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 reliés aux besoins futurs d'information

LA NORME DDI DANS LE MONDE RÉEL

A. Michelle Edwards[1] et Marie-Josée Bourgeois[2]

RÉSUMÉ

La norme DDI (Data Documentation Initiative) est une norme internationale d'élaboration de métadonnées. L'Initiative de démocratisation des données (IDD) ainsi que des universités partenaires, dont l'Université de Guelph, s'emploient à créer les métadonnées de toutes les enquêtes de Statistique Canada accessibles à la communauté de l'IDD. On a sélectionné dans le livre de code DDI plus de cent balises qu'on a appliquées aux enquêtes contenues dans le dépôt de l'IDD. L'Université de Guelph et l'équipe de l'IDD utilisent le logiciel Nesstar pour créer les métadonnées et le serveur Nesstar pour mettre les données et les métadonnées à la disposition des utilisateurs. Dans le présent exposé, nous passons en revue les balises sélectionnées et le processus de création des métadonnées et présentons l'exemple d'une enquête maintenant entièrement accessible à la communauté de l'Université de Guelph.

MOTS CLÉS : DDI; IDD

1. INTRODUCTION

1.1 Le monde de l'enseignement supérieur

Dans les années 80, devant l'augmentation des coûts associés aux données de Statistique Canada, les chercheurs, les étudiants et les enseignants des établissements d'enseignement postsecondaire canadiens se sont tournés vers d'autres sources de données. En effet, les données issues des États-Unis, de la Grande-Bretagne et de la Chine leur étaient offertes à meilleur marché. Toutefois, celles-ci ne rendaient pas toujours compte de la situation canadienne et pouvaient présenter des lacunes. En outre, de nombreux établissements d'enseignement supérieur ne fournissaient pas le soutien technique dont les chercheurs avaient besoin pour traiter des fichiers de données complexes.

Afin d'atténuer quelque peu l'effet des coûts plus élevés du recensement de 1986, l'Association canadienne des utilisateurs de données publiques (ACUDP) et l'Association des bibliothèques de recherche du Canada (ABRC) créaient en 1989 un consortium d'achat spécial. L'expérience a démontré qu'il était possible pour Statistique Canada et les établissements d'enseignement supérieur canadiens de conclure avec succès ce genre d'entente.

En 1993, un groupe de travail parrainé par la Fédération canadienne des sciences sociales (FCSS) élaborait un projet qui allait être accepté à la fois par Statistique Canada et les milieux universitaires. Statistique Canada et le Programme des services de dépôt ont joué un rôle de premier plan à cet égard. En février 1996, le Conseil du Trésor donnait le feu vert à un projet pilote de cinq ans sur l'Initiative de démocratisation des données (IDD) et, en mars de la même année, le gouvernement fédéral intégrait celle-ci à sa Stratégie scientifique et technologique. Le projet a connu un succès foudroyant et, en avril 2001, on le consacrait programme officiel. Le programme opère depuis la Bibliothèque et Centre d'information de Statistique Canada.

1.2 L'Initiative de démocratisation des données

L'Initiative de démocratisation des données (IDD) a été créée afin d'améliorer l'accès aux données par les établissements d'enseignement postsecondaire canadiens. Il s'agit d'une démarche rentable qui a donné d'excellents

[1] A. Michelle Edwards, Academic Services, University of Guelph, Guelph, Ontario, N1G 5H8.
[2] Marie-Josée Bourgeois, Initiative de démocratisation des données, Statistique Canada, Ottawa (Ontario) K1A 0T6.

résultats à ce jour. Avant l'introduction de l'IDD, les universités et collèges du Canada devaient acheter les données de Statistique Canada un fichier à la fois. Aujourd'hui, grâce à ce programme, et moyennant une cotisation annuelle, les membres du corps professoral et les étudiants des établissements d'enseignement postsecondaire participants bénéficient d'un accès illimité à un grand nombre de fichiers de microdonnées, de bases de données et de fichiers géographiques de Statistique Canada. Ce programme équitable et à coût abordable offre aux chercheurs universitaires la possibilité de consulter les statistiques les plus récentes – et d'autres moins récentes –, lesquelles s'avèrent de puissants outils d'analyse pour leurs études sur la société canadienne.

L'IDD a contribué à l'avènement d'une culture de données au Canada. Depuis son introduction, les bibliothécaires de données et des noyaux de centres de données poussent comme des champignons. Ces bibliothécaires de données se sont assurés une formation entre eux et travaillent ensemble au service de la communauté. Ils travaillent également de concert avec les professeurs pour amener les étudiants à s'intéresser aux statistiques sociales.

L'IDD est un exemple d'une mise en valeur importante de la technologie de l'autoroute de l'information au Canada puisqu'elle permet aux universités d'offrir pour la première fois une gamme complète de services de données aux étudiants comme aux professeurs. Il appert aussi de plus en plus que l'Initiative contribue de manière importante à l'enseignement et à la recherche au Canada. Des cours sont remaniés pour encourager les étudiants à utiliser les données, et des bourses ont été remportées pour des propositions liées directement à la disponibilité de données par l'intermédiaire de l'IDD.

Les chercheurs qui, auparavant, devaient compter principalement sur les sondages d'opinion pour obtenir des données de source canadienne, peuvent maintenant compléter celles-ci au moyen des microdonnées à grande diffusion de Statistique Canada.

Aujourd'hui, 68 universités et collèges participent à l'IDD.

L'IDD compte maintenant plus de 250 enquêtes dans sa collection de fichiers agrégés, de fichiers de microdonnées à grande diffusion, de fichiers SPSS ou SAS et de documents connexes en Word ou en PDF. Ces fichiers sont accessibles aux établissements d'enseignement supérieur par le biais d'un site FTP qui réside sur un serveur de Statistique Canada réservé à l'IDD.

2. DOCUMENTATION D'ENQUÊTE

2.1 Aujourd'hui – le monde réel

La documentation d'enquête comprend souvent plusieurs parties : un livre de code décrivant le codage des questions d'enquête; un guide de l'utilisateur expliquant comment on a mené l'enquête, calculé les poids et recueilli les données, ainsi qu'un fichier de données, souvent un fichier plat ASCII contenant une série de nombres. Pour interpréter le fichier de données, les utilisateurs ont besoin soit d'un fichier de clichés d'enregistrement qui décrit la position des questions ou des variables et le contenu du fichier, soit d'un fichier de syntaxe pour progiciel statistique, comme SAS ou SPSS.

Avant la mise en œuvre de l'IDD, les établissements d'enseignement supérieur recevaient des fichiers de données brutes (en format ASCII) accompagnés d'un fichier de clichés d'enregistrement. Quant à la documentation d'enquête connexe (livre de code ou guide de l'utilisateur), ils la recevaient sous forme d'imprimés dans une reliure. Les chercheurs qui voulaient analyser des résultats d'enquête devaient emprunter un exemplaire de la documentation auprès de la bibliothèque de l'établissement. En effet, un fichier contenant des nombres n'était pas très utile pour déterminer les variables et les codes nécessaires à leur analyse.

À mesure qu'Internet s'est développé et que les utilisateurs ont mieux maîtrisé l'informatique, les fichiers d'enquête ont été mis à leur disposition au moyen d'interfaces Web centralisées. Des fichiers de syntaxe pour progiciel statistique ont aussi été mis à la disposition des utilisateurs pour leur permettre de consulter les fichiers dans l'environnement de leur choix. Dans bien des cas, les utilisateurs pouvaient désormais effectuer une analyse

provisoire sans emprunter la documentation imprimée auprès de la bibliothèque; au cours de leur recherche, toutefois, ils devaient consulter cette documentation pour connaître la méthodologie d'enquête.

Aujourd'hui, de nombreux livres de code et guides de l'utilisateur sont disponibles sous forme électronique. Les utilisateurs peuvent maintenant obtenir électroniquement la documentation dont ils ont besoin. Toutefois, pour obtenir l'information nécessaire à leur travail de recherche et d'analyse, ils doivent encore accéder à plusieurs fichiers : un fichier de syntaxe pour lire le fichier de données, un livre de code pour interpréter les codes utilisés dans le fichier de données et un guide de l'utilisateur pour comprendre la méthodologie d'enquête.

2.2 Demain – le monde parfait

Naguère accessibles sous forme de fichiers électroniques et d'imprimés, les données d'enquête et la documentation sont maintenant entièrement disponibles sous forme électronique. Toutefois, pour obtenir un tableau complet de l'enquête, les utilisateurs doivent encore accéder à un certain nombre de fichiers différents. Un livre de code ou un fichier électronique contenant toute la documentation d'enquête, y compris le fichier de données, constituerait la prochaine génération de moyens d'accès à l'enquête. Les utilisateurs pourraient alors obtenir des détails sur la méthodologie, la définition des codes, des renseignements sur les données et le contexte de l'enquête, le tout dans un seul fichier – un véritable guichet unique. Imaginez si ce fichier était aussi consultable! Les utilisateurs pourraient chercher de l'information au niveau des questions ou des variables avant de rassembler les données, ce qui réduirait le temps nécessaire pour effectuer une analyse provisoire.

Dès lors, comment les fournisseurs de données peuvent-ils faciliter la transition entre plusieurs fichiers électroniques et un seul fichier ou livre de code? Une nouvelle initiative, appelée Data Documentation Initiative, semble offrir la réponse : un fichier en langage XML composé de balises décrivant toute la documentation associée à une enquête donnée.

3. DATA DOCUMENTATION INITIATIVE

3.1 Contexte

Data Documentation Initiative (DDI) est une initiative internationale consistant à établir une norme de documentation technique pour décrire les données en sciences sociales (http://www.icpsr.umich.edu/DDI/). Les objectifs de DDI consistent à faciliter l'échange et le transport de documents en créant une norme en langage XML, à favoriser la préservation de la documentation concernant des ensembles de données et à améliorer la documentation d'enquête en conservant toutes les capacités du livre de code électronique, mais en augmentant considérablement la portée et la rigueur de l'information qu'il contient (http://www.icpsr.umich.edu/DDI/codebook/index.html).

Élaborée par le Inter-University Consortium for Political and Social Research, la spécification de métadonnées DDI est devenue le projet d'une alliance de 25 établissements de l'Amérique du Nord et de l'Europe. L'alliance se réunit environ deux fois par année et comporte plusieurs groupes de travail.

3.2 Le livre de code DDI

Le livre de code DDI est un fichier en langage XML composé d'un certain nombre de balises et d'attributs à l'intérieur des balises. Chaque balise décrit un aspect de l'enquête; par exemple, le titre d'une enquête apparaîtrait comme suit : <titl>Enquête sur les finances des consommateurs</titl> ou le diffuseur des données apparaîtrait comme suit : <distrbtr abbr="IDD" affiliation="Statistique Canada">Initiative de démocratisation des données</distrbtr>. Les balises sont <titl> et <distrbtr>, alors que les attributs seraient les métadonnées indiquées à l'intérieur de la balise, comme abbr="IDD" et affiliation="Statistique Canada" figurant à l'intérieur de la balise de diffuseur <distrbtr>.

Les concepteurs d'un fichier XML disposent de plusieurs options pour rendre l'information accessible aux utilisateurs. À l'aide d'un navigateur Internet, on peut aisément consulter le fichier sous forme de liste de balises et de leur contenu. Toutefois, les utilisateurs préfèrent consulter le fichier XML dans un format familier, soit HTML. On peut utiliser des logiciels comme Saxon et Nesstar, entre autres, pour consulter des fichiers XML en format HTML. Le groupe de l'IDD et l'Université de Guelph ont choisi d'utiliser la gamme de produits Nesstar pour élaborer et consulter les livres de code DDI, principalement parce que le logiciel Nesstar a été spécialement mis au point pour la norme DDI.

Le livre de code DDI est réparti en cinq sections : Description du document, Description de l'enquête, Description des fichiers, Description des variables et Description des autres documents relatifs à l'enquête. Chaque section contient un certain nombre de balises et d'attributs servant à décrire le contenu. À l'intérieur de chaque section, les balises sont considérées comme obligatoires (elles sont appariées aux balises du Dublin Core utilisées par les catalogueurs des bibliothèques), recommandées ou répétables. Les balises répétables sont celles qui peuvent figurer plus d'une fois pour montrer qu'il existe plusieurs occurrences de certaines métadonnées, par exemple des variables. Comme la balise <var> sert à décrire chaque variable figurant dans l'enquête, elle est considérée comme une balise répétable.

3.2.1 Description du document

La première section du livre de code décrit le fichier XML ou le livre de code conforme à la norme DDI. Il peut contenir 71 balises, dont quatre sont obligatoires.

Cette section contient les renseignements bibliographiques qui décrivent le fichier du livre de code conforme à la norme DDI. Les balises de cette section comprennent des éléments comme le titre du document, le numéro d'identification, le nom de l'auteur du document et le lieu de stockage.

3.2.2 Description de l'enquête

Cette section concerne l'enquête ou l'étude proprement dite. Elle contient notamment les renseignements suivants : qui a mené l'enquête; qui a recueilli les données; comment et quand on a recueilli les données; les unités d'analyse; la couverture géographique; un résumé de l'enquête; les mots-clés servant à la recherche. Cette section peut contenir 243 balises, dont 26 sont obligatoires.

3.2.3 Description des fichiers

La troisième section du livre de code conforme à la norme DDI décrit le fichier de données accompagnant l'enquête en question. Les balises comprennent, entre autres, la structure des fichiers, le nombre de variables, le nombre d'observations et le format des fichiers. Cette section peut contenir 34 balises, dont aucune n'est jugée obligatoire.

3.2.4 Description des variables

Cette section du livre de code est l'une des plus précieuses pour l'utilisateur. À chaque variable du fichier de données correspond une balise contenant des renseignements comme le nom et les valeurs de la variable, le libellé de la question associée à la variable ainsi que les directives précédant et suivant la question. À cela s'ajoutent une fréquence ou valeur moyenne et un intervalle pour chaque variable. L'utilisateur peut chercher des renseignements sur les variables et les questions avant d'extraire des données pour les analyser.

Cette section peut contenir 91 balises, dont quatre seulement sont obligatoires. Les balises de cette section étant répétables, on peut s'en servir pour chaque variable d'un ensemble de données.

3.2.5 Description des autres documents relatifs à l'enquête

La dernière section du livre de code DDI permet aux concepteurs du livre de code de lier la documentation créée par les auteurs de l'enquête aux données. Le guide de l'utilisateur et le livre de code de l'enquête constituent des

exemples typiques de la documentation qui figure dans cette section. Grâce à l'ajout de ces fichiers au livre de code DDI, les utilisateurs ont maintenant accès à un seul fichier contenant à la fois les métadonnées et les données d'une enquête. Cette section peut contenir 49 balises, dont huit sont obligatoires et répétables pour chaque document inscrit dans le livre de code.

3.3 Choix des balises DDI

L'utilisateur dispose de près de 500 balises pour baliser ou créer un livre de code conforme à la norme DDI. Comment détermine-t-on quelles balises il faut inclure ou non? Au congrès de l'IASSIST (Association internationale pour les services et techniques d'information en sciences sociales) tenu en 2000 à Amsterdam, un groupe de personnes chargées de créer des livres de code DDI a proposé que les créateurs commencent par les 30 balises qui étaient alors obligatoires. Ces balises comprenaient des éléments comme le titre, le producteur et la date de création. En nous attaquant à la tâche à l'Université de Guelph, nous nous sommes rendu compte que le nombre de 30 balises était insuffisant. À chaque ensemble de données monté au moyen du système d'extraction Web de l'Université de Guelph correspondait déjà un fichier Lisez-moi. Les métadonnées contenues dans ces fichiers Lisez-moi étaient trop nombreuses pour les 30 balises de base. Nous avons donc commencé par les 30 balises de base et ajouté les balises supplémentaires nécessaires pour inclure les métadonnées des fichiers Lisez-moi.

En poursuivant nos travaux à l'Université de Guelph et en nous attaquant à certaines enquêtes plus compliquées et à d'autres dont le titre a changé au fil des ans (Enquête sur les habitudes de fumer des Canadiens, Enquête sur la population active, fichiers du Recensement), nous avons vu augmenter le nombre de balises que nous jugions essentielles. Nous avons ajouté les balises pour faciliter aux utilisateurs l'interprétation des métadonnées. À ce jour, nous utilisons environ 150 balises et attributs pour documenter les ensembles de données dans la collection de l'Université de Guelph.

4. LA NORME DDI À L'UNIVERSITÉ DE GUELPH

À l'Université de Guelph, nous créons des livres de code conformes à la norme DDI pour chacun des quelque 770 ensembles de données compris dans notre fonds de données. Notre fonds se compose d'enquêtes de l'IDD, d'enquêtes de l'ICPSR et de nombreux ensembles de données de sources internationales, dont le Fonds monétaire international (FMI) et les Nations Unies (ONU).

Le projet TriUniversity Data Resources (TDR) regroupe l'Université de Guelph, l'Université de Waterloo et l'Université Wilfrid Laurier. Le projet TDR offre l'accès aux ensembles de données au moyen d'une page Web centralisée. Au début du semestre de l'hiver 2006 (janvier 2006), tous les ensembles de données que détient l'Université de Guelph étaient dotés de livres de code conformes à la norme DDI et étaient accessibles à nos partenaires du projet TDR qui utilisent le logiciel Nesstar. Les utilisateurs peuvent désormais chercher dans un ou plusieurs ensembles de données l'information nécessaire à leurs projets de recherche. À ce jour, ils apprécient l'interface Web Nesstar et la capacité accrue de consulter la totalité de la documentation d'enquête.

5. LA NORME DDI À L'IDD

Notre objectif consiste à rendre toute la collection de l'IDD conforme à la norme DDI dans les deux langues officielles pour offrir une documentation plus complète et plus cohérente des fichiers et en assurer l'accès et la préservation à long terme.

Tous les fichiers mensuels et annuels de l'Enquête sur la population active, en anglais et en français (de 1976 à 2005, soit 720 fichiers) sont maintenant conformes.

L'IDD a choisi d'acheter le logiciel Nesstar afin d'offrir aux établissements d'enseignement supérieur des fichiers DDI complets.

Le logiciel Nesstar est un logiciel perfectionné de gestion des données, conforme à la norme DDI, qui répond à nos besoins. Il comprend des outils de conversion et d'édition de données et de métadonnées. Il peut même servir à

publier des données et la documentation connexe dans un catalogue sur un serveur Nesstar à partir duquel on peut mettre ces ressources à la disposition de la communauté de l'IDD au moyen du logiciel Nesstar WebView.

Comme la création de métadonnées prend parfois du temps, l'utilisation de la norme DDI par le logiciel Nesstar rend le processus plus simple, plus efficace et plus souple. À l'intérieur du logiciel Nesstar, la structure DDI (ou un sous-ensemble de cette dernière) peut servir à importer de l'information, à intégrer les données et les métadonnées et à personnaliser le processus de production de métadonnées à l'intérieur du cadre de travail global de la norme DDI.

On peut ainsi produire des métadonnées selon les modèles DDI établis, ou les personnaliser selon les besoins des utilisateurs.

Pour créer les métadonnées, nous prenons un fichier portable SPSS auquel nous appliquons le modèle d'ensembles de balises (Université de Guelph et IDD) et nous recueillons la documentation provenant des guides de l'utilisateur, des livres de code, des questionnaires, du catalogue en ligne de Statistique Canada et de la Base de métadonnées intégrée (BMDI). Après avoir terminé un ensemble de données, nous le publions sur le serveur NESSTAR pour le mettre à la disposition de nos utilisateurs.

6. CONCLUSIONS

Au cours des dix dernières années, la documentation d'enquête sous forme d'imprimés dans des reliures, accompagnés de fichiers de données brutes et de clichés d'enregistrement, a fait place au livre de code d'aujourd'hui, conforme à la norme DDI et contenant toute la documentation d'enquête, les données et la documentation connexe. Les chercheurs peuvent maintenant utiliser des logiciels comme Nesstar en guise de guichet unique pour chercher des données d'enquête et l'information qui les accompagne.

Plus nous créons de livres de code DDI, plus nous nous éloignons du contexte du « monde réel » pour atteindre lentement celui du « monde parfait ». Plus les créateurs de données seront conscients de l'avantage de créer un seul fichier contenant toute la documentation d'enquête, plus les fournisseurs de données seront en mesure d'offrir à leurs utilisateurs des livres de code DDI et des données de manière efficace et rapide.