Complexité liée à la variabilité sémantique des statistiques socio-économiques

Christine Plumejeaud*, Jérôme Gensel*

*Laboratoire d'Informatique de Grenoble, équipe STEAMER,
681 rue de la Passerelle, 38400 St Martin d'Hères
{prenom.nom}@imag.fr,
http://steamer.imag.fr/

Résumé. Cette recherche concerne le développement de méthodes pour l'analyse de similarité entre valeurs statistiques, issues de sources multiples, sur des échelles géographiques et des périodes temporelles variables. Les valeurs des indicateurs statistiques, comme les effectifs des différentes catégories socio-professionelles, mesurés en un certain lieu à une certaine période, représentent des réalités complexes et difficilement comparables. Les éléments textuels (mots clés, thèmes et résumés) présents dans les métadonnées (profil ISO19115 ou SDMX) de l'indicateur définissent une part de la signification de chaque valeur. Cependant, les catégories qu'une valeur représente sont fortement hétérogènes dans l'espace et le temps, et la valeur fait l'objet de processus d'estimation qui rendent difficile l'analyse de similarité d'une valeur avec une autre. Ce travail expose en détail cette problématique, en expliquant pourquoi les métadonnées sont nécessaires mais encore insuffisantes ou sous-exploitées, et explore les solutions existantes pour faciliter la comparaison de valeurs.

1 Introduction

L'information statistique disponible aujourd'hui sur tous les espaces géographiques est un potentiel de richesse encore largement inexploité du fait de l'hétérogénéité et de l'évolutivité des sources d'information. En Europe, particulièrement, la quantité d'indicateurs socio-économiques produits, comme le chômage, le Produit Intérieur Brut (PIB), mais également la pyramide des âges, n'a cessé d'augmenter depuis 1950 jusqu'à nos jours, pour couvrir les différentes échelles géographiques (des communes aux États) et proposer une vue plus exhaustive de l'espace européen. Par ailleurs, les méthodes de production de ces indicateurs se sont rationalisées et perfectionnées : les collectes de données, issues de comptage sur des zonages territoriaux, se sont régularisées, et la publication de métadonnées accompagnant ces données est devenue systématique, suivant en cela les recommandations d'INSPIRE, directive publiée par le Parlement européen (2007), mais aussi de différents travaux de recherche, tels ceux de Shoshani (1982); Dean et Sundgren (1996), relayés par les instances internationales sous forme de guides méthodologiques à l'intention des producteurs de données, UN/ECE (1995).

Dans ces conditions, il serait envisageable d'exploiter cette richesse pour, par exemple, estimer des valeurs manquantes pour un certain indicateur à une certaine échelle et période en

utilisant la valeur d'autres indicateurs, identiques ou équivalents, connus à d'autres échelles, d'autres périodes. Il apparaît cependant qu'un même indicateur peut avoir une sémantique différente suivant le producteur de données, et les méthodes de mesure et réajustement employées. Par exemple, les données du chômage produites par l'Institut National des Statistiques et des Études Économiques (INSEE) en France diffèrent par leur valeur de celles produites par Eurostat, alors que les deux instituts déclarent s'appuyer sur la même définition du chômage, celle du Bureau International du Travail (BIT). Ce problème de cohérence des données statistiques dans le domaine social, économique ou agricole est connu depuis longtemps des statisticiens qui manipulent ces données, qui, comme Wilks (1939), soulignent le manque d'homogénéité des échantillons, la non-comparabilité et l'inexactitude de ces données. La question est de savoir si les métadonnées *dans leur format actuel* peuvent expliquer ces divergences, mais également être utilisées pour rendre comparables des données qui ne le sont pas, et réaliser ainsi une véritable *intégration statistique*, selon Colledge (1998).

Les métadonnés sont définies comme « des données sur les données ». Selon l'ONU, UN/ECE (1995), qui complète cette définition dans le cas des données statistiques, les métadonnées doivent répondre à deux besoins. Il s'agit, d'une part de définir le contenu des données (en fournissant des définitions, mais également en décrivant le processus de production des données), et, d'autre part, d'expliquer pour quel usage les données ont été produites. Il apparaît que les métadonnées sont nécessaires, mais aujourd'hui insuffisantes pour réaliser complètement l'intégration statistique. Parmi les standards utilisés pour décrire des données statistiques (le Dublin-Core, la norme ISO 19115, le standard SDMX), aucun ne semble capable de restituer, dans un formalisme exploitable par des automates, les processus de transformation agissant sur les données. En effet, si ces informations sont généralement décrites dans des rubriques structurées, le cœur de l'information est encore décrit par du texte écrit en langage naturel. C'est un problème qui a déjà été décrit par Comber et al. (2005) pour l'analyse de données d'occupation du sol. Aujourd'hui, un nombre croissant de travaux s'orientent vers l'usage d'ontologies, tels ceux de Pattuelli et al. (2003); Comber et al. (2010), pour l'extraction semi-automatisée de métadonnées des textes, mais également pour l'alignement sémantique de données issues de sources hétérogènes.

Dans cet article, nous exposons plus en détail le problème que pose la recherche d'équivalence entre valeurs statistiques, qui s'inscrit dans la problématique plus générale de l'intégration statistique. La première section décrit quelles sont les spécificités de l'information statistique, et les causes de la variabilité sémantique. La seconde section formule des critiques à l'endroit des standards de métadonnées actuels vis à vis du problème exposé. La troisième section expose les nouvelles pistes qui sont explorées pour remédier à ce problème. La dernière section délivre nos conclusions et les perspectives qui sont entrevues.

2 Les causes de la variabilité sémantique

L'information statistique se présente sous la forme d'une série de nombres ou de variables qualitatives, collectées sur des découpages territoriaux, via des recensements ou des enquêtes datées, et qui donnent, par exemple, la mesure de la population, du nombre de ventes d'une marque de voiture, ou des préférences électorales. Ainsi, chaque valeur est associée au moins à une unité territoriale et une date. La production de ces données est opérée par différents acteurs, qui sont les instituts statistiques nationaux, des agences européennes comme Eurostat

et l'Agence Européenne de l'Environnement, ou encore des groupes de recherche, et il s'en suit une certaine hétérogénéité dans les définitions, classifications, et méthodes de production de ces indicateurs.

2.1 L'aspect multi-dimensionnel de l'information statistique

Un nombre conséquent d'indicateurs statistiques socio-économiques se présentent sous la forme de tableaux de contingence, qui associent des valeurs à des catégories croisées. C'est le cas notamment des données démographiques, qui sont publiées par sexe et par tranche d'âge, ou de la population active telle que les publie l'INSEE sur son site[1] et comme reproduit dans le tableau 1.

Population active (en milliers)	hommes	femmes	ensemble
15 ans ou plus	14 806	13 463	28 269
15-64 ans	14 702	13 394	28 096
15-24 ans	1 487	1 224	2 712
25-49 ans	9 576	8 756	18 332
50-64 ans	3 639	3 413	7 052
dont : 55-64 ans	1 801	1 677	3 478
65 ans ou plus	104	69	173

TAB. 1 – *Population active selon le sexe et l'âge en 2009.*

Ce caractère multi-dimensionnel de l'information statistique est exposé dans Rafanelli et Shoshani (1990), qui présentent l'objet statistique comme étant un quadruplet $< N, C, S, f >$ où :

- N est le nom de l'indicateur statistique ;
- C est un ensemble fini de catégories (ou dimensions) $C_1, C_2, ..., C_n$, qui ont chacune leur unité de mesure, et un domaine spécifique ;
- S est un attribut résumant la variable quantitative mesurée, qui possède un domaine de valeur, et une unité de mesure ;
- f est la fonction d'agrégation utilisée pour résumer les valeurs (la somme, le compte, le minimum, le maximum ou la moyenne).

L'objet statistique défini dans l'exemple précédent porte sur N, la « population active » utilise le compte comme fonction d'agrégation f, et S correspond au décompte de personnes actives, qui sont classées suivant deux dimensions, e.g. par tranche d'âge (C_1), et sexe (C_2).

Cependant, ces catégories sont rarement homogènes sur l'ensemble des données collectées, car elles sont spécifiques au lieu et à la période étudiés. Les catégories socio-professionnelles, les recensements à caractère ethnique ou les tranches d'âge sont des exemples de classifications instables. Les recensements à caractère ethnique font débat, et l'histoire du recensement de la population aux États-Unis (déroulée par Gauthier (2002)) démontre comment les transformations politiques et sociales d'un pays peuvent amener à reconsidérer les classifications

[1] Les données sont disponibles sur http ://www.insee.fr/fr/themes/tableau.asp ?reg_id=0&ref_id= NATCCF03170.

officielles employées. De même, en France, les actifs sont classés en fonction de leur statut professionnel (salarié, chef d'entreprise, indépendant), de la taille de l'entreprise dans laquelle ils travaillent, du secteur de l'activité (primaire, secondaire ou bien tertiaire), du niveau d'études requis pour pratiquer leur profession, etc. Mais ce mode de classification de la population en catégories socio-professionnelles n'a pas d'équivalent européen comme l'explique Kieffer et al. (2002), parce que chaque pays construit ces catégories en fonction de son histoire et de théories spécifiques. Ce problème d'instabilité des catégories, déjà relevé dès 1982 par Shoshani (1982), n'a toujours pas trouvé de réponse, et rend difficile la comparaison de valeurs à travers l'espace et le temps.

2.2 Les processus de transformation à l'oeuvre

L'exemple du chômage montre à quel point ces processus de transformation peuvent faire diverger les résultats produits, même lorsque les recensements se basent sur des définitions communes. Ainsi, en dépit d'une tentative d'harmonisation européenne symbolisée par le partage d'une définition commune définie par l'Organisation Internationale du Travail, l'INSEE et Eurostat publient des chiffres de chômage différents pour la même unité (la France) à la même date : ainsi, le taux de chômage publié par l'INSEE en Février 2008 (8,4 %) diffère de celui estimé par Eurostat (8,8 %). Pour les deux instituts, un chômeur est une personne qui n'a pas eu d'activité rémunérée supérieure à une heure pendant une semaine, et qui peut prouver sa recherche d'emploi. Cependant, les méthodes de calcul, de pondération et de correction des chiffres à partir de l'enquête emploi trimestrielle diffèrent entre l'INSEE, et Eurostat. Une explication détaillée de ces méthodes est publiée dans des documents publics [2], mais le format de ces documents ne permet d'exploiter leur contenu par un système informatique. Par exemple, Eurostat explique qu'il s'appuie sur les chiffres de l'enquête emploi publiée par l'INSEE, mais qu'il les réajuste ensuite :

> Les séries mensuelles sur le chômage et l'emploi sont calculées dans un premier temps au niveau de quatre catégories (hommes et femmes de 15 à 24 ans, hommes et femmes de 25 à 74 ans) pour chaque État membre. Ces séries sont ensuite corrigées des variations saisonnières et tous les agrégats au niveau national et européen sont calculés. [...] Pour la Suède et la Finlande, la tendance-cycle a été utilisée à la place des données corrigées des variations saisonnières jugées trop volatiles. [...]

Il faut noter ici que le texte explique comment sont transformées les données (par l'usage de la tendance-cycle en Suède par exemple), sans pour autant produire la formule exacte permettant de revenir au niveau des données brutes. Le chômage illustre aussi l'évolution des méthodes de calcul et de mesure dont sont l'objet les indicateurs. Goux (2003) explique que l'INSEE fait évoluer régulièrement sa méthodologie de calcul du chômage, décrite et justifiée dans des documents accessibles en ligne [3].

Par ailleurs, les indicateurs statistiques sont parfois le produit de réflexions théoriques qui visent à produire une représentation synthétique d'un ensemble de facteurs mesurés par des données brutes. Une étude récente de l'UMS 2414 RIATE et al. (2008) portant sur le déclin démographique en Europe, et présentée devant le parlement européen en 2008, donne un

[2] INSEE : http ://www.insee.fr/fr/methodes/sources/pdf/eeencontinu.pdf
EUROSTAT : http ://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-29012010-AP/FR/3-29012010-AP-FR.PDF

[3] http ://www.insee.fr/fr/methodes/sources/pdf/estimations_chomageBIT_enquete_emploi.pdf

exemple de ce type d'indicateur. Dans le texte de l'étude se trouve la définition d'un indicateur synthétique de vieillissement :

> Il suffit alors d'effectuer le rapport entre l'âge moyen d'une population et son espérance de vie en bonne santé pour en déduire un *indicateur synthétique de vieillissement* exprimé sous la forme d'un pourcentage du potentiel d'activité de la population qui a été consommé.

A travers cet exemple, il apparaît que la représentation du calcul d'un indicateur composite à partir d'indicateurs de base est possible (ici, c'est le ratio de deux indicateurs de base qui fournit l'indicateur). En effet, si la relation mathématique entre l'indicateur synthétique de vieillissement (noté V) et l'espérance de vie en bonne santé est connue, il devient possible d'estimer des valeurs manquantes de V en employant des indicateurs bruts (espérance de vie, âge moyen). Par ailleurs, la connaissance de la formule de calcul permettrait de déduire que l'indicateur V est un bon *proxy*[4] pour calculer l'espérance de vie d'une population, lorsque son âge moyen est connu. Dans ces conditions, si les transformations sont clairement exprimées, ainsi que les éléments sur lesquels elles opèrent, l'analyse de similarité est quasiment directe.

Cependant, les formules permettant de fabriquer des indicateurs composites sont rarement si simples. Un guide méthodologique résumant les bonnes pratiques est publiée par l' Organisation pour la Coopération et le Développement Economique (2008), pour l'ensemble des opérations qui doivent être mises en oeuvre : sélection, normalisation, estimation de données manquantes, pondération et agrégation des données. Bien qu'effectivement les opérations soient plus complexes, l'étude de ce document fait apparaître l'usage récurrent d'une certaine terminologie pour ces opérations : les mots « correlation », « moyenne », « variance », « hypothèse » reviennent très fréquemment. La description des transformations recourt généralement à un vocabulaire statistique, qu'il s'agit de référencer, afin de construire un dictionnaire des transformations possibles (comme simplement une pondération, une moyenne, une différence). Ce dictionnaire pourrait servir de support à une structuration des métadonnées, dont l'usage est fortement recommandé car, de plus en plus, la production de ce type d'indicateur vise un certain niveau de qualité, non seulement en termes de méthode de construction, mais également en termes de documentation.

3 Usage des métadonnées

Dans cette section, nous examinons les différents standards de métadonnées en profondeur, pour vérifier s'ils répondent complètement à la problématique de l'intégration statistique.

3.1 Les normes de l'information géographique

Les données statistiques étant à références spatiale et temporelle, il semble naturel d'étudier les standards de métadonnées dédiés aux données géographiques. Le premier, le Dublin-Core, créé à l'initiative des États-Unis d'Amériques en 1995, insiste particulièrement sur les aspects légaux (droits de propriété et d'usage) concernant les données. Le standard ISO 19115, promu

[4] *proxy* est un des termes employé en statistiques comme synonyme de variable auxiliaire aidant à retrouver une variable inconnue, qui serait ici l'espérance de vie d'une population.

par la norme INSPIRE pour la diffusion de données géographiques, inclut les 15 éléments proposés dans le Dublin-Core. Les éléments communs au Dublin-Core concernent :

- la description du contenu : le titre, et la date de publication, le sujet et une description des données ;
- la désignation des responsables : l'identification des contributeurs et responsables de cette publication ;
- les informations nécessaires pour l'utilisation des données : les droits de propriété, le format de distribution ;
- les informations relatives à la qualité des données : mode et fréquence de maintenance, couverture spatio-temporelle.

La norme ISO 19115 intègre des éléments spécifiques à l'information géographique permettant de reconnaître leur extension spatiale et le mode de représentation utilisé. Notre intérêt se concentre sur les éléments disponibles pour décrire la *qualité interne* de données spatio-temporelles, regroupés dans l'élément $DQ_Quality$, qui se mesure à l'aune de sept critères (le lignage, l'exactitude de la position spatiale, et temporelle, la précision des attributs thématiques, la complétude, la cohérence logique et sémantique), d'après Servigne et al. (2006). Tous ces critères sont quantifiables, à l'exception du lignage qui doit permettre de retracer les procédures d'acquisition, les sources, et les méthodes employées pour transformer les données brutes, et obtenir par dérivation la donnée décrite. A travers ce critère, deux objectifs sont visés : assurer que les méthodes de production respectent les normes en vigueur, et donc s'assurer que les données sont comparables, mais également rassurer l'utilisateur sur la nature des sources employées. La figure 1 représente la structure de l'élément $DQ_Lineage$ de la norme, qui utilise de façon conjointe un élément $LI_ProcessStep$ pour décrire une étape de transformation, et LI_Source pour décrire les « ingrédients » de la transformation. L'élément $LI_Lineage$ s'applique à tout ou partie du jeu de données : l'emprise spatio-temporelle des éléments concernés par cette transformation est décrite dans l'élément EX_Extent.

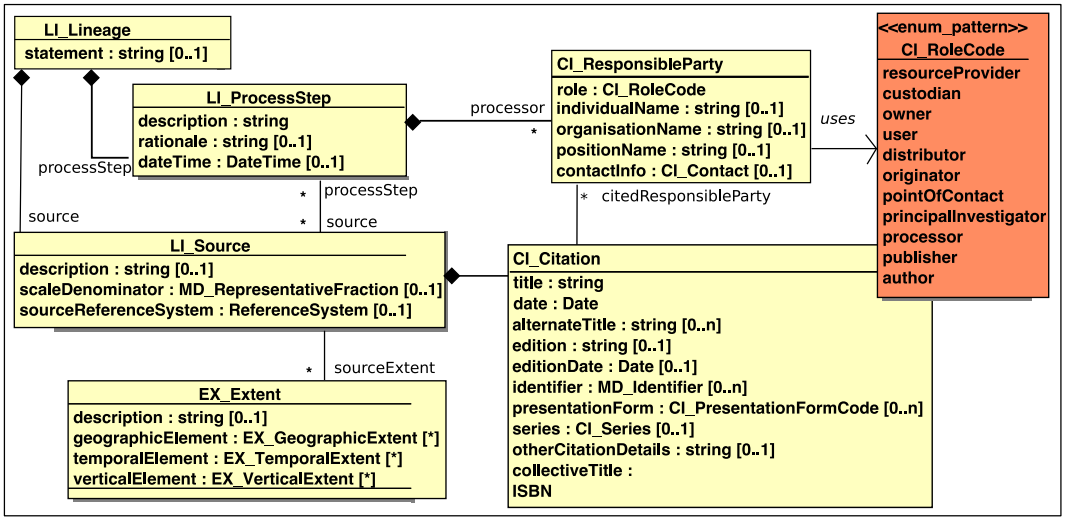

FIG. 1 – *L'élément Lineage défini par la norme ISO 19115.*

Cette norme extensible, a fait l'objet d'une adaptation pour les données statistiques afin de la rendre plus opérationnelle, Plumejeaud et al. (2010). Il s'agissait de simplifier la saisie de métadonnées pour des données statistiques. Par exemple, la référence spatiale correspond à un code d'unité territoriale, et donc le mode de définition des éléments EX_Extent devait être modifié. De même, l'identification de chaque indicateur a été enrichie avec son code et son unité de mesure. Enfin, contrairement à un jeu de données géographiques plus classique, chaque valeur du jeu de données possède un lignage spécifique, puisque les valeurs de certaines unités territoriales du jeu de données sont issues de sources spécifiques. La spécialisation de l'élément $LI_Lineage$ de la norme simplifie la représentation du lignage. En effet, la saisie de chaque processus de transformation et de la source associée est trop fastidieuse pour les utilisateurs, et donc il a été convenu de décrire soit par un champs textuel, soit par l'adjonction d'un document multi-média les processus de transformation. Mais, en réalité, à moins d'extraire les informations depuis les documents, (ce qui suppose de créer un corpus spécialisé, et d'inventer un format pour structurer ces informations), cet élément dérivé de $LI_Lineage$ ne propose que l'URL de la source originelle des données, et ne permet pas de retranscrire les transformations des données. Par ailleurs, dans cette extension, l'aspect multi-dimensionnel de l'information statistique a été négligé, et il reste à créer des éléments de type *MD_Catégorie* décrivant la nature et le domaine des catégories associée aux indicateurs.

3.2 SDMX, un modèle pour l'échange de données statistiques

Les langages semi-structurés ont été avancés depuis quelques années comme une solution au problème de non-interopérabilité entre Systèmes d'Information Statistiques (SIS), Meyer et al. (2004). La raison étant que ces langages (basés sur XML et l'emploi de schémas XSD) permettent d'embarquer une description du format des données dans les données. Ainsi, la flexibilité et la souplesse s'en trouveraient accrue. C'est la raison pour laquelle le Statistical Data Model eXchange (SDMX) est aujourd'hui promu par l'OCDE, mais également Eurostat, ou d'autres organismes de production de données statistiques. Les auteurs de SDMX se sont concertés pour définir une liste de concepts[5], termes[6], et codes[7] identifiés et partagés dans un registre, en vue de faciliter l'intégration statistique. Les concepts, référencés par un identifiant unique, quelque soit la langue, peuvent être valués soit par des codes, indépendants de la langue, soit par du texte. Par exemple, le statut (OBS_STATUS) d'une valeur statistique est décrit par une liste de codes (A, B, E, F, I, M, P, S) qui signale si la valeur est normale (A), manquante (B), estimée (E), etc., et le registre définit exactement la signification de ces codes. L'échange de données statistiques nécessite la définition du format des données dans un fichier externe, le *Data Structure Definition (DSD) file*, qui structure l'information à l'aide de ces codes et ces concepts harmonisés. La liste des concepts et des codes qui leurs sont associés peut également être étendue dans le fichier DSD. Par exemple une catégorie TRANCHE_AGE sera définie comme concept, avec la liste des différentes tranches d'âges codifiées, décrites chacune par un champs texte. Le fichier de données fait référence à cette description au niveau par exemple la valeur observée, lorsque sont présentées les données :

```
<Obs TIME_PERIOD="2009" OBS_VALUE="14 702" OBS_STATUS="A"
TRANCHE_AGE="15-64" SEX="M"/>
```

[5]http ://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf
[6]http ://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf
[7]http ://sdmx.org/wp-content/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf

La prise en charge dans ce standard de l'aspect multi-dimensionnel de l'information statistique est donc meilleure que dans le format ISO 19115. En ce qui concerne la définition des transformations, le concept DOC_METHOD s'utilise pour décrire les méthodes de mesure, la définition et les transformations opérées sur les données, et le concept COMPARABILITY pour fournir un commentaire sur l'équivalence de cette donnée avec une autre. Néanmoins, ces champs sont valués par du texte, et présentent donc la même difficulté d'analyse de leur contenu que celle décrite pour la norme ISO 19115. Ce format ne résout pas non plus le problème d'hétérogénéité des catégories, comme le rapporte l'expérience à grande échelle menée par Oakley et al. (2005), membres du Bureau Australien des Statistiques, producteur officiel en Australie, dans le cadre d'un projet visant à exporter au format SDMX les statistiques économiques nationales. Ils rapportent en particulier le cas de concepts de SDMX qui ne trouvent pas leur pendant dans leur base de données statistiques, parce qu'ABS utilise une classification nationale spécifique, qui, par exemple, codifie les activités de pêche, agriculture et d'exploitation forestière avec la lettre A, alors que SDMX classe séparément les activités de pêche (avec le code AYB) et les activités d'agriculture, chasse et exploitation forestière, codées AYA.

4 Pour aller plus loin que les métadonnées

Les problèmes qui viennent d'être exposés méritent a priori l'intervention d'experts pour leur résolution. En effet, comment réussir à aligner des catégories ou bien réaliser un graphe des transformations des données sans une acquisition manuelle de ces connaissances ? Les deux problèmes peuvent être traités séparément : dans le premier cas, il s'agit de tirer un meilleur parti des textes contenus dans les métadonnées, et dans le second d'étendre les métadonnées avec des formalismes peut-être plus adaptés à la retranscription fine d'un lignage. Nous examinons ici un certain nombre de travaux qui offrent des pistes intéressantes.

4.1 Alignement sémantique des indicateurs

Les travaux de Wadsworth et al. (2006) et de Comber et al. (2005, 2010) étudient plusieurs approches pour l'alignement des différentes catégories d'usage de sol, en vue de comparer des cartes d'occupation du sol produites par différents organismes, à dix ans d'intervalles. Ces catégories, qui sont qualitatives et non-ordonnées présentent le même niveau d'hétérogénéité que les catégories socio-professionnelles. Parmi les différentes approches étudiées pour leur alignement, il ressort qu'une analyse détaillée de la description de la catégorie (lorsqu'elle comporte plus de 100 mots) par une technique de fouille de données textuelle permet d'établir une matrice de recouvrement entre catégorie, que les auteurs démontrent supérieure aux autres techniques qui nécessitent l'intervention d'experts. La technique consiste à établir la liste des mots employées dans la description de chaque catégorie, puis à calculer leur poids sémantique dans chaque catégorie à l'aide d'une mesure de fréquence inverse dans le document (*Inverse Document Frequency*, IDF, telle que discutée par Robertson (2004)). Le niveau de recouvrement entre chaque catégorie est ensuite calculé via l'usage de la théorie de l'analyse sémantique latente probabiliste introduite par Hofmann (1999), qui stipule que la similarité sémantique entre deux concepts (ici les catégories) peut-être mesurée par la quantité d'information qu'ils partagent (les mots). Cette approche permet également d'identifier les concepts qui structurent une classification données, et les termes qui s'y rapportent.

Cette approche pourrait être appliquée à l'ensemble des métadonnées collectées sur les indicateurs statistiques, pour l'analyse de similarité entre indicateurs, mais également pour l'alignement des catégories associées aux valeurs des indicateurs, à la condition que les textes descriptifs produits soient suffisamment longs. Il est ainsi possible d'établir une méta-classification des classifications d'indicateurs ou de catégories, qui peut être structurée sous la forme d'une ontologie de concepts, liées par des relations de subsomption,« is-a »(ou héritage), ou bien des relations méréologiques (ou composition), « Is_part_of », mais également une relation sémantique de domaine indiquant dans quelle mesure une catégorie recouvre l'autre, telle que « intersects », valuée par l'indice de recouvrement. Par exemple, les concepts de chômeur, population active et population peuvent être identifiés via l'analyse des définitions des indicateurs statistiques, qui sont des instances de cette ontologie. Il apparaît que le nombre de chômeurs défini et mesuré par l'INSEE ou l'EUROSTAT se rapporte à un concept de « chômeur » identique, et qu'un chômeur est un membre de la population active, qui elle-même fait partie de la population totale :

- *chômeur_INSEE* is-a *chômeur*
- *chômeur_Eurostat* is-a *chômeur*
- *chômeur* is_part_of *population-active*
- *population-active* is_part_of *population*

L'analyse des métadonnées peut donc servir à établir puis peupler une ontologie. Pattuelli et al. (2003) montrent que l'existence d'une ontologie statistique facilite la compréhension des termes statistiques, et la création d'un glossaire interactif. Cependant, identifier que les chômeurs suivant l'INSEE ou EUROSTAT se rattachent à un même concept "chômeur" ne résoud pas le problème d'équivalence des valeurs que nous avons pointé. L'ontologie sert essentiellement à résoudre et à raisonner sur les problèmes d'équivalences entre catégories. Il s'agit aussi de pouvoir ensuite raisonner sur les valeurs au niveau de leur transformations (modalités de calcul, réajustement, estimation).

4.2 La description des transformations

Pour raisonner au niveau des transformations, il faut se doter de formalismes de représentation. La description des transformations en vue de retrouver des données originelles à partir de données transformées est une ambition affichée par Woodruff et Stonebraker (1997), qui proposent un formalisme sous forme de graphe (direct et acyclique) du flot de données dans une base de données, chaque nœud modélisant une fonction de transformation, et chaque arc correspondant à une donnée particulière. Le graphe est défini par l'utilisateur via une interface graphique. L'utilisateur spécifie chaque fonction f de transformation par son nom et son type, ainsi que les paramètres d'entrée (des attributs de tuples), par leur type et leur nom, et l'indicateur produit par leur type. Brilhante et al. (2006) structurent les indicateurs dans une ontologie et associent les indicateurs à leurs formules de calcul dans une base de connaissance. Cependant les modalités de construction et d'acquisition de ces opérateurs ne sont pas mentionnées, et les formules ne sont pas rédigées suivant un formalisme mathématique standardisé.

L'usage d'un graphe de flots de données semble adapté à la problématique du lignage, mais si la description des transformations nécessite la contribution d'un utilisateur, il faut absolument proposer des outils pour faciliter cette saisie. L'interface graphique pourrait par exemple être enrichie avec un dictionnaire des fonctions de transformation usuelles dans le domaine statistique, et un dictionnaire des données utilisables comme entrée de la transformation. Ainsi,

l'utilisateur éviterait la saisie du type, et du nom de l'indicateur par un simple « *drag and drop* » depuis une liste d'indicateurs recensés dans l'ontologie. L'ontologie des indicateurs précédemment évoquée est réutilisable pour retrouver rapidement l'instance d'un indicateur utilisé dans la formule comme ingrédient, et ceci a déjà été souligné dans le travail de Brilhante et al. (2006). Il s'agit aussi d'établir le dictionnaire des transformations, le plus souvent des opérandes mathématiques, qui peuvent être simples comme la division, ou plus complexes comme une formule de normalisation. La saisie des formules peut être envisagée à l'aide de langages comme MathML[8], (pour Mathématiques Markup Language), ou bien OpenMath[9]. Par exemple, la formule de l'indicateur synthétique de vieillissement pourrait s'écrire (en utilisant Amaya[10] comme éditeur WYSIWYG de MathML) comme dans la figure 2 :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE math PUBLIC "-//W3C//DTD MathML 2.0//EN" "http://www.w3.org/TR/MathML2/dtd/mathml2.dtd">
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mi>IndicateurSynthétiqueVieillisement</mi>
  <mo>=</mo>
  <mfrac bevelled="true">
    <mi>ageMoyenPopulation</mi>
    <mi>espéranceVieBonneSanté</mi>
  </mfrac>
</math>
```

FIG. 2 – *Formule d'un indicateur composite, exprimée en MathML.*

Ces formules peuvent être rattachées aux indicateurs dans l'ontologie, ou à leurs instances. Par exemple, les formules d'ajustement des variations saisonnières pourraient être accolées aux instances de l'indicateur *chômeur* : *chômeur_INSEE* et *chômeur_Eurostat*.

5 Conclusion et perspectives

Cet article expose en détail les problèmes que pose la recherche d'équivalences entre valeurs d'indicateurs socio-économiques lorsque ces statistiques sont produites par différentes sources, en différents lieux et dates. Les métadonnées, qui avaient été recommandées comme une solution à l'intégration statistique, permettent effectivement de donner du sens aux données, mais cette solution *a minima* ne résout pas les conflits de catégorie, ni les changements de sémantique liées aux transformations de certaines données. En effet, les métadonnées restent essentiellement une description textuelle qui n'autorise pas un traitement automatisé de l'information. Il apparaît néanmoins que l'emploi de techniques de fouille de texte appliquées sur les métadonnées peut faciliter le calcul d'alignement des classifications, mais également aider à construire une ontologie statistique, en mettant en relation dans un graphe les concepts représentés par les indicateurs. Cette ontologie serait utilisable également pour établir une base de connaissance des transformations.

Pour exploiter cette base de connaissances, et par exemple calculer l'équivalence entre deux valeurs, il faudrait l'associer à un raisonneur capable de résoudre automatiquement les liens entre valeurs d'indicateurs par l'interprétation des formules de transformation. Dans cette perspective, nous allons travailler sur la définition d'un formalisme plus avancé des transforma-

[8]http ://www.w3.org/TR/MathML3/
[9]http ://www.openmath.org/documents/bibliography.html
[10]http ://www.w3.org/Amaya/Overview.html

tions, tel qu'un Langage de Modélisation Algébrique (LMA), et à la construction de l'ontologie des indicateurs statistiques associée.

Références

Brilhante, V., A. Ferreira, J. Marinho, et J. S. Pereira (2006). Information integration through ontology and metadata for sustainability analysis. In *International Environmental Modelling and Software Society (iEMSs) Third Biannual Meeting "Summit on Environmental Modelling and SoftwareThird Biennial Meeting"*, Burlington, USA.

Colledge, M. J. (1998). Statistical integration through metadata management. *International Statistical 67*(1), 79–98.

Comber, A., P. F. Fisher, et R. A. Wadsworth (2005). A comparison of statistical and expert approaches to data integration. *Journal of Environmental Management 77*, 47–55.

Comber, A., A. Lear, et R. Wadsworth (2010). A comparison of different semantic methods for integrating thematic geographical information : the example of land cover. In *AGILE'2010*.

Dean, P. et B. Sundgren (1996). Quality aspects of a modern database service (position paper). In P. Svensson et J. C. French (Eds.), *SSDBM*, pp. 156–161. IEEE Computer Society.

Gauthier, J. G. (2002). *Measuring America : the decennial censuses from 1790 to 2000*. U.S. Census Bureau.

Goux, D. (2003). Une histoire de l'enquête emploi. *Economie et Statistique* (362).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In H. M., G. F., et T. R. (Eds.), *Proceedings of 22nd International Conference on Research and Development in Information Retrieval*, Berkeley, Californie, US, pp. 50–57.

Kieffer, A., M. Oberti, et E. Preteceille (2002). Enjeux et usages des catégories socio-professionnelles en europe. *Sociétés contemporaines 45-46*, 157–185.

Meyer, D., T. Hothorn, F. Leisch, et K. Hornik (2004). Statdataml : An xml format for statistical data. *Journal of Computational Statistics 19*(3), 493–509.

Oakley, G., A. Hamilton, et J. Michel (2005). Experiences and plans of the australian bureau of statistics related to data and metadata exchange. Technical report, Australian Bureau of Statistics.

OCDE, O. (2008). Handbook on constructing composite indicators : Methodology and user guide. Technical Report ISBN 978-92-64-04345-9, OCDE.

Parlement européen (2007). directive 2007/2/ce établissant une infrastructure d'information géographique dans la communauté européenne (inspire) (http ://inspire.jrc.ec.europa.eu/).

Pattuelli, M. C., S. W. Haas, S. W. Haas, et J. Wilbur (2003). The govstat ontology. In *Proceedings of the 2003 annual national conference on Digital Government research (DG.O)*.

Plumejeaud, C., J. Gensel, et M. Villanova-Oliver (2010). Opérationnalisation d'un profil iso 19115 pour des métadonnées socio-économiques. In *Actes du XXVIIIème congrès INFORSID*, pp. 25–41.

Rafanelli, M. et A. Shoshani (1990). Storm : A statistical object representation model. In Z. Michalewicz (Ed.), *In Proc. of Statistical and Scientific Database Management, 5th In-*

ternational Conference SSDBM, Charlotte, NC, USA, pp. 14–29.

Robertson, S. (2004). Understanding inverse document frequency : On theoretical arguments for idf. *Journal of Documentation 60*(503-520), 2004.

Servigne, S., N. Lesage, et T. Libourel (2006). Spatial data quality components, standards and metadata. In *Fundamentals of Spatial Data Quality*, Number ISBN 1905209568, pp. 179–210. International Scientific and Technical Encyclopedia.

Shoshani, A. (1982). Statistical databases : Characteristics, problems, and some solutions. In *VLDB '82 : Proceedings of the 8th International Conference on Very Large Data Bases*, San Francisco, CA, USA, pp. 208–222. Morgan Kaufmann Publishers Inc.

UMS 2414 RIATE, UMR 8504 Géographie-cités, LIG, IGEAT, Umeå University, Università l'Orientale, et Université "Alexandru Ioan Cuza" Iaşi. (2008). Régions en déclin : un nouveau paradigme démographique et territorial. Etude pour le Parlement Européen - Structural and cohesion policies - Juillet 2008 IP/B/REGI/IC/2007-044, Parlement européen.

UN/ECE (1995). Guidelines for the modelling of statistical data and metadata. Technical report, UN/ECE, New York, Geneva.

Wadsworth, R. A., A. Comber, et P. F. Fisher (2006). Expert knowledge and embedded knowledge or why long rambling class descriptions are useful. In G. E. Andreas Riedl, Wolfgang Kainz (Ed.), *Progress in Spatial Data Handling, Proceedings of SDH*, pp. 197 – 213. Springer Berlin / Heidelberg.

Wilks, S. S. (1939). The rise of modern statistical science. In *MIT Industrial Statistics Conference*, New York, NY, USA, pp. 283–310. Pitman Publ. Corp.

Woodruff, A. et M. Stonebraker (1997). Supporting fine-grained data lineage in a database visualization environment. In *Proceedings of the Thirteenth International Conference on Data Engineering*, Number Report n °UCB/CSD-97-932, Birmingham, U.K., pp. 91–102.

Summary

This research aims at developing new methods for similarity analysis of statistical values, coming from various sources, at different geographical scales and validity periods. The values associated to statistical indicators, such as "social and occupational group workforce", measured on given locations and for given dates, have not an homogeneous semantic. The indicator metadata (using a ISO 19115 profile, or inside the SDMX format) provide a part of the value' meaning through textual elements (keywords, themes, abstract). However, a value is often a combination of various categories (sex, age, group) that change over time and space; besides, it can be the result of an estimation process. In those conditions, values are not equivalent: this is the semantic variability of statistical indicators. We explain why the metadata, although necessary, are not sufficient, and we dress the list of the various existing solutions for this problem.