QUESTIONS DE DONNÉES

BULLETIN TRIMESTRIEL DU RCCDR

CRDCN
RCCDR

DEUXIÈME TRIMESTRE - JUIN 2024

Bienvenue au nouveau numéro de *Questions de données* !

Pour cette édition, nous avons convié nos collègues de Coalition Publica à une séance de questions-réponses sur leur organisation, qui est la seule autre initiative scientifique majeure de la Fondation canadienne pour l'innovation en sciences humaines et sociales.

Nous avons également interrogé Grant Gibson, directeur adjoint, recherche et évaluation du RCCDR, sur le module de formation Stata qu'il a développé, sur le logiciel lui-même et sur les projets de formation du RCCDR.

Enfin, nous revenons sur quelques recherches qui ont retenu notre attention lors de la présentation du Centre de données de recherche qui s'est tenue dans notre institution hôte, l'Université McMaster.

À l'approche de la belle saison, nous vous souhaitons un été serein et productif, et nous vous remercions de lire *Questions de données*.

Ryan Murphy
Rédacteur en chef, *Questions de données*

Si vous avez des commentaires ou des idées pour de futurs thèmes ou articles, n'hésitez pas à nous contacter à info@crdcn.ca.

Le RCCDR est le fruit d'une collaboration entre son personnel central, basé à l'Université McMaster, Statistique Canada et 42 institutions de recherche et d'enseignement supérieur dans tout le pays. Le RCCDR est reconnaissant pour les fonds de fonctionnement principal fournis par la Fondation canadienne pour l'innovation (FCI), en tant qu'initiative scientifique majeure, et par les Instituts de recherche en santé du Canada (IRSC) et le Conseil de recherches en sciences humaines (CRSH), en tant que plateforme de recherche stratégique. Le RCCDR remercie également de leur soutien financier Statistique Canada, les partenaires provinciaux et ses 42 institutions collaboratrices.

Crédit de photo de couverture : Canva

Mieux connaître nos collègues : entrevue avec Coalition Publica

Seule autre initiative scientifique majeure du Canada en sciences humaines et sociales, Coalition Publica s'efforce de bâtir et de gérer une infrastructure nationale ouverte dédiée à la diffusion de la recherche et à l'édition savante numérique.

Nous avons organisé une séance de questions-réponses pour permettre à nos lectrices et lecteurs d'en savoir un peu plus sur Coalition Publica.

Q1. Comment est né Coalition Publica ?

A. Coalition Publica est une initiative conjointe du Public Knowledge Project (PKP) et d'Érudit. Tous deux sont fondés en 1998 lorsque le passage à l'édition numérique a rendu la recherche universitaire plus accessible et ont joué un rôle crucial de soutien à la publication des revues en fournissant à leurs éditeurs les moyens de les diffuser auprès d'un public plus vaste.

Acteurs de longue date dans le paysage numérique canadien, Érudit et PKP ont collaboré sur de nombreux projets en sciences humaines et sociales, dont Synergies, un projet subventionné en 2011 pour stimuler la production de revues savantes canadiennes en format numérique.

Avec l'importance croissante de l'édition en libre accès, PKP et Érudit ont uni leurs forces en 2017 pour demander conjointement un financement en tant que Coalition Publica, dans le but de mieux soutenir les revues canadiennes dans la transition vers un libre accès pérenne.

Q2. De quelle manière Coalition Publica aide et accompagne les chercheuses et chercheurs ?

A. Coalition Publica est un projet de recherche financé par la Fondation canadienne pour l'innovation (FCI) et le Conseil de recherches en sciences humaines du Canada (CRSH), et à ce titre, joue un rôle de soutien à la recherche dans les disciplines de son champ de compétence. En combinant la plateforme de diffusion Érudit et Open Journal System (OJS), le logiciel phare de PKP, Coalition Publica peut offrir aux revues canadiennes des services de production et de diffusion numériques à faible coût, de même qu'un accompagnement professionnel pour l'adoption des meilleures pratiques d'édition savante en libre accès.

La survie d'une revue dépend en partie de sa capacité à financer ses activités, et pour celles qui participent à Coalition Publica, un facteur clé de viabilité à long terme est le Partenariat pour le libre accès. Ce modèle de financement collectif lancé par Érudit et le Réseau canadien de documentation pour la recherche permet aux bibliothèques d'apporter un soutien financier permanent aux revues.

En participant à Coalition Publica, les revues contribuent à augmenter la capacité de recherche en sciences humaines et sociales au Canada, car le contenu des revues et articles diffusé sur la plateforme Érudit constitue la base d'un corpus de données textuelles.

Enrichi de collections de périodiques provenant entre autres des riches fonds de Bibliothèque et Archives Canada et de Bibliothèque et Archives nationales du Québec, ce corpus est mis à la disposition des chercheuses et chercheurs pour l'étude de la recherche en tant qu'objet social, ce qui permet d'examiner la manière dont elle est développée, communiquée et construite.

Q3. Pourquoi le libre accès est-il si important ?

A. Le libre accès pérenne est au cœur de la mission de Coalition Publica, car la suppression des obstacles à l'accès à la recherche, tant pour les lectrices et lecteurs que pour les autrices et auteurs, peut grandement favoriser un système équitable de diffusion de la connaissance dont bénéficiera la société.

Coalition Publica promeut la diffusion en libre accès, qui permet de consulter des publications savantes sans frais et assure ainsi la circulation des connaissances au sein des communautés de recherche et dans l'ensemble de la société. Le libre accès est de plus en plus considéré comme une pratique de diffusion essentielle pour préserver l'écosystème dynamique et innovant de la publication savante au Canada.

À travers le monde, de plus en plus d'universitaires et de chercheuses et chercheurs encouragent et utilisent le libre accès pour que les savoirs issus de la recherche puissent remplir leur fonction essentielle, qui est de servir le bien commun.

Q4. Qu'est-ce que le libre accès apporte aux chercheuses et chercheurs et aux politiques publiques ?

A. Tout d'abord, Coalition Publica contribue à cet apport en soutenant les revues par un travail sur les normes ouvertes et les meilleures pratiques, la découvrabilité et l'indexation, les identifiants pérennes et la qualité des métadonnées, ainsi que par un soutien en matière de financement et de formation.

Ce travail en faveur du libre accès entraîne plusieurs conséquences importantes pour les chercheuses et chercheurs et les politiques publiques. Par exemple, il améliore l'accessibilité, ce qui démocratise l'information et permet aux communautés de la recherche et de la pratique, aux dirigeants politiques et à la population de tous les milieux et de toutes les régions de profiter des derniers résultats de la recherche.

Coalition Publica s'efforce d'élargir l'accès aux connaissances, ce qui accélère l'avancement de la recherche et stimule l'intégration de ses résultats dans l'élaboration des politiques publiques. La visibilité accrue, le plus grand potentiel de collaboration et la publication sans entraves favorisent l'élaboration de politiques publiques fondées sur des données probantes qui permettent des prises de décision éclairées notamment dans les domaines de l'éducation, des soins de santé, de l'environnement, de la technologie et des infrastructures.

Une autre raison d'être de Coalition Publica est de permettre aux chercheuses et chercheurs et aux institutions de réaliser des économies. Grâce à la réduction ou à la suppression des frais, les chercheuses et chercheurs peuvent réorienter les sommes dédiées vers d'autres activités de recherche. Pour leur part, les institutions peuvent optimiser l'allocation de leur financement.

Q5. Quels sont les prochains projets de Coalition Publica ?

A. L'avenir de Coalition Publica s'articule autour de trois priorités : (1) la poursuite des services d'édition et du soutien aux revues canadiennes en sciences humaines et sociales ; (2) le développement technologique partagé de logiciels et de plateformes d'édition ; et (3) la recherche sur l'écosystème de l'édition savante.

Avec l'importance croissante du libre accès immédiat dans les politiques gouvernementales (comme en attestent l'annonce en 2022 par le Fonds de recherche du Québec (FRQ) de son intention de s'aligner sur le Plan S et la révision en cours de la Politique des trois organismes sur le libre accès aux publications), une grande partie du travail de soutien aux revues de Coalition Publica consistera à s'assurer qu'elles sont équipées pour fonctionner dans un environnement de libre accès immédiat et qu'elles respectent les meilleures pratiques en matière de licences ouvertes et d'accessibilité.

En outre, Coalition Publica poursuivra ses travaux sur l'interopérabilité entre OJS et Érudit, ainsi que sur la diffusion et la découvrabilité des revues. À ce propos, nous prévoyons des avancées en matière de licences de diffusion, de protocoles, de normes, de métadonnées, d'interfaces utilisateur et d'expérience utilisateur (UI/UX), de multilinguisme et d'accessibilité. L'élaboration de mesures, le recueil de statistiques et la création de rapports à partir de données d'utilisation, de tableaux de bord et du Code de bonnes pratiques COUNTER, version 5 sont également des priorités, tout comme les collaborations internationales pour améliorer les infrastructures, les normes et la cybersécurité.

Enfin, Coalition Publica prévoit de continuer à renforcer ses collaborations avec la communauté de la recherche. Le développement des corpus du dépôt de données textuelles est en cours pour assurer une évolution cohérente des collections et préserver leur car-

actère national. Grâce à une subvention de l'Initiative pancanadienne d'accès aux connaissances (PCKAI), Coalition Publica est aussi en mesure de soutenir, avec des programmes de bourses, des jeunes chercheuses et chercheurs qui, par l'entremise de leurs projets de recherche, fournissent des informations essentielles sur l'écosystème de la communication savante. Ainsi, nous continuons à alimenter en travaux de qualité les chercheuses et chercheurs et les revues qui se joignent à la communauté du libre accès.

Ce qu'il faut retenir :

- Coalition Publica est une initiative conjointe du Public Knowledge Project et d'Érudit.
- Coalition Publica offre aux revues canadiennes des services de production et de distribution numériques à coût réduit ou nul et favorise les meilleures pratiques d'édition savante en libre accès.
- Coalition Publica et le RCCDR sont les seules initiatives scientifiques majeures en sciences humaines et sociales.

La recherche de McMaster mise en valeur au showcase du CDR

En mai, le Centre de données de recherche (CDR) de l'Université McMaster, l'institution hôte du RCCDR, a organisé son showcase annuel.

L'événement est proposé gratuitement aux étudiantes et étudiants, au personnel et aux enseignantes et enseignants de McMaster, afin de mieux faire connaître le CDR et les recherches des étudiantes et étudiants, des post-doctorantes et post-doctorants et des professeurs qui y travaillent.

Outre un aperçu des activités du CDR et des fichiers de données qu'il héberge, des présentations ont été faites sur les ressources de l'Initiative de démocratisation des données et sur les services du système d'information géographique (SIG) mis à la disposition des chercheuses et chercheurs par la bibliothèque de l'Université McMaster.

Boris Kralj, du département d'économie de McMaster, a présenté les résultats de son étude sur les effectifs des médecins canadiens basée sur des microdonnées de Statistique Canada. Par ailleurs, Grant Gibson, directeur adjoint, recherche et évaluation du RCCDR, a fait une présentation sur les politiques et procédures de réplication des données.

Le showcase a également permis à des étudiantes et étudiants de présenter des recherches menées avec des microdonnées accessibles grâce au CDR, d'illustrer les types de recherches possibles avec diverses microdonnées et de recueillir de précieux commentaires des participants.

Jhalok Talukdar, du département des méthodes, des preuves et de l'impact de la recherche en santé, a utilisé le fichier de données de l'Enquête sur la vie après le service militaire pour examiner l'augmentation des douleurs chroniques chez les vétérans canadiens.

Utilisateur du CDR de McMaster depuis environ cinq mois, Jhalok Talukdar affirme que sans l'accès à ces données, il n'aurait pas pu mener à bien son étude, qu'il a l'intention d'approfondir avec des fichiers de données longitudinales.

Jhalok Talukdar, du département des méthodes, des preuves et de l'impact de la recherche en santé. Photo: RCCDR

Il en esquisse les implications possibles pour les politiques publiques : « Les résultats de cette étude mettent en évidence le besoin urgent d'interventions ciblées pour faire face à la prévalence croissante des maladies chroniques chez les anciens combattants et pour améliorer leur santé et leur bien-être en général. »

« En permettant de mieux comprendre les facteurs des douleurs chroniques chez les anciens combattants, notre étude jette les bases de futures recherches et interventions visant à améliorer leur qualité de vie, ajoute-t-il. Les résultats soulignent le besoin de programmes de soins de santé complets qui répondent aux besoins de santé physique et mentale des anciens combattants, en particulier des plus âgés et de ceux qui souffrent d'obésité ou de problèmes de santé mentale. »

À l'aide de l'Enquête sur les dépenses des ménages (EDM), Annika Bauer, du département de kinésiologie, a examiné les achats d'aliments enrichis en vitamine D par groupes sociodémographiques. Elle a utilisé des microdonnées pour analyser des données sociodémographiques granulaires et des données sur les dépenses, ce qui lui a permis d'obtenir de précieuses informations sur la consommation de ces aliments.

Annika Bauer envisage déjà la suite de sa recherche : « la prochaine étape consistera à comparer nos résultats avec le nouveau cycle de l'EDM afin de mieux comprendre l'impact de l'inflation et de l'augmentation du coût de la vie sur l'accessibilité des aliments enrichis en vitamine D et leur adoption par les Canadiens. Comprendre comment la consommation de ces aliments est répartie au Canada permettra d'évaluer l'équité des efforts actuels en matière d'enrichissement des aliments. »

Maria Nicula, également du département des méthodes, des preuves et de l'impact de la recherche en santé, a présenté une étude utilisant l'Enquête sur la santé dans les collectivités canadiennes (ESCC) pour examiner l'incidence de la discrimination fondée sur le poids – et de ses interactions avec d'autres formes de discrimination – sur l'état de santé autodéclaré.

« Sans les microdonnées [accessibles via] le CDR, explique-t-elle, je n'aurais pas été en mesure d'établir un lien entre mes résultats et les données prédictives. J'espère que ce travail enrichira la base de connaissances [qui plaide en faveur de la] décision de faire de la discrimination fondée sur le poids un délit au même titre que d'autres formes de discrimination, compte tenu de ses effets délétères sur les personnes stigmatisées. »

Nous attendons avec impatience de connaître la suite de ces recherches menées par les membres de notre Réseau et espérons partager bientôt d'autres résultats lors de nos webinaires et événements, ainsi que dans nos instantanés recherches-politiques.

Pour mieux connaître le CDR le plus proche de chez vous, consultez le site web du RCCDR.

Ce qu'il faut retenir :

- Le Centre de données de recherche de l'Université McMaster a organisé en mai son showcase annuel démontrant le potentiel des recherches utilisant le Réseau pour accéder à des microdonnées.
- Des chercheuses et chercheurs ont présenté leurs travaux et leurs implications possibles pour les politiques publiques.
- Vous pouvez vous renseigner sur tous les CDRs du Canada à l'adresse suivante : www.crdcn.ca.

Retour aux fondamentaux avec Stata

Les chercheuses et chercheurs travaillant dans les Centres de données de recherche (CDR) du RCCDR ont le choix entre plusieurs programmes pour manipuler des données. L'un des plus populaires est Stata, qu'environ deux tiers des chercheuses et chercheurs du RCCDR utilisent dans le cadre de leurs travaux.

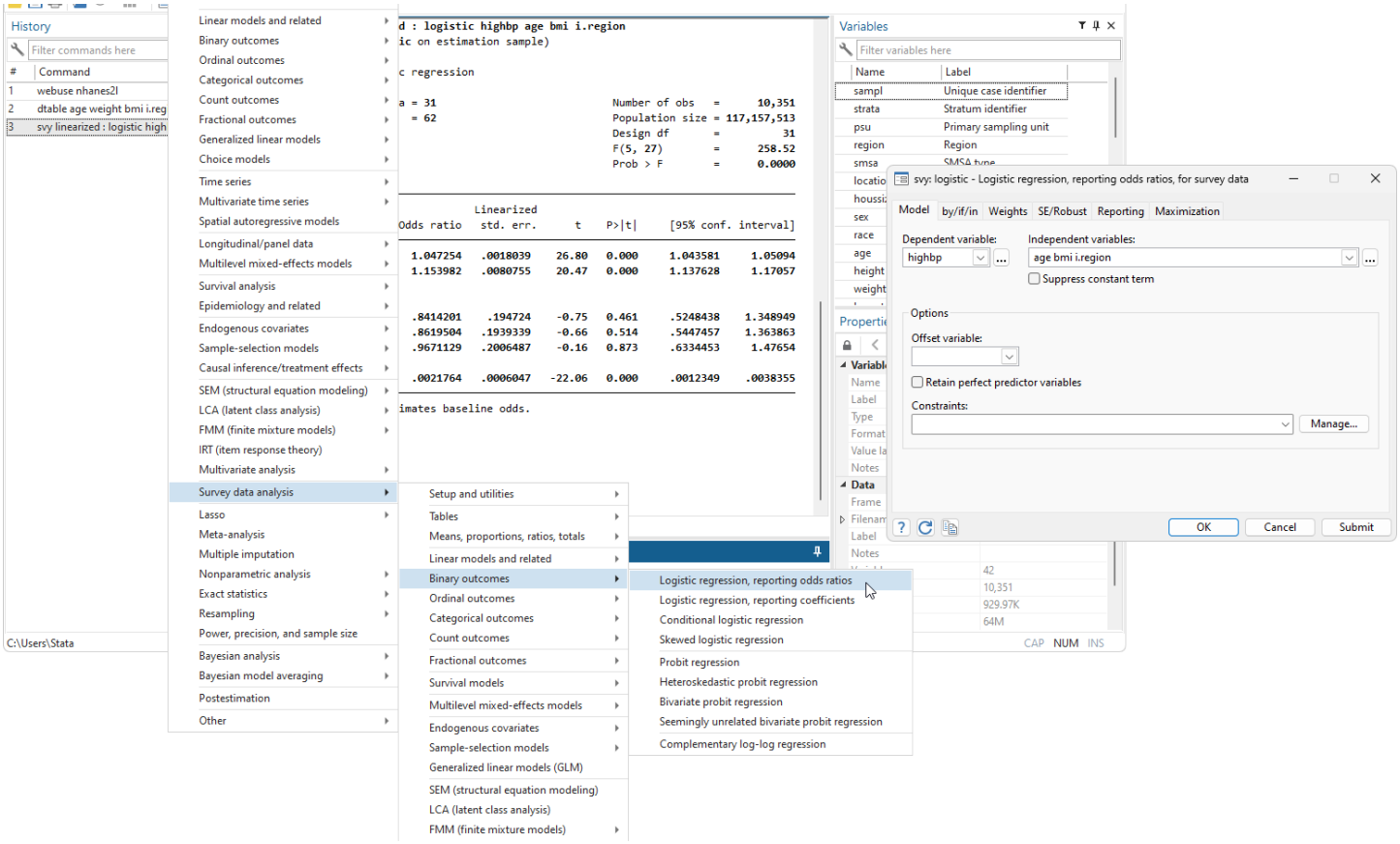
Grant Gibson, directeur adjoint, recherche et évaluation du RCCDR, a récemment créé un module de formation à Stata destiné aux chercheuses et chercheurs du RCCDR et à ceux qui souhaitent travailler dans un CDR mais n'ont pas d'expérience avec ce logiciel. Nous avons discuté avec lui de ce module, du logiciel et des avenues possibles de formation.

Grant Gibson raconte que le module Stata est né d'une demande des directeurs académiques du RCCDR, mais aussi de la volonté de rendre les CDRs plus accessibles aux utilisateurs sans connaissances en programmation statistique.

« Certaines personnes souhaitent utiliser les CDRs, mais ne savent pas comment travailler avec un logiciel de programmation statistique, explique-t-il. Microsoft Excel permet d'analyser des fichiers de données de l'ordre d'un millier d'observations, mais ce n'est plus possible avec une base de données un peu volumineuse ou pour effectuer davantage qu'un calcul ou une régression de base. »

Il constate également un écart entre ce qui est enseigné dans les cours de statistiques et les compétences requises en recherche : « Les cours de statistique se concentrent à juste titre sur l'interprétation des modèles, sur les biais et la variance entre estimateurs, et sur la résolution des problèmes d'estimation qui contredisent les hypothèses des estimateurs ».

Cependant, ajoute-t-il, « dans les CDRs, les cher-



cheuses et chercheurs consacrent une grande partie de leur temps à la manipulation des données et à la constitution d'échantillons, à la fusion de données provenant de sources multiples, à l'exportation des résultats des modèles et à la gestion des poids d'échantillonnage. J'espère que ce module de formation aidera les utilisateurs à combler d'éventuelles lacunes à ces égards. Et comme il n'est pas destiné aux utilisateurs experts des CDRs, certaines de ses sections pourront être utiles même aux débutants. »

L'acquisition de compétences sur Stata peut aussi s'avérer indispensable pour des chercheuses et chercheurs qui souhaitent collaborer avec une co-autrice ou un co-auteur utilisant principalement ce logiciel, qui est très prisé. Selon Grant Gibson, cette popularité est due en partie au fait qu'il est très intuitif, qu'il intègre des fonctionnalités pour les séries temporelles et les données de panel, et qu'il offre une capacité inégalée à obtenir un flux de travail reproductible.

Bien entendu, comme n'importe quel outil, Stata ne convient pas à toutes les tâches. Il a par exemple tendance à charger toutes les données dans la mémoire lorsqu'il est ouvert ou importé, de sorte que les chercheuses et chercheurs travaillant avec des fichiers de données administratives le trouveront lent, faute d'une préparation adéquate. « Dans de tels cas, conseille Grant Gibson, il est préférable d'utiliser un logiciel plus adapté aux grands fichiers de données ou, si la recherche ne porte que sur une partie d'un fichier volumineux, d'utiliser StatTransfer, SAS ou Polars pour créer un fichier de travail avant de le charger dans Stata. »

La popularité de Stata a également des causes démographiques, car il s'agit du logiciel de prédilection des économistes, qui sont nombreux à travailler dans les CDRs.

« Une autre raison de sa très large adoption est que, comme avec SAS, les choses fonctionnent de manière très autonome, avance Grant Gibson. Dans un CDR, il peut en effet être fastidieux d'installer systématiquement des plugins logiciels et des bibliothèques externes pour travailler sur les données. Avec le lancement du CDRv, plusieurs de ces irritants seront atténués ou complètement éliminés, et je m'attends à ce que certains utilisateurs abandonnent Stata au profit de R et de Python. »

En ce qui concerne l'avenir de la formation, Grant Gibson indique son intention d'ajouter du contenu au module Stata et invite quiconque souhaite y contribuer à le contacter.

« J'aimerais que ce soit un point de départ à partir duquel nous pourrons procéder à un élargissement, à un approfondissement et à des mises à jour. Si vous ne connaissez pas encore Stata et souhaitez voir figurer un élément dans le module, n'hésitez pas à me contacter pour que je l'ajoute », lance-t-il, avant d'ajouter qu'il créera du contenu sur le flux de travail et la recherche reproductible, ainsi que pour le logiciel R si la demande est suffisante.

Si vous avez des questions ou pour contribuer à la formation, vous pouvez contacter Grant Gibson à grant.gibson@crdcn.ca.

Ce qu'il faut retenir :

- Stata est un logiciel de programmation statistique utilisé par environ deux tiers des chercheuses et chercheurs dans les CDRs.
- Le module de formation conçu par Grant Gibson s'adresse aux personnes qui apprennent à utiliser ce logiciel pour travailler avec des données dans un CDR.
- Il est prévu d'enrichir le module et d'offrir de la formation sur le flux de travail et la recherche reproductible.