Recueil du Symposium 2001 de Statistique Canada
La qualité d'un organisme statistique : une perspective méthodologique

RÔLE DE LA BASE DE MÉTADONNÉES INTÉGRÉE À STATISTIQUE CANADA

Paul Johanis

RÉSUMÉ

La base de métadonnées intégrée renferme des informations concernant toutes les enquêtes de Statistique Canada. Elle inclut une description des sources de données, la méthodologie, des définitions des concepts et des variables mesurées de même que des indicateurs de la qualité. Il s'agit d'un outil efficace pour diffuser des informations traitant de la qualité aux utilisateurs de données. Elle couvre l'ensemble des données recueillies par Statistique Canada. L'information portant sur la qualité est conforme à la Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie de Statistique Canada. Cette information est présentée de façon cohérente et systématique.

Mots-clefs : Base de métadonnées intégral; définition des sources de données

1. INTRODUCTION

La Base de métadonnées intégrée (BMDI) est un dépôt centralisé de renseignements sur chacune des quelque 400 enquêtes réalisées par Statistique Canada. Ces enquêtes représentent les activités de base du Bureau et la BMDI, qui est la principale source d'information à leur sujet, constitue donc une ressource documentaire essentielle pour les responsables de la gestion d'ensemble des données du Bureau et pour les utilisateurs des données.

Les métadonnées pouvent soutenir au moins trois grandes fonctions d'un organisme statistique, à savoir la diffusion des données, la production de ces dernières, y compris leur collecte et la gestion du système statistique. La BMDI a été conçue tout spécialement pour faciliter la diffusion des données, autrement dit pour fournir aux utilisateurs les renseignements dont ils ont besoin pour interpréter les données statistiques que nous diffusons. Pour réaliser cet objectif, nous avons élaboré le contenu de la BMDI en nous fondant en grande partie sur les exigences de la Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie (PIUQDM).

La base de données réside sur un serveur central. Les métadonnées ont été recueillies à partir de divers dépôts préexistants de métadonnées qui ont été reformatées, validées et chargées dans la nouvelle base de métadonnées. Régulièrement, un générateur HTML lit la base de données et produit des pages formatées en HTML qui peuvent être consultées sur le site Web de Statistique Canada. Ces pages sont conçues de façon à répondre aux exigences de la PIUQDM. Elles peuvent être consultées, grâce à des hyperliens, à partir de notre base de données de sortie en ligne, appelée CANSIM 2, à partir de notre catalogue en ligne ou à partir des tableaux statistiques diffusés sur le site Web. Les pages peuvent aussi être consultées directement au moyen du moteur de recherche offert à la section sur les métadonnées du site Web. La base de données est tenue à jour grâce à un système de saisie de données déployé sur l'Intranet du Bureau. Les mises à jour sont soumises à un contrôle de la qualité et enregistrées avant qu'il ne soit possible de les utiliser pour générer des pages HTML externes.

2. INFORMATION SUR LA QUALITÉ DES DONNÉES ET LA MÉTHODOLOGIE DANS LA BMDI

Conformément à la PIUQDM susmentionnée, un ensemble précis de renseignements sommaires sur la qualité des données doit être fourni aux utilisateurs, ou mis alors à leur disposition, pour chaque produit statistique. La documentation sommaire qui doit être fournie conformément à la politique doit être présentée sous plusieurs rubriques types, chacune correspondant à un ensemble généralement cohérent de renseignements. Les pages Web

produites à partir de la BMDI reflètent exactement ces rubriques et les lignes directrices sur le contenu de chaque section.

Lorsqu'ils accèdent au point d'entrée dans la BMDI sur le site Web de SC, les utilisateurs voient s'afficher un message d'introduction normalisé qui présente les renseignements sur la qualité des données et la méthodologie, et insiste sur l'importance qu'il y a à tenir compte de ces renseignements. La méta-information est présentée en fonction d'une entité appelée enquête. Une enquête peut être une enquête directe, un programme statistique fondé sur des données administratives ou une activité d'intégration de données. Le Bureau réalise environ 400 enquêtes qui sont chacune décrites dans la BMDI. Chaque enquête peut compter une ou plusieurs éditions qui représentent chacune un cycle de l'enquête. Par exemple, une enquête mensuelle compte 12 éditions, ou cycle, par année. Les métadonnées sont recueillies pour chaque édition de l'enquête.

C'est au niveau de l'édition de l'enquête que les renseignements sur la qualité des données et sur la méthodologie sont fournis aux utilisateurs des données. Le défi que pose la définition des renseignements à fournir pour chaque édition de l'enquête consiste à trouver un juste équilibre entre la complétude et l'excès de détail. On est constamment tenté d'énumérer chaque attribut possible mais une base de données aussi détaillée risque de ne pas pouvoir être tenue à jour, car sa maintenance représenterait un fardeau trop important pour les gestionnaires d'enquête qui doivent fournir les renseignements nécessaires. Notre principe directeur est de nous en tenir à l'objectif fondamental du projet, c'est-à-dire fournir aux utilisateurs des données diffusées les renseignements nécessaires pour interpréter ces données et les placer dans le contexte approprié.

Sous la rubrique Sources des données et méthodologie, un paragraphe d'introduction décrit le but, les objectifs et la nature générale de l'enquête et précise la période de référence des données. Suit une section décrivant l'univers conceptuel, la population cible de l'enquête et l'unité statistique. En plus de ces renseignements élémentaires, les utilisateurs ont besoin de renseignements sur les méthodes utilisées pour réaliser les activités statistiques. Ces renseignements, conjugués à l'information sur la qualité des données produites, leur permettent de déterminer dans quelle mesure la source de données répond à leurs besoins.

Dans la BMDI, nous avons défini une entité méthodologique qui peut prendre la forme de plusieurs catégories méthodologiques. Ces catégories incluent le plan d'échantillonnage, la méthode de collecte et de saisie des données, les procédures de dépistage des erreurs, la méthode d'imputation, la méthode d'estimation, les processus concernant les séries chronologiques et la méthode de contrôle de la divulgation de renseignements confidentiels. Elle inclut aussi les méthodes d'évaluation de la qualité, à partir desquelles des liens peuvent être établis avec divers rapports et études sur les sources d'erreur et d'autres aspects de la qualité des données. De cette façon, il est possible d'obtenir une description normalisée des méthodes d'enquête. Il est aussi possible d'établir un lien entre ces descriptions normalisées de la documentation plus détaillée concernant la méthode d'enquête appropriée.

La grande rubrique suivante requise aux termes de la PIUQDM est celle intitulée Concepts et variables mesurées. Pour chaque programme d'enquête, la base de données contient la liste des variables pour lesquelles des données sont produites, accompagnée de leur définition et de leur classification. Pour modéliser cet aspect de la base de données, nous nous sommes conformés à la norme ISO 11179 dont nous avons retenu les composantes fondamentales, à savoir le concept d'élément de donnée, la classe d'objets, l'élément de donnée et le domaine de valeurs. Toutefois, nous avons utilisé une terminologie mieux comprise par les utilisateurs de données statistiques. À Statistique Canada, le concept d'élément de donnée est simplement appelé concept, la classe d'objet est une unité statistique, l'élément de donnée est une variable, tandis que le domaine de valeur, accompagnée de sa définition, correspond à ce que nous appelons une classification. Néanmoins, nous avons maintenu dans le modèle la structure et les relations fondamentales implicites de la norme. Nous avons défini chacun de ces éléments dans la Politique sur les normes du Bureau. En vertu de cette politique, des normes peuvent être établies en ce qui concerne la définition des concepts, des variables, des classifications, des unités et des populations. Les programmes statistiques qui ont adopté ces définitions normalisées n'ont donc pas à fournir de renseignements supplémentaires pour les variables concernées, puisqu'elles seront déjà décrites conformément aux normes qui seront mémorisées dans la base de données.

La dernière rubrique obligatoire aux termes de la PIUQDM est celle intitulée Exactitude des données. Sous cette rubrique, plusieurs mesures de la qualité ont été définies au niveau de l'édition d'une enquête. Elles incluent les composantes nécessaires pour calculer le taux de réponse, l'erreur de couverture, le taux d'imputation et l'erreur

d’échantillonnage pour les variables clés. L’erreur d’échantillonnage est exprimée sous forme de coefficient de variation. Les c.v. ne sont enregistrés dans la base de métadonnées que pour certaines variables importantes. De nouveau, il est possible d’établir des liens avec de la documentation plus détaillée sur la qualité des données.

Le modèle complet de données figure à l’annexe 1.

3. CONCLUSION

La base de métadonnées intégrée est un outil efficace pour diffuser des informations traitant de la qualité aux utilisateurs de données. Elle couvre l’ensemble des données recueillies par Statistique Canada. L’information portant sur la qualité est conforme à la Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie de Statistique Canada. Cette information est présentée de façon cohérente et systématique. Maintenant que ce système est en place, le défi à venir consistera à maintenir et améliorer la qualité des informations qu’elle contient.

Annexe 1 : Rubriques concernant la méta-information dans la BMDI

Activité statistique
Enquête
- Univers
- Concept d'élément de donnée
  - Concept
  - Unité statistique
  - Classification
- Base de sondage

Édition d'une enquête
- Conception du questionnaire
- Instrument d'enquête
- Plan d'échantillonnage
- Échantillon
- Méthode de collecte et de saisie des données
- Procédures de vérification des données
- Méthode d'imputation
- Procédures de contrôle de la qualité
- Méthode d'estimation
- Processus concernant les séries chronologiques
- Fichiers de données
- Méthode de contrôle de la divulgation de renseignements confidentiels
- Méthode d'évaluation de la qualité
- Mesures de la qualité
  - Erreur de couverture
  - Taux de réponse
  - Taux d'imputation
  - Erreur d'échantillonnage pour les variables importantes
  - Autre évaluation de la qualité

Pour chaque élément de cette liste, nous recueillons les renseignements génériques suivants :

- Période de référence
- Thème
- Sujet
- Mot clé
- Organisme
- Personne-ressource
- Documentation supplémentaire