dépôts de données, normes qui couvriraient la sécurité, le respect de la vie privée et la protection des données ainsi que le partage ouvert et la normalisation des métadonnées.
- ▶ Proposer du financement pour soutenir d'autres études sur les lacunes – et les moyens de les éliminer – des politiques relatives aux données et des exigences techniques connexes afin de répondre aux besoins présents et futurs en matière de stockage, de découverte, de gestion et d'accessibilité des données de recherche restreintes au Canada.
- ▶ Élaborer des programmes de formation pour les intendantes et intendants des données restreintes afin de créer de la documentation et des cadres de travail réutilisables qui maximiseront le potentiel de réutilisation des données restreintes, lorsqu'il est approprié de le faire.
- ▶ Créer une communauté de pratique regroupant les curatrices et curateurs ainsi que les intendantes et intendants de données pour faciliter l'adoption des normes et processus mentionnés ci-dessus.
- ▶ Proposer plus de financement pour la curation et la gestion des données à accès restreint, particulièrement pour la prise en charge des métadonnées et de la documentation.
- ▶ Élaborer des lignes directrices sur la gestion et l'octroi d'accès aux données restreintes conformément aux politiques des trois organismes.

Pour les bibliothèques et les dépôts nationaux :

- ▶ Former les chercheuses et chercheurs qui recueillent des données restreintes à optimiser la repérabilité, l'accessibilité et la réutilisabilité.
- ▶ Encourager les chercheuses et chercheurs à communiquer rapidement avec leur comité d'éthique de la recherche pour que le consentement des personnes participantes et l'approbation éthique reflètent les plans de découverte, d'accessibilité et de réutilisabilité des données.



- ▶ Sensibiliser à l'importance de la repérabilité et de l'accessibilité des données restreintes pour la communauté de recherche.
- ▶ Plaider auprès des organismes nationaux, dont ceux mentionnés ci-dessus, pour le développement des infrastructures de soutien pour les données à accès restreint.
- ▶ Faire part aux organismes nationaux des difficultés et de l'expérience des chercheuses et chercheurs qui veulent faciliter la découverte de leurs jeux de données à accès restreint.
- ▶ Encourager le recours à des normes minimales en ce qui a trait à la découverte des données de recherche, dont les métadonnées ouvertes ainsi que les renseignements et les procédures d'accès relatifs aux données restreintes.
- ▶ Encourager les descriptions de données et la documentation ouverte, des outils essentiels pour la compréhension, la découverte et la réutilisation des données de recherche.
- ▶ Rédiger de la documentation et fournir des conseils et des services institutionnels aux chercheuses et chercheurs ainsi qu'aux gestionnaires de données des divers établissements.

Conclusion

Dans le cadre de ce projet, nous avons étudié comment 48 sources de données de santé à accès restreint du Canada ont rendu leurs jeux de données repérables et accessibles pour les chercheuses potentielles et chercheurs potentiels, et avons extrait des renseignements et des métadonnées sur les jeux de données, dont les descriptions, les restrictions d'accès et les procédures de demande d'accès, pour recenser les éléments de métadonnées communs aux différentes sources de données. Au Canada, le potentiel de découverte et de réutilisation des données de santé à accès restreint est limité par l'absence de métadonnées sur les jeux de données et sur les méthodes d'accès, l'incapacité de l'infrastructure actuelle à soutenir les données à accès restreint, l'absence de lignes directrices à l'intention des intendantes et intendants des données qui voudraient adopter les pratiques exemplaires en matière de découverte des données, et le manque de clarté, de la part des bailleurs de fonds, sur la manière de rendre les données à accès restreint repérables et accessibles conformément aux politiques de gestion des données.

Nous recommandons fortement aux diverses parties prenantes mentionnées dans ce rapport d'établir des priorités concernant les données à accès restreint et de collaborer pour améliorer leur repérabilité. Les parties prenantes doivent être consultées de manière continue pour répondre aux besoins et aux préoccupations du grand public. Sans les métadonnées, l'infrastructure et la formation pour les intendantes et intendants de données à accès restreint au pays, ces précieuses données de recherche resteront cachées et inaccessibles.



Ressources connexes

Alliance de recherche numérique du Canada. (s.d.). *Réseau d'experts*. Consulté le 25 mars 2024. <https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts>

Read, K. B., G. Gibson, A. Leahey, L. Peterson, S. Rutley, V. Smith et K. Stathis (13 juillet 2022). « Access-Limited Data Source Grading Rubric ». <https://doi.org/10.17605/OSF.IO/KC4U9>

Read, K. B., G. Gibson, A. Leahey, L. Peterson, S. Rutley, V. Smith, K. Stathis et J. Shi (9 août 2022). « Canadian data source identification and evaluation datasets ». <https://doi.org/10.17605/OSF.IO/UBZN2>

Read, K. B., G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (19 janvier 2024). « Datasets Exploring Metadata Commonalities Across Restricted Health Data Sources in Canada ». <https://doi.org/10.17605/OSF.IO/TXRVE>

Read, K. B., G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (2024). « Understanding the challenges associated with finding and accessing restricted data in Canada: a mixed methods study ». *FACETS*, 9: 1-9. <https://doi.org/10.1139/facets-2023-0102>

Read, K. B., G. Gibson, A. Leahey, L. Peterson, S. Rutley, J. Shi, V. Smith et K. Stathis (16 août 2024). « Identifying metadata commonalities across restricted health data sources: A mixed methods study exploring how to improve the discovery of and access to restricted datasets ». *Journal of eScience Librarianship*, 13(2) :e907. <https://doi.org/10.7191/jeslib.907>