No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2005 : Défis méthodologiques reliés aux besoins futurs d'information

2005

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2005 de Statistique Canada
Défis méthodologiques pour les besoins futurs d'information

DÉCOUVRIR LES VARIABLES DANS LES FICHIERS DE MICRODONNÉES : COMPARAISON DE LA DOCUMENTATION CONFORME À LA NORME DDI À UN REGISTRE DE MÉTADONNÉES ISO/IEC 11179

Tim Dunstan[1], Charles Humphrey[2]

1. INTRODUCTION

Il a toujours été important de documenter les microdonnées des enquêtes de Statistique Canada, que les données proviennent d'un fichier principal confidentiel ou d'un fichier à grande diffusion[3]. Ces deux types de fichier comportent un ensemble de variables qui doivent faire l'objet d'une description complète et précise pour qu'on puisse comprendre et analyser les données. Les utilisations secondaires des microdonnées dépendent fortement d'une description très détaillée des variables, dont le cliché d'enregistrement sur le support physique d'information, le code constituant l'ensemble de réponses de chaque variable, l'origine de chaque variable dans l'instrument d'enquête initial, le traitement des non-réponses et la répartition des réponses pour chaque variable. Les renseignements contextuels, comme les directives concernant l'utilisation d'une variable de pondération en fonction du plan d'enquête ou les lignes directrices concernant la variabilité due à l'échantillonnage, fournissent également des renseignements essentiel à l'analyse des données d'une enquête. Collectivement, tous ces renseignements font partie intégrante de l'aspect global de la documentation et brossent un tableau complet des données et des circonstances de leur création. Dans le monde de l'archivage de données, on envisage à l'heure actuelle d'appliquer à la documentation des données un modèle de cycle de vie, selon lequel l'information serait recueillie à toutes les étapes d'une enquête. Cela nous amène à poser, à juste titre, la question suivante : quelle information doit entrer dans la documentation des données?

La majorité des cadres de travail utilisés pour définir les éléments acceptés de la documentation des données sont fondés sur la convention ou les meilleures pratiques. Tout récemment, cependant, deux initiatives internationales ont redéfini le contenu et la structure de la documentation des données grâce à la formulation de deux normes de métadonnées. Dans le présent exposé, nous comparons la façon dont ces deux normes, ISO/IEC 11179 et DDI, soutiennent la documentation des variables dans les fichiers de microdonnées. À partir d'un modèle global de l'information la plus recherchée par les chercheurs qui effectuent une analyse secondaire des données, nous évaluons la mesure dans laquelle chaque norme englobe et décrit cette information.

2. INCIDENCE DE LA TECHNOLOGIE SUR LES MÉTADONNÉES

Les métadonnées constituent une information organisée systématiquement pour décrire une autre information au moyen de définitions, d'illustrations, de classifications, de conceptualisations et d'autres façons d'exprimer les relations entre des objets. Dans le domaine de la technologie de l'information, les métadonnées sont traitables par ordinateur pour faciliter l'accès à l'information qu'elles décrivent ou pour offrir de nouvelles façons de consulter et

[1] Statistique Canada
[2] University of Alberta
[3] Le fichier principal contient des données suffisamment détaillées sur les participants à une enquête pour qu'on puisse identifier certains répondants. Ces données ne sont pas accessibles au public. Par contre, les fichiers de microdonnées à grande diffusion sont conçus afin de réduire au minimum le risque de divulguer des renseignements sur un répondant. Toutefois, le Comité de la diffusion des microdonnées de Statistique Canada doit approuver les mesures prises pour assurer la confidentialité des données contenues dans un fichier afin que ce dernier ne soit rendu public.

de comprendre l'information. Dans ce contexte, les métadonnées dépendent à la fois de la technologie informatique existante et des processus intellectuels utilisés pour organiser l'information.

Par exemple, les enregistrements CANSIM I ont été structurés dans des champs de longueur fixe et organisés en fonction d'un traitement séquentiel, qui correspondait aux techniques informatiques prédominantes des années 1960 et 1970, lorsque l'information était stockée sur bande magnétique et traitée séquentiellement par un ordinateur central. L'habitude d'entasser les métadonnées CANSIM dans des champs fixes a entraîné la prolifération d'abréviations utilisées pour consigner le plus d'information possible dans des champs de taille fixe. Dans bien des cas, on s'est retrouvé avec plus d'une abréviation pour le même concept. La longueur d'une abréviation dépendait de l'espace qui restait dans le champ. Il est clair que la technologie informatique de l'époque a eu une forte incidence sur la teneur des métadonnées de ce temps-là.

CANSIM II a été mis au point pendant les années 1990, lorsque la forme dominante de technologie de l'information consistait en un réseau réparti utilisant une mémoire à disques magnétiques avec accès direct à l'information. Le concept de base de données relationnelles a remplacé la structure des enregistrements CANSIM I, libérant ainsi les emplacements d'enregistrement fixes de la contrainte d'un espace limité réservé à l'information. Cette nouvelle structure permettait également de coupler l'information contenue dans un champ à celle d'un autre champ et d'éliminer une grande quantité d'information redondante d'un enregistrement à l'autre, ce qui a accru le contrôle réel sur les concepts et leurs abréviations.

L'exemple de CANSIM montre l'incidence du contexte technologique sur l'évolution des métadonnées. De même, les deux normes de métadonnées qui font l'objet de notre étude reflètent le contexte informatique dans lequel elles sont apparues. Leur pertinence découle de la technologie d'aujourd'hui, fondée sur l'informatique répartie alliée à Internet. ISO/IEC 11179 est un schéma qui fonctionne bien avec des bases de données relationnelles réparties; DDI emploie un langage de balisage de textes codés, fondé sur la technologie Web. Si les deux normes de métadonnées visent une information identique, leur mise en œuvre mène cependant à des flots de traitement différents, ce qui soulève une question à laquelle la présente étude ne répond pas : est-il plus avantageux d'organiser les métadonnées concernant les microdonnées selon la structure d'une base de données relationnelles ou au moyen de la gestion des métadonnées dans le cadre du Web?

3. MICRODONNÉES OU DONNÉES AGRÉGÉES : LES DIFFÉRENCES AU NIVEAU DES MÉTADONNÉES

Un autre aspect dont il faut traiter ici est la distinction entre les microdonnées et les données agrégées, et les différences que présentent les métadonnées entre ces structures de données. Comme nous l'avons mentionné plus haut, les microdonnées constituent une information recueillie au niveau d'observation et contiennent des détails sur des individus ou des membres spécifiques de l'unité d'observation. Elles sont souvent appelées données brutes parce qu'en général elles n'ont pas été traitées et qu'elles restent au niveau auquel on a d'abord observé l'information.

Les données agrégées, par contre, ont été portées (après traitement) à un niveau supérieur à celui auquel on a observé les données. Autrement dit, il s'agit de microdonnées résumées. Elles sont habituellement agrégées selon des critères géographiques, chronologiques ou sociaux. Par exemple, les microdonnées d'un recensement sont couramment résumées selon les critères géographiques habituels : divisions de recensement, subdivisions de recensement, régions métropolitaines de recensement, secteurs de recensement, etc. L'agrégation des microdonnées exige une définition précise de la variable pour laquelle on calcule les résumés ainsi que les définitions des variables à résumer. Ces définitions font partie intégrante des métadonnées décrivant le fichier de données agrégées.

Pourquoi aborder les différences entre les données agrégées et les microdonnées? Parmi les deux normes de métadonnées examinées dans la présente étude, il se peut que l'une convienne mieux aux microdonnées et l'autre, aux données agrégées. Comme les données agrégées et les microdonnées possèdent des caractéristiques différentes, il est possible que des normes de métadonnées différentes correspondent mieux à cette unicité. Les enregistrements d'un fichier de microdonnées sont créés au niveau d'observation de l'enquête initiale, alors que les enregistrements de données agrégées ne le sont pas (par exemple, un enregistrement peut résumer de l'information sur bien des individus au sein d'une division de recensement). D'une part, les métadonnées concernant des microdonnées doivent comprendre des renseignements plus détaillés sur les données non traitées. D'autre part, les métadonnées concernant

des données agrégées doivent comprendre des renseignements sur les concepts utilisés pour procéder à l'agrégation. Sur le plan du contenu, il existe donc des différences qualitatives qui doivent être décrites dans les métadonnées de ces deux types de structure de données. Nous n'explorons pas ces différences dans la présente étude, mais cette distinction nous a incités à nous concentrer sur une seule des deux structures de données, soit celle des microdonnées.

4. L'OBJET DE L'ÉTUDE ET SA MÉTHODOLOGIE

L'objet de notre étude consistait à comparer les normes de métadonnées ISO/IEC 11179 et DDI à partir d'un modèle des besoins en information d'un chercheur menant une analyse secondaire d'un fichier de microdonnées. Nous avons cerné les éléments ou attributs de chaque norme de métadonnées qui pouvaient contenir l'information recherchée par le chercheur si la documentation était pleinement établie au moyen de la norme de métadonnées. Nous n'avons pas tenté strictement d'apparier une norme à l'autre. La comparaison consistait plutôt à appliquer la norme pour répondre aux besoins en information de l'utilisateur final.

Nous avons adopté cette approche en raison des différences susmentionnées entre les caractéristiques des microdonnées et celles des données agrégées. Au lieu de laisser les normes de métadonnées déterminer le domaine d'information pour la documentation des microdonnées, nous avons choisi d'utiliser un modèle externe, soit les besoins en information d'un chercheur menant une analyse secondaire. À partir des éléments d'information tirés de ce modèle externe, nous avons alors tenté de repérer des éléments semblables dans chaque norme de métadonnées.

Une autre raison pour laquelle nous avons choisi le point de vue de l'utilisateur final tient au fait que le domaine des métadonnées, défini par le producteur des microdonnées, ne répond pas nécessairement à tous les besoins en information de l'utilisateur final. Dans le cas de Statistique Canada, le producteur des données est aussi le créateur des métadonnées; aussi la conception des métadonnées est-elle liée au point de vue du producteur plutôt qu'à celui de l'utilisateur final des données. Adopter le point de vue de l'utilisateur final, c'est comme utiliser à rebours le télescope de la documentation des données.

Ni le point de vue de l'utilisateur final, ni celui du producteur des fichiers de microdonnées, ne représente nécessairement le modèle de documentation des données le plus exhaustif. L'information à inclure dans la documentation des données constitue toujours un enjeu de recherche. Comme nous l'avons mentionné plus haut, le monde de l'archivage de données étudie à l'heure actuelle un modèle de cycle de vie pour résoudre cette question. Une approche fondée sur le cycle de vie permet d'intégrer un processus au modèle de documentation et de déterminer l'information importante à documenter à toutes les étapes d'une enquête.

Nous avons également choisi le point de vue de l'utilisateur final pour neutraliser les effets de la technologie dans la comparaison des deux normes de métadonnées; Peu nous importait de savoir si une structure de base de données relationnelles était préférable à un langage de balisage de textes codés. Face à ces deux structures de métadonnées, nous voulions savoir dans quelle mesure elles répondaient aux besoins en information de l'utilisateur final.

Nous nous sommes concentrés sur les microdonnées en raison du niveau de détail associé à ces données. La nature des microdonnées et celle des données agrégées, abordées plus haut, peuvent être juste assez différentes pour qu'une norme soit plus utile que l'autre à l'égard des microdonnées. Nous voulions aussi examiner dans quelle mesure ISO/IEC 11179 documente les données au niveau des variables. La Base de métadonnées intégrée (BMDI) de Statistique Canada emploie déjà la norme ISO/IEC 11179 et a été mise en œuvre avec des séries de données chronologiques, qui sont une forme de données agrégées. Nous voulions voir dans quelle mesure on pouvait employer cette norme pour décrire les variables au niveau des microdonnées dans la BMDI. Enfin, notre étude avait un objectif très pragmatique : celui de voir dans quelle mesure il était aisé de traduire d'une norme à l'autre les métadonnées relatives aux variables.

Figure 1
Tâches nécessitant de l'information à l'étape de la découverte de données

Trouver des variables selon des titres de sujet ou des thèmes communs.
Trouver des données en utilisant une unité d'observation pertinente pour l'analyse à effectuer.
Parcourir les brèves descriptions des variables d'intérêt.
Lire la description détaillée d'une variable.
Cerner le niveau de mesure et l'ensemble de réponse d'une variable.
Examiner l'intervalle des valeurs d'une variable à l'intérieur d'un échantillon.
Examiner les étiquettes des variables correspondant à une mesure catégorique.
Examiner la répartition des réponses pour une variable.
Déterminer le type et le taux de non-réponse à une variable.
Déterminer qui, dans l'échantillon, a fourni de l'information pour cette variable.
Déterminer la nécessité de poids d'échantillonnage.
Évaluer la qualité des données et déterminer s'il y a lieu de recourir à l'imputation et aux réponses par procuration.
Établir le contexte de l'information contenue dans le questionnaire.
Déterminer la source de l'information contenue dans cette variable.

La méthodologie utilisée pour comparer ces deux normes de métadonnées comportait deux étapes. Premièrement, nous avons défini les diverses étapes de l'exécution d'une analyse secondaire et cerné les besoins en information à chaque étape. Nous avons ensuite identifié les éléments de chaque norme qui contiendraient cette information. Deuxièmement, nous avons choisi un ensemble de variables tirées du cycle 17 de l'Enquête sociale générale (ESG 17) pour représenter la diversité des variables qu'on trouve dans les fichiers de microdonnées. Dans la mesure du possible, nous avons ensuite produit pour ces variables des métadonnées selon chaque norme. Enfin, nous avons comparé l'exhaustivité de la représentation des métadonnées concernant ces variables.

5. LES BESOINS EN INFORMATION DE L'UTILISATEUR FINAL

Nous avons catégorisé l'information nécessaire pour mener une analyse secondaire de données en trois grandes étapes : la découverte, l'extraction et l'analyse des données. Pour chaque étape, nous avons distingué plusieurs tâches, dont chacune présentait ses propres besoins en information. Par exemple, l'étape de la découverte des données comportait quatorze tâches distinctes (voir la figure 1). Nous avons ensuite décrit le type d'information recherché pour chacune des tâches. Par exemple, nous avons ventilé le contenu des métadonnées de la première tâche de la figure 1 sous forme de liste de titres de sujet, de mots-clés ou de thèmes couramment acceptés pour catégoriser chaque variable. Les éléments de la norme ISO/IEC 11179 pouvant contenir cette information sont THEME, TOPIC, KEYWORD; ceux de la norme DDI sont <varGRP><labl> et <varGRP><txt>. La liste complète des tâches, du contenu des métadonnées et des éléments des normes ISO/IEC 11179 et DDI est disponible auprès des auteurs.

Après avoir cerné ces éléments des deux normes de métadonnées, nous avons choisi, dans le fichier principal de l'ESG 17, treize variables représentatives des types de variable qu'on trouve habituellement dans les fichiers de microdonnées. Nous les avons choisies pour répondre à l'un des critères suivants :

- des variables tirées de questions donnant lieu à des ramifications ou à des instructions « passez à »;
- des variables pouvant servir à calculer de nouvelles variables;
- des variables dotées des mêmes étiquettes de catégorie mais utilisant des échelles différentes;
- des variables tirées d'une question ouverte;
- des variables apparaissant dans des fichiers de microdonnées pour lesquels on a établi des catégories types et qui permettent de comparer la répartition des réponses.

Figure 2

Enquête sociale générale, cycle 17 – Variables utilisées pour mettre à l'essai la mise en œuvre des métadonnées

ACMYR	*Activité principale du répondant au cours des 12 derniers mois*
EDUSTAT	*Études à temps plein ou à temps partiel du répondant*
MAR_Q125	*Cherchez-vous un travail rémunéré?*
MAR_Q130	*Au cours des 12 derniers mois, avez-vous travaillé à un emploi rémunéré ou à votre propre compte?*
AGE_LSTPDWK	*Âge du répondant lorsqu'il a effectué un travail rémunéré pour la dernière fois (supprimée dans le fichier à grande diffusion)*
AGE_LSTPDWKC	*Âge du répondant lorsqu'il a effectué un travail rémunéré pour la dernière fois*
MAR_Q140M	*En quel mois avez-vous effectué un travail rémunéré pour la dernière fois? (supprimée dans le fichier à grande diffusion)*
MAR_Q140Y	*En quelle année avez-vous effectué un travail rémunéré pour la dernière fois? (supprimée dans le fichier à grande diffusion)*
MAR_Q140A	*Quel âge aviez-vous lorsque vous avez effectué un travail rémunéré pour la dernière fois? (supprimée dans le fichier à grande diffusion)*
LS_Q210	*En utilisant la même échelle [de 1 à 10], quel sentiment éprouvez-vous maintenant à l'égard de votre vie?*
LS_Q120	*Veuillez évaluer ces aspects… Que diriez-vous de votre emploi ou votre activité principale?*
LS_Q130	*Veuillez évaluer ces aspects… Que diriez-vous de l'emploi de votre temps libre?*
EDUC10	*Niveau d'études le plus élevé atteint par le répondant – 10 groupes.*

Nous avons ensuite décrit chacune de ces variables en appliquant les deux normes de métadonnées. Dans le cas de la norme DDI, nous avons utilisé Data Publisher® de NESSTAR pour produire ces métadonnées[4]. La source de l'information provenait du dictionnaire de données de l'ESG 17 diffusé par la division auteure, du questionnaire de l'ESG 17, de la version SPSS® du fichier de données ainsi que de classifications personnelles des noms et des mots-clés des groupes de variables.

6. RÉSULTATS DE L'ÉTUDE

Toutes les variables d'essai tirées de l'ESG 17 ont été décrites en détail selon la norme DDI et selon les besoins en information du modèle de l'analyse secondaire de données à l'étape de la découverte des données. Les métadonnées

[4] Data Publisher est une application MS Windows brevetée par NESSTAR Inc. qui produit un fichier XML de métadonnées conformes à la norme DDI. Cette application utilise des formules pour identifier les champs d'information associés aux éléments DDI. L'avantage de cet outil est qu'on peut remplir la documentation sans avoir à baliser directement l'information. Data Publisher peut aussi lire des fichiers internes SPSS pour récupérer l'information de nombreux éléments utilisés pour décrire les variables.

concernant quatre des cinq tâches à l'étape de l'extraction des données ont également été saisies selon la norme DDI. Toutefois, les deux tâches à effectuer à l'étape de l'analyse des données se situaient au-delà des deux normes de métadonnées.

On peut utiliser des objets ISO ou des combinaisons d'objets ISO pour produire les métadonnées variables requises du modèle de l'analyse secondaire de données. On peut apparier la norme ISO 11179 pour produire une documentation équivalente conforme à la norme DDI.

7. CONCLUSIONS

Dans la présente étude, nous avons utilisé la version 2 de la norme DDI, qui comprend une bibliothèque d'étiquettes en cinq parties pour baliser le texte des livres de codes traditionnels en sciences sociales[5]. Les éléments de la partie IV, qui décrivent les variables, convenaient le mieux au modèle d'information de l'utilisateur final. La mise en œuvre de la version 2 donne un volume important de codes redondants pour des variables présentant une information identique. Par exemple, les variables portant les mêmes étiquettes de valeur sont toutes codées avec l'ensemble de balises et d'étiquettes. Ce codage répétitif accroît de beaucoup la taille du fichier XML contenant les métadonnées, tout en soulevant des problèmes de traitement lorsqu'il s'agit d'apparier les renseignements entre les deux normes examinées dans la présente étude. La propagation d'étiquettes de valeur balisées selon la norme DDI à partir du champ ISO/IEC 11179 contenant cette information est beaucoup plus directe que le processus inverse, soit le passage de DDI à ISO/IEC 11179. Cette asymétrie est attribuable à la nature des modèles d'élément de ces deux normes.

À la fin de l'automne 2005, DDI Alliance, organisme de surveillance de la norme DDI, a publié la version 3 de cette norme[6]. Cette nouvelle version représente une restructuration importante de la norme : la conception hiérarchique initiale, celle d'un livre de codes en cinq parties, a fait place à une conception modulaire qui permet des relations plus complexes entre les éléments des métadonnées. Un avantage de cette nouvelle version orientée objet de la norme DDI tient à sa capacité d'intégrer et de comparer des métadonnées externes. Par exemple, on peut coupler des vocabulaires et des thésaurus externes au niveau des variables sans doubler ni reproduire le contenu. Par conséquent, des bases de données de classification contenant de l'information plus abstraite à un niveau conceptuel ou définitionnel peuvent exister en dehors des métadonnées DDI. On peut utiliser ces registres externes pour comparer de manière uniforme les similitudes et les différences entre les variables, les concepts et les définitions. Par exemple, la BMDI est un registre de métadonnées ISO/IEC 11179 qui tient à jour des définitions et des concepts courants. On peut coupler ces éléments à la documentation de microdonnées au niveau des variables pour les enquêtes de Statistique Canada, qui se situent habituellement à un niveau moins abstrait et qui nécessitent, dans bien des cas, des liens répétitifs et redondants pour décrire les variables de microdonnées. DDI et ISO/IEC 11179 sont des normes complémentaires qui peuvent remplir des fonctions distinctes mais apparentées. Ensemble, elles enrichissent grandement le potentiel des métadonnées à l'égard des microdonnées.

---

[5] Ces cinq parties comprennent des éléments servant à décrire le document de métadonnées, l'enquête ou l'étude dans laquelle on a créé les données, les fichiers de données de l'enquête, les variables de chaque fichier et d'autres documents liés à l'étude. Un fichier de métadonnées conformes à la norme DDI est un gros document ASCII contenant des étiquettes tirées de chacune de ces cinq parties et le contenu décrivant une enquête. Rédigés en langage XML, ces documents peuvent être traités sur Internet et affichés à l'aide d'une variété infinie de feuilles de style.

[6] DDI Alliance est une association mutuelle dont les membres interviennent dans l'élaboration de la norme DDI. On trouvera de plus amples renseignements au sujet de l'Alliance sur le site www.icpsr.umich.edu/DDI/org/index.html