

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

DOCUMENTATION SUR LES ÉLÉMENTS DE DONNÉES DANS LES ORGANISMES STATISTIQUES

Paul Johanis¹

RÉSUMÉ

Les renseignements explicatifs qui accompagnent les données statistiques sont appelés métadonnées, et leur présence est essentielle à la bonne compréhension et à la bonne interprétation des données. Au fil des ans, les travaux de recherche se sont poursuivis afin d'établir un moyen normalisé de structurer et de présenter ce genre de renseignements. Le présent article décrit l'expérience de Statistique Canada en ce qui concerne la conceptualisation, la dénomination et l'organisation des variables au sujet desquelles sont produites des données.

1. INTRODUCTION

La principale fonction des organismes statistiques nationaux est de produire et de diffuser des données statistiques sur la situation économique et sociale de leur pays. Les données statistiques prennent la forme de chiffres de divers types contenus dans des fichiers de données, des tableaux statistiques ou des textes comme les communiqués de presse et les articles. Sans ces éléments, ces chiffres ne peuvent pas être compris. Les renseignements explicatifs qui les accompagnent sont appelés métadonnées et leur présence est essentielle à la compréhension et à la bonne interprétation des données statistiques.

Au niveau le plus fondamental, ces renseignements explicatifs doivent couvrir au moins la description des données. Une norme utile, et utilisée par Statistique Canada, pour structurer et présenter ce genre de métadonnées est l'ISO/CEI 11179, *Technologies de l'information — Spécifications et normalisation des éléments de données*². En terminologie statistique, les éléments de données sont souvent dénommés variables. Par conséquent, cette norme fournit des lignes directrices pour la structuration et la présentation de l'information descriptive élémentaire au sujet des variables. Le processus même de création d'information descriptive conformément à cette norme a toutefois également l'effet de rendre plus cohérentes et rigoureuses la conceptualisation, la dénomination et l'organisation des variables pour lesquelles les données sont produites. Le présent article fait part de l'expérience de Statistique Canada à cet égard.

2. LA NORME

ISO/CEI 11179 est une norme pour la création et la tenue des registres d'éléments de données. Elle est présentée en six parties. La partie 3 fournit le cadre pertinent pour la documentation sur les éléments de données. Statistique Canada a en grande partie appliqué la version 2003 de cette partie de la norme (deuxième édition) dans son registre de métadonnées.

Les aspects fondamentaux de cette partie de la norme sont illustrés dans le graphique qui suit. Les éléments de données sont des expressions des concepts d'éléments de données et sont représentés au moyen de domaines de valeurs. Grossièrement, en termes statistiques, cela revient à dire que les variables sont des expressions de leur concept sous-jacent et qu'elles sont représentées au moyen de classifications.

¹ Paul Johanis, Statistique Canada, Canada, K1A 0T6

² ISO/CEI 1994, 1995, 1997, 2000, 2001, 2002. NOTA : La norme ISO/CEI 11179 est révisée à l'heure actuelle sous le titre général *Technologies de l'information — Registres de métadonnées (RM)*.



Figure 3 – Métamodèle de haut niveau

Chaque composante de ce métamodèle de haut niveau est à son tour décomposée davantage dans la norme. Le concept d'élément de données est composé d'une classe d'objets, qui, en jargon statistique, fait référence à l'unité d'observation, et d'une propriété. L'élément de données est composé d'un concept d'élément de données et d'une classe de représentation. Les domaines de valeurs peuvent être non énumérés, c'est-à-dire qu'ils peuvent prendre toute valeur dans un intervalle continu, ou énumérés, auquel cas ils peuvent prendre un ensemble de valeurs permises et de significations de valeur. Nous présenterons ces décompositions en détail dans le contexte de la mise en application de la norme à Statistique Canada.

3. APPLICATION AU REGISTRE DE MÉTADONNÉES DE STATISTIQUE CANADA

Statistique Canada a créé un registre central de métadonnées, qui est appelé à l'interne Base de métadonnées intégrée (BMDI). Celle-ci contient de l'information sur les plus de 400 programmes statistiques actifs du Bureau. Le genre d'information fournie englobe les sources et les méthodes utilisées pour produire les données publiées d'après ces programmes, les indicateurs de la qualité des données, ainsi que les noms et les définitions des variables, et les classifications connexes, publiés par la voie de la base de données en ligne du Bureau, dénommée CANSIM.³

Pour chacune de ces variables, nous avons créé des métadonnées entièrement spécifiées conformément à la norme. Par conséquent, dans chaque cas, la classe d'objets, la propriété et la classe de représentation ont été spécifiées, nommées et définies, de même que leurs domaines de valeurs. Ces composantes sont indépendantes les unes des autres et, par conséquent, réutilisables dans la construction d'autres éléments de données. La stratégie générale avait donc pour objectif d'être parcimonieux dans la création de ces parties constitutives et d'utiliser des combinaisons et

³ L'ensemble complet de métadonnées au niveau des variables ne peut être consulté qu'à l'interne à Statistique Canada pour le moment. À partir du 6 juillet 2005, les métadonnées au niveau des variables révisées et validées seront progressivement ajoutées sur le site Web de Statistique Canada consultable de l'extérieur pour toutes les variables publiées dans CANSIM.

des permutations de ces composantes élémentaires pour représenter la diversité des variables pour lesquelles les données sont publiées par le Bureau.

Le processus suivi pour créer ces métadonnées a débuté par l'analyse de chaque tableau de données statistiques présenté aux utilisateurs de CANSIM sur le site Web de Statistique Canada. Il existe plus de 1 500 tableaux de ce type. Pour chaque tableau, nous avons commencé par identifier l'unité d'observation pour les variables présentées, ce qui nous a menés à la spécification, à la dénomination et à la définition des classes d'objets. Ceci fait, l'étape suivante consistait à cerner les propriétés de l'unité qui étaient mesurées, puis de les nommer et de les définir. Enfin, nous avons identifié le type de représentation (un dénombrement, un indice, une valeur, un type, etc.). Ces renseignements de base nous ont fourni les éléments nécessaires pour nommer et définir les éléments de données, sous la forme *Classe de représentation* de *Propriété* de *Classe d'objets*. Ce processus donne des noms de variables telles que Type d'industrie d'établissement, Catégorie d'âge de personne, ou Indice des prix à la production du produit.

Pour les variables représentées par des valeurs dans une classification, les domaines de valeurs appropriés ont ensuite été déterminés, nommés et définis. Dans la norme, les domaines de valeurs équivalent à ce que l'on considérerait comme étant l'ensemble des classes dans un niveau donné d'une classification statistique. Ces domaines sont également indépendants et réutilisables dans le contexte de nombreux éléments de données.

Toutes les composantes de base et les liens qui les unissent dans le contexte de tableaux statistiques particuliers ont été mémorisés dans la BMDI, à partir de laquelle il est donc possible de produire, de façon dynamique et sur demande, la définition complète de chaque variable, conformément aux spécifications de la norme. En utilisant la variable Type de dépenses de l'emplacement commercial comme exemple, le résultat final est présenté dans le format suivant :

Type de dépenses de l'emplacement commercial

« Dépenses » s'entend de la diminution des bénéfices économiques ou du potentiel de service durant la période de déclaration, sous la forme de sorties ou de consommation d'éléments d'actif ou de création d'éléments de passif qui donnent lieu à une diminution des capitaux propres, autre que celle reliée aux distributions aux propriétaires.

« Emplacement commercial » s'entend d'une unité statistique définie comme étant une unité productrice à un emplacement géographique unique, pour laquelle a lieu une activité économique et pour laquelle il existe, au moins, des données sur l'emploi.

« Type » s'entend de la déclaration des « Dépenses de l'emplacement commercial » en utilisant la ou les classifications suivantes :

- [Catégories de dépenses, Enquête annuelle des manufactures \(EAM\)](#)

À mesure qu'a progressé l'analyse, de nombreuses situations se sont présentées où il a fallu établir et appliquer des conventions ou des approches cohérentes pour faire face à la multiplicité des moyens utilisés par les producteurs de données pour présenter les données statistiques. Nous décrivons dans le présent article les choix pratiques qui ont été faits en appliquant la norme, y compris les conventions adoptées pour traiter certaines des situations rencontrées le plus fréquemment.

4. UNITÉ STATISTIQUE – CLASSE D’OBJETS

Dans la Politique concernant les normes de Statistique Canada, la définition de l’unité statistique est la suivante : « l’unité d’observation ou de mesure pour laquelle on recueille ou calcule des données »⁴. Si l’on se réfère à la norme ISO, cette définition correspond à un type statistiquement pertinent de classe d’objets, définie dans la norme comme étant : « un ensemble de notions, d’abstractions ou de choses du monde réel qu’on peut cerner et définir explicitement et dont les propriétés et le comportement suivent les mêmes règles »⁵. En appliquant la norme, il a été décidé d’essayer de limiter le nombre de classes d’objets identifiées en utilisant dans la mesure du possible les unités statistiques fondamentales. Ces dernières sont, par définition, celles qui ne sont pas des types de toute autre unité statistique et ne peuvent être dérivées par groupement de toute autre unité statistique⁶. Nous avons identifié les types suivants d’unités statistiques fondamentales :

- Agents : entités qui agissent et dont les actions sont déclarées à Statistique Canada. En statistique sociale, une « personne » peut être considérée comme un agent.
- Événements : actions des agents telles que déclarées à Statistique Canada. Les événements sont discrets dans le temps (surviennent durant une période donnée) et finis (peuvent être dénombrés). En statistique sociale, une « naissance » peut être considérée comme un événement.
- Articles : choses qui sont généralement produites ou gérées par les agents. En statistique économique, le « produit » peut être considéré comme un article.

Les unités statistiques fondamentales ont été identifiées de façon à réduire au minimum le nombre de classes d’objets. Cependant, certaines unités statistiques qui peuvent être dérivées d’une certaine façon à partir d’unités statistiques fondamentales sont utilisées si couramment que nous les avons identifiées en tant que classes d’objets distinctes. Les unités dérivées courantes des unités statistiques fondamentales comprennent les ensembles suivants :

- Sous-ensembles d’unités statistiques fondamentales basés sur une caractéristique inhérente. Mentionnons, par exemple, « Personnes de 15 ans et plus » utilisé comme une classe d’objets. Il s’agit d’un sous-ensemble de la classe d’objets « Personne » basé sur la propriété d’âge. De cette façon, il est possible de définir des éléments de données tels que « Type de profession des personnes de 15 ans et plus ».
- Sous-ensembles fondés sur les rôles que peuvent assumer les unités statistiques. Par exemple, « Étudiant », « Mère » et « Employé » sont des sous-ensembles de la classe d’objets « Personne » basés sur diverses propriétés de rôle. De cette façon, il est possible de définir des éléments de données tels que Catégorie de domaine principal d’études de l’étudiant. Les sous-ensembles fondés sur les rôles diffèrent de ceux fondés sur les caractéristiques inhérentes en ce sens qu’une même unité statistique peut assumer plus d’un rôle à la fois. Elle peut aussi assumer ou cesser de remplir un rôle au cours du temps.
- Super-ensembles d’unités statistiques fondamentales. Par exemple, « Famille » est un groupe de personnes conformément à certaines règles de groupement.

En prenant pour point de départ les unités statistiques fondamentales et en n’identifiant que les unités statistiques supplémentaires définies conformément à certaines propriétés lorsque ces dernières étaient observées fréquemment, nous avons abouti à la détermination d’un ensemble relativement limité d’environ 80 classes d’objets.

Aux termes de la norme ISO, presque n’importe quoi peut être une classe d’objets. En pratique, il convient de faire certains choix. L’un des tests fréquemment utilisés pour essayer d’isoler la classe d’objets qui nous intéressait consistait à nous poser la question « Que dénombre-t-on ici? ». Par exemple, nous étions souvent tentés de considérer « Industrie » comme une classe d’objets. Cependant, les organismes statistiques ne publient pas de statistiques significatives sur le nombre d’industries (par exemple, en 2004, le Canada comptait 212 industries actives). En général, nous publions plutôt le nombre et la taille des entreprises, classées par industrie. Par conséquent, nous avons toujours considéré « Industrie » comme une propriété, habituellement attribuée à une classe d’objets, telle que les établissements ou les emplacements commerciaux. Nous avons également choisi de considérer « Économie » comme une classe d’objets, qui a été utilisée fréquemment pour les variables figurant dans les

⁴ Statistique Canada, Politique concernant les normes, http://www.statcan.ca/francais/about/policy/standards_f.htm

⁵

⁶ Voir Mechanda, K., Johanis, P., et Webber M. (2003) *Conceptual Model for the Definitional Metadata of a Statistical Agency*, présenté à l’Open Forum 2003 on Metadata Registries, à Santa Fe, au Nouveau-Mexique.

comptes nationaux. Ici, nous n'avons pas appliqué le test « Que dénombre-t-on ici? ». Les organismes statistiques nationaux ne publient pas de statistiques sur le nombre d'économies. Cependant, si l'on considère l'ensemble des pays, il existe ce que l'on peut considérer comme une population d'économies, ayant chacune ses propres caractéristiques de taille, composition et comportement. Par conséquent, nous avons défini de nombreux éléments de données en choisissant Économie comme classe d'objets, telle que la Valeur du PIB de l'Économie.

L'ISO/CEI 11179 est une norme souple. Comme nous l'avons mentionné plus haut, pratiquement n'importe quoi peut être considéré comme une classe d'objets par ajout d'un nombre de plus en plus grand de propriétés à une classe d'objets fondamentale. Si l'on considère l'espace définitionnel disponible pour inclure les composantes Classe d'objets – Propriété – Domaine de valeurs, il est possible aux termes de la norme d'emballer toute la signification d'un élément donné dans la classe d'objets et de réduire le nombre de propriétés à une (Occurrence de), avec une valeur de domaine égale à Oui ou Non. Il s'agit en fait de la façon dont les variables catégoriques sont représentées sous forme de « variables muettes » dans les processus et logiciels d'analyse statistique. Par exemple, la classe d'objets « Personne occupée francophone » peut être définie à l'aide de la propriété « Occurrence » et d'une valeur de domaine Oui ou Non. Cependant, on aboutit ainsi à un très grand nombre d'éléments de données (dans ce cas, un élément de données pour chaque langue d'intérêt), ce qui rend leur gestion plus difficile et la tâche de trouver les données, plus coûteuse pour les utilisateurs. Nous avons donc décidé de considérer plutôt la personne employée comme la classe d'objets, la langue parlée comme la propriété et les catégories de langue comme le domaine de valeurs. Nous obtenons ainsi l'élément de données Catégorie de langue parlée de la personne employée. En choisissant de limiter le nombre de classes d'objets, nous emmagasinons plus de signification dans la propriété et l'espace définitionnel du domaine de valeurs. Nous estimons que cette approche équilibrée améliore la gestion des éléments de données et la facilité de recherche, et favorise la réutilisation des composantes définitionnelles élémentaires, et donc soutient l'harmonisation des données.

5. PROPRIÉTÉS

Les propriétés sont simplement les caractéristiques d'intérêt de l'unité d'observation (classe d'objets). Une fois que la classe d'objets est définie, il est relativement simple de déterminer quelles sont les caractéristiques mesurées. Il existe toutefois quelques conditions spéciales pour lesquelles des règles uniformes ont été élaborées en pratique. Ces règles couvrent les propriétés particulières de temps, de lieu et d'occurrence, et l'utilisation de propriétés combinées ou composées. Nous allons les examiner l'une après l'autre.

Chaque observation d'une unité statistique se rapporte à l'état de cette unité à un point particulier dans le temps ou une période particulière, que l'on appelle souvent date de référence ou période de référence de l'observation. On pourrait considérer cette référence temporelle comme une propriété ou une caractéristique de l'unité statistique et, par conséquent, l'utiliser dans la construction de la variable. Bien qu'il soit sans aucun doute indispensable à la bonne interprétation des données, dans notre application de la norme, nous n'avons pas considéré le temps comme une propriété de la classe d'objets. Plutôt, nous le considérons comme une propriété de l'observation proprement dite, liée au processus d'observation plutôt qu'à la signification intrinsèque de l'élément de données. Par conséquent, l'information sur la période de référence est mémorisée dans la Région d'administration et d'identification du métamodèle, plutôt que dans celle du concept d'élément de données.⁷ En plus de la date de création, de la date d'entrée en vigueur et de la date du dernier changement, qui sont proposées dans la norme pour les dates ou périodes de référence, nous avons créé un type de cadre temporel appelé « Période de référence », qui contient l'information temporelle pertinente pour une observation donnée ou un ensemble d'observations. Habituellement, les observations sont faites dans le contexte d'une exécution donnée d'une enquête (par exemple, l'exécution de juillet 2005 de l'Enquête mensuelle sur la population active) et la même période de référence est attribuée à toutes les observations faisant partie de cet ensemble.

Une autre propriété qui s'applique à la plupart, voire toutes les observations, est une sorte de référence géographique. Cependant, contrairement au temps, cette propriété ne découle pas du processus d'observation proprement dit. Il s'agit réellement d'une propriété de la classe d'objets. Par conséquent, inous o é it i p

l'emplacement géographique de la classe d'objets pour traiter la variable « géographie ». Cette représentation Nom peut alors utiliser divers domaines de valeurs de noms géographiques, comme [Provinces et territoires du Canada de la Classification géographique type de 2001](#).

Nous avons également adopté une convention normalisée pour une troisième propriété particulière, que nous appelons la propriété d'« occurrence ». Elle est utilisée lorsque l'élément de données pour lequel on produit la documentation est le résultat d'une opération statistique, comme une compilation. Un processus de compilation produit un nouvel élément de données, un dénombrement de l'incidence de l'unité d'observation pour une valeur donnée dans un domaine de valeurs. Cet élément de données prend la forme Dénombrement de l'Occurrence de l'Objet, où Dénombrement est la classe de représentation, Occurrence est la propriété et Emplacement commercial est la classe d'objets. Par conséquent, cette propriété d'occurrence a été définie uniformément dans tous les cas où des dénombrements d'incidence ont été déclarés.

Cet exemple illustre également une autre convention adoptée lors de l'identification des éléments de données. Dans notre application de la norme, lorsque les données statistiques sont présentées sous forme d'un tableau, les éléments de données ne sont pas identifiés et définis pour chaque cellule de ce tableau. La conception et l'étiquetage du tableau communiquent de l'information qu'il n'est pas nécessaire de stocker dans la définition de l'élément de données. Seules les dimensions du tableau doivent être définies en tant qu'éléments de données et documentées.

À l'heure actuelle, nous avons chargé 220 propriétés dans la base de données opérationnelle, et 280 autres attendent, dans notre zone de transfert, d'être révisées et validées. Il semble que la spécification, la dénomination et la définition d'environ 500 propriétés suffiront à couvrir toutes les données statistiques publiées dans CANSIM.

6. TYPE DE REPRÉSENTATION

Les métadonnées concernant un élément de données ne sont pas complètes sans la description de la représentation de ce dernier. En effet, il s'agit de la forme que prendra l'élément de données dans un fichier de données ou dans un tableau statistique. Dans notre application, nous commençons par identifier un type de représentation. Celui-ci décrit la forme particulière de la représentation et est utilisé dans le nom de l'élément de données. Des termes tels que dénombrement, valeur et nombre sont utilisés comme types de représentation dans le cas d'éléments de données à domaine de valeurs non énuméré, appelés en termes statistiques variables continues. La Valeur des Dépenses de l'Emplacement commercial est un exemple de ce genre d'élément de données, où le type de représentation utilisé est « Valeur ». Dans le cas d'éléments de données à domaine de valeurs énuméré, nous avons utilisé des types de représentation tels que le nom, le type ou la catégorie, comme dans l'élément de données Catégorie d'âge de la Personne. En dernière analyse, nous constatons qu'un assez petit ensemble d'environ 25 types de représentation suffit à couvrir tous les produits statistiques du Bureau.

7. ÉLÉMENTS DE DONNÉES

Une fois que ces composantes de base sont définies, il est possible, grâce à la combinaison et à la permutation de Classe d'objets par Propriété par Type de représentation, de spécifier, de dénommer et de définir tous les éléments de données produits par un bureau national de la statistique. C'est à ce stade que l'application uniforme de la norme ISO/CEI 11179 pourrait accroître considérablement l'harmonisation entre les divers organismes statistiques nationaux. La liste de classes d'objets et de types de représentation présentée ici peut presque certainement être appliquée par tous les organismes statistiques nationaux. Par conséquent, les efforts d'harmonisation se concentreraient sur l'adoption de noms et de définitions communs des propriétés, ce qui est un objectif réalisable. Cela permettrait au moins aux utilisateurs des données de repérer les éléments de données communs aux divers dépôts de données des organismes statistiques nationaux et d'être certains que les définitions sous-jacentes sont les mêmes. Par contre, il existe des différences importantes et, dans de nombreux cas, justifiées, entre les domaines de valeurs, ou classifications, utilisés pour représenter ces éléments de données.

8. DOMAINES DE VALEURS

Selon le type de représentation, les éléments de données ont des domaines de valeurs qui sont énumérés ou non énumérés. Les domaines de valeurs non énumérés sont utilisés pour représenter les variables continues, c'est-à-dire des variables qui peuvent prendre toute valeur numérique dans un intervalle. Par exemple, le domaine de valeurs non énuméré pour l'élément de données Valeur de revenu de personne pourrait être l'ensemble de nombres entiers compris entre zéro et l'infini. La seule information nécessaire pour comprendre et interpréter une telle valeur est l'unité de mesure (c.-à-d. dollars canadiens) et la précision (par exemple, deux décimales). En revanche, plus d'information est nécessaire pour interpréter la valeur prise d'après un domaine de valeurs énuméré. Un tel domaine est un ensemble de catégories représentées par des codes ou des étiquettes, ou les deux, ayant une signification non liée à la valeur réelle. Par exemple, le nombre 325410 est un code du SCIAN signifiant Fabrication de produits pharmaceutiques et de médicaments, ce qu'il est impossible de savoir sans se reporter aux métadonnées. Par conséquent, dans la norme, les domaines de valeurs énumérés sont constitués de paires de valeurs permises (codes) et de significations de valeur (étiquettes), qui peuvent également avoir une définition.

Les domaines de valeurs peuvent également être considérés comme des composantes de base indépendantes pouvant être associées à des éléments de données appropriés, selon le besoin. En fait, de gérer, enregistrer et tenir à jour ces domaines de valeurs est une tâche ordinaire des organismes statistiques nationaux, où ces domaines prennent habituellement la forme de classifications statistiques. Dans notre application de la norme, nous avons recréé des classifications statistiques à partir de domaines de valeurs spécifiés conformément à la norme. Comme celle-ci ne prévoit pas d'entité « Classification », nous avons élaboré un certain nombre de conventions à cette fin⁸.

En premier lieu, nous avons attribué à chaque domaine de valeurs un domaine de valeur suprême, ne contenant qu'une seule valeur permise et une seule signification de valeur, qui est la valeur mère de toutes les valeurs permises subordonnées. Ce domaine de valeur supérieure est une marque de réservation, c'est-à-dire un dispositif organisationnel conçu pour être le conteneur de la classification d'intérêt. On lui donne le nom de la classification qu'il est destiné à représenter (par exemple, SCIAN Canada 2002). Sous ce domaine de valeur supérieure sont identifiés hiérarchiquement un ou plusieurs domaines de valeurs. Chacun est un ensemble de valeurs permises, avec les significations de valeur connexes, qui sont mutuellement exclusives et exhaustives de l'univers des observations qui doivent être classées. À chaque domaine de valeurs est attribué un niveau dans la hiérarchie. À chaque valeur permise est attribuée une valeur permise provenant d'un domaine de valeurs de niveau plus élevé et son ordre parmi les valeurs filles est enregistré. Chaque valeur permise peut être la fille d'une, et uniquement une, valeur permise mère et est donc exclusive en agrégation. Ces conventions permettent de reconstruire la structure complète de n'importe quelle classification.

Les classifications originales, qui pourraient être des classifications types (et enregistrées en tant que telles dans le champ d'état d'enregistrement — voir Région d'administration et d'identification de la norme) et leurs variantes sont traitées systématiquement de cette façon dans la BDMI. La classification originale est considérée comme un domaine de valeurs « général » et est marquée en tant que tel dans le registre de métadonnées. De cette façon, il est facile de repérer les cibles éventuelles d'une future harmonisation.

Certains domaines de valeurs possèdent un grand nombre de variantes. Ainsi, à l'heure actuelle, Statistique Canada publie des données conformément à 13 variantes de la Classification des industries de l'Amérique du Nord. Certains domaines de valeurs changent également au cours du temps, mais on parle alors de versions plutôt que de variantes. Par exemple, le SCIAN Canada 1997 était la classification type pour le type d'industrie. Elle comportait de nombreuses variantes sous le même domaine de valeurs général. Lorsque l'original a été remplacé par le SCIAN Canada 2002, ce dernier a été considéré comme une nouvelle version du même domaine de valeurs. De même, toute variante du SCIAN Canada 1997 qui a été mise à jour aux termes de la version 2002 a été considérée comme une nouvelle version de la variante en question. Toujours en gardant le même exemple, il existe également d'autres classifications utilisées pour le type d'industrie, dont la Classification internationale type par industrie (CITI) et la

⁸ La partie 2 de l'ISO/CEI 11179 traite de la classification, mais ce traitement se rapporte à la classification des éléments de données et de leurs composantes de base dans un registre de métadonnées en vue de faciliter l'organisation et la recherche, sujet qui sera abordé à la section 9 du présent article, et non de la classification des observations dans un domaine de valeurs énuméré, la question qui nous intéresse ici.

Nomenclature statistique des activités économiques dans les communautés européennes (NACE). Ces classifications sont tout apparentées, ce que l'on représente dans le modèle en veillant à ce que les domaines de valeurs constituant toutes ces classifications apparentées utilisent des significations de valeur extraites d'un ensemble commun de significations de valeur, qui est le domaine conceptuel des industries. Par conséquent, dans notre application de la norme, les domaines conceptuels sont les conteneurs des significations de valeur qui sont réutilisés dans les domaines de valeurs qui représentent des éléments de données construits d'après une même classe de représentation et propriété (par exemple, type d'industrie).

À l'heure actuelle, nous avons chargé dans la BDMI 1 145 domaines de valeurs qui sont utilisés individuellement ou conjugués à d'autres pour former 550 classifications.

9. ORGANISATION DES ÉLÉMENTS DE DONNÉES

L'application de la norme ISO/CEI 11179 pour produire la documentation sur les variables publiées par Statistique Canada a produit environ 900 éléments de données nommés et définis. Afin d'aider les utilisateurs à parcourir une liste contenant autant d'éléments, il est souhaitable de fournir un certain cadre d'organisation. Les éléments de données sont organisés d'après leurs parties constituantes, par exemple, tous les éléments de données utilisant la même classe d'objets ou la même propriété ou ceux qui sont représentés en utilisant le même domaine de valeurs. Ce genre de présentation pourrait être pratique pour les utilisateurs. Cependant, la norme contient également des dispositions en vue d'associer des termes de classification à chaque article administré dans un registre, comme un élément de données. À Statistique Canada, nous avons dressé une liste de 27 sujets de haut niveau et en avons associé un à chaque élément de données.

10. CONCLUSION

L'utilisation de la norme ISO/CEI 11179 pour décrire les éléments de données à Statistique Canada a fourni un cadre rigoureux permettant d'identifier, de nommer et de définir une grande quantité d'information statistique au moyen de termes et de concepts de métadonnées établis avec une relative parcimonie. La puissance de la norme tient à la façon dont les concepts de données complexes sont ventilés en composantes qui peuvent être rassemblées et réutilisées pour produire des métadonnées définitionnelles et contextuelles complètes. Cette granularité offre aussi un moyen puissant de normalisation et d'harmonisation, car il est plus facile d'arriver à un accord sur des composantes spécifiquement délimitées, comme des classes d'objets et des classes de représentations, que sur des concepts de données complexes mal pensés. Dans l'ensemble, l'application de la norme a joué un rôle-clé en aidant le Bureau à atteindre un objectif visé depuis longtemps, à savoir fournir aux utilisateurs des données un inventaire des éléments de données qu'il produit et des métadonnées connexes.