

ACCÈS AUX FICHIERS DE MICRODONNÉES DE STATISTIQUE CANADA

Jeanine Bustros ¹

RÉSUMÉ

Un environnement en transformation, des nouvelles technologies et un besoin continu pour de l'information a augmenté la pression sur les systèmes statistiques pour développer des nouveaux moyens par lesquels les chercheurs pourraient avoir accès aux données recueillies. Sous ces nouvelles règles, Statistique Canada a dû innover pour permettre l'accès aux données détaillées et non masquées qui viendraient compléter l'accès déjà offert par le biais des fichiers de microdonnées publics. Ces nouvelles méthodes ont été développées en tenant compte de la Loi sur la statistique qui empêche la divulgation de toute information qui pourrait mener à l'identification d'une personne. Ce document résume le cadre de travail ainsi que la mise sur pied des méthodes qui permettent l'accès aux fichiers de microdonnées à Statistique Canada; soient, l'Initiative de démocratisation des données, l'Accès à distance et les Centres de recherche.

MOTS CLÉS : Confidentialité, fichier de microdonnées, accès aux données, initiative de démocratisation des données, centre de recherche, accès à distance.

ABSTRACT

The fast paced environment, the new technology and the continuing need for information has increased the pressure on the statistical systems to develop new means by which the data collected is made accessible to the research community. Under these new rules, Statistics Canada had to complement the access to their traditional Public Use Microdata Files and aggregate tables by allowing access to unscreened microdata files. These access methods have been developed taking into consideration the Statistics Canada Act, which prevents the disclosure of any information that could possibly be related to any individual person. This paper focuses on the framework and the implementation of new ways to access microdata files within Statistics Canada, namely the Data Liberation Initiative, the Remote Access and the Research Data Centres.

KEY WORDS: Confidentiality, Microdata file, Data access, Data Liberation Initiative, Research Data Centre, Remote access.

1. INTRODUCTION

Le présent document résume les différentes initiatives en cours à Statistique Canada qui permettent aux utilisateurs d'avoir accès aux fichiers microdonnées de Statistique Canada. Trois initiatives principales sont en cours à Statistique Canada, soient : l'Initiative de démocratisation des données, l'Accès à distance, et les Centres de recherche. Toutes ces initiatives permettent aux chercheurs d'avoir accès aux données de Statistique Canada sous formes plus ou moins détaillées tout en respectant la loi sur la Statistique.

2. CONTEXTE

2.1 Loi sur la statistique

La Loi sur la statistique stipule dans l'article 17 b :

« Aucune personne ne peut révéler ni sciemment faire révéler, par quelque moyen que ce soit, des renseignements obtenus en vertu de la présente loi de telle manière qu'il soit possible, grâce à ces révélations, de rattacher à un particulier, à une entreprise ou à une organisation identifiables les détails obtenus dans un relevé qui les concerne exclusivement.

¹ Jeanine Bustros, division de la diffusion, Statistique Canada, 8 C Parc Tunney, Ottawa, Canada, K1A 0T6, bustros@statcan.ca

La loi est non seulement restrictive mais n'indique aucune limite de temps. Les données recueillies doivent être protégées à jamais.

En plus, quel que soit le processus mis en place pour la protection des données, on doit toujours tenir compte des données existantes disponibles. Autant du côté de Statistique Canada, par exemple les fichiers publics de microdonnées diffusés, les tableaux, etc.; ou à l'extérieur de Statistique Canada, par exemple les données administratives, les bilans financiers, etc.

2.2 Changements sociaux et économiques

Ces dernières années, l'économie et la société canadienne ont fait face à une évolution rapide. Afin de comprendre tous les éléments qui sous-tendent cette transformation, il est important d'avoir de l'information analytique objective et opportune sur l'économie et les conditions sociales. Ce processus analytique permettra d'établir la base à un débat politique éclairé. Ces besoins analytiques sont bien couverts dans le document intitulé : « Rapport final du groupe de travail mixte du Conseil de recherches en Sciences Humaines du Canada et de Statistique Canada sur l'avancement de la recherche utilisant les statistiques sociales, décembre, 1998. »

Dans un certain sens, le Canada est bien pourvu pour répondre à ce besoin analytique. Statistique Canada a entrepris, ces dernières années, plusieurs enquêtes longitudinales qui suivent les individus sur une longue période de temps. Ces nouvelles enquêtes permettent de jeter un nouveau regard sur plusieurs sujets : la santé, la dynamique du travail, l'environnement des enfants, la formation, etc.

Cependant, la nature même de ces enquêtes fait en sorte qu'il est difficile de diffuser les données de façon détaillées et d'en encourager l'analyse pour laquelle ces données ont été recueillies. En effet il est difficile, sinon impossible de créer des fichiers publics de microdonnées conformes à la Loi sur la Statistique et qui répondent entièrement aux besoins des chercheurs.

De plus, la technologie utilisée dans la préparation de la plupart des enquêtes de Statistique Canada, ces dernières années, a permis la collecte d'information beaucoup plus complexe. Par conséquent beaucoup plus d'informations sont recueillies, autant pour les enquêtes transversales que longitudinales, par contre la diffusion de ces données est restreinte afin de respecter la confidentialité de l'information.

Malgré la disponibilité de ces informations, peu de chercheurs au Canada ont la possibilité d'analyser les données recueillies par ces enquêtes. Deux facteurs principaux expliquent cet état : le manque de chercheurs engagés dans la recherche quantitative, spécialement dans le domaine exigeant des méthodes statistiques avancées; ainsi que la difficulté d'accéder l'information.

Statistique Canada a mis sur pied, ces dernières années, trois initiatives qui améliorent l'accès aux microdonnées. Ces initiatives répondent à des besoins variés exprimés par les différents intervenants impliqués dans la recherche. Sans nécessairement résoudre tous les problèmes reliés à la recherche, elles permettent un meilleur accès aux données recueillies.

3. INITIATIVE DE DÉMOCRATISATION DES DONNÉES (IDD)

L'Initiative de Démocratisation des Données est un partenariat entre les universités canadiennes et Statistique Canada qui permet un accès facile à toutes les données électroniques diffusées par Statistique Canada. L'IDD est un nouveau modèle de diffusion destinée à l'enseignement et à la recherche.

L'IDD a débuté comme un projet pilote en 1996, pour une durée de cinq ans.

3.1 Objectifs de l'IDD

- Donner accès aux étudiants et aux professeurs d'université à un vaste ensemble de bases de données diffusées par Statistique Canada.
- Fournir un accès abordable et équitable aux données partout au pays; et cela quel que soit l'emplacement de l'institution.
- Établir une nouvelle culture à l'égard des données canadiennes. En effet cette initiative a permis l'utilisation de données canadiennes dans les salles de classe universitaire.

3.2 Organisation de l'IDD

- L'IDD est basée sur le partenariat entre Statistique Canada et les universités et collèges canadiens. Ce partenariat passe par les bibliothèques de ces institutions et se reflète dans la composition du Comité consultatif externe dont les membres sont : un administrateur, un chercheur, des bibliothécaires et bien sûr du personnel de Statistique Canada. En plus d'exprimer les besoins de la communauté, ce comité très actif conseille

sur la direction future afin de mieux répondre aux membres.

- L'équipe de l'IDD voit au travail quotidien : l'acquisition des fichiers des données ainsi que de leur documentation, la maintenance du site FTP et du site web, le contrôle de la qualité ainsi que du service offert aux membres aux moyens de communications par serveurs de liste.
- Le Comité consultatif interne de Statistique Canada (SC) est constitué de membres des divisions de SC qui fournissent les données au groupe de l'IDD. Ce groupe conseille l'équipe de l'IDD sur les nouvelles enquêtes et des changements majeurs qui pourraient survenir. D'autre part, l'équipe de l'IDD transmet les besoins de sa clientèle surtout dans le cadre de la documentation des fichiers.
- Le Conseil de gestion est formé par les autres ministères fédéraux qui ont permis la mise sur pied de cette initiative.

3.3 Types de données disponibles

Toutes les données numériques produites par SC font parti de la collection de l'IDD :

- ⇒ les fichiers publics : données masquées tirées de données d'enquête;
- ⇒ les bases de données agrégées - p. ex. CANSIM, tableaux du recensement, E-STAT;
- ⇒ les fichiers géographiques et les fichiers du recensement.

Les données sont accédées par protocole de transfert de fichier (FTP), Internet ou CD-ROM.

3.4 Processus

- Les institutions post secondaires qui ont convenu de participer déboursent des frais (3 000 \$ canadiens ou 12 000 \$ canadiens selon la taille de l'établissement).
- Elles signent un accord de licence qui limite l'utilisation à l'enseignement et à la recherche.
- Elles nomment une personne-ressource de l'IDD qui est chargée de mettre en œuvre l'initiative au sein de l'établissement.

3.5 Responsabilités

Responsabilités de l'université :

- ⇒ contrôler l'utilisation des données conformément à la licence
- ⇒ établir l'infrastructure nécessaire pour permettre la diffusion des données au sein de l'université

- ⇒ désigner la personne-ressource de l'IDD et les autres ressources nécessaires pour gérer l'accès aux données.

Responsabilités de l'équipe de l'IDD à l'intérieur de Statistique Canada :

- ⇒ établir l'infrastructure de l'IDD.
- ⇒ gérer la collecte des données de l'IDD.
- ⇒ préparer la documentation nécessaire pour répondre aux besoins de la collectivité de l'IDD.
- ⇒ contrôler les communications.

3.6 La licence

- La licence donne accès aux fichiers, mais ne donne pas de droit de propriété.
- Les fichiers doivent servir à des fins d'enseignement et de recherche seulement.
- Les fichiers de l'IDD sont réservés aux utilisateurs autorisés, qui sont en général les professeurs et étudiants de l'université membre.
- La confidentialité des données doit être protégée : p. ex. couplage des données interdit. Les utilisateurs doivent être informés des modalités de la licence.

3.7 Communications

- Les communications internes et avec les membres de l'IDD sont une partie inhérente du projet. En plus de répondre et de fournir le service de soutien à la communauté, l'équipe de l'IDD est en contact constant avec les divisions auteurs pour obtenir les fichiers, pour donner de la rétroaction venant des membres ou tout simplement pour chercher des réponses précises reliées au déroulement et contenu de l'enquête dont les données sont tirées.
- L'IDD modère deux serveurs de liste :
 - ⇒ DLILIST : tribune générale de discussion avec 200 membres.
 - ⇒ DLIORDER : réservé aux personnes-ressources de l'IDD pour contrôler les commandes.

3.8 Les 4 premières années

L'IDD, comme tout projet a consacré à ses débuts beaucoup de temps à former son équipe, à bâtir les liens avec la communauté universitaire et à se vendre auprès des divisions internes de SC. Étant donné la nature centralisée de l'IDD, il a aussi fallu mettre en place des méthodes de vérification et de correction des lacunes liées à la documentation des produits. En effet, il arrive que les fichiers publics soient en format

ASCII avec un cliché d'articles de base. Dans ces cas là, en plus d'écrire les commandes SPSS pour répondre aux besoins des chercheurs universitaires, l'équipe de l'IDD les utilisent pour valider la documentation du fichier public.

3.9 Où en sommes-nous ?

- Le programme débute la 5^e année et dernière année du projet pilote.
- 66 universités et collèges sont membres du programme. Ceci dépasse toutes les attentes.
- Les membres ont effectué plus de 100 000 téléchargements à partir des sites FTP
- La collection comporte plus de 100 titres (ou plus de 10 000 fichiers)
- L'équipe de l'IDD et le Comité consultatif externe sont sur le point d'élaborer un plan d'action à partir de l'évaluation du projet effectuée au cours de la 4^e année afin de stabiliser le programme au-delà du projet pilote.

3.10 Défis majeurs

- Les progrès rapides et l'évolution du projet ont fait en sorte que tous les partis; soient les contacts de l'IDD, l'équipe de l'IDD au sein de SC et les chercheurs, ont dû réviser les pratiques et le processus entourant la gestion de la collection.
- Des deux côtés, on a dû apprendre à se connaître et à connaître les données. La formation continue de la personne ressource à l'université et du personnel de SC, constitue un élément important du projet.

4. ACCÈS À DISTANCE

L'accès à distance est une procédure grâce à laquelle les chercheurs reçoivent des fichiers synthétiques à partir desquels ils peuvent effectuer une «quasi» analyse des données, afin d'en vérifier la logique des programmes informatiques et de déterminer la faisabilité de l'analyse des données. Par la suite, ces mêmes programmes transmis via Internet sont soumis au fichier principal contenant les données confidentielles. Les résultats sont vérifiés pour s'assurer qu'ils répondent aux critères de SC sur la confidentialité, ensuite ils sont transmis via Internet au chercheur.

4.1 Structure de fichier synthétique

Afin de permettre la «quasi» analyse :

- ⇒ le fichier synthétique doit avoir le même cliché d'articles que le fichier principal : les mêmes variables, les mêmes codes, les mêmes positions, etc.
- ⇒ il faut préserver la cohérence des données par bloc de variables ou section du questionnaire.
- ⇒ il faut avoir un nombre suffisant d'enregistrements.
- ⇒ Il faut protéger la confidentialité des données.

La mise sur pied d'un tel fichier doit être rapide et peu coûteuse.

4.2 Méthode

Le principe de l'accès à distance est basé sur la création du fichier synthétique qui est un sous-ensemble du fichier principal. Les étapes importantes de la mise sur pied du fichier synthétique sont :

- Tout d'abord on choisit les blocs de variables ayant besoin d'être permutés au fin de la confidentialité. En effet toutes les données sur un fichier ne sont pas sensibles nécessairement à la confidentialité. En général, les variables démographiques et géographiques sont plus à risque du point de vue de la confidentialité que les variables reliées aux attitudes et comportements.
- Ensuite on décide de la stratification nécessaire et possible pour la permutation. À date, on a considéré les variables géographiques, le sexe et l'âge pour la stratification.
- Par la suite, on identifie le nombre d'enregistrements complets du fichier synthétique en fonction de la confidentialité et des besoins de l'analyse. À cette étape, le fichier synthétique a un nombre moindre d'enregistrements comparé au fichier public ou le fichier principal.
- Enfin, on permute par bloc en ayant un donneur différent par bloc pour chaque strate. Ce faisant, on contrôle la valeur manquante, i.e., on s'assure qu'on n'utilise pas des données manquantes lors de la permutation.
- Le tout est terminé par la vérification des distributions de fréquence et rapports globaux entre les variables reproduisant ceux du fichier principal.

4.3 Protocole pour l'accès

- Le chercheur présente un protocole d'analyse (besoins, détails).
- Une fois approuvé par la division auteur, on lui envoie le fichier synthétique et la documentation.

- Le chercheur élabore son ou ses programmes informatiques et les vérifie au moyen du fichier synthétique.
- Les programmes, SAS, SPSS ou STATA, sont envoyés par courrier électronique à SC.
- SC exécute les programmes, produit les résultats, s'assure de la protection de la confidentialité. S'il existe un problème de confidentialité, les résultats peuvent être supprimés en totalité ou partiellement.
- On transmet par courrier électronique les résultats, y compris le registre, le programme et le résumé des résultats du point de vue de la confidentialité.
- S'il y a une erreur dans le programme, SC ne corrigera pas l'erreur, mais renverra le programme au chercheur.
- Le programme de recherche est administré par SC et les universités.
- Le système informatique des CDR est autonome, il n'a aucun lien avec celui de SC ou celui de l'université.
- L'accès est limité aux chercheurs dont les projets ont été approuvés et qui ont été assermentés en vertu de la Loi sur la statistique.
- Un employé de SC est sur place pour :
 - ⇒ assurer la confidentialité, de concert avec les chercheurs (tâche principale);
 - ⇒ faire respecter les exigences en matière de sécurité/de confidentialité;
 - ⇒ fournir de l'aide technique sur les méthodes et les données;
 - ⇒ contribuer à l'examen des documents;
 - ⇒ s'occuper de la paperasserie : assermentation, etc.; et,
 - ⇒ effectuer des recherches à l'occasion.

4.4 Où en sommes-nous ?

Le service est offert dans le cadre de l'Enquête nationale sur la santé de la population et de l'Enquête longitudinale nationale sur les enfants et les jeunes.

Les défis sont importants :

- Délai d'exécution : en effet dans le domaine de la recherche, on s'attend à avoir un temps de réponse rapide, ce qui n'est pas nécessairement évident surtout dans le cadre de la vérification de la confidentialité.
- La courbe d'apprentissage pour les chercheurs relativement à un aspect particulier et à l'ensemble du contenu d'ensemble de données complexes.
- Manque de normes à l'intérieur de SC à l'égard : de la création de fichiers, des codes, et de la présentation de la documentation. Ceci rend difficile l'apprentissage du chercheur s'il travaille avec plus d'un fichier de données.

5. CENTRES DE DONNÉES DE RECHERCHE (CDR)

En partenariat avec le Conseil de recherches en sciences humaines (CRSH), les CDR représentent un système national grâce auquel les chercheurs peuvent avoir accès à des microdonnées détaillées aux fins de la recherche, tout en respectant les dispositions relatives à la confidentialité de la Loi sur la statistique. Les CDR se retrouvent dans les campus universitaires partout au pays. Leur emplacement est sécuritaire pour les données confidentielles. Ils représentent un bureau officiel de SC avec un niveau de sécurité équivalent.

5.1 Environnement

5.2 Processus

Les CDR donnent l'accès à tous les chercheurs des universités et d'autres institutions régionales (gouvernement, ONG) qui ont besoin de microdonnées confidentielles pour effectuer un projet de recherche dont le rapport sera diffusé publiquement (et non pas seulement la production de tableaux transversaux à des fins internes).

Critères d'admissibilité :

- ⇒ Le projet doit être réalisable.
- ⇒ Le chercheur doit faire preuve d'une connaissance raisonnable de la question traitée.
- ⇒ Les données peuvent appuyer les travaux proposés.
- ⇒ La méthodologie doit être raisonnable.
- ⇒ Le document de recherche initial doit correspondre au mandat de SC.
- ⇒ Une autre version pouvant contenir des commentaires peut être publiée à une date ultérieure.
- Des propositions courtes (2 à 4 pages) traitant des questions soulevées ci-dessus et l'historique de la recherche doit être envoyées au comité d'examen du CRSH.
- Des présentations peuvent être soumises de façon continue.
- On prend en compte de la capacité des CDR lors de l'acceptabilité des demandes.
- La proposition du chercheur doit être acceptée par le comité d'examen du CRSH avant que le chercheur puisse avoir accès aux CDR.

- Le chercheur doit être assermenté en vertu de la Loi sur la statistique.
- Le chercheur doit produire des produits/des services pour SC qui pourraient être un article d'analyse qui sera oublié dans une des publications de SC.
- Le chercheur doit participer à une demi-journée de formation sur tous les aspects de la confidentialité définis par SC.

5.3 Plan à court terme

- 2 CDR ouvriront d'ici l'été 2000 : un à l'université de McMaster et l'autre à celle du Nouveau-Brunswick.
- Les fichiers principaux de 3 enquêtes longitudinales : l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), l'Enquête nationale sur la santé de la population (ENSP) et l'Enquête sur la dynamique du travail et sur le revenu (EDTR) seront disponibles.
- Les fichiers disponibles sont en format : ASCII, SAS et SPSS.

5.4 Plan à long terme

- 9 CDR à l'intérieur d'universités : Nouveau-Brunswick, Dalhousie, Montréal, Toronto, McMaster, Waterloo, Calgary, Alberta, Colombie-Britannique.
- 5 enquêtes longitudinales : outre les 3 enquêtes mentionnées ci-dessus, on ajoutera deux nouvelles enquêtes : l'Enquête sur le lieu de travail et les employés (ELTE) et l'Enquête auprès des jeunes en transition (EJET).
- On rajoutera le logiciel STATA ou autre selon le besoin, à SAS et SPSS.

5.5 Défis

Comme tout nouveau projet les CDR font face à plusieurs défis, en voici quelques-uns :

- ⇒ La mise en œuvre des CDR dans des universités en tant que bureau officiel de SC est une première pour SC. La logistique entourant la mise sur pied d'un tel bureau se base sur des nouveaux paramètres : le nombre d'employés requis, la dotation de ses employés, etc.
- ⇒ Comment les chercheurs situés à l'extérieur des régions des CDR pourraient-ils avoir accès aux données de SC? La recherche est un processus continu. Il se peut très bien qu'un chercheur pour un même projet ait à se déplacer plusieurs fois. Il

n'est pas certain que les CDR peuvent répondre aux besoins des chercheurs situés en région.

- ⇒ La confidentialité fait partie de la culture de SC. La formation et la connaissance des contraintes relatives à la confidentialité sont des éléments clés du succès de cette initiative. La formation et la sensibilisation des chercheurs aux critères de SC reliés à la confidentialité sont des aspects importants de cette initiative.

6. CONCLUSION

Ce document a présenté un survol de trois initiatives relativement jeunes à SC qui ont pour but de donner accès à des données détaillées de SC. Ces trois initiatives font face à des défis communs :

- Les besoins de documentation et de métadonnées sont importants, surtout pour des enquêtes complexes comme celles de SC. Avec l'avènement de l'Internet la diffusion de la documentation, même volumineuse, est de plus en plus facile. Le site de SC contient déjà plusieurs guides et documentations de fichiers publics. On s'attend que la documentation de fichiers confidentiels soit bientôt disponible aussi. De plus, un nouveau projet de métadonnées permettant la recherche au niveau de la variable devrait être disponible d'ici 2 ans.
- La formation relative au contenu des fichiers des données est essentielle; en effet il faut comprendre le contenu de ces enquêtes complexes. Cela demande un investissement de la part du chercheur. Comme aucune norme n'existe à l'interne, la documentation ainsi que les codes utilisés ne sont pas uniformes d'une enquête à l'autre. Cela rend l'apprentissage de ces fichiers plus difficiles dans le cas de chercheurs intéressés par plus d'un fichier.
- Le modèle de l'IDD donne une place prépondérante à la communication. Le succès de l'initiative se base sur l'entraide entre universités. Il sera intéressant de voir l'évolution de la communication surtout dans les CDR : entre les centres et entre les chercheurs.
- Ces trois initiatives relèvent de trois services différents à l'intérieur de SC. Pourtant, tous les trois visent une clientèle dont les besoins sont semblables et cela pour des enquêtes communes. La communication interne entre les gestionnaires de ces trois initiatives est importante afin

d'identifier les besoins de leur clientèle et ainsi d'être en mesure d'influencer collectivement les divisions auteurs lors de la production de la documentation et des métadonnées entourant les enquêtes.

REMERCIEMENTS

Je tiens à remercier Gustave Goldmann et Jean-Louis Tambay pour leurs commentaires.