Opérationnalisation d'un profil ISO 19115 pour des métadonnées socio-économiques

Christine Plumejeaud* — **Jérôme Gensel*** — **Marlène Villanova-Oliver***

**Laboratoire d'Informatique de Grenoble*
681 rue de la Passerelle, BP. 72, 38402 St Martin d'Hères, France
{christine.plumejeaud, jerome.gensel, marlene.villanova-oliver}@imag.fr

RÉSUMÉ. *La documentation des données recueillies et utilisées dans les systèmes d'information géographique est essentielle pour rendre possible leur découverte, l'échange et l'interopérabilité entre différents systèmes. Cette importance est soulignée par la directive INSPIRE, qui préconise l'usage de la norme ISO19115 pour la création de métadonnées, avec la mise en place de dispositifs pour la découverte des données spatiales via les métadonnées. Notre recherche montre que les spécificités de l'information statistique (en tant qu'information à référence spatiale et temporelle) nécessite une adaptation de cette norme, et définit une extension permettant de capturer le lignage de chaque valeur d'un ensemble de statistiques socio-économiques. Nous décrivons, de plus, un processus développé dans le cadre du projet européen ESPON DB 2013, pour l'acquisition, le stockage et la restitution des métadonnées associées aux indicateurs statistiques dans un système d'information.*

ABSTRACT. *The documentation of data used in any Geographical Information System (GIS) is essential for the share, the exchange, and the interoperability between various systems. This fact has been highlighted by the INSPIRE directive at the European level, whose recommendations are to create and maintain metadata based on the ISO19115 standard, and to permit the discovering of spatial data through out metadata. In this context, our research shows that statistical data, (which is a kind of information attached to spatial and temporal references), have some specificities requiring the adaptation of the standard. This work defines an extent of the ISO 19115 standard allowing for the complete definition of the lineage of each value of a statistical dataset. Furthermore, we describe a process for the acquisition, the storage and distribution of data with their metadata that have been fully tested and implemented inside a European project called ESPON DB 2013, building thus an active metainformation system.*

MOTS-CLÉS : *métadonnées – norme ISO 19115 – lignage d'indicateurs statistiques.*

KEY WORDS: *metadata – ISO 19115 standard – statistical information – lineage – genealogy*

1. Introduction

L'information spatiale est entrée dans la vie quotidienne des usagers du Web, via les premières applications de calcul d'itinéraires (routiers, de randonnées), et toutes les applications associées à l'API GoogleMap. L'exploitation de cette information, généralement issue de sources hétérogènes, nécessite une description précise via ce qu'on appelle les métadonnées, ou « données qui renseignent sur certaines données et qui permettent leur utilisation pertinente » (Bergeron, 1993). Les métadonnées donnent donc à l'utilisateur des éléments pour comprendre si les données sont en adéquation avec ses besoins, en décrivant à la fois le contenu de l'information, sa fiabilité, et sa disponibilité, et permettent d'établir la qualité des données au sens le plus large du terme (Servigne *et al.*, 2006). Ces métadonnées doivent reposer sur un format structuré et partagé, pour autoriser un usage interopérable entre différents systèmes d'information géographique. INSPIRE[1], une directive européenne pour le partage et la diffusion de l'information géographique, préconise à cet effet l'usage d'un standard pour les données à référence temporelle et spatiale : la norme ISO 19115[2].

Parce qu'elles se présentent le plus souvent comme des ensembles datés de nombres associés à des unités territoriales, les statistiques socio-économiques sont des informations à références spatiales et temporelles qu'il serait légitime de décrire par des métadonnées au standard ISO 19115. Dans le domaine de la statistique, le besoin de produire des métadonnées a été éprouvé très tôt (Mac Carthy, 1982) et réitéré plusieurs fois : (ONU, 1995), (Dean *et al.*, 1996), (Kent *et al.*, 1997). Alors que la norme ISO 19115 a été largement étudiée et adoptée par les différents producteurs et usagers de données à références spatiales tels que l'Agence Européenne de l'Environnement[3] et le Joint Research Center[4], on constate que les acteurs du système de collecte de données socio-économiques se tiennent à l'écart de cette norme. Preuve en est l'absence de citation de cette norme dans les comptes-rendus des dernières conférences internationales[5,6] sur les systèmes de gestion de l'information statistique. Une des raisons est sans doute le manque d'adéquation de la norme ISO 19115 vis à vis des spécificités de l'information statistique. Parmi elles, la structuration particulière, composite et hétérogène, de l'information statistique est à souligner. On raisonne souvent avec un *jeu de données statistiques* regroupant généralement une collection *d'indicateurs* différents, *mesurés* chacun de façon spécifique sur un ensemble d'unités territoriales, à différentes dates. Il existe donc une réelle complexité dans le fait d'associer des métadonnées expressives et exploitables à une telle structure.

[1] http://inspire.jrc.ec.europa.eu/
[2] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020
[3] http://www.eea.europa.eu/fr
[4] http://ec.europa.eu/dgs/jrc/index.cfm
[5] http://www.unece.org/stats/documents/2007.05.msis.htm
[6] http://www.unece.org/stats/documents/2009.05.msis.htm

Un standard et un modèle pour l'échange de données statistiques, appelé SDMX (pour *Statiscal Data Model eXchange*)[7], commence toutefois à se dessiner, à l'initiative du domaine bancaire et de celui des assurances. Si SDMX prend mieux en compte certaines spécificités de l'information statistique que l'ISO 19115, il n'offre cependant pas de fonctionnalités aussi avancées que l'ISO 19115 pour la gestion de l'information spatiale. Ce standard qui mixe, dans une même structure, données et métadonnées, et exige des utilisateurs de concevoir et publier le schéma structurant données et métadonnées, est d'une part encore émergeant, et, d'autre part, peu opérationnel pour des utilisateurs qui ne souhaitent pas reformater leurs données dans une structure complexe.

Nous avons donc étudié, dans un premier temps, l'adéquation de la norme ISO 19115 pour l'information statistique. Si elles confirment l'intérêt d'utiliser cette norme pour décrire l'information statistique à composantes spatiale et temporelle, nos conclusions (ESPON, 2009) mettent aussi en exergue le besoin de l'adapter pour mieux en gérer les spécificités. Nous en proposons donc dans un second temps une extension, qui prend la forme d'un profil de la norme, pour la rendre opérationnelle, et ainsi promouvoir son utilisation auprès des acteurs du système d'information pour les données statistiques : producteurs, distributeurs, usagers, etc. Le profil proposé est intégré dans un processus d'acquisition, de traitement, de stockage et de restitution des données et métadonnées statistiques. Ce processus, qui vise une plus grande utilisabilité des données statistiques, est implémenté et testé dans le cadre d'un système d'information développé pour le projet *ESPON 2013 DataBase*[8], dédié à la collecte et la documentation de données statistiques produites sur tout le territoire européen et sur une longue période de temps (de 1950 à 2050) (Espon, 2009)

L'article est organisé comme suit : la section 2 étudie la compatibilité de l'information statistique avec la norme ISO 19115, et présente le profil créé. La section 3 décrit le processus de gestion de l'information statistique qui a été développé au sein du projet *ESPON 2013 DataBase*. La section 4 délivre quelques critiques de ce profil ISO 19115 relatives à l'exploitation des données. Enfin, la section 5 conclut et donne les perspectives de ces travaux.

2. Compatibilité de l'information statistique avec la norme ISO 19115

Publiée en 2003, la norme ISO 19115 est issue de travaux internationaux sur le partage des données de nature principalement environnementale. Un nombre important de systèmes dédiés à la mutualisation de l'information géographique de nature environmentale mettent en œuvre la norme et développent des catalogues de métadonnées disponibles sur le Web (Barde, 2005). L'avantage d'utiliser la norme ISO 19115 pour les données statistiques ayant des références temporelle et géographique réside donc dans l'interopérabilité assurée entre les systèmes de

[7] http://www.sdmx.org/

[8] http://www.espon.eu/main/Menu_Projects/Menu_ScientificPlatform/espondatabase2013.html

distribution de l'information de nature environnementale et l'information socio-économique. L'emploi de cette norme est, de plus, une obligation légale dans le cadre européen, avec la directive INSPIRE, pour toutes les données à référence spatiale et temporelle. Il s'agit de démontrer qu'une adaptation de la norme est nécessaire mais aussi possible pour prendre en compte les particularités de l'information statistique.

2.1. *Description de l'information statistique*

L'information statistique circule généralement sous la forme de jeux de données, qui regroupent chacun une collection d'indicateurs différents, mesurés chacun de façon spécifique sur un ensemble d'unités territoriales, à différentes dates. Un jeu de données présente de manière schématique trois niveaux d'information différents :
- *Premier niveau* : Ce niveau d'information est commun au ***jeu de données*** : il renseigne sur son nom, son responsable, son créateur, ses modalités de distribution, et sa maintenance (régularité, fréquence).
- *Deuxième niveau* : Le second niveau d'information décrit chaque ***indicateur***. En effet, un indicateur est désigné dans le jeu de données par un code, qui ne suffit pas à sa compréhension : il est aussi nécessaire de lui associer un nom, une description textuelle, l'unité de mesure employée, et une classification thématique. La sémantique est aussi donnée par la méthode de mesure (méthodologie) qui est spécifique à chaque indicateur. Un indicateur comme la consommation d'eau par ménage peut-être recueilli par un sondage ou une enquête auprès d'un échantillon représentatif des ménages, puis estimé à partir de cet échantillon, alors que le nombre de naissances par commune est un chiffre enregistré au niveau des mairies, sans estimation.
- *Troisième niveau* : Le troisième niveau d'information décrit les ***valeurs*** des indicateurs, sur chacune des unités statistiques. Pour un même indicateur, on constate que sa sémantique varie souvent d'un producteur de données à l'autre, et d'une époque à l'autre. Nous illustrons ce problème avec l'indicateur « chômage ». En dépit d'une tentative d'harmonisation européenne symbolisée par le partage d'une définition commune définie par l'Organisation Internationale du Travail, l'Institut National de la Statistique et des Etudes Economiques (INSEE) et Eurostat publient des chiffres différents pour la même unité (la France) : 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[9], et Eurostat[10]. L'exemple du chômage illustre aussi l'évolution des méthodes de calcul et de mesure dont sont l'objet les indicateurs. Par exemple,

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

l'INSEE fait évoluer régulièrement sa méthodologie de calcul du chômage (Goux, 2003), en le justifiant dans des documents accessibles en ligne[11]. Il est donc nécessaire d'accompagner les valeurs des indicateurs d'une information de provenance (le lignage) indiquant la source, la date et la méthode de collecte et de traitement des données, afin de pouvoir les comparer et les interpréter plus justement. Cette information est souvent complexe : elle comporte à la fois des formules de calcul, des définitions, et est le plus souvent compilée dans des documents externes au jeu de données. Le guide de l'OCDE sur la construction des indicateurs composites (OECD, 2008) montre la difficulté de résumer ces manipulations (pondérations, réajustements, etc.) en une simple formule. En effet, dès qu'elles impliquent plusieurs indicateurs, il faut s'assurer que les composants sont désignés par un code connu et documenté quelque part, et que cette documentation est accessible à tous.

Enfin, suivant les sources utilisées, le niveau de restrictions d'usage sur les données peut varier. Par exemple, des données spécifiques sur l'emploi en Pologne peuvent avoir été collectées par des organismes de recherche qui ne souhaitent pas les diffuser au grand public, mais seulement à un nombre restreint d'utilisateurs, alors que généralement ces statistiques sur l'ensemble du territoire européen sont disponibles librement. C'est pourquoi ces contraintes doivent être également associées au niveau des valeurs dans le jeu de données.

Il faut noter également que le lignage des valeurs peut être commun à un ensemble d'indicateurs, en particulier dans le cas de tableaux de contingences. Un tableau de contingence correspond à la désagrégation d'un indicateur (la population par exemple) en fonction de catégories (classes d'âge, catégories d'emploi, etc.) pour former de nouveaux indicateurs. Ces indicateurs (population active en milliers par tranche d'âge et sexe, par exemple, cf. tableau 1) partagent alors le même lignage. Le code de l'indicateur est alors un pointeur vers une cellule du tableau de contingence, par exemple « actifs_15_64_m » pour population active âgée entre 15 et 65 ans, de sexe masculin.

Age	Hommes	Femmes	Ensemble
15 ans ou plus	actifs_15_m	actifs_15_f	actifs_15
15-64 ans	actifs_15_64_m	actifs_15_64_f	actifs_15_64
15-24 ans	actifs_15_24_m	actifs_15_24_f	actifs_15_24
25-49 ans	actifs_25_49_m	actifs_25_49_f	actifs_25_49
50-64 ans	actifs_50_64_m	actifs_50_64_f	actifs_50_64
dont : 55-64 ans	actifs_55_64_m	actifs_55_64_f	actifs_55_64
65 ans ou plus	actifs_65_m	actifs_65_f	actifs_65

Tableau 1. *Exemple de codes d'indicateurs créés pour une table de contingence extraite du site INSEE, « population active en milliers selon le sexe et l'âge en 2008 en France»*[12]

[11] http://www.insee.fr/fr/methodes/sources/pdf/estimations_chomageBIT_enquete_emploi.pdf
[12] http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATCCF03170

2.2. Structure de la norme ISO 19115

La norme ISO 19115 décrit un ensemble d'informations (obligatoires ou facultatives) qui peuvent être associées aux données dites « brutes » et propose une grammaire XML pour la structuration de ces informations. Les informations s'organisent dans différentes rubriques. Elles apparaissent dans le schéma simplifié de la norme, figure 1, créé d'après la lecture de la spécification ISO 19115, et qui ne détaille que les attributs obligatoires. En effet, la norme offre la possibilité d'adapter le niveau de détails et de richesse des informations à ses besoins et seules les informations relatives à l'identification du jeu de données et à la personne responsable de la fiche de métadonnées sont obligatoires : c'est le noyau de la norme (en foncé dans la figure 1).

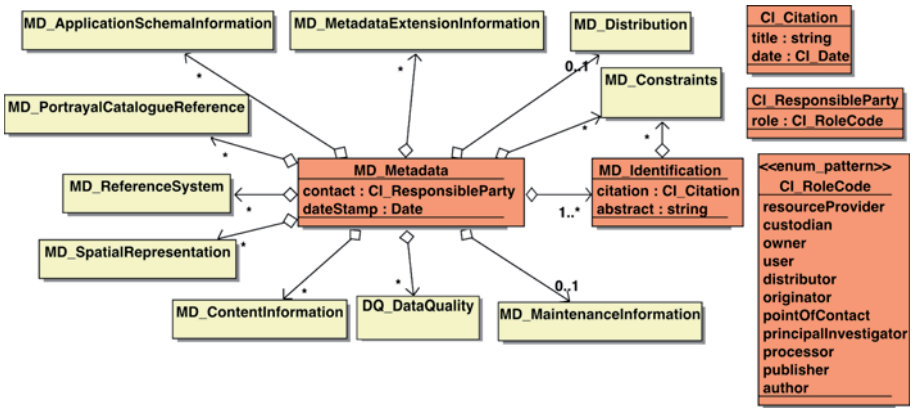

Figure 1. *Ensemble des rubriques de la norme ISO 19115*

Cet ensemble de données peut être ensuite étendu à volonté dans un profil, qui est une extension des éléments de base (par ajouts de nouveaux éléments, et/ou spécialisation des éléments existants), mais il doit forcément inclure le noyau.

La rubrique **MD_*ApplicationSchemaInformation***, bien que peu utilisée et souvent mal comprise (Barde, 2005), permet de spécifier des niveaux d'agrégation récursifs. On peut proposer une fiche de métadonnées pour un agrégat de données (*DS_Aggregate*), et chaque agrégat de données est composé de jeux de données (*DS_Dataset*) qui possèdent chacun leur propre fiche de métadonnées (*MD_Metadata*) (voir figure 2).

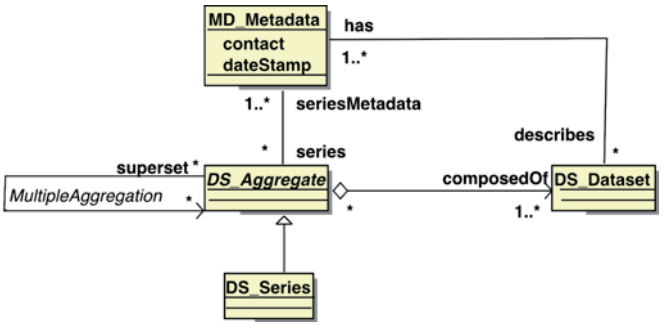

Figure 2. *Composition simplifiée de MD_ApplicationSchemaInformation*

La norme est le plus souvent utilisée au niveau de *MD_Metadata* pour décrire des données thématiques rattachées à un support spatial de type quelconque. Ceci s'accompagne d'une simplification extrême de la description thématique, qui présuppose que les données thématiques ont une description commune, et la même provenance, comme l'illustre l'exemple suivant. Les données du Corine Land Cover, (décrivant la nature d'occupation des sols), disponibles en ligne sur le site de l'Agence Européenne de l'Environnement, sont associées à une fiche de métadonnées du niveau *MD_Metadata*, conforme à la norme ISO 19115, mais qui ne comporte aucune description détaillée des 45 classes d'usage du sol que contient cette base de données. Les différentes classes d'usage du sol partagent effectivement le même lignage, étant donné qu'elles ont été déduites à partir de traitements de la même information : l'image satellitaire. Ceci est une observation qui se vérifie quel que soit le format (vectoriel[13] ou raster[14]) de ce jeu de données. Cependant, cette pratique de simplification de l'information thématique se vérifie pour d'autres jeux de données. Même lorsqu'ils contiennent plusieurs indicateurs, les jeux de données sont associés à des métadonnées de niveau *MD_Metadata*, où l'identification ne détaille pas chacun des indicateurs mesurés.

2.3. *Création d'un profil de la Norme ISO 19115*

Le profil ***esponMD*** qui définit l'extension de la norme ISO 19115 pour l'information socio-économique et que nous proposons a pour objectif d'être opérationnel, et de tenir compte des spécificités de l'information statistique : il doit décrire les données suivant trois niveaux différents (le jeu de données, les indicateurs, et les valeurs).

2.3.1. *Gestion des différents niveaux d'information*

Pour décrire l'information statistique, il est nécessaire de produire une information thématique détaillée du jeu de données, et cette description est

[13] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2000-clc2000-seamless-vector-database-1

[14] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-clc1990-100-m-version-12-2009

nécessairement portée par l'élément *MD_Identification* qui dans la norme contient aussi les informations de classification. *MD_Identification* pourrait donc permettre de décrire un indicateur. Cependant *MD_Identification* ne porte ni les informations sur l'usage (*MD_Constraint*) ni les informations de qualité (*DQ_DataQuality*), qui renseignent sur la provenance et les méthodes de mesure des valeurs, qui sont rattachées à *MD_Metadata*. Or la description de l'information statistique nécessite de rattacher ces informations à la description de l'indicateur, et de plus l'élément *MD_Identification* de *MD_Metadata* doit être unique. Pour décrire plusieurs indicateurs, la logique exige donc de multiplier le nombre de fiches *MD_Metadata*, qui contiendront alors les informations pour un indicateur donné.

Notre profil utilise le mécanisme d'agrégation de fiches décrit dans la figure 2, car l'élément *DS_Dataset* permet de disposer de plusieurs fiches de *MD_Metadata*, une par indicateur. Comme il est aussi nécessaire de représenter une information commune à l'ensemble du *DS_Dataset*, celle qui décrit les auteurs des métadonnées (*CI_ResponsibleParty*), la description générale du jeu de données (*MD_Identification*), la distribution (*MD_Distribution)* et la maintenance (*MD_Maintenance)* éventuelle des données, nous utilisons la racine *MD_Metadata*. Ainsi, *MD_Metadata* décrit le jeu de données, et possède un attribut *describes* de type *DS_Dataset*. L'élément *DS_Dataset* contient (*has*) plusieurs fiches *MD_Metadata* décrivant séparément les indicateurs.

2.3.2. *Adaptation de l'élément MD_Identification*

L'identification d'un indicateur se fait avec ***IndicatorInformation*** qui étend *AbstractMD_Identification*, comme l'illustre la figure 3; (sur cette figure, les éléments originaux de la norme apparaissent en rose, ceux étendus en jaune clair).

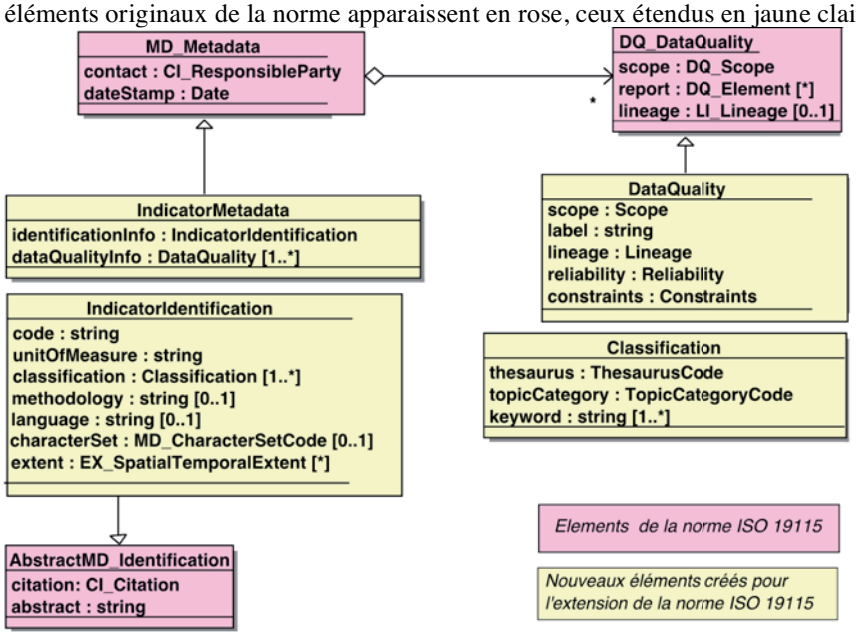

Figure 3. *Informations redéfinies au niveau indicateur dans l'extension.*

Cet élément rend obligatoires les informations concernant l'indicateur : *citation* (pour le nom), *abstract* (pour le résumé), *code*, *unitOfMeasure* (unité de mesure), et *classification*. La classification de type ***Classification*** propose à l'utilisateur de choisir un thésaurus (parmi ceux référencés par le système d'information), et un des thèmes ou sous-thèmes que propose ce thésaurus, auquel il peut associer un à plusieurs mots clés (des chaînes de caractères).

2.3.3. *Simplification de l'élément DataQuality*

L'utilisateur doit ensuite mentionner des informations renseignant sur la qualité des valeurs associées à l'indicateur de façon obligatoire, mais simplifiée : l'élément ***DataQuality*** qui étend *DQ_DataQuality* est prévu à cet effet. Le champ *label* référence des étiquettes placées dans le jeu de données en face de chaque valeur et permet de dispenser l'utilisateur de l'énumération fastidieuse des éléments géographiques concernés par cette qualité (via le champ *scope*), ou bien du calcul de la couverture spatio-temporelle (*extent*) de cette rubrique. Une partie des informations sera calculée lors de l'acquisition des données, comme on peut le faire aujourd'hui pour d'autres données (Diaz *et al*., 2007) et l'utilisateur ne doit fournir qu'un ensemble minimal d'information. La figure 4 illustre l'ensemble des informations que porte l'élément ***DataQuality***, et les nouveaux éléments du profil apparaissent en couleur claire (jaune).

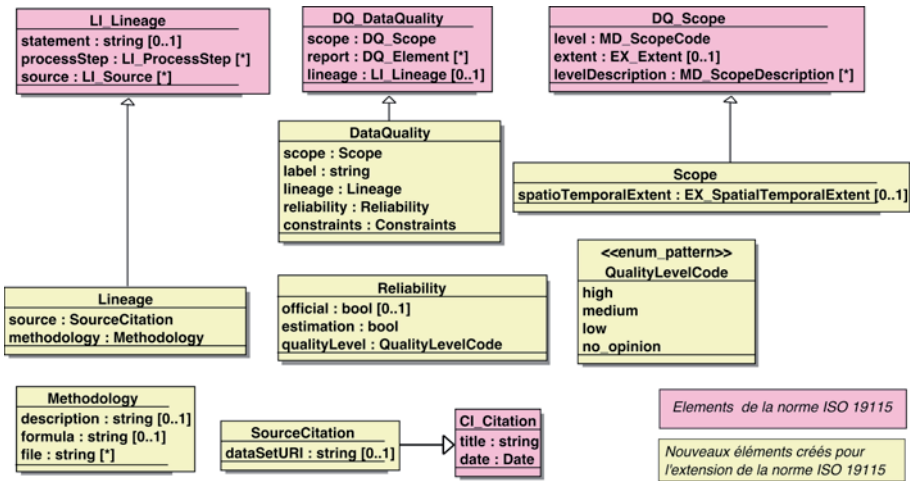

Figure 4. *Informations portées par l'élément DataQuality au niveau des valeurs*.

Les informations sur la qualité sont largement simplifiées par rapport à l'élément *DQ_DataQuality*. Le **rapport** (*report*), qui donne des éléments complets et objectifs concernant l'évaluation de la qualité des données est facultatif dans la norme comme dans l'extension, mais dans l'extension, l'élément de type *Reliability,* qui exprime la confiance que l'auteur des métadonnées accorde à ces données, de façon subjective,

est obligatoire. Il indique pour cela si les données sont issues d'une estimation, (*estimation* vaut vrai), et donne son opinion (*qualityLevel*) sur une échelle de valeur codifiée (*QualityLevelCode*). Le fait de savoir si la donnée est issue d'un fournisseur de données institutionnel (*official*) est déduit lors de l'acquisition de la fiche de métadonnées, en comparant le nom de la source des données (*SourceCitation.title)* à la liste des fournisseurs officiels reconnus par le projet. Le **lignage** (*lineage*) des données doit mentionner une source (*source*) et la méthode de mesure/collecte des données (*methodology*). La source précise obligatoirement le nom du fournisseur (*CI_Citation.title*) et la date de récupération des données (*CI_Citation.date)*. Le champ *methodology* est un équivalent simplifié de *LI_ProcessSteps,* qui est présent dans la rubrique *DQ_DataQuality* de la norme, et permet de décrire les transformations ou les méthodes de calcul des valeurs. Ceci est fait par divers moyens : soit, avec un champs textuel qui en donne une description (*description*), qui peut aussi être une URL, soit avec une formule (*formula*) qui peut s'exprimer dans un langage semi-structuré (comme MathML[15]), soit avec un ensemble de fichiers et documents multimédias que l'utilisateur pourra adjoindre à la fiche de métadonnées simplifiées. Cet expédient, en attendant des moyens plus automatisés de traiter l'information, permet au moins de collecter la connaissance, pour les utilisateurs humains qui accèderont à cette documentation, lors de la restitution des métadonnées.

2.3.4. *Contraction de l'élément Constraints*

Les contraintes associées à l'usage des données et des métadonnées ont aussi été redéfinies pour les rendre plus opérationnelles. Elles sont attachées au niveau des valeurs du jeu de données, dans la rubrique *DataQuality*, et sont obligatoires. *Constraints* est une composition de contraintes sur les métadonnées (*MetadataConstraints*) et sur les données (*DataConstraints*), (voir figure 5).

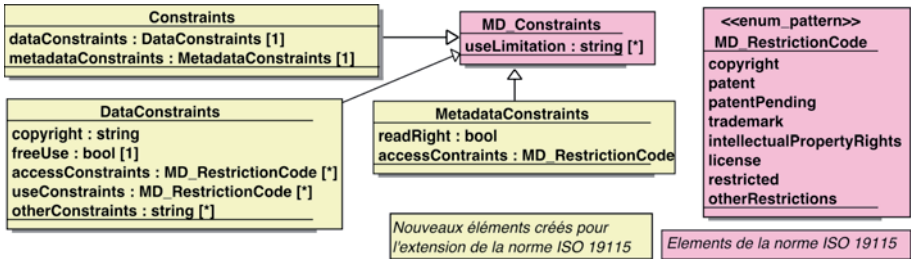

Figure 5. Définition de contraintes d'usage pour des valeurs du jeu de données.

La norme prévoit, en effet, un mécanisme permettant de diffuser des métadonnées, sans les données, ou bien de cacher complètement les métadonnées (et les données ne sont alors pas accessibles non plus) : *readRight* vaut faux. Les

[15] http://www.w3.org/Math/

données peuvent être diffusées librement à tout public (*freeUse* vaut vrai) ou bien être réservées à une communauté d'utilisateurs enregistrés auprès du distributeur des données (*freeUse* vaut alors faux). Les données sont toujours accompagnées d'un *copyright*, champ textuel équivalent de *useLimitation*, mais obligatoire, et qui vaut « *Espon 2013 Program* » par défaut. L'auteur des métadonnées est encore libre de réutiliser les codes spécifiques pour les mentions légales que propose la norme avec les champs *accessConstraints*, et *useConstraints*.

L'opérationnalisation de la norme ISO 19115 signifie donc la création d'un profil plus simple, mais aussi plus contraignant : l'utilisateur est forcé de renseigner certains champs qui, auparavant, étaient optionnels. En contrepartie, l'information qu'il doit livrer est simplifiée au maximum, et une partie sera calculée après acquisition des données. En effet, ce profil est utilisé pour l'acquisition des données statistiques dans un système d'information implémenté par le projet *ESPON 2013 database*, qui peut ensuite diffuser des métadonnées complétées.

3. Intégration du profil ISO 19115 dans un Système d'Information Statistique

L'objectif de nos travaux est de rendre compte de la qualité des données compilées par le projet ESPON 2013 DB. Cet objectif ne peut être accompli sans la collecte et la restitution de métadonnées avec les données. Nous décrivons ici le flot de données.

3.1. *Structuration des données et métadonnées*

Le flot de données s'appuie sur la définition d'un modèle de document, produit par un tableur tel qu'Excel, pour des données socio-économiques. Ce formatage a été facilement adopté par les partenaires du projet car il s'apparente aux fichiers que les acteurs pouvaient déjà s'échanger : des fichiers Excel, présentant une première colonne avec la liste des unités statistiques, codifiées suivant la version de nomenclature utilisée, puis un nombre non limité de colonnes d'indicateurs, représentés par leur code dans la base de données du fournisseur de données, et, au croisement de la ligne et de la colonne, la valeur de l'attribut associé. Ce modèle est présenté dans la figure 6.

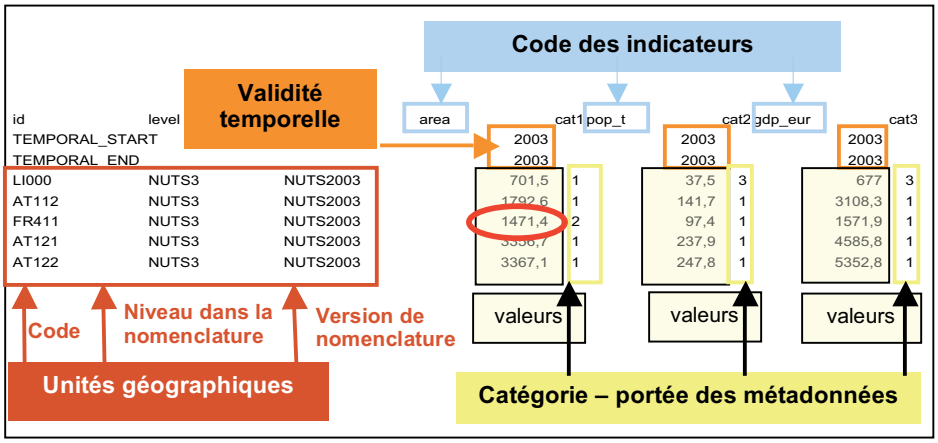

Figure 6. *Modèle de document Excel pour les données*

Il formalise certaines informations implicites: la version de la nomenclature, le niveau d'information pour l'unité géographique codifiée (équivalent d'une échelle géographique), les dates de validité des colonnes. De plus, pour chaque colonne d'attributs, une colonne adjacente a été ajoutée (*catégorie*) afin de spécifier la provenance de chaque valeur, via une étiquette unique qui référence des informations de lignage qui sont précisées dans un autre fichier. Par exemple, la valeur 1471,4 entourée en rouge dans la figure 6 correspond à l'aire (area) de l'unité FR411 mesurée en 2003, dans le niveau 3 de la Nomenclature des Unités Territoriales Statistiques (NUTS), version du NUTS datant de 2003, et les informations de provenance, qualité et contraintes de cette valeur sont associées à l'étiquette « 2 ». Cette étiquette pointe sur les informations qui sont développées dans le fichier de métadonnées.

Un second fichier, appelé « fichier de métadonnées », accompagne le premier fichier, dans lequel sont développées les informations descriptives sur les trois niveaux d'information définis, *jeu de données, indicateur,* et *valeur*. Ce second fichier, au format XML, respecte le schéma de l'extension que nous avons définie et il doit être produit par les fournisseurs de données du projet, à l'aide d'un éditeur que nous avons adapté et mis en ligne pour les utilisateurs. Ce fichier peut être accompagné d'un ensemble de documents qui explicitent les procédures de production et de transformation des données, et sont référencés dans la section lignage par leur nom. Ces fichiers seront conservés dans un répertoire, et référencés depuis la base de données qui conserve le reste des informations.

3.2. *Flot des données et métadonnées*

La figure 7 donne une vue d'ensemble du flot de données mis en place. Un éditeur de métadonnées facilite l'édition d'un fichier de métadonnées conforme à l'extension de la norme ISO 19115 définie. Ensuite, l'acquisition des données dans le système se fait par l'analyse et le traitement de la paire de fichiers données/métadonnées dans le système que l'utilisateur dépose par le biais d'un portail Web. Cette opération a pour conséquence le stockage dans une base de données spatio-temporelle de toutes les informations, données et métadonnées. Ces données (dites thématiques) se rattachent à une information spatiale qui présente une structure hiérarchique et évolutive (Plumejeaud *et al.*, 2009). L'ensemble constitue un système d'information *actif* (tel que défini par (ONU, 1995), permet de recomposer un nouveau jeu de données correspondant à une certaine requête spatio-temporelle, à la volée, et de l'accompagner des métadonnées reconstruites et complétées. La complétion concerne, par exemple, le calcul automatique de la couverture spatio-temporelle d'un indicateur, de la portée d'une information sur la qualité, et de l'ajout d'information sur la qualité des valeurs disséminées. L'interface de requête spatio-temporelle respecte les directives INSPIRE et propose des critères d'interrogation sur le lieu, la date, les acteurs, et la nature des données. Le schéma de la base de donnée, les vues de l'interface d'exportation, et de l'éditeur de métadonnées sont publiées dans deux rapports techniques du projet ESPON (ESPON TR1, 2010), (ESPON TR2, 2010), et le système est accessible en ligne depuis mars 2010.

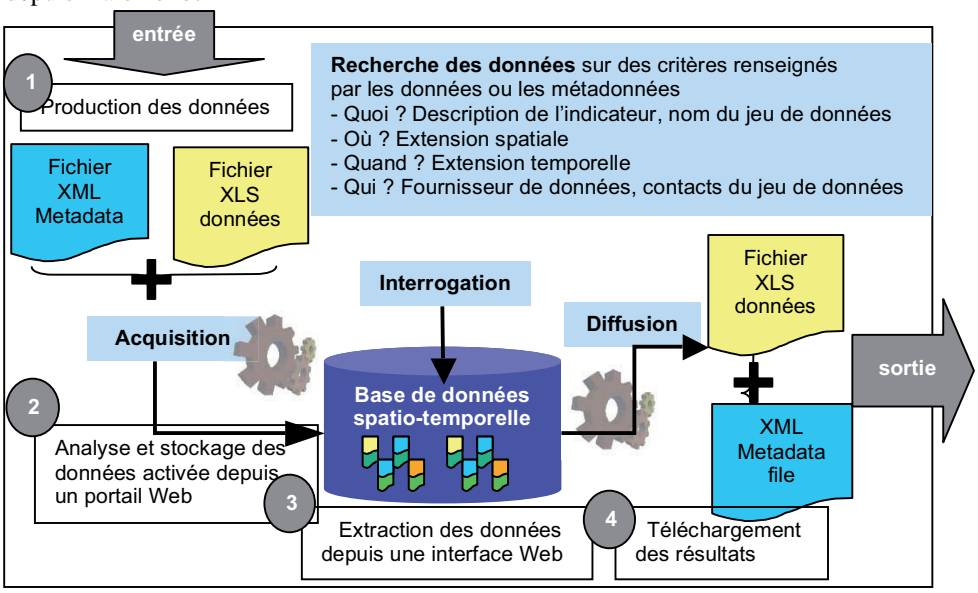

Figure 7. *Schéma du flot de données.*

3.3. *L'édition des métadonnées*

Pour éditer les métadonnées, nous avons choisi d'utiliser et d'adapter un outil *open-source* réputé : *Geonetwork (v2.4.2)*. Son adaptation nécessite l'ajout de notre profil de métadonnées (c'est-à-dire du schéma XSD correspondant) à l'ensemble des profils déjà supportés. Il s'agit de suivre une procédure documentée[16] et de manipuler des feuilles de styles XSL s'appliquant aux différents éléments de la grammaire, pour obtenir la présentation désirée de modèle pré-rempli (ou *template*) de métadonnées, et sa validation automatique. L'éditeur est en ligne sur l'adresse suivante : http://espon04.infeurope.lu:8081/geonetwork.

Dans ce *template*, l'utilisateur est invité à travers un formulaire HTML à ne remplir que les champs absolument nécessaires au stockage des métadonnées, et tous ceux qui peuvent être déduits du fichier de données, tels que la couverture spatiale et temporelle, sont absents de ce formulaire. Le formulaire présente dans l'ordre les informations associées au jeu de données, puis aux indicateurs, puis les informations de qualité (ou lignage) qui sont associées aux étiquettes de la colonne « catégorie » du fichier de données, via le champ *label*. Lorsque les données correspondent à des mises à jour d'indicateurs déjà présents dans la base de données, l'utilisateur peut sélectionner le code ou le nom de l'indicateur dans une liste de choix, et les champs définissant l'indicateur sont pré-remplis.

L'utilisateur doit valider son formulaire, et tous les éléments obligatoires mais manquants du profil lui sont signalés. *Geonetwork* permet aussi de vérifier la cohérence de certains champs, par l'implémentation de règles logiques faciles à spécifier par les concepteurs. Il peut ensuite sauver et exporter directement sa fiche XML sur son poste, qui sera prête pour accompagner le fichier de données pour la seconde phase du traitement : l'acquisition des informations.

4. Discussion du profil ISO 19115

Le profil *esponMD* de la norme ISO 19115 est utilisé pour la collecte de métadonnées et leur restitution dans un système opérationnel. Cependant, cette opérationnalisation de la norme a, dans un souci de simplicité, ignoré certains problèmes liés à la complexité de l'information statistique, lorsque l'on s'intéresse à l'expertise des données présentes dans le système d'information, en vue de détecter les valeurs anormales, ou bien de compléter les jeux de données manquants.

4.1. *Évaluation de la qualité des données*

Le premier problème est qu'une grande partie des informations détaillées et codifiées de lignage des données, et d'évaluation relative à la qualité ont été

[16] http://geonetwork-opensource.org/documentation/how-to/geonetwork-v2-2-shema-template-howto

élaguées du profil. Notamment, les rapports d'évaluation de la qualité des données se sont réduits à une simple expression d'opinion. Ceci peut être considéré comme une perte non négligeable d'information mais s'explique par le fait que très généralement l'intégration de ce type d'information dans un schéma semi-structuré, tel que propose l'ISO 19115, est véritablement fastidieuse, voire impossible, car les rapports restent externes au jeu de données, et doivent donc utiliser un champ « scope » pour désigner l'ensemble des valeurs concernées par l'évaluation.

En première solution, la conservation automatique et obligatoire des sources de données rend possible l'accès à ces rapports, s'ils existent. Il s'avère aussi qu'ils sont généralement couplés aux documents de méthodologie collectés par le système. Mais, une solution plus performante pour l'analyse automatique ultérieure des données serait l'emploi du modèle de données SDMX. En effet, le modèle de SDMX est plus évolué que celui de l'ISO 19115 en ce qui concerne la gestion de la qualité des données, puisqu'il peut associer chaque donnée avec une ou plusieurs métadonnées (comme le statut de l'observation, la fréquence de mesure, la source, ou tout autre information définie dans la grammaire du fichier d'échange de données). Le rapport est donc directement « embarqué » dans la donnée. On peut envisager d'étendre le modèle du système d'information pour capturer ce type d'information dans un champ semi-structuré, basé sur l'emploi des balises et des codes que propose SDMX, et de conserver un rapport associé à chaque valeur dans le système d'information, puis même d'exporter les données au format SDMX.

4.2. *Gestion des tableaux de contingence*

La gestion des indicateurs issus de tableaux de contingence n'est pas entièrement satisfaisante. Ces indicateurs partagent le même lignage, et le document décrivant la méthodologie peut éventuellement expliciter les catégories. Mais il manque, au niveau du jeu de données, la spécification formalisée et obligatoire du tableau. En effet, dans la perspective de l'expertise croisée des jeux de données, certains contrôles simples peuvent être menés à partir de la connaissance de ces catégories, (comme la somme du nombre de femmes et d'hommes doit être égale la population). D'autre part, ces modes de classification et d'organisation des données ne sont homogènes ni dans l'espace ni dans le temps. Par exemple, 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'étude requis pour pratiquer leur profession, etc. Mais ce mode de classification de la population en catégories socioprofessionnelles n'a pas d'équivalent européen (Kieffer *et al.*, 2002), parce que chaque pays construit ces catégories en fonction de son histoire. De précédents travaux (Haas *et al.*, 2003) proposent une ontologie des indicateurs, qui exploite les informations de classification thématique, les noms, les résumés et les codes, pour calculer les synonymes d'un indicateur donné. De la même façon, une définition précise et formalisée dans les métadonnées des tables de contingence,

dans un élément dédié, permettrait de mettre en œuvre des alignements de tableaux de contingence à l'aide d'ontologie.

5. Conclusions et perspectives

Dans cette étude, nous nous sommes intéressés à la possibilité de mettre en œuvre la norme ISO 19115 pour établir des métadonnées pour des données statistiques de type socio-économiques, à référence spatiale et temporelle. Nous répondons à la question de la compatibilité de ce type d'information avec la norme ISO 19115 de façon positive, moyennant une adaptation de la norme dans une extension, c'est-à-dire un profil. Le profil créé a pour le but de prendre en compte les trois niveaux d'information identifiés et de simplifier l'acquisition des métadonnées dans un profil de la norme. L'article présente aussi le cadre opérationnel dans lequel ce profil est utilisé, et décrit le flot de données associé. Sans entrer dans certains détails (schéma de stockage, et interface d'extraction des données de la base, édition des métadonnées), qui sont expliqués dans des rapports techniques (Espon TR1, 2010), (Espon TR2, 2010), nous montrons que ce profil peut-être utilisé pour la collecte de métadonnées simples et suffisantes pour les premiers niveaux d'usage des métadonnées (la découverte et l'exploration). Le profil a déjà fait l'objet d'une évaluation, car bien que l'éditeur de métadonnées ne soit en ligne que depuis mars 2010, les utilisateurs ont produit des données dans le format Excel décrit, et des métadonnées dans un formulaire textuel qui ressemblait en tout point au formulaire Web qui vient d'être mis en ligne. Cet usage de métadonnées opérationnelles, à l'échelle d'un programme européen, qui implique une trentaine de partenaires produisant des données spécialisées dans le domaine socio-économique, peut être considéré comme un succès.

Comme perspective, nous souhaitons développer un système pour expertiser la qualité des données stockées, qui pourrait ensuite donner lieu à l'estimation de données manquantes, ou jugées peu conformes à la réalité observée. Deux des principaux problèmes que pose l'estimation sont la description du processus d'évaluation (le lignage) de la donnée manquante, et l'exploitation de la qualité des données. Le profil *esponMD* de la norme ISO 19115 est une première étape pour la résolution de ce problème, car l'élément de lignage permet au pire d'adjoindre un document, sans format structuré, décrivant pour un humain le processus de production de la donnée. Mais, en s'astreignant à l'emploi d'un formalisme semi-structuré qui pourrait s'inspirer de SDMX, un automate pourrait décrire et interpréter le lignage successif des données et en donner une représentation sous la forme d'une formule, modélisant la généalogie de l'information.

Enfin, la dissémination des jeux de données en fichiers XML respectant la structuration du format SDMX faciliterait l'interopérabilité du système avec les autres systèmes d'information statistique, comme celui d'Eurostat, qui participe activement au développement du standard SDMX. Inversement, l'intégration de fichiers transmis dans le format SDMX sera certainement possible.

6. Bibliographie

Barde, J., Mutualisation de Données et de Connaissances pour la Gestion Intégrée des Zones Côtières. Application au Projet SYSCOLAG. Thèse de doctorat - Université Montpellier II , 2005

Bergeron M., Vocabulaire de la géomatique. office de la langue française. 1993.

Dean P., Sundgren B., Quality Aspects of a Modern Database Service. In *Proceedings of the 8th International Conference on Scientific and Statistical Database Management, SSDBM'96,* pp. 156-161, 1996

Díaz L., Martín C., Gould M., Granell C., Manso M.A., Semi-Automatic Metadata Extraction from Imagery and Cartographic Data. *IGARRS'07*, 2007, 3051-3052,

Haas SW, Pattuelli MC, Brown RT. Understanding Statistical Concepts and Terms in Context: The GovStat Ontology and the Statistical Interactive Glossary. In: *Proceedings of the Annual Meeting of the American Society of Information Science and Technology* Vol. 40 pp. 193-199, 2003

Espon, Espon First Interim Report, February 2009, 162 p. http://www.espon.eu/export/sites/default/Documents/Projects/ScientificPlatform/ESPON Database2013/fir_espondb_2013_27-02-09.pdf

Espon TR1, Espon Technical Report "Acquisition and Storage of data and metadata in ESPON 2013 DB", February 2010, 30 p., *to appear*

Espon TR2, Espon Technical Report "Web Application Report", February 2010, 22 p. *to appear*

Goux D., Une histoire de l'Enquête Emploi, *Economie et Statistique*, N° 362, 2003

Kent J-P. and Schuerhoff M. Some Thoughts About a Metadata Management System. In *Proc. 9th International Conference on Scientific and Statistical Database Management*, Olympia, Washington, 1997, pp. 174–185

Kieffer A., Oberti M., Preteceille E., Enjeux et usages des categories socioprofessionnelles en Europe, *Sociétés Contemporaines,* n°45-46, 2002, pp. 5-15

McCarthy J., Metadata Management for Large Statistical Databases, In *Proc. 8th International Conference on Very Large Database,* 1982, pp. 234-243

OECD, Handbook on construction composite indicators : methodology and user guide. ISBN 978-92-64-04345-9, 2008

ONU-ECE, Guidelines for the modeling of statistical data and metadata, Geneva, 1995, www.unece.org/stats/publications/metadatamodeling.pdf

Plumejeaud C., Gensel J., Villanova-Oliver M., Ben Rebah M., Vergnaud G., Modélisation de hiérarchies territoriales multiples, *Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2009)*, Paris, France, November 25-27 2009

Servigne S., Lesage N., Libourel T. Spatial data quality components, standards and metadata. Fundamentals of Spatial Data Quality. International Scientific and Technical Encyclopedia. ISBN 1905209568. pp. 179-210, 2006.