Research Group on Human Capital
Working Paper Series

Point de vue sur l'accessibilité aux données des administrations publiques

Working Paper No. 19-04

Catherine Haeck et Marie Connolly

Décembre 2019

Groupe de recherche sur le
CAPITAL HUMAIN
ESG UQÀM

https://grch.esg.uqam.ca/serie-de-cahiers-de-recherche/

Point de vue sur l'accessibilité aux données des administrations publiques

9 décembre 2019

Catherine Haeck[1], Professeure et Fellow CIRANO
Marie Connolly, Professeure et Fellow CIRANO

Groupe de recherche sur le capital humain
Département des sciences économiques
Université du Québec à Montréal

Cet article dresse un portrait de l'accessibilité des données des administrations publiques en portant une attention particulière aux données fiscales, ainsi qu'aux données des deux plus grands postes de dépenses du gouvernement du Québec, soit la santé et l'éducation. Nous ne sommes certainement pas les premières à parler du potentiel des données de source administrative : nommons, parmi d'autres, les écrits de Card et al. (2010), Einav et Levin (2014), Statistique Canada (2009) et Connelly et al. (2016). Les vertus de l'analyse quantitative pour outiller les décideurs étaient déjà mises de l'avant par Amos Tverysky et Daniel Kahneman (prix Nobel d'économie) il y a de cela 40 ans. Mais notre contribution ici est de présenter le point de vue des chercheurs québécois et discuter de leur accès aux données des administrations publiques canadiennes et québécoises. Dans ce point de vue, nous mettons l'accent sur les microdonnées administratives anonymisées sur les individus. Les données agrégées sont plus facilement accessibles, mais ces données ne permettent pas de répondre à un vaste ensemble de questions permettant de mieux comprendre le fonctionnement de notre société. Ce point de vue dresse l'état de nos connaissances sur le sujet à l'heure d'écrire ces lignes sachant très bien que l'accès aux données évolue en continu à travers le Canada, et que nous ne sommes pas en mesure de couvrir l'ensemble des initiatives à travers chaque province.

Au Canada, l'accès aux microdonnées fiscales a permis des avancées importantes dans l'étude de la distribution des revenus et de la mobilité intergénérationnelle du revenu. Au niveau de la santé, l'utilisation des données sur les individus provenant des systèmes d'information est entamée depuis plusieurs années,

[1] Auteure de correspondance, Catherine Haeck : haeck.catherine@uqam.ca

mais la démocratisation de l'accès à ces données pour des fins de recherche pourrait engendrer d'importantes retombées pour améliorer la santé des collectivités. Quant aux données en éducation, bien qu'elles existent, elles tardent à être mises à la disposition des chercheurs malgré l'importance des retombées que pourrait procurer l'accès à ces données. Au Québec, plus de 25 pour cent des n'obtiennent pas leur diplôme d'études secondaires en sept ans. Cette statistique à elle seule est inacceptable. Il faut exploiter le pouvoir des microdonnées pour identifier plus rapidement les futurs décrocheurs ainsi que les meilleures pratiques en enseignement, et ce travail doit être réalisé par des individus qui ne sont pas impliqués dans la création des programmes. L'accès et la disponibilité des données fiscales, de santé et d'éducation sont abordés tour à tour en dressant un portrait de la position du Québec relativement aux développements d'autres provinces canadiennes, notamment le Nouveau-Brunswick, l'Ontario et la Colombie-Britannique.

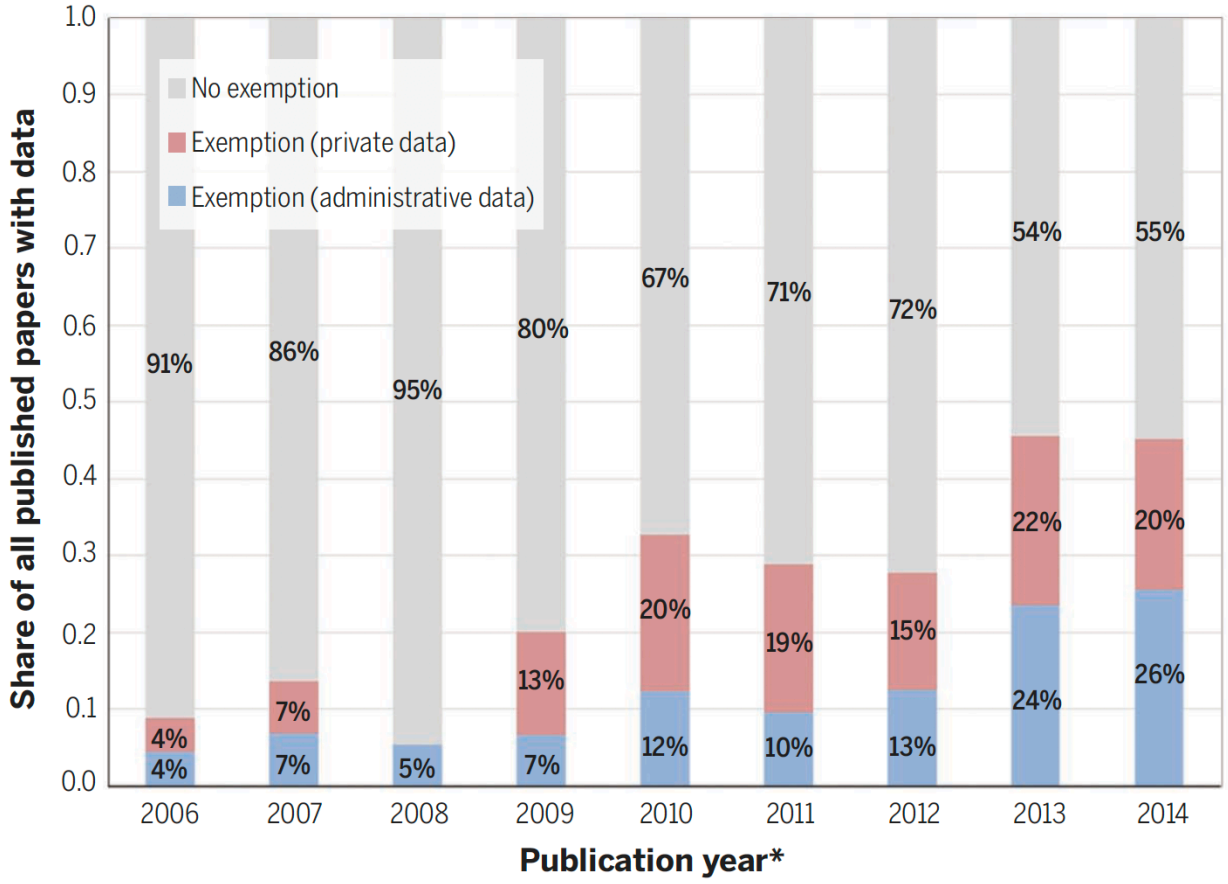
Pourquoi l'accès aux données administratives ?

L'accès aux microdonnées permet, par exemple, de faire un suivi détaillé de l'évolution de la pauvreté, des inégalités, de l'état de santé et de la mobilité socioéconomique. Ces données permettent aussi d'identifier les impacts réels sur la vie des gens des divers programmes et politiques que met en place la société. Ces microdonnées constituent une source d'information essentielle à la recherche appliquée. Grusky et al. (2019) font un compte rendu sommaire pour illustrer les bénéfices d'exploiter les données pour évaluer les politiques et les programmes telles que les allocations aux familles et les interventions durant la petite enfance. Pour nous, le message est clair : si l'on souhaite pouvoir informer les gouvernements en se fondant sur des faits (ce qu'on appelle en anglais *evidence-based*, ou *evidence-informed*, *decision making*), encore faut-il avoir des preuves sur lesquelles se baser, et des preuves qui sont valides pour l'ensemble des individus et non pas pour un petit groupe sélectionné de manière non aléatoire.

Les sources de données les plus utilisées traditionnellement sont les enquêtes, par exemple les grandes enquêtes menées par Statistique Canada, telles que l'Enquête sur la population active ou l'Enquête sociale générale. Avec la montée en puissance des ordinateurs, de leur capacité de stockage et de leur pouvoir computationnel, il existe de plus en plus de de données administratives pouvant être utilisées à des fins de recherche. En effet, les diverses administrations publiques génèrent des données qui, une fois accessibles aux chercheurs, permettent de repousser les limites de notre savoir et d'informer les dirigeants. De plus en plus d'articles publiés dans les meilleures revues scientifiques en sciences économiques sont basés sur des données de source administrative. La Figure 1, reproduite de Einav et Levin (2014), montre qu'en 2016, 26 % des articles empiriques publiés dans la *American Economic Review*, une revue évaluée par les pairs de premier plan dans le monde, utilisent de telles données, une hausse marquée depuis 2006 quand seulement 4 % des articles en utilisaient. L'accès à ces données stimule la recherche et offre à nos chercheurs une position enviable à l'échelle internationale. À ce chapitre, les pays scandinaves sont sans

contredit des chefs de file : en permettant à leurs chercheurs d'accéder à des données sur l'état civil, l'impôt, la santé ou l'éducation, ils ont réussi à susciter l'intérêt des éditeurs des meilleures revues pour des études sur le Danemark, la Norvège ou la Suède. Les meilleurs chercheurs américains sont même souvent en partenariat avec des chercheurs scandinaves afin de pouvoir accéder à ces informations pour faire des études de meilleure qualité et publier dans les meilleures revues.

Figure 1 : Parts des articles publiés dans la *American Economic Review* selon la source de données



Source : reproduite de Einav et Levin (2014)

Parmi les données administratives ayant le pouvoir d'informer nos politiques, pensons aux données fiscales recueillies chaque année par l'Agence du revenu du Canada et Revenu Québec, qui contiennent des informations précises sur les diverses sources de revenus des individus. Pensons également aux données de facturation et de prescription de médicaments de la Régie de l'assurance maladie du Québec ou encore aux résultats aux examens du ministère de l'Éducation et de l'Enseignement supérieur que les élèves passent chaque année. Quelle est la trajectoire de revenus des femmes suite à la naissance d'un enfant ? Quelle est l'efficacité d'un nouveau médicament pour traiter une maladie ? Le réseau des Centres de la petite enfance et les services de garde subventionnés ont-ils mieux préparé les enfants à l'école, et cela

se répercute-t-il en de meilleurs résultats scolaires ? Pour répondre à ces questions, il faut des microdonnées accessibles aux chercheurs, et les données des administrations publiques offrent trois grands avantages (Card et al. 2010).

Le premier avantage des données administratives, c'est la taille. Les enquêtes sont coûteuses et ne peuvent couvrir l'ensemble d'une population. En revanche, les données administratives sont recueillies auprès de chacune des personnes concernées par le programme ou l'initiative gouvernementale pour lesquels les informations sont recueillies. Le second avantage est que ces données sont par définition de type longitudinal, offrant donc un suivi à travers le temps. Puisqu'une déclaration de revenus doit être produite chaque année, les données fiscales permettent de suivre un individu d'année en année. Le troisième avantage est la qualité des données. Il faut reconnaître que les données administratives ne sont pas créées à des fins de recherche, mais plutôt à des fins administratives, de sorte qu'elles ne sont pas toujours parfaitement adéquates. Néanmoins et comparativement aux données d'enquête, les données administratives offrent l'avantage de ne pas être soumises à des problèmes comme des taux de réponse faibles — d'ailleurs en baisse ces dernières années — ou de l'attrition dans le cadre d'études longitudinales ou encore à de la sous déclaration. Ces problèmes reliés aux données d'enquête peuvent être partiellement résolus en créant des poids d'enquête, mais ce processus est coûteux et sa fiabilité dépend lourdement des informations dont les méthodologues disposent pour créer ces poids. Soyons claires : nous ne pensons pas que les enquêtes sont inutiles et devraient disparaître. En fait, leur rôle est important, surtout pour recueillir de l'information qui n'est pas disponible autrement à grande échelle ou pour cibler des populations d'intérêt bien précises. Mais nous avons tout intérêt à ce que, en parallèle, l'accès aux données de source administrative soit élargi pour mieux nous informer et pour stimuler la recherche sur les grands enjeux de notre société.

Accès aux données, environnement sécurisé

Le point de départ de l'accessibilité, c'est un environnement d'accès aux microdonnées anonymisées et hautement sécurisé. Sur ce point, le Canada est certainement un leader dans le monde. Créé en 2001, le Réseau canadien des Centres de données de recherche (RCCDR) héberge les microdonnées confidentielles de Statistique Canada dans des laboratoires informatiques hautement sécurisés. Cette infrastructure de recherche assure la protection des renseignements personnels. Les chercheurs n'ont pas les données en leur possession. Les données ne sont accessibles qu'à l'intérieur des laboratoires. Les laboratoires sont sous surveillance constante et ne sont accessibles qu'aux chercheurs autorisés. Les ordinateurs ne sont pas connectés à l'internet et il est impossible d'utiliser des clefs USB pour sauvegarder des informations. Les résultats tirés de l'analyse des données doivent toujours être validés par un analyste de Statistique Canada qui s'assure que l'information divulguée ne permet pas d'identifier un individu. De plus, la structure des ensembles de données ne permet pas d'obtenir en seulement quelques clics

l'information sur un individu en particulier. Les ensembles de données se comptent en centaines de sous-ensembles et tous les sous-ensembles ne sont pas accessibles aux chercheurs. Seules les données pertinentes au projet le sont.

En pratique, le RCCDR réunit plus de 32 laboratoires à travers le Canada situés en milieu universitaire, dont cinq au Québec dans le réseau du Centre interuniversitaire québécois de statistiques sociales (CIQSS). Ces laboratoires permettent l'accès à plusieurs types de données, incluant les données d'enquêtes, les données de recensement et certaines données des administrations publiques. Il est aussi possible d'effectuer des couplages de données grâce au soutien de Statistique Canada, ce qui permet d'enrichir des bases de données existantes. Ces couplages sont réalisés par Statistique Canada car les clés d'appariement ne sont jamais rendues disponibles aux chercheurs. Durant la dernière décennie, un effort soutenu de la part de Statistique Canada et du RCCDR a permis la création d'un large éventail de données des administrations publiques rendues disponibles aux chercheurs. Bien que les progrès faits soient importants, certains types de données et de certaines provinces sont plus accessibles que d'autres.

Au Québec, l'Institut de la statistique du Québec (ISQ) offre l'accès à des données d'enquête et certains ensembles de données administratives dans ses centres d'accès aux données de recherche (CADRISQ). Le fonctionnement des laboratoires de l'ISQ est extrêmement similaire à celui des laboratoires de Statistique Canada. Au Nouveau-Brunswick, le *New Brunswick Institute for Research, Data and Training* (NB-IRDT) est un chef de file au pays en termes de développement de l'utilisation de données de nature administrative, particulièrement en matière de santé. Ce développement a été rendu possible par plusieurs modifications des lois entourant la protection des renseignements personnels pour des fins de recherche, ainsi que la passation de deux projets de loi (projet de loi 57 en mai 2017 et projet de loi 29 en juin 2019) permettant à NB-IRDT de recevoir des données anonymisées, mais ayant un identifiant unique permettant le couplage entre les diverses sources de données.

Le réseau national ou les réseaux provinciaux ont chacun leurs avantages. Les données de Statistique Canada ont l'avantage d'offrir un portrait pancanadien, ce qui permet de mener des recherches comparatives entre les provinces et d'apprendre en observant ce qui se fait ailleurs. Les ensembles de données de l'ISQ, en contrepartie, abordent des thématiques délaissées par Statistique Canada. Par exemple, alors que Statistique Canada n'a plus aucune source de données récentes sur les enfants à travers le Canada, l'ISQ mène plusieurs enquêtes sur cette population. Dans un monde idéal, ces deux réseaux travailleraient en étroite collaboration afin d'offrir des données à la fois représentatives des différences locales et représentatives de l'ensemble du pays.

Il est évident que les infrastructures permettant un accès sécurisé existent déjà. Les avancées récentes dans l'accès aux données du Québec suggèrent une certaine ouverture du gouvernement, mais pour

obtenir un réel accès, le gouvernement devra en faire plus, commençant par une réforme de la Loi sur l'accès à l'information pour des fins de recherche.

Données fiscales pancanadiennes

Des microdonnées fiscales longitudinales sont disponibles à l'échelle canadienne notamment via la Banque de données administratives longitudinales (DAL) ainsi que la Base de données sur la mobilité intergénérationnelle du revenu (BDMIR) et l'Étude longitudinale et internationale des adultes (ELIA). Ces ensembles de données sont disponibles dans l'ensemble des centres de données de recherche du RCCDR. Les données fiscales reposent sur les données des « fichiers T1 des familles », provenant du fichier de données fiscales recueillies chaque année auprès de tous les contribuables canadiens dans le cadre de leur déclaration de revenus. Ces données fiscales sont disponibles à partir de 1982. Ainsi, il est possible de suivre les trajectoires de revenus de milliers de Canadiens et d'étudier les facteurs influençant les inégalités de revenus ou bien la mobilité économique d'une génération à l'autre.

Nos propres travaux basés sur la BDMIR ont permis de démontrer que, parallèlement à une hausse des inégalités de revenus depuis le début des années quatre-vingt, nous avons connu une hausse du degré auquel ces inégalités se perpétuent d'une génération à l'autre (Connolly, Haeck et Lapierre 2019). Que ce soit pour l'ensemble du Canada ou pour chacune des provinces, la transmission intergénérationnelle du revenu, soit le pouvoir explicatif du revenu parental sur le revenu d'un jeune à l'âge adulte, a augmenté entre les cohortes nées au début des années soixante et celles du milieu des années quatre-vingt. Autrement dit, la mobilité socioéconomique intergénérationnelle est en baisse : si l'on visualise la distribution des revenus comme une échelle, non seulement les échelons sont devenus plus distants l'un de l'autre, mais il est plus difficile de gravir ces échelons. Un tel travail de recherche, qui est basé sur des millions de jeunes et leurs parents, n'est réalisable que grâce à l'accès aux données fiscales. Les données fiscales fournissent également une source d'information fiable quant aux revenus, et leur couplage avec d'autres données est très utilisé par Statistique Canada. Un bon exemple est celui du Recensement canadien de la population : plutôt que de demander aux répondants « quel est votre revenu » comme c'est le cas pour une enquête, Statistique Canada couple les données du Recensement avec des données de l'Agence du revenu du Canada afin d'avoir des informations beaucoup plus exactes sur le revenu. Ceci réduit le fardeau de réponse et augmente la précision des résultats.

Santé, un grand chantier en cours

À notre connaissance, la première province à avoir créé un centre de recherche pour étudier des questions de santé à partir de données médico-administratives (réclamation des médecins et des hôpitaux) est le Manitoba. Marchessault (2011) documente que dès les années soixante-dix, Dr. Noralou Roos et Dr. Paul

Henteleff se rencontrent pour discuter et commencer le développement de ce qui va devenir en 1991 le Centre des politiques de santé du Manitoba (*Manitoba Centre for Health Policy*). Dès le départ, leur modèle utilise les données du monde réel (*real-world data*) pour étudier des questions de santé. On parle aujourd'hui de *pragmatic clinical trial*, mais cette idée ne date pas d'hier. Plusieurs facteurs ont contribué au développement du modèle manitobain. Les données étaient colligées à un seul endroit par une seule organisation. Ceci a permis d'établir plus facilement une relation de collaboration et de partage d'information entre les chercheurs et les administrateurs publics. De plus, la petite taille de la province, à une époque où les capacités d'analyse des données administratives n'étaient pas celles que l'on connaît aujourd'hui, a aussi facilité ce développement. La compétence des personnes engagées dans ce projet novateur, ainsi qu'un financement stable, ont grandement contribué au succès du projet.

Depuis plus de 25 ans, l'Ontario a développé un centre de recherche pour étudier des questions de santé à partir de données administratives : ce centre est aujourd'hui connu sous le nom de *Institute for Clinical Evaluative Sciences* (ICES). On considère aujourd'hui l'ICES comme un leader non seulement au Canada, mais aussi à l'international. Ce regroupement de plus de 250 chercheurs-cliniciens de haut calibre mène des recherches sur des thèmes variés en santé des populations avec les données médico-administratives de l'Ontario. L'ICES est un dépositaire légal des données. Les données sont obtenues à la source et l'institut a un financement de base du ministère de la Santé de l'Ontario qui lui assure une stabilité financière et lui permet de préparer les données pour des fins de recherche. L'institut travaille en partenariat avec le ministère afin de répondre à certaines questions. Depuis près de cinq ans, les données médico-administratives sont complémentées par des données sociales, incluant des données des autres ministères. L'ICES gère aussi les données de santé de certains groupes autochtones avec qui ils travaillent en étroite collaboration. De manière générale, les données de l'ICES sont accessibles à tout chercheur faisant une demande d'accès. En pratique, les coûts d'accès sont très élevés et peuvent constituer un frein au développement de la recherche pour les chercheurs ne faisant pas directement partie du regroupement.

Une autre initiative provinciale est celle du Nouveau-Brunswick. NB-IRDT permet l'accès et la fusion d'un remarquable éventail de données longitudinales : *Citizen Database, Physician Billing, Vital Statistics, NB Suicide Registry, NB Cancer Registry, Healthy Toddler Assessment, NB Prescription Drug Program*, et la liste continue (voir la liste complète). La base de données *Citizen Database* inclut des données démographiques et géographiques (p.ex. le code postal) sur tous les résidents du Nouveau-Brunswick. La base de données *Physician Billing* inclut tous les paiements faits aux médecins par patient, ainsi que les paiements réciproques et les salaires versés aux médecins. Cette base de données est similaire à celle de la Régie de l'assurance maladie du Québec, à la différence qu'elle est entièrement accessible dans le laboratoire NB-IRDT et peut-être facilement jumelée avec d'autres sources de données (jumelage réalisé par l'analyste sur place). La base de données *Vital Statistics* est similaire à la base de données des fichiers de naissances accessibles dans les laboratoires de l'ISQ. Mais encore une fois, au Nouveau-Brunswick il

est possible de lier la base de données avec d'autres sources, ce qui n'est pas le cas à l'ISQ, sauf à grands frais et de manière moins automatisée. Finalement, le *Healthy Toddler Assessment* inclut les résultats de l'évaluation volontaire des enfants de 18 mois à propos de la santé visuelle, dentaire et auditive, du développement langagier, de la croissance, des habiletés physiques et des relations sociales. Ces données rappellent celles collectées par les CLSC durant la petite enfance, mais qui ne sont pas rendues accessibles à des fins de recherche. L'accès aux données de NB-IRDT et le couplage de données sont gratuits pour les chercheurs universitaires, et ce même pour les données médico-administratives.

À l'échelle pancanadienne, il existe une plateforme pour les données sur les médicaments, C-NODES. Le futur est dans le développement d'une plateforme pancanadienne des données médico-administratives. Les Instituts de recherche en santé du Canada (IRSC) financent C-NODES et aussi le développement d'une plateforme pancanadienne pour les microdonnées de santé via sa Stratégie de recherche axée sur le patient (SRAP). Plusieurs unités de SRAP existent, dont une au Québec. Par contre, la fusion de ces données à l'échelle canadienne est loin d'être réalisée et il semble bien que ce projet de couplage ne se concrétisera pas avant plusieurs années.

Au Québec, certains groupes de chercheurs ont des accès privilégiés aux données administratives de la santé, quoique rien de l'ampleur de l'ICES. L'Institut de la statistique du Québec a récemment développé une entente avec le ministère de la Santé pour gérer l'accès aux microdonnées, mais le processus vient d'être mis en application. De nouvelles modalités ont été mises en œuvre suite au lancement en juin 2019 du nouveau Guichet d'accès aux données de recherche. Les données sont maintenant accessibles dans les Centres d'accès aux données de recherche de l'ISQ (CADRISQ) dont certains sont situés en milieu universitaire. Ce modèle va cependant devoir évoluer pour permettre un accès plus simple aux chercheurs en balisant le rôle de la CAI dans l'approbation des demandes d'accès. Certains doutent de la qualité des données administratives, puisque celles-ci ne sont pas collectées pour des fins de recherche, mais bien pour des fins administratives : par exemple, pour le paiement des médecins. Lorsque l'objectif est purement administratif, le montant facturé doit être juste, l'intervention facturée doit aussi l'être puisqu'elle est associée au montant facturé, mais le diagnostic associé à l'intervention, par exemple, n'entre pas dans la facturation et peut donc être une donnée de moindre qualité. Bien entendu, si ces données commencent à être utilisées par les chercheurs-cliniciens pour suivre l'évolution de certaines conditions et faire des analyses sur des données réelles (*real-world data*), alors l'incitatif pour redresser la qualité des données va naître. On sait que les expériences contrôlées ont par ailleurs aussi leurs limites. Obtenir des résultats dans l'ensemble de la population est la voie de l'avenir. Le Québec doit se réveiller et utiliser de manière démocratique le fort potentiel de ses données administratives à travers différents ministères pour répondre à des questions qui touchent directement le bien-être de sa population.

Éducation et petite enfance, l'enfant pauvre des données

Pour ce qui est des études postsecondaires, la Plateforme de liens longitudinaux entre l'éducation et le marché du travail (PLEMT) est un excellent exemple de ce qui peut être fait à l'échelle canadienne. Cette source de données est une plateforme qui permet le couplage de plusieurs sources d'information. Au cœur de cette plateforme se trouvent le Système d'information sur les étudiants postsecondaires, qui contient des données sur chaque étudiant dans tous les établissements postsecondaires canadiens (principalement, les universités), ainsi que le Système d'information sur les apprentis inscrits, lequel fournit des informations similaires, mais pour les programmes de formation de qualification professionnelle et d'apprentissage. Ces informations sont disponibles depuis 2009 pour l'ensemble des provinces canadiennes et permettent donc le suivi du nombre de diplômés par établissement et par programme. La plateforme contient également des clefs de couplage permettant de faire le lien avec les données fiscales contenues dans les fichiers de déclarations de revenus. Il est donc possible de suivre les jeunes diplômés sur le marché du travail et de répondre à une foule de questions. Par exemple, Frenette (2019) exploite ces données et compare les revenus d'emploi des jeunes ontariens selon qu'ils aient un diplôme de niveau universitaire, collégial, ou aucun diplôme postsecondaire, en plus de différencier selon le revenu parental du jeune. Il trouve qu'un diplôme universitaire procure un rendement important sur le marché du travail, et que les jeunes dont les parents avaient un faible revenu en bénéficient encore plus. Ce qui est remarquable de cette plateforme est que toutes les institutions postsecondaires sans exception y participent. Grâce à un degré de coopération nationale peu observé dans d'autres domaines, les chercheurs et les décideurs auront accès à des informations d'une portée jusqu'ici inégalée.

En revanche, il existe un grand vide concernant les données sur l'éducation primaire et secondaire, ou même de la petite enfance. Pourtant, le discours du budget fédéral prononcé le 22 mars 2016 par l'honorable Bill Morneau faisait explicitement référence à l'importance des données sur les enfants : « Il est impossible de mettre en œuvre des politiques efficaces sans prendre appui sur des données rigoureuses. Si nous souhaitons sortir les enfants de la pauvreté, nous devons d'abord en comprendre la cause ». Or les données pancanadiennes sur les enfants et les jeunes sont rares. Les données provenant des systèmes d'information des ministères le sont encore plus.

Nous avons tenté d'obtenir les microdonnées des systèmes d'information des Commissions scolaires du Québec afin d'étudier l'impact de la réforme des congés parentaux sur la réussite scolaire. Nous avions mené un projet pilote avec succès auprès d'une commission scolaire afin de valider que l'information dont nous avions besoin était disponible. Cependant le processus à la Commission de l'accès à l'information s'est avéré une barrière demandant un trop fort investissement en temps. Nous avons donc abandonné le projet. Ces données avaient un fort potentiel analytique pour répondre à une variété de questions

pertinentes pour le secteur de l'éducation, mais aussi pour les familles elles-mêmes. Les données que nous cherchions à obtenir sont disponibles dans d'autres provinces.

Sur ce point, le Nouveau-Brunswick et la Colombie-Britannique semblent devancer l'ensemble des provinces canadiennes. Les microdonnées du système d'éducation primaire et secondaire sont récemment devenues accessibles au Nouveau-Brunswick, et leur documentation est en train d'être développée. À l'heure actuelle, ces données longitudinales individuelles peuvent facilement être couplées avec tous les ensembles de données administratives de NB-IRDT dont nous avons discuté plus haut. Les clefs de couplage existent déjà. L'accès aux données et le couplage des données sont gratuits pour les chercheurs universitaires. Ce modèle en faveur de la recherche est différent du modèle québécois. En admettant que les données deviennent disponibles, le modèle québécois n'a pas créé des environnements de couplage entre les sources administratives, et quand des couplages sont effectués, les chercheurs doivent assumer des coûts importants pour financer ces couplages et les clefs de couplage doivent être détruites. Ce modèle ne favorise donc pas la recherche et engendre une redondance de coûts pour les contribuables québécois.

En Colombie-Britannique, les politiciens ont décidé de mettre les enfants au cœur de leurs priorités, et pour ce faire ils se sont donné des outils pour pouvoir évaluer leur parcours de vie durant le primaire et le secondaire. Depuis plus de 10 ans, les chercheurs ont accès aux microdonnées du ministère de l'Éducation de la province pour la période allant de 1995 à 2016. Ces données incluent l'ensemble des enfants inscrits à l'école en Colombie-Britannique. Un identifiant unique est disponible pour chaque enfant afin de pouvoir le suivre durant tout son parcours primaire et secondaire. Chaque année, pour chaque élève, plusieurs informations sont recueillies, dont les suivantes : (1) des variables démographiques concernant l'élève (sexe, langue parlée à la maison, identité autochtone, date de naissance, code postal), (2) des informations sur le programme d'études (année d'étude, langue seconde, éducation spécialisée, immersion française, cours suivis au secondaire, nombre de reprises de cours), (3) des informations sur l'école (identifiant unique banalisé de l'école au 30 septembre, école publique ou indépendante), (4) les résultats aux tests provinciaux et les notes par matière (participation, rang centile, etc.) en 4^{e}, 7^{e}, 10^{e} et 12^{e} année, le nombre d'écoles où l'enfant a été inscrit durant son parcours et l'état d'obtention du diplôme et (5) des caractéristiques du quartier de résidence (valeur moyenne des maisons, pourcentage de personnes vivant sous le seuil de faible revenu, pourcentage de personnes par type de diplôme, etc.). Ces données sont accessibles aux chercheurs universitaires par l'entremise de PopData BC et éventuellement Statistique Canada. L'accès est conditionnel à faire un projet qui est dans l'intérêt public. Pour environ 20 pour cent des enfants de 4^{e}, 7^{e}, 10^{e} et 12^{e} année, il est possible d'accéder aux informations collectées en 2016-17 lors d'une enquête de satisfaction auprès des jeunes et leurs parents.

Ces données font partie d'un projet plus vaste qui vise à comprendre le parcours des enfants et des jeunes. En pratique, ces données peuvent être couplées avec un large éventail de données administratives aussi

sous la responsabilité de PopData BC. Ces données peuvent ainsi être fusionnées avec les données du système d'éducation postsecondaire, les données des fichiers de naissance et les données du système d'assurance sociale. En théorie, elles peuvent aussi être couplées aux données du système de la santé, mais il semblerait qu'en pratique ça ne sera pas possible à court terme. Le couplage des données n'est pour l'instant pas gratuit en Colombie-Britannique.

Il est évident que les provinces doivent collaborer pour parvenir à instaurer un système d'information comparable à travers le Canada. Nous avons besoin de microdonnées comparables et représentatives des enfants de chaque province. Au-delà des données scolaires, nous avons besoin de données sur leur parcours de vie, leur bien-être, leur réussite à l'école, leur développement comportemental et cognitif, leur milieu familial, leur milieu scolaire, etc. Ces données maintenues en continu permettent aux chercheurs d'évaluer les programmes et les politiques qui touchent les enfants et de voir s'ils en sortent gagnants ou perdants. Historiquement, le Canada avait l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) de Statistique Canada, une enquête d'une richesse peu commune, couvrant plusieurs cohortes d'enfants de zéro à cinq ans. La variété de questions couvertes et d'informations recueillies permettait d'aborder plusieurs questions concernant les enfants. Cette enquête a débuté en 1994 et a été discontinuée en 2008. Aucune autre enquête pancanadienne n'offre la richesse de l'ELNEJ depuis.

Les données de l'ELNEJ ont permis l'analyse du programme québécois des services de garde à contribution réduite. Plusieurs études — mentionnons Baker et al. (2008), Lefebvre et Merrigan (2008), Haeck et al. (2015) et Haeck et al. (2018) – montrent que la transformation de l'offre de services de garde a eu un impact important sur la participation au marché du travail des mères, mais a eu peu d'effets, voire même des effets négatifs à court terme, sur le développement des enfants. Une combinaison de facteurs, incluant la qualité variable des services et l'intensité de garde accrue, peut expliquer ces résultats chez les enfants. Le trouble déficitaire de l'attention avec ou sans hyperactivité (TDAH) et l'anxiété chez les jeunes sont des sujets dont on parle de plus en plus. Déjà en 2006 et 2008, les données de l'ELNEJ nous informaient d'une tendance à la hausse de l'anxiété (Haeck et al. 2018). Mais depuis, plus d'information. Ce type de données permet de veiller au grain, d'avoir un portrait représentatif de nos enfants, et non pas un portrait basé sur des opinions ou des situations possiblement réelles. Il nous permet aussi de voir émerger des changements et d'essayer d'en comprendre les causes.

Au Québec, l'Institut de la statistique du Québec a mené des enquêtes portant sur les jeunes, l'une des plus connues étant l'Enquête longitudinale sur le développement des enfants du Québec (ELDEQ). Cette enquête longitudinale se concentre sur une cohorte d'enfants nés en 1998. Certaines enquêtes, telles que l'Enquête québécoise sur le développement des enfants à la maternelle (EQDEM), offrent aussi des portraits statistiques transversaux de différentes cohortes d'enfants à un moment de leur vie. Bien que ces

données soient riches et pertinentes, elles ne nous permettent pas de se comparer le Québec aux autres provinces canadiennes.

Nous avons besoin d'un effort concerté de collecte de données de nature administrative couplées avec des données d'enquête. Plusieurs ensembles de données existent déjà. Les fichiers de naissances sont pancanadiens et accessibles dans les laboratoires de Statistique Canada. Ils offrent une première mesure du développement de l'enfant, soit son poids au moment de la naissance. Quelques informations socioéconomiques sont aussi disponibles. L'Agence du revenu du Canada possède les relevés fiscaux de tous les Canadiens : il serait donc possible d'avoir un profil prénaissance et postnaissance de la situation économique des parents. Statistique Canada a fusionné ces données ensemble. Les données de l'EQDEM pourraient être fusionnées avec ces données administratives. Si on y ajoutait les microdonnées administratives du ministère de la Famille et celles du ministère de l'Éducation et de l'Enseignement supérieur, on aurait un portrait détaillé du parcours de vie de nos enfants sur le plan économique et éducatif.

En effet, au Québec, les résultats des étudiants, leur parcours dans le système scolaire, la présence ou non d'un plan d'intervention et leur lieu de résidence sont déjà colligés dans le système d'information du ministère de l'Éducation et de l'Enseignement supérieur et celui des commissions scolaires, mais ces informations ne sont que très partiellement accessibles à des fins de recherche. Les données de santé sont colligées par les différentes agences de santé à travers le pays. La Direction de la protection de la jeunesse a aussi des données sur les enfants et les jeunes. Le ministère de la Famille a des données sur les parcours en service de garde. Mises ensemble, ces données offriraient un portrait représentatif de plusieurs dimensions déterminantes de la vie d'un enfant. Ces données pourraient être complétées par les informations présentes dans les recensements canadiens, ou par des enquêtes ciblées sur des thèmes tels que le bien-être et les comportements.

Ce type de fusion multi-sources n'est pas possible à court terme à l'Institut de la statistique du Québec à coûts raisonnables, principalement parce que la Loi sur l'accès à l'information rend l'exercice très difficile, que les clefs de couplages n'existent pas et que les données fiscales longitudinales, par exemple, ne sont pas disponibles à l'Institut et encore moins fusionnées avec celles des naissances. Devant l'impossibilité de réformer nos lois, la meilleure avenue est de collaborer avec Statistique Canada. On limite ainsi les coûts et on augmente la possibilité que des chercheurs utilisent les données.

Enfin, la proposition de création d'un Institut en éducation est possiblement un pas dans la bonne direction si cet organisme aura comme mandat, notamment, de faire de la recherche basée sur des microdonnées québécoises en éducation et des approches statistiques permettant d'identifier des liens de causalité. Si l'institut d'avère être uniquement un agrégateur de recherche effectuée au Québec ou de recherches descriptives pour valider l'implantation de certaines approches sans en mesurer l'impact causal sur les

enfants, les avancements seront limités. Les analyses descriptives et qualitatives ont leur valeur, mais elles doivent être appuyées par des analyses quantitatives permettant de mesurer l'impact de court et long terme sur les enfants et les enseignants. Le domaine de l'éducation est submergé par des analyses de corrélations qui rendent l'identification des pratiques réellement efficaces peu évidente. Nous avons la capacité au Québec d'innover et d'évaluer nos innovations pour choisir celles qui doivent être privilégiées. Pour ce faire, il faut que les outils d'évaluation, dont les microdonnées longitudinales provenant de sources administratives, soient accessibles aux chercheurs.

Conclusion

Le gouvernement fédéral a fait des avancées importantes en matière d'accès aux données administratives dans les cinq dernières années. L'environnement de couplage des données créé par Statistique Canada permet de limiter le coût des fusions de données et d'augmenter les possibilités de recherche. Un exemple frappant est la possibilité de couplage avec des données fiscales (comme pour le Recensement) et les bases de données dérivées des fichiers fiscaux (comme la Banque de données administratives longitudinales). Par contre, les données portant sur la santé et l'éducation sont de propriété provinciale.

La quantité d'information collectée au Québec par les administrations publiques, incluant celles en santé et en éducation, est phénoménale. Le manque d'accessibilité pour des fins de recherche menée par des chercheurs universitaires non financée par le secteur privé freine la recherche et la capacité des chercheurs à informer nos dirigeants. La recherche se fait donc souvent à partir de données collectées par les chercheurs via des enquêtes payées par des fonds publics, mais qui souvent demeurent la propriété des chercheurs ayant obtenu la subvention. Il y a donc multiplication des enquêtes aux frais des contribuables. Cette situation n'est pas justifiée dans un environnement où les données sont anonymisées et sont disponibles uniquement dans des laboratoires sécurisés comme ceux du CIQSS.

De plus, si le Québec ne participe pas aux initiatives pancanadiennes pour la création d'ensembles de données à grande échelle, il devient impossible pour les chercheurs de faire des analyses comparatives en santé et en éducation. Pour rendre ces données accessibles et utiles, il faut une volonté politique au niveau provincial et il faut surtout réformer la Loi sur l'accès à l'information. Le Québec est la seule province au Canada qui exige des chercheurs de passer par deux paliers d'approbation pour accéder à des données : premièrement une approbation de la Commission d'accès à l'information et deuxièmement une approbation du ministère concerné et détenteur des données. Même l'Institut de la statistique du Québec doit se plier à cet exercice, ce que Statistique Canada n'a pas besoin de faire. La Commission d'accès à l'information n'est pas outillée pour faire face à des demandes utilisant des méthodologies statistiques complexes impliquant plusieurs sources de données. Le processus d'accès est un labyrinthe de formulaires et de démarches qui ne semble pas avoir de fin lorsque les sources de données sont multiples. La Loi sur

l'accès à l'information devrait être réformée pour permettre à l'Institut de la statistique du Québec de faire avancer le Québec en suivant le modèle du Nouveau-Brunswick. Le budget 2018-2019 mentionnait explicitement la promotion de l'accès aux données de recherche et des sommes ont été octroyées en ce sens à l'ISQ, ce qui témoigne d'une certaine ouverture sur cette question. Mais il reste encore du travail à faire.

Bien entendu, il faut absolument un environnement sécurisé où le couplage des différentes bases de données est contrôlé par divers processus pour assurer la confidentialité. Statistique Canada et l'Institut de la statistique du Québec ont une vaste expertise dans ce domaine acquise sur plusieurs décennies. L'accès à leurs données est contrôlé et exclut l'information personnelle (nom, prénom, adresse de résidence, numéro d'assurance sociale). La réglementation et l'infrastructure physique pour accéder à des données confidentielles existent déjà. Il faut maintenant que le gouvernement réforme la Loi sur l'accès à l'information et donne le mandat aux différents ministères de rendre leurs microdonnées accessibles dans les environnements sécurisés existants et seulement dans des environnements sécurisés.

Le potentiel de retombées est vaste. L'accès à des données de qualité permet à nos chercheurs de tirer le maximum de leurs activités de recherche, favorisant ainsi un rayonnement à l'échelle canadienne et internationale. Cet accès permet également la formation de la relève en recherche au Québec en techniques de pointe. Les nouveaux diplômés ont ainsi accès à de meilleures perspectives professionnelles et peuvent davantage contribuer à notre société et à la croissance économique. Finalement, sans données, nous ne pourrons jamais apprendre de nos erreurs et tirer profit de nos bons coups, afin de mieux guider nos politiques et d'améliorer le bien-être de notre société.

Bibliographie

Baker, M., Gruber, J., & Milligan, K. (2008). Universal Child Care, Maternal Labor Supply, and Family Well-being. *Journal of Political Economy*, *116*(4), 709-745.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding Access to Administrative Data for Research in the United States. *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Social Science Research*, *59*, 1-12.

Connolly, M., Haeck, C., & Lapierre, D. (2019). Social Mobility Trends in Canada: Going up the Great Gatsby Curve. Cahier de recherche numéro 19-03, Groupe de recherche sur le capital humain, mai 2019 (version révisée).

Einav, L., & Levin, J. (2014). Economics in the Age of Big Data. *Science*, *346*(6210), 1243089.

Frenette, M. (2019). Do Youth from Lower-and Higher-income Families Benefit Equally from Postsecondary Education? Statistique Canada, Analytical Studies Branch Research Paper Series, Analytical Studies Branch Research Paper Series, Catalogue no. 11F0019M — No. 424, 26 avril 2019.

Grusky, D.B., Hout, M., Smeeding, T.M., et Snipp, M. (2019). The American Opportunity Study: A New Infrastructure for Monitoring Outcomes, and Advancing Basic Science. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5(2): 20-39.

Haeck, C., Lefebvre, P., & Merrigan, P. (2015). Canadian Evidence on Ten Years of Universal Preschool Policies: The Good and the Bad. *Labour Economics*, *36*, 137-157.

Haeck, C., Lebihan, L., & Merrigan, P. (2018). Universal Child Care and Long-term Effects on Child Well-being: Evidence from Canada. *Journal of Human Capital*, *12*(1), 38-98.

Lefebvre, P., & Merrigan, P. (2008). Child-care Policy and the Labor Supply of Mothers with Young Children: A Natural Experiment from Canada, *Journal of Labor Economics*, *26*(3), 519-548.

Marchessault, G. (2011). The Manitoba Centre for Health Policy: A Case Study. *Healthcare Policy* 6 (Special Issue) February: 29-43. doi:10.12927/hcpol.2011.22117

Statistique Canada (2009). Statistique Canada : lignes directrices concernant la qualité. Cinquième édition — octobre 2009. Catalogue no. 12-539-X.