

## Titre : **L'accès aux données de recherche au Canada : acquis et défis**

Jean Poirier, Centre interuniversitaire québécois de statistiques sociales, Université de Montréal ([jean.poirier@umontreal.ca](mailto:jean.poirier@umontreal.ca))

Céline Le Bourdais, Centre interuniversitaire québécois de statistiques sociales, Université McGill

Séance « Liberté, secret statistique, éthique, déontologie »

### **Introduction**

Grâce aux grandes opérations de collecte menées notamment par ses instituts nationaux de statistique, le Canada et le Québec disposent d'un large éventail de bases de données pour la recherche. Pour en tirer profit, les chercheurs doivent pouvoir accéder aux fichiers de micro-données sans contrevenir aux dispositions des lois protégeant le respect de la confidentialité des informations.

Après un bref survol des grandes sources de données statistiques canadiennes existantes, nous présenterons les solutions mises de l'avant pour permettre aux chercheurs d'exploiter ces données dans le respect des lois sur la protection des renseignements personnels. Ces solutions combinent, de façon originale, différentes approches en regard de l'accès aux données : large accès à des fichiers masqués ; accès restreint aux micro-données détaillées par le biais de centres sécurisés établis en milieu universitaire ; traitement de données en différé ; obtention du consentement des répondants à partager leurs informations avec les chercheurs. Après avoir mis en évidence les avantages et les limites de ces approches, nous présenterons rapidement quelques initiatives récentes visant à favoriser l'utilisation des données de recherche et mentionnerons certains défis auxquels nous sommes présentement confrontés.

### **Quelles données pour la recherche au Canada?**

#### ***La part importante des données collectées par Statistique Canada***

Statistique Canada joue un rôle dominant dans le système d'information statistique au Canada. Contrairement à ce que l'on observe aux États-Unis par exemple, la part des données produites par le milieu universitaire canadien est relativement faible dans l'ensemble de la production des données statistiques.

Créé en 1918, Statistique Canada a comme mandat de fournir l'information statistique sur les caractéristiques et le comportement des ménages, des entreprises, des institutions et des gouvernements canadiens aux fins de recherche, d'élaboration des politiques, d'administration de programmes, de prise de décision et d'information générale conformément à la *Loi sur la statistique*. À cette fin, il réalise les recensements (obligation inscrite dans la Constitution) et

mène un grand nombre d'enquêtes (~ 400). Statistique Canada effectue également certaines opérations de couplage de données de différentes sources. Statistique Canada joue enfin un important rôle de coordination. Il est notamment chargé de collaborer avec les ministères à la collecte et à la publication des renseignements statistiques, de veiller à prévenir le double emploi dans la collecte de renseignements par les ministères, de donner des avis sur des sujets concernant les programmes statistiques des ministères et organismes et de conférer avec eux à cet égard. (Potvin et Latraverse, 2005). Statistique Canada jouit en général d'une bonne réputation en vertu de sa capacité à collecter de l'information qui alimente les débats sur des enjeux émergents et sur les questions relatives aux politiques publiques (Voyer, 2005).

La statistique officielle est beaucoup plus centralisée au Canada qu'au Québec, seule province à disposer de son institut de statistique, l'Institut de la statistique du Québec (ISQ), créé en 1998 de la fusion de quatre organismes dont le Bureau de la statistique. Alors que le modèle québécois s'apparente davantage à celui des États-Unis, le modèle canadien correspond davantage à celui qui a cours en Australie, en Grande Bretagne et, dans une moindre mesure, en France.

### **Les données de recensement**

Les données de recensement se sont imposées depuis les années 1960 comme source privilégiée dans plusieurs domaines de la recherche sociale. La possibilité qu'elles offrent de mener les analyses tant au niveau micro que macro explique largement cette popularité (Gaffield, 2005).

Au Canada, les chercheurs sont relativement favorisés en ce qui concerne l'existence de données de recensement. Les premiers recensements remontent en effet au milieu du XVII<sup>ème</sup> siècle et la pratique de réaliser des recensements « modernes » s'est imposée à partir du milieu du XIX<sup>ème</sup> siècle. De 1881 à 1961, le nombre de questions augmente de façon constante. À partir de 1956, les recensements sont quinquennaux et en 1971, l'approche des deux formulaires est adoptée : un formulaire long pour un échantillon des répondants (un tiers en 1971 et un cinquième à partir de 1996) ; un formulaire court pour la majorité restante.

### **Les données d'enquêtes**

Au début des années 1980, le manque d'information sur différentes thématiques sociales et de santé et l'incapacité des instruments statistiques permettant de les aborder sont mis en évidence. Cela conduit Statistique Canada à lancer en 1985 le *Programme de l'Enquête sociale générale* (ESG) qui consiste à réaliser chaque année une enquête transversale sur un thème précis (santé, familles, victimisation, soutien social, éducation). Certains des thèmes ont par la suite fait l'objet d'un programme particulier d'enquêtes (santé, éducation). Les autres thématiques (famille, victimisation, emploi du temps) sont reprises à tous les cinq ans approximativement.

La décennie 90 est celle des enquêtes longitudinales prospectives. Le Canada découvre, après plusieurs autres pays, les avantages de la puissance analytique de telles enquêtes. De 1993 à 2001, six enquêtes longitudinales sont ainsi lancées, qui s'ajoutent à l'Enquête nationale auprès des diplômés, initiée en 1978. Il s'agit de :

- L'Enquête sur la dynamique du travail et du revenu (EDTR), lancée en 1993
- L'Enquête nationale sur la santé de la population (ENSP), lancée en 1994

- L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), lancée en 1994
- L'Enquête sur le milieu de travail et les employés (EMTE), lancée en 1999
- L'Enquête auprès des jeunes en transition (EJET), lancée en 2000
- L'Enquête longitudinale auprès des immigrants du Canada (ELIC), lancée en 2001.

Comme le soulignent Picot et Webber (2006), ces nouvelles enquêtes ont été élaborées « dossier par dossier », principalement en réponse aux besoins formulés par les ministères chargés de l'élaboration de politiques. Le programme des enquêtes longitudinales de Statistique Canada n'a donc pas été développé dans le cadre d'une stratégie intégrée et cohérente de production de l'information statistique.

Parmi les nombreuses autres enquêtes menées par Statistique Canada, mentionnons les grandes enquêtes post-censitaires suivantes qui fouillent certaines questions effleurées par les recensements :

- L'Enquête auprès des peuples autochtones (EAPA)
- L'Enquête sur la participation et les limitations d'activité (EPLA)
- L'Enquête sur la diversité ethnique (EDE)

L'importance accrue des questions liées à la santé dans les préoccupations de la population et des gouvernements mène à la mise en œuvre, au début des années 2000, du programme des Enquêtes sur la santé dans les collectivités canadiennes (ESCC). Chaque cycle de l'ESCC comprend deux collectes : une enquête générale sur un gros échantillon (~130,000) conçue pour fournir des estimations fiables à l'échelle régionale (région socio-sanitaire) ; une enquête thématique sur un plus petit échantillon (~30,000) conçue pour fournir des données à l'échelle provinciale sur des thématiques variées.

Plusieurs jugent insuffisante la participation des chercheurs universitaires dans la réalisation des enquêtes de Statistique Canada. En effet, le mode de gestion des enquêtes adopté par Statistique Canada confine les chercheurs à un rôle strictement consultatif dans la phase de conception de l'enquête alors que les décisions en matière d'échantillonnage et de contenu sont souvent prises ultérieurement par les représentants des clients (les ministères) et le personnel de Statistique Canada.

Statistique Canada crée également à partir des années 90, plusieurs fichiers de données longitudinales reposant principalement sur des données administratives. Mentionnons notamment :

- La base de données administratives longitudinales (DAL), créée à partir des données fiscales sur la dynamique des gains et du revenu de 20% des familles canadiennes depuis 1982
- La base de données longitudinales sur les immigrants (BDIM), créée à partir des dossiers d'immigration et de fiscalité des cohortes d'immigrants depuis 1980.

## **Autres sources de données**

### **Les données de l'Institut de la statistique du Québec**

La contribution de l'Institut de la statistique du Québec (ISQ) à la production de données de recherche, bien que d'une ampleur nettement moindre que celle de Statistique Canada, mérite d'être soulignée pour trois raisons :

1. le caractère novateur de certaines des enquêtes menées par l'ISQ, uniques au Canada
2. un mode de gestion des enquêtes qui a souvent fait une plus large place aux chercheurs universitaires, considérés comme de véritables partenaires
3. des modes d'accès aux fichiers de micro-données qui diffèrent de ceux privilégiés par Statistique Canada.

Le portefeuille d'enquêtes de l'ISQ se compose d'une trentaine d'enquêtes transversales portant principalement sur divers aspects de la santé et de ses déterminants (santé physique, santé mentale, violence à l'égard des enfants, nutrition, tabagisme, consommation d'alcool et de drogues chez les jeunes, limitations d'activité) et d'une enquête longitudinale. Il s'agit de *l'Étude longitudinale du développement des enfants du Québec (ELDEQ)*, qui suit un échantillon représentatif d'environ 2000 enfants nés au Québec en 1997-1998.

### **Les données issues des chercheurs académiques**

Bien que les enquêtes menées par les chercheurs universitaires soient proportionnellement beaucoup moins importantes que dans d'autres pays, elles représentent cependant des sources extrêmement importantes d'information sur certains thèmes. Mentionnons par exemple, l'Enquête sur la fécondité du Canada de 1984, réalisée par des chercheurs de trois universités, dont Évelyne Lapierre-Adamcyk ; et l'Enquête sur l'Établissement des Nouveaux Immigrants, dirigée par Jean Renaud de l'Université de Montréal, qui a suivi pendant dix ans une cohorte d'immigrants dans leur processus d'insertion au Québec et dont s'est largement inspiré Statistique Canada pour la conception de l'ELIC.

Les chercheurs académiques sont également à l'origine de la création d'un grand nombre de fichiers de micro-données produits à partir des recensements anciens. En effet, la Loi sur la protection des renseignements personnels permet aux Archives nationales de communiquer des renseignements personnels recueillis dans le cadre d'un recensement 92 ans après sa réalisation. Depuis les années 60, plusieurs chercheurs ont ainsi créé des fichiers de micro-données à partir des recensements du XIX<sup>ème</sup> siècle et ceux de 1901 et de 1906. Ces fichiers concernent différentes régions et périodes. Mentionnons par exemple le projet de recherche sur les familles canadiennes, initié à l'Université de Victoria, qui s'appuie sur la création d'une base de micro-données d'un échantillon national de 5% du recensement de 1901 ; et le projet du laboratoire de démographie urbaine de l'Université Laval qui exploite les données de six recensements s'échelonnant de 1851 à 1901 pour la ville de Québec.

### **Les données des fichiers administratifs**

Quant aux données administratives, un nombre important de ministères et d'organismes sont engagés dans la production de telles données, tant au niveau fédéral que provincial. « De nombreuses bases de données ont été ainsi constituées dans le réseau de la santé à travers les années, et elles sont présentement utilisées principalement à des fins de gestion et de

surveillance du réseau. On estime que la qualité de ces données est habituellement assez bonne, et les BDAL (bases de données administratives longitudinales ) peuvent donc constituer une source de données longitudinales importante pour la recherche en santé publique et en santé des populations ». (RRSPQ, 2006)

## **Quel accès aux micro-données pour la recherche?**

Examinons d'abord l'accès aux données sous la responsabilité de Statistique Canada en raison de la place importante qu'elles occupent pour la recherche.

### ***L'accès aux données de Statistique Canada***

Au centre des dispositions de la *Loi sur la statistique* (amendée en 1971 et 1985) « figure un contrat social implicite avec les répondants en vertu duquel Statistique Canada peut imposer aux répondants le fardeau des demandes de données et, dans certains cas, exiger une réponse, afin de produire des renseignements qui revêtent un intérêt évident pour le grand public. Ce contrat prévoit toutefois un engagement absolu d'assurer la confidentialité des réponses pouvant permettre d'identifier des personnes ». (Potvin et Latraverse, 2005 : 49)

Statistique Canada doit donc assurer un équilibre entre les avantages des données statistiques pour le public, qui découlent notamment du travail des chercheurs, et la protection des renseignements personnels.

### **Des tableaux sur mesure à la création des fichiers de micro-données à grande diffusion**

À partir des années 1960, la demande d'accès aux micro-données de Statistique Canada s'intensifie de la part du milieu académique mais également des analystes œuvrant au sein d'organismes gouvernementaux (voir Halliwell, 2005, pour les facteurs qui sous-tendent cette demande). Pour y répondre, Statistique Canada mise d'abord sur un service spécial chargé de produire des tableaux sur mesure à partir des micro-données. Les limites d'une telle solution apparaissent rapidement : le processus est lourd, les délais sont longs, les coûts pour les chercheurs sont relativement élevés et l'absence d'interaction entre les chercheurs et les données constitue un sérieux handicap au travail de recherche. Mais jusqu'au début des années 70, Statistique Canada peut difficilement faire plus. En effet, avant qu'elle soit amendée en 1971, la Loi sur la statistique n'autorise la diffusion d'aucune partie du questionnaire d'un répondant. Les amendements apportés en 1971 permettent plus de flexibilité à cet égard. En effet, « l'identifiabilité » remplace alors l'individualité comme critère déterminant la confidentialité. En vertu de ce changement, la diffusion de renseignements personnels sur les individus devient possible à la condition qu'on ne puisse identifier un répondant.

La création de fichiers de micro-données à grande diffusion (FMGD) devient alors possible sous réserve du respect des deux conditions suivantes : 1) La diffusion augmente la valeur analytique des données collectées et 2) les précautions raisonnables ont été prises pour éviter l'identification des répondants. Le Canada devient ainsi le deuxième pays, après les États-Unis, à produire de tels fichiers. Les premiers FMGD sont créés en 1974 à partir d'un échantillon du recensement de 1971. Trois fichiers sont ainsi produits : un fichier individu, un fichier famille et un fichier ménage et logement. Chaque fichier contient un grand nombre de

variables sur les unités d'analyse ainsi qu'un résumé de l'information contextuelle sur les deux autres unités. La taille de l'échantillon passe de 1% en 1971 à 2% dans les années 1980 à 3% à partir de 1991. Des FMGD sont également créés à partir des enquêtes que réalise Statistique Canada. Accessibles aux chercheurs à un prix relativement abordable jusqu'au milieu des années 80, ces fichiers se sont imposés comme source privilégiée de données en sciences sociales, surtout en démographie et en économie.

L'arrivée au pouvoir d'un nouveau gouvernement en 1984 dans un contexte de lutte aux déficits change la donne. Face aux pressions du milieu de la recherche, le gouvernement revient sur sa décision de supprimer le recensement de 1986 mais plusieurs des services de Statistique Canada dont la création des FMGD sont soumis à une nouvelle politique de recouvrement des coûts.

En conséquence, le prix des FMGD explose, ce qui entrave fortement l'accès aux données par la communauté scientifique. Les chercheurs canadiens se tournent alors davantage vers les micro-données américaines ou européennes, accessibles à moindre coût (ou même gratuitement) et qui offrent par ailleurs de plus grandes possibilités de publication.

Un consortium se forme au début des années 90, associant Statistique Canada aux chercheurs universitaires (représentés par l'Association des bibliothèques de recherche du Canada) et à cinq ministères fédéraux, pour financer l'accès aux FMGD des recensements de 1986 et 1991 et des enquêtes sociales générales.

### **L'initiative de démocratisation des données**

*L'Initiative de démocratisation des données (IDD) (« Data Liberation Initiative »)* prend le relais au milieu des années 90 pour faciliter l'accès des FMGD à la communauté scientifique. Il s'agit d'une entente de coopération entre les universités canadiennes, Statistique Canada et six départements ministériels du gouvernement fédéral visant à donner aux universités un accès privilégié aux données de Statistique Canada dans le cadre de leur missions d'enseignement et de recherche. Moyennant une cotisation annuelle abordable dont le montant varie selon le type d'institution (universités de recherche ou universités d'enseignement), les chercheurs et les étudiants des universités participantes bénéficient d'un accès illimité à l'ensemble des FMGD créés par Statistique Canada ainsi qu'aux métadonnées correspondantes. L'IDD connaît rapidement un immense succès, près de soixante-dix institutions canadiennes d'enseignement supérieur y adhérant. Au sein de chaque institution, un employé de la bibliothèque, habituellement responsable des données numériques, est désigné pour offrir un support professionnel et technique aux utilisateurs et pour s'assurer du respect des termes de la convention entre son institution et Statistique Canada.

La création de FMGD et leur diffusion par le biais de l'IDD constituent des avancées majeures en ce qui concerne la formation et la recherche dans le domaine des statistiques sociales au Canada. La création de FMGD est cependant loin de satisfaire toutes les demandes d'accès aux micro-données à des fins de recherche. En effet, la création de FMGD :

- empêche de mener à bien certaines recherches du fait du masquage de données qu'elle implique. Pensons notamment aux recherches nécessitant l'utilisation de données fines sur l'âge, la religion, l'ethnie, l'occupation, le lieu de résidence.

- ne permet pas de préserver les liens entre les différentes unités du recensement (individu, famille, ménage-logement). En d'autres mots, un FMGD ne peut être hiérarchique alors que la demande pour ce type de fichier ne cesse de croître.
- n'est pas possible pour les enquêtes longitudinales. La grande quantité d'information collectée dans ce type d'enquêtes, qui explique leur utilité pour la recherche et la prise de décision, en compromet l'accès en facilitant l'identification des répondants. La quantité d'information à masquer pour assurer la protection de la confidentialité est telle qu'elle priverait les données de leur potentiel analytique.
- Ne permet pas de prendre en compte l'effet de plan dans l'analyse de la variance, les informations sur le plan de sondage étant masquées dans les FMGD.

À ces contraintes majeures s'ajoutent des irritants comme les délais de production des FMGD, qui varient de 6 à 18 mois après la publication des premiers résultats à partir du fichier-maître.

### **La création d'un Réseau de Centres de données de recherche dans la foulée de l'Initiative canadienne pour les statistiques sociales**

En 1998, un Groupe de travail conjoint de Statistique Canada et du Conseil de recherche en sciences humaines du Canada identifie trois obstacles majeurs au développement de la capacité de recherche du Canada en statistiques sociales : 1- le manque de chercheurs ayant une formation adéquate ; 2- la difficulté que présente l'accès aux données ; 3- l'insuffisance des moyens de communication entre chercheurs et utilisateurs des résultats de recherche. Pour lever ces obstacles est lancée l'*Initiative canadienne pour les statistiques sociales* dont l'un des éléments-clés est la mise en place d'un Réseau national de *Centres de données de recherche* (CDR). Le concept des «centres de données de recherche» s'inspire de l'expérience réussie des *U.S. Census Bureau Research Data Centres*, lancée il y a plusieurs années, afin de rendre accessibles aux chercheurs universitaires les micro-données confidentielles des recensements américains détenues à Washington.

Créé en 2000 à l'occasion de l'obtention d'une importante subvention d'infrastructure de la *Fondation canadienne de l'innovation*, le Réseau s'est rapidement étendu, passant de six centres fondateurs à treize aujourd'hui.

Concrètement, chaque CDR, établi en milieu universitaire, fournit un environnement physique sécuritaire (c'est-à-dire un laboratoire informatique à accès réglementé) où sont logés les micro-données détaillées des enquêtes de Statistique Canada, ainsi que les ordinateurs et les logiciels requis pour analyser ces données. Peuvent accéder au CDR tous les chercheurs (y inclus les étudiants) :

- dont le projet de recherche a au préalable été approuvé, soit par un comité de pairs, pour les chercheurs, soit par un comité de mémoire ou de thèse, pour les étudiants
- et qui ont été assermentés comme « employé réputé » de Statistique Canada. La *Loi sur la statistique* permet en effet à Statistique Canada de conférer le statut de « personne réputée être à l'emploi de Statistique Canada » à certains chercheurs universitaires, leur donnant ainsi accès à des données contenant de l'information confidentielle dans certaines conditions prescrites et contrôlées.

Les chercheurs mènent leurs analyses à l'intérieur de la zone sécuritaire. Pour sortir des résultats, ceux-ci doivent être examinés par un professionnel de Statistique Canada présent dans le centre, qui s'assure qu'ils ne posent aucun risque à la confidentialité des répondants.

Chaque CDR a son propre mode d'organisation. Le *Centre interuniversitaire québécois de statistiques sociales* (CIQSS), un des six membres fondateurs du Réseau national et seul CDR au Québec, a opté dès le début pour une approche réseau. Localisé à l'Université de Montréal, le CIQSS est ouvert à l'ensemble des chercheurs québécois et il réunit formellement sept universités partenaires qui contribuent financièrement à son fonctionnement : l'Institut national de la recherche scientifique, l'Université Concordia, l'Université Laval, l'Université McGill, l'Université de Montréal, l'Université du Québec à Montréal, ainsi que l'Université de Sherbrooke qui a joint le regroupement en 2004.

Une entente conclue avec l'Institut de la statistique du Québec (ISQ) permet également au CIQSS d'offrir l'accès aux données des enquêtes de l'ISQ dans un laboratoire spécialement aménagé à cet effet. Le regroupement des universités québécoises et de l'ISQ autour d'un projet mobilisateur en statistiques sociales a joué un rôle central dans la venue à Montréal et l'installation sur le campus de l'Université de Montréal de l'Institut de la statistique de l'UNESCO. La mise en commun de l'ensemble des contributions reçues des différents organismes et universités permet au CIQSS d'offrir une gamme de services et d'activités ouverts à l'ensemble des étudiants et chercheurs issus des milieux académique, gouvernemental et communautaire.

Voulant rapprocher les bases de données des chercheurs sans pour autant compromettre son fonctionnement en réseau, le CIQSS s'est fait l'ardent défenseur au sein du Réseau canadien de l'ouverture d'antennes dans ses institutions membres. Concrètement, les antennes sont une copie fidèle à plus petite échelle de l'infrastructure du CIQSS à une exception près : l'analyse de divulgation des résultats ne se fait pas sur place.

Les grands ensembles de données suivants sont présentement accessibles dans les CDR :

- Les données des différents cycles des sept grandes enquêtes longitudinales.
- Les données des différents cycles des programmes des Enquêtes sociales générales et des Enquêtes sur la santé dans les collectivités canadiennes
- Les données des trois grandes enquêtes post-censitaires
- Les données d'autres grandes enquêtes transversales

Au total, une cinquantaine de bases de données sont en cours d'exploitation dans le Réseau et cinquante autres devraient s'y ajouter au cours des quatre prochaines années.

Si cette approche d'accès aux données rencontre un succès indéniable comme en témoigne la multiplication du nombre de projets (plus de 600 dans le Réseau dont environ 110 au CIQSS), de chercheurs (1200 dans le Réseau et 350 au CIQSS), d'institutions universitaires partenaires (plus de 40), et d'organismes subventionnaires, elle n'est cependant pas exempte d'inconvénients. Ceux-ci concernent principalement :

- La lourdeur de la procédure d'accès aux CDR (délai moyen d'approbation des projets de 17 jours ouvrables) et de la procédure d'analyse des résultats pour vérification des risques de divulgation (délai qui varie de 24 à 48 heures), qui se répercute sur le processus de recherche.
- Les contraintes d'ordre logistique. En général, les CDR ne sont ouverts qu'aux heures habituelles de bureau. Ces contraintes sont particulièrement importantes pour les chercheurs rattachés à des institutions qui ne disposent pas de CDR ou d'antennes et elles sont bien sûr dissuasives pour les chercheurs hors du Canada.

- La non-disponibilité de certains ensembles de données dans les CDR : les données de recensement, les bases de données administratives, les bases de données issues d'un couplage entre des données administratives et des données d'enquête ne sont pas accessibles dans le Réseau.
- Les coûts élevés d'opération et de fonctionnement des CDR.

Parallèlement à la création du Réseau des CDR, Statistique Canada a expérimenté d'autres approches pour permettre l'accès aux fichiers de micro-données non masquées à des fins de recherche et d'évaluation des politiques et des programmes tout en préservant la confidentialité des informations

### **La création de fichiers partagés et l'accès en différé**

Une première approche, qui opère en amont de la collecte, consiste à obtenir des répondants l'autorisation de partager leur information avec les employés d'autres organismes que Statistique Canada. Un fichier est alors créé (fichier « partagé ») contenant l'information non masquée de tous les répondants ayant consenti à ce partage. Seuls les employés des organismes dûment identifiés dans la demande de consentement ont accès à ce fichier partagé dans des conditions précisées dans l'entente de partage. Cette approche a été utilisée dans certaines enquêtes longitudinales pour permettre aux ministères ayant financé ces enquêtes d'avoir accès aux données : Santé Canada pour l'Enquête nationale sur la santé de la population (ENSP) ; le ministère des Ressources humaines et du développement des compétences pour l'Enquête longitudinale nationale auprès des enfants et des jeunes (ELNEJ). Seuls des organismes gouvernementaux ont conclu de telles ententes de partage avec Statistique Canada à des fins d'évaluation de programmes et de politiques. Dans ces enquêtes, environ 95% des répondants ont donné leur consentement.

En aval de la collecte des données, la mise en place d'un service d'accès indirect ou d'accès en différé constitue une deuxième approche mise de l'avant par Statistique Canada. Il s'agit pour les chercheurs de programmer leurs analyses à partir des métadonnées d'une enquête (questionnaire, dictionnaire, documents méthodologiques) et dans certains cas d'une base de données factice (« dummy file ») et de soumettre leur programme par courrier électronique à Statistique Canada qui l'exécute à partir du fichier-maître. Les résultats sont ensuite analysés du point de vue des risques de divulgation d'informations confidentielles. Les résultats sont ensuite communiqués aux chercheurs. Ce service, qui n'est offert que pour certaines enquêtes, suscite plusieurs critiques : les délais d'obtention des résultats alourdissent considérablement le processus de recherche ; les directions d'enquêtes qui offrent ce service supportent une gamme très restreinte de logiciels ; les bases de données factices ne conviennent pas toujours pour l'élaboration des programmes.

Mentionnons enfin que Statistique Canada a récemment développé une base de métadonnées intégrée (BMDI) qui centralise les renseignements sur chacune des enquêtes qu'il réalise. Conçue spécifiquement pour faciliter la diffusion des données, cette base, accessible gratuitement à partir du site Internet de Statistique Canada, vise à fournir aux utilisateurs les renseignements dont ils ont besoin pour analyser les données (Johannis, 2001).

## **L'accès aux autres sources de données**

### **Les données de l'ISQ**

Confronté à la même exigence que Statistique Canada d'assurer un équilibre entre la valorisation des données par la recherche et les droits à la protection des renseignements personnels des répondants, l'ISQ a privilégié des approches innovatrices à plusieurs égards :

- La création de fichiers publics n'est pas la solution privilégiée par l'ISQ pour faciliter l'accès à ses données d'enquête, de tels fichiers publics n'ayant été créés que pour deux ou trois enquêtes.
- On a plutôt privilégié la création de fichiers légèrement masqués, les fichiers aux fins d'analyse et de recherche externe (fichiers FARE) et l'octroi de licence d'utilisation de tels fichiers à des organismes, notamment des centres et des groupes de recherche. De telles licences précisent les mesures physiques et informatiques de stockage des données ; requièrent la signature d'un engagement au respect de la confidentialité des renseignements personnels de la part des chercheurs ; prévoient des visites de vérification et des mesures légales en cas de non respect des conditions de la licence ; exigent la destruction des données à la fin du projet.
- L'ISQ rend également accessibles ses fichiers-maîtres dans un laboratoire sécurisé en milieu universitaire, le Centre d'accès aux données de recherche de l'ISQ (CADRISQ), situé au CIQSS. Si le mode de fonctionnement du CADRISQ suit le même modèle que celui des CDR, la procédure d'accès est cependant plus souple et plus rapide. Les projets ne sont pas soumis à une approbation des pairs et ne font l'objet que d'une revue institutionnelle par un comité de l'ISQ.

Les modes d'accès aux données de l'ISQ présentent des avantages certains mais ne sont pas sans inconvénients. Par exemple, si le recours aux fichiers FARE favorise sans nul doute la recherche en offrant un contenu informationnel plus riche que les fichiers publics et moins de contraintes d'accès aux données que le travail dans un centre sécurisé, les dispositions de la licence d'utilisation de ces fichiers (entente avec des organismes et non des individus, conditions de stockage des données) limitent cependant fortement leur utilisation à des fins de formation (enseignement, réalisation de mémoires et de thèses). Notons également l'absence de base intégrée de métadonnées qui centraliserait toute l'information nécessaire à l'exploitation des enquêtes de l'ISQ.

### **Les données administratives**

Les bases de données administratives de Statistique Canada (dont la DAL et la BDIM) ne sont accessibles que dans les locaux de Statistique Canada à Ottawa. Elles ne sont donc pas accessibles dans les CDR.

La multitude d'organismes gouvernementaux canadiens produisant des données dans le cadre des programmes qu'ils administrent n'ont pas de mandat explicite d'appuyer la recherche. Le contexte d'accès et d'utilisation des données varie d'une province à l'autre. Les dispositions régissant la conservation, la divulgation et l'utilisation des données sont complexes et parcellaires, et ne permettent qu'à certains chercheurs seulement d'avoir accès à de telles données.

Trois centres de données établis dans des universités canadiennes, qui sont le fruit de partenariat entre le secteur de la santé (ministères et organismes de santé publique) et les chercheurs universitaires, offrent un accès centralisé dans un environnement sécuritaire à de nombreuses bases de données administratives à des fins de recherche, ainsi que l'expertise technique et professionnelle requise pour leur exploitation. Il s'agit du :

- Manitoba Centre for Health Policy (MCHPP) de l'Université du Manitoba
- Centre for Health Services and Policy Research (CHSPR) de l'Université de Colombie Britannique
- Population Health Research Unit (PHRU) de l'Université Dalhousie, en Nouvelle-Écosse.

Le Québec ne dispose pas d'infrastructure de ce type. Pour utiliser les données, les chercheurs doivent franchir plusieurs étapes d'un processus qui peut s'avérer long et complexe (RRSPQ, 2006) :

1. identification des sources de données pertinentes, de leurs fournisseurs et des conditions de production des données ;
2. démarches d'accès aux données impliquant notamment des demandes à la Commission d'accès à l'information et au responsable de l'accès aux documents et de la protection des renseignements personnels du propriétaire ou gestionnaire des données
3. évaluation de la qualité des données, définition des opérations de validation et de traitement.

Le soutien technique et professionnel, la documentation, les coûts et les délais varient selon l'organisme propriétaire ou dépositaire des bases de données.

Il est cependant possible aux chercheurs d'accéder à plusieurs bases de données administratives et de les apparier, une fois toutes les étapes précédentes franchies. En raison de la Loi sur la protection des renseignements personnels, tous les fichiers créés à partir de données administratives doivent être détruits à la fin du projet et ne peuvent être réutilisés aux fins d'un autre projet.

### **Les fichiers de micro-données issus de recherches académiques**

L'accès aux fichiers de micro-données créés par les chercheurs universitaires est fort limité, voire impossible. En effet, le Canada ne dispose pas d'une organisation chargée de la conservation des données de recherche produites à partir des fonds publics et il n'y a pas de stratégie ou de politiques nationales en matière d'accès aux données de recherche.

On assiste ainsi à la perte de données produites à grands coûts suite à la dégradation des supports de stockage, à la perte des métadonnées, à l'obsolescence des logiciels et du matériel utilisé pour la production des données, mais aussi comme conséquence des politiques en matière de protection des renseignements personnels et d'un manque de planification ou d'attention porté à la conservation d'un tel patrimoine.

## Initiatives, défis et perspectives

### ***Nouvelles initiatives***

Sans qu'ils s'inscrivent dans un plan d'ensemble, plusieurs projets novateurs en cours de réalisation visent à accroître la disponibilité et l'accès aux données de recherche. En voici une liste partielle :

- Projet de création de fichiers de micro-données détaillées à partir d'un très large échantillon des recensements de 2001 et 2006 à des fins d'analyse approfondie. L'information contenue dans de tels fichiers serait beaucoup plus grande que celle des FMFG. Par contre, la création de fichiers hiérarchiques à partir des données de recensement n'est toujours pas à l'ordre du jour.
- Projet d'*Infrastructure de recherche sur le Canada au XX<sup>ème</sup> siècle* (IRCS). Il s'agit d'une initiative pancanadienne dont l'objectif est de développer des bases de micro-données à partir d'échantillons des recensements de 1911 à 1951 à des fins de recherche. Les FMGD seront accessibles par le biais de l'IDD et les fichiers analytiques plus détaillés seront accessibles dans le Réseau des CDR.
- Projet visant à améliorer de façon significative la documentation électronique des fichiers de données accessibles dans les CDR. Cette initiative du Réseau canadien des CDR a comme objectif de documenter une centaine de fichiers de micro-données accessibles dans les CDR au cours des quatre prochaines années en tirant profit de la norme DDI (Data Documentation Initiative) qui tend à s'imposer en Europe et aux États-Unis. De telles bases évolutives de métadonnées vont constituer des outils dynamiques facilitant l'analyse de fichiers de données complexes.
- Deux projets pilotes sont en cours visant à explorer la possibilité de rendre accessibles dans les CDR des bases de données administratives. Le premier concerne les bases de données sur la criminalité et la justice du *Centre canadien de statistiques juridiques* (CCSJ). Le second porte sur des fichiers issus du couplage de données administratives de santé (sous juridiction provinciale) avec les données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) relevant de Statistique Canada.
- L'ISQ explore actuellement, dans le cadre de projets pilotes, différentes modalités d'accès à distance aux données. Un premier projet expérimente un mode d'accès à distance à l'une de ses enquêtes sociales et de santé par des équipes de recherche hors Québec. Un second concerne la mise en place d'une plate-forme d'accès à des données administratives (projet EPSEBE).

### ***Défis***

Les questions relatives à l'accès aux données de recherche ont récemment fait l'objet de plusieurs grands forums de discussion et de consultations menées au niveau national. Nous en avons retenu deux qui illustrent bien, selon nous, quelques-uns des principaux défis auxquels fait face le système de production de l'information statistique au Canada.

En janvier 2006, Statistique Canada et le Conseil de recherche en sciences humaines du Canada (CRSH), en collaboration avec le CIQSS, organisaient à Montréal une *Conférence sur « Les enquêtes longitudinales sociales et sur la santé dans une perspective internationale »*.

Réunissant des experts d'une douzaine de pays (chercheurs, gestionnaires d'enquêtes, bailleurs de fonds et planificateurs), cette conférence visait à dresser un bilan critique de l'expérience canadienne en matière d'enquêtes longitudinales à la lumière de ce qui se fait ailleurs, et à identifier les défis que pose la réalisation de telles enquêtes au plan des capacités de collecte et d'analyse, de l'accès aux données, de la diffusion des résultats et de l'élaboration de politiques. Trois points majeurs sont ressortis des travaux de cette conférence :

1. la possibilité de mener des études comparatives internationales est devenue une préoccupation importante qui orientera la conception des nouvelles opérations de collecte. Cela présuppose que les chercheurs d'autres pays puissent avoir accès relativement facilement aux données détaillées des enquêtes canadiennes. Différentes approches sont à l'étude par Statistique Canada : consentement des répondants à partager leurs informations avec des chercheurs internationaux ; création de fichiers légèrement masqués à l'intention de chercheurs internationaux ; aménagement de centres sécurisés dans d'autres pays ; accès à distance ; octroi de licences d'utilisation des données à des chercheurs individuels. L'approche (ou les approches) retenue devront bien sûr respecter les règles de la Loi sur la statistique.
2. Statistique Canada s'engage à revoir le mode de gestion de ses enquêtes longitudinales de façon à impliquer activement les chercheurs universitaires dans toutes les étapes associées à la réalisation des enquêtes. Cette participation soutenue des chercheurs constitue le moyen le plus efficace d'éviter que la poursuite des objectifs scientifiques à long terme de l'enquête ne soit compromise par des décisions ponctuelles risquant d'introduire des discontinuités dans la chronologie des informations recueillies.
3. comme dans plusieurs pays, les enquêtes longitudinales initiées par Statistique Canada l'ont été « dossier par dossier », selon les possibilités de financement liées aux besoins en données de différents ministères. Sur la base de l'expérience des dix dernières années, une réflexion s'est engagée sur la possibilité d'adopter une approche plus intégrée, axée sur la mise en place d'un véritable portefeuille d'enquêtes qui assurerait une meilleure articulation entre les différentes enquêtes (les répondants de l'ELNEJ pourraient passer à l'EJET, par exemple) et qui permettrait de mieux combler les lacunes du système d'information statistique.

En 2004 s'est tenue sous l'égide du Conseil national de recherches du Canada une *Consultation nationale sur l'accès aux données de la recherche scientifique* (CNADRS). Le rapport final, présenté en janvier 2005, attire l'attention sur l'absence d'infrastructure chargée de la conservation nationale des données de même que sur l'absence d'une stratégie ou de politiques nationales d'accès aux données. Devant les sérieuses préoccupations que soulève la perte de données « tant comme bien national que comme élément de base longitudinal devant permettre la mesure du changement à travers le temps » (Canada, 2005, p. 2), le Groupe de travail recommande notamment :

1. que soient initiées des consultations avec les principaux intervenants afin de déterminer quels sont les obstacles juridiques à l'accès aux données scientifiques afin d'aboutir à des propositions de modification des lois sur la protection des renseignements personnels ou de leur interprétation juridique
2. d'entreprendre un examen des textes législatifs pour y déceler les incohérences entravant le partage de données sur le plan international
3. que les bases de données d'importance nationale qui sont « à risque » soient identifiées et mises en sécurité à Bibliothèque et Archives Canada.

4. que les organismes subventionnaires prévoient un financement spécifique pour permettre la formation de tous les chercheurs principaux aux pratiques exemplaires en matière de gestion de bases de données, de gestion des droits associés à la conservation des données, des normes concernant les métadonnées afin d'assurer l'accès et la conservation des données.

## ***Perspectives***

Le Canada dispose d'un large éventail d'ensembles de données qui présentent un immense potentiel pour la recherche. Cependant, les particularités du système de production de l'information statistique, alliées à la sensibilité croissante de l'opinion publique canadienne en ce qui concerne la protection des renseignements personnels, ont imposé, peut-être plus qu'ailleurs, de fortes contraintes aux chercheurs pour accéder aux données.

Paradoxalement, ces contraintes semblent avoir eu des effets bénéfiques importants. Elles sont en effet à l'origine d'initiatives majeures pour la recherche au Canada : l'Initiative de démocratisation des données, l'Initiative canadienne pour les statistiques sociales, l'Infrastructure de recherche sur le Canada au XX<sup>ème</sup> siècle. Les partenariats fructueux établis entre Statistique Canada et le milieu de la recherche universitaire dans le cadre de telles initiatives ont débouché sur la constitution de réseaux dynamiques impliquant des chercheurs d'horizons disciplinaires variés, les responsables des données numériques en poste dans les bibliothèques universitaires, les analystes et les méthodologues des instituts de statistique, les représentants d'organismes subventionnaires de même que les responsables de la formulation des politiques publiques.

Malgré l'ampleur des défis à relever, il nous semble que cette « structuration » du champ des statistiques sociales canadiennes ouvre des perspectives stimulantes pour le développement de la recherche et de la formation dans le domaine. Au CIQSS, c'est le pari que nous faisons.

## Références bibliographiques

Currie, Raymond et Byron Spencer, 2005. « Les Centres de données de recherche : un progrès considérable dans le renforcement de la capacité de recherche en sciences sociales », *Horizons*, vol. 8, numéro 1, pages 38 à 41.

Gaffield, Chad, 2005. Ethics, Technology and Confidential Research Data: The Case of the Canadian Century Research Infrastructure Project. Communication présentée au *World History Conference*, Sydney, Australia, 3 au 9 juillet 2005.

Halliwell, Cliff, 2005. « Recherche des données désespérément », *Horizons*, vol. 8, numéro 1, pages 31 à 37.

Picot, Garnett et Marian Webber, 2005. « L'avenir des enquêtes longitudinales : l'état des lieux », *Horizons*, vol. 8, numéro 1, pages 16 à 22.

Potvin, Maryse et Sophie Latraverse, 2004. Projet MEDIS, Étude comparative de la collecte de données visant à mesurer l'étendue et l'impact de la discrimination dans certains pays – Rapport final Canada, mai 2004.

Réseau de recherche en santé des populations du Québec – RRSPQ, 2006. Les bases de données administratives longitudinales en santé des populations : vers un accès et une utilisation facilitée. Document de travail. 7 mars 2006.

Canada, 2005. Groupe de travail de la Consultation nationale sur l'accès aux données de recherche scientifique. Consultation nationale sur l'accès aux données de recherche scientifique : rapport final. 31 janvier 2005.

Voyer, Jean-Pierre, 2005. « L'Initiative visant les lacunes statistiques à la croisée des chemins », *Horizons*, vol. 8, numéro 1, pages 4 à 7.