Apport automatisé de sémantique lors de manipulations de documents géographiques

Michel Mainguenaud, Souissi Nissrine

Apport automatisé de sémantique lors de manipulations de documents géographiques

Mainguenaud Michel [*] **– Souissi Nissrine** [**]

[*] *Institut National des Sciences Appliquées de Rouen*
Place Emile Blondel
F-76131 Mont-Saint-Aignan Cedex - France

[**] *Université Paris XII - Val de Marne*
61, avenue du Général de Gaulle
94010 Créteil Cedex – France

michel.mainguenaud@insa-rouen.fr
nissrine.souissi@insa-rouen.fr

RÉSUMÉ. *Comment donner des informations riches sémantiquement à un décideur pour prendre une décision? Telle est la question à laquelle nous essayons de répondre au travers de l'étude de trois approches de production de documents sémantiquement riches, cohérents et exploitables. L'originalité de notre étude repose sur une méthode dirigée à la fois par la représentation spatiale et par une approche conceptuelle, ce qui nous permet de l'appliquer à différents problèmes thématiques.*

ABSTRACT. *How to provide data with a high semantic to a decision planner ? We define three approaches to produce documents from intial one(s) in order to increase their semantics. The originality of the approach is based on a method driven at the same time by the spatial representations and a conceptual approach. This method can be used in several other fields.*

MOTS-CLÉS : *Modèle de données, Opérateur de manipulation, représentation spatiale.*

KEYWORDS : *Data model, Manipulation operators, spatial representation*

1. Introduction

L'approche participative est re-apparue avec la volonté assez récente des gouvernements d'intégrer la dimension « implication des populations » aux politiques de développement socio-économique. Il ne s'agit pas d'une fin en soi, mais d'un ensemble méthodologique (utilisant une série d'outils) qui contribue à ce développement.

L'approche participative tend en réalité à modifier la perception du rôle de chacun des intervenants (état, services techniques, populations...). Elle propose un partage de responsabilités entre les différents partenaires. Elle favorise la prise de décision et la prise en charge par les populations des différentes actions menées. L'objectif est d'associer et d'impliquer étroitement les populations aux différents niveaux et étapes du processus (élaboration du diagnostic, analyse des diverses contraintes et priorités, conception et programmation des actions à entreprendre…). Ce processus est vu comme une séquence d'opérateurs conduisant à la réalisation de chacune de ces étapes. Avancer dans ce processus, consiste à passer d'une étape à une autre par la filiation des documents. Ces derniers génèrent de nouveaux documents (dits produits) qui nous intéressent ici tout particulièrement. Ils représentent le cœur de notre processus. Il est nécessaire de créer des documents produits synthétiques à forte valeur ajoutée, outils de réflexion et d'action pour les décideurs.

Nous proposons une approche visant à produire de l'information cohérente et à la rendre plus riche et exploitable. Ces deux actions sont très différentes et ne font pas appel aux mêmes compétences. Ainsi de nouveaux modes de collaboration ont été inventés entre les communautés de *producteurs d'information* et les *spécialistes de l'information*.

L'étude présentée dans cet article, s'inscrit dans le cadre du projet JOYSTIC (appel d'offre Société de l'Information du CNRS). Ce projet vise à intégrer l'approche participative aux politiques de développement territorial. Les documents produits sont des représentations spatiales (cartes, croquis, images, photographies…) élaborées au cours du diagnostic de territoire. Elles sont à la fois support d'analyse et objets de dialogue permettant aux acteurs du territoire (élus, acteurs locaux…) d'énoncer leur vision de l'espace.

Dans la section 2, nous présentons un état de l'art de travaux actuels ayant les mêmes objectifs que nous mais n'abordant pas forcément la même problématique. Dans la section 3, nous définissons les constantes sémantiques que nous associons aux attributs alphanumériques dans notre modèle de données. Dans la section 4, 5 et 6, nous présentons respectivement les trois approches de production de documents par enrichissement : interprétée, compilée et hybride. Nous terminons par une conclusion et des perspectives.

2. Etat de l'art

Nous exposons, dans cette section, trois méthodes différentes ayant l'objectif de produire un document riche (complet) et exploitable. Dans une première partie, nous présentons le principe d'utilisation de la clause GROUP BY dans une requête SQL. Dans une deuxième partie, nous présentons une méthode de réécriture d'un catalogue de situations cliniques dans un formalisme exploitable. Dans une troisième partie, nous présentons une méthode de construction semi-automatique d'une ontologie de domaine. Nous terminons cette section en donnant, dans une quatrième partie, l'apport et l'originalité de notre proposition par rapport à celles proposées dans la littérature.

2.1. *Requête SQL : clause GROUP BY*

Comme c'est le cas dans la plupart des langages relationnels modernes, SQL (Date *et al*., 1997) est fondé sur le calcul relationnel de tuples et sur l'algèbre relationnelle. SQL offre toutefois des possibilités dépassant celles de l'algèbre de base et du calcul relationnels tels les opérateurs d'agrégation. Ces opérateurs s'appliquent à l'ensemble des valeurs, en ignorant les valeurs nulles, d'une colonne (à l'exception de l'opérateur *count* qui peut s'appliquer au niveau tuple) et génèrent une seule valeur en sortie.

SQL admet notamment la partition des tuples d'une table dans des groupes. L'un des opérateurs d'agrégation fournit par SQL peut alors leur être appliqué. La valeur de l'agrégat n'est plus calculée sur toutes les valeurs de la colonne spécifiée mais sur toutes les valeurs d'un groupe. L'opérateur est ainsi évalué pour chaque groupe. Le partitionnement des tuples dans les groupes est réalisé en utilisant la clause GROUP BY suivi d'une liste d'attributs qui les définissent.

Suivant le niveau de détail souhaité, les attributs traités par GROUP BY peuvent aussi apparaître dans la liste de sélection. Tous les attributs supplémentaires qui n'apparaissent pas dans la clause GROUP BY, ne peuvent être sélectionnés qu'en utilisant une fonction d'agrégation.

La plupart du temps, le moteur de requêtes avertit l'utilisateur d'une *probable incohérence* de calcul des agrégats à l'aide d'un message d'erreur (e.g. avant exécution de la requête SELECT <Attribut$_1$>, <Attribut$_2$>, …, <Attribut$_n$> FROM … WHERE … GROUP BY <Attribut$_1$>). Le message est de la forme : « La colonne <Attribut$_2$> est incorrecte dans la liste de sélection parce qu'elle n'est pas contenue dans une fonction d'agrégation et ne figure pas dans la liste des attributs de la clause GROUP BY ». Cette requête est non valide en SQL puisque le Système de Gestion de Bases de Données (SGBD) ne gère pas les dépendances fonctionnelles. Elle ne sera donc pas exécutée. Si une dépendance fonctionnelle existe entre l'attribut de partitionnement et chacun des attributs présents dans la clause SELECT mais absent de la clause GROUP BY, la requête est sémantiquement correcte. Autrement elle

génèrerait des résultats incohérents (la relation résultat n'est pas en 1ère forme normale).

Le SGBD arrive à identifier les attributs manquants mais n'est pas en mesure de les intégrer automatiquement. L'utilisateur doit les ajouter manuellement. A ce sujet, nous pouvons dire finalement que le moteur de requêtes SQL génère des résultats cohérents mais pas forcément pertinents et exploitables. Les requêtes SQL potentiellement valides sont rejetées.

2.2. *Réécriture d'un catalogue de situations cliniques*

Edités dans un format documentaire, les Guides de Bonnes Pratiques (GBP) (Sackett *et al*., 1996) se présentent sous la forme d'un catalogue de situations cliniques théoriques. Chacune d'elles est associée aux conduites à tenir préconisées. Les versions numériques de ces documents textuels sont actuellement largement accessibles sur les réseaux. Néanmoins, elles n'ont aucun impact sur le comportement des médecins (Matillon *et al*., 2000). Seul l'encadrement des pratiques avec la suggestion d'une recommandation centrée-patient au moment de la décision médicale permet de modifier la décision du médecin.

Du fait d'un processus éditorial linéaire et du style littéraire adopté, les GBP apparaissent comme des *documents textuels* souvent incomplets, toujours ambigus (Tierney *et al*., 1995). La traduction formelle reste délicate. Les récents travaux de Shiffman (Shiffman *et al*., 2000) - centrés sur le document - sur l'élaboration d'un modèle documentaire ou *Guideline Elements Model* (GEM) permettent la structuration logique des documents textuels représentant les GBP. Le modèle GEM, basé sur une *Document Type Definition* (DTD) XML, vise à développer un filtre logique des éléments d'information des GBP textuels afin de faciliter la construction des bases de connaissances.

Des travaux explorant l'implémentation d'un système de guidelines à partir de l'instance, sont actuellement en cours d'élaboration. Dans ce sens, l'objectif de (Georg *et al*., 2005) consiste à créer une instance GEM du GBP canadien (Feldman *et al*., 1999) sous contrainte de la DTD. La première étape de la création de cette instance est le marquage du texte via des éléments : *decision.variable* pour décrire les éléments de la décision, *action* pour décrire le traitement recommandé… Pour la création de l'instance, une étape de normalisation des termes caractérisant l'état du patient et les traitements était nécessaire. A partir de l'instance GEM du GBP textuel, l'objectif a été ensuite de construire une base de règles. Cette base permet la réécriture du GBP textuel dans un formalisme exploitable. La forme de référence est : SI X et Non Y ALORS Z AVEC NP. X correspond à l'état du patient, au traitement courant et à la réponse au traitement courant. Y correspond aux pathologies ne devant pas être présentes dans le profil clinique. Z correspond au traitement proposé et NP correspond au niveau de preuve de la recommandation.

La construction de la base de règles repose sur une étape préliminaire d'identification des éléments *decision.variable*, *action*... dans l'instance GEM. L'objectif est donc de localiser ces éléments et d'en extraire le contenu. Ce traitement est réalisé par un *parser* SAX pour *Simple API for XML* (http://www.megginson.com/SAX/).

Le document produit (ensemble de règles) est certes exploitable mais *incomplet*. Le GBP initial se représente comme un ensemble de documents incomplets, mais *rien n'a été fait dans ce sens pour les enrichir*. L'approche présentée fait de la réécriture du GBP textuel initial dans un formalisme exploitable.

2.3. Construction semi-automatique d'ontologies

La construction manuelle d'une ontologie, même assistée par des outils conviviaux, est un travail de modélisation long et difficile. Chercher à automatiser ce processus de construction est primordial. Dans ce sens, (Giraldo *et al.*, 2002) propose de construire une ontologie en deux étapes. Une version initiale très simple de l'ontologie est construite à la main pour fixer le point de vue à adopter pour classer les concepts du domaine étudié. Elle est simple dans la mesure où elle ne correspond qu'à deux niveaux d'une hiérarchie de concepts. Cette ontologie initiale est ensuite semi-automatiquement enrichie en trois phases : une phase d'extraction, une phase de structuration et une phase de représentation.

La phase d'extraction consiste à extraire un ensemble de termes-classes, de termes-propriétés et de relations par application d'heuristiques. Une classe est vue comme une représentation abstraite d'un ensemble d'objets « complexes » repérés, dans les DTDs, par le fait qu'il s'agit d'éléments décomposables. L'approche repose sur l'exploitation de nombreuses DTDs représentatives du domaine. Leurs composants devraient apparaître dans au moins une des DTDs de l'échantillon. En suivant ce raisonnement, une méthode de repérage des termes associés aux propriétés dans les DTDs a été obtenue. Il s'agit des éléments non décomposables. Ce processus d'extraction s'accompagne par ailleurs de traitements éliminant les doublons, extrayant les différents mots d'expressions composées, remplaçant les abréviations par leur signification en clair et éliminant certains termes jugés non pertinents.

Le processus complet d'extraction est semi-automatique. Les termes extraits automatiquement sont présentés à l'utilisateur. Il juge, après une lecture rapide, de l'opportunité d'enrichir les fichiers de termes « non pertinents » et des abréviations. Il peut alors demander à exécuter de nouveau le processus d'extraction automatique pour prendre en compte les nouvelles versions de ces fichiers. Une fois le contenu de ces fichiers arrêté, le concepteur analyse de façon plus fine les résultats produits par le logiciel. Cette opération peut conduire à éliminer des termes, modifier certains noms, transformer certains termes-classes en termes-propriétés.

La phase de structuration consiste à construire, pour chacun des concepts composant la hiérarchie initiale, un réseau de relations les liant à d'autres classes du domaine. Chacune des classes est caractérisée par des propriétés. Le problème de structuration correspond alors à un problème de fusion de hiérarchies partielles extraites automatiquement avec une ébauche de hiérarchies de classes construites manuellement. Les classes qui ne peuvent être placées dans la hiérarchie initiale sont des éléments candidats pour faire partie d'autres hiérarchies.

Cette approche permet d'enrichir partiellement l'ontologie initiale par de nouveaux termes, dont la pertinence pour l'ontologie doit être validée manuellement par le concepteur. Elle fait intervenir également l'utilisateur pour fournir une liste d'abréviations et leurs explications pour compléter l'information. Actuellement, les techniques de construction des autres hiérarchies non existantes dans l'ontologie, ne sont pas réalisées. Ainsi, si aucune classe ne correspond à une hiérarchie existante elle est systématiquement rejetée.

2.4. *Synthèse*

Les trois travaux décrits, ont souligné des insuffisances observées sur les données de base issues des documents produits. Mais aucun d'eux ne propose une méthode qui permet de compléter ces documents. Dans le premier cas, le SGBD ne dispose pas d'une fonction lui permettant l'intégration automatique d'attributs manquants pour compléter la requête SQL. Néanmoins, il garantit la cohérence des résultats. Dans le deuxième cas, l'approche proposée n'enrichit ou complète en aucun cas le Guide des Bonnes Pratiques initial. Ainsi, même si le nouveau GBP est exploitable, il est incomplet. Il serait plus intéressant d'enrichir d'abord le GBP structuré avant de le réécrire dans un formalisme exploitable. Dans le troisième cas, des termes-classes peuvent être rejetés et ne pas enrichir l'ontologie initiale si celle-ci ne comporte aucune hiérarchie qui correspond à ces termes-classes.

A la différence de ces travaux, notre approche permet l'enrichissement dynamique et transparent de documents afin de diminuer ces insuffisances. Le principe retenu est l'intégration automatique des données ignorées lors de la génération des documents produits. Notre objectif est de répondre de manière satisfaisante à différents problèmes thématiques et d'aider les décideurs dans leur prise de décision. Le décideur n'est pas impliqué dans ce processus d'enrichissement de documents. Nous définissons la notion de constante sémantique et formalisons son usage dans le processus d'enrichissement de documents. L'adjonction d'attributs (plus-value sémantique) dépend de la sémantique de l'opérateur et des valeurs des constantes sémantiques définies sur les opérandes.

3. Formalisation des constantes sémantiques

Le processus d'enrichissement des documents s'appuie sur la notion de constante sémantique. Nous définissons une constante sémantique comme une information supplémentaire associée à un attribut alphanumérique. Chaque représentation spatiale, aussi structurée soit elle, n'est pas une information isolée et suffisante du système d'information car par nature elle est incomplète. Une autre catégorie d'information souvent négligée est aussi disponible: aux représentations spatiales sont en effet associés des attributs alphanumériques. Bien que de composition hétérogène par rapport à la représentation, ils présentent un complément d'information particulièrement riche.

Les documents manipulés sont étiquettés. Une étiquette est formalisée par un ensemble d'attributs associés à un document. Chaque attribut suit le même formalisme que précédemment : un couple (attribut, domaine). Nous ne nous attachons pas au formalisme utilisé pour gérer cet ensemble d'attribut (relationnel, orienté objet, …). Nous nous plaçons au niveau conceptuel. La justification de cette constante sémantique s'appuie sur le fait que les opérateurs de manipulation ne connaîtront pas *a priori* les documents manipulés. L'opérateur de projection (relationnel) est l'exemple de l'opérateur permettant de retenir à partir d'un ensemble d'attributs, un sous-ensemble pertinent pour une requête. La démarche présentée par cette constante est similaire en ayant pour objectif d'automatiser cette opération de détermination du sous-ensemble des attributs pertinents. La définition d'un attribut devient donc un triplet (attribut, domaine, constante sémantique). L'approche d'augmentation du schéma a été initialement introduite dans (Barrera *et al*. 1981) dans le domaine des images.

Pour illustrer notre approche, nous utilisons l'objet géographique Ville décrit par les attributs alphanumériques : Nom, Population, Surface, Densité et MurExtérieur. A l'objet géographique Ville est associée une représentation spatiale gérée par un attribut défini sur un type spatial. La notion de type abstrait nous permet de faire complètement abstraction du modèle de représentation. Ce type peut être lui même structuré avec des attributs. Nous présentons dans cette section les différentes classes de ces constantes sémantiques par la définition et le rôle de chacune d'elles.

3.1. *Les constantes sémantiques spatiales*

L'objectif est d'établir des liens entre les attributs alphanumériques et la représentation spatiale. Ces liens peuvent se présenter sous la forme de dépendances fonctionnelles, multivaluées ou peuvent être basés sur la sémantique du type spatial. Deux constantes indépendantes : Topologie et Granularité sont définies pour établir des liens basés sur la sémantique spatiale entre attributs alphanumériques et représentation spatiale. Les valeurs de ces constantes sémantiques sont définies à

partir des propriétés topologiques d'une représentation spatiale classique sans région composite ni trous. Les valeurs d'instances sont fixées par le concepteur-BD.

La Topologie précise la validité conceptuelle d'un attribut pour la limite (frontière), l'intérieur ou la globalité (limite et intérieur) de la représentation spatiale.

La Granularité indique la validité conceptuelle d'un attribut pour un sous-ensemble ou l'intégralité de cette représentation.

Les constantes de types sont donc pour Topologie (Limite –L-, Intérieur –I- ou Globalité –G-) et pour Granularité (Sous-Ensemble, Intégralité).

Le tableau 1 illustre l'instanciation des constantes Topologie et Granularité avec les attributs Nom et Surface de l'objet Ville. A titre d'exemple, le nom est une information valide pour n'importe quelle partie (sous-ensemble) de la représentation spatiale de la ville (a fortiori pour la représentation spatiale complète). La surface (considérée ici comme un attribut et non un attribut calculé) ne sera pertinente que pour l'intégralité de la représentation spatiale (la surface d'un sous-ensemble quelconque de la représentation spatiale serait inférieure ou égale à la valeur disponible).

Attribut	**Domaine**	**Constante sémantique**	
		Topologie	**Granularité**
Nom	Chaîne	Globalité	Sous-Ensemble
Surface	Réel	Intérieur	Intégralité

Tableau 1. *Schéma associé aux attributs Nom et Surface de Ville*

3.2. *Les constantes sémantiques métiers*

L'objectif est de permettre au concepteur-métier de relier sémantiquement les attributs alphanumériques entre eux par des relations de pertinence indépendamment de la représentation spatiale. Deux constantes sémantiques AttacheIntra et AttacheInter sont définies. Les valeurs d'instance sont fixées par le concepteur-métier.

La constante sémantique AttacheIntra indique si un (ou plusieurs) attribut(s) apporte une information supplémentaire et pertinente à un autre attribut, à condition que ces attributs appartiennent à la même étiquette. Ceci est exprimé par une relation de pertinence unidirectionnelle entre deux attributs a et b de l'étiquette d'un document « a -> b ». Notons ici, que plusieurs relations de pertinence peuvent éventuellement être définies pour a. Si à un attribut a, AttacheIntra a la valeur « *Nulle* » alors il n'existe aucun attribut du document source qui soit pertinent pour a.

La constante sémantique AttacheInter fait intervenir la notion de document dérivé. Un document dérivé d'un document source est un document dont l'étiquette contient en plus de l'ensemble des attributs de l'étiquette du document source, d'autres attributs supplémentaires (à l'image des sous-classes dans les modèles orienté-objet). Cette constante indique si un (ou plusieurs) attribut(s), des documents dérivés du document source, apporte une information supplémentaire et pertinente à un autre attribut (appartenant au document source).

Cette notion est importante dans le cadre de la généricité d'une approche par opérateur. Dans le monde orienté objet, la signature d'un opérateur peut être utilisée avec une sous-classe d'une classe appartenant en partie gauche de la signature sans pour autant que la méthode ait été redéfinie. Cette constante sémantique permettra de s'affranchir de la redéfinition sur le plan du schéma. Bien évidemment aucune action sur le plan de la dynamique ne sera possible (manipulations effectuées sur les attributs de la sous-classe - seules les valeurs de l'étiquette du document source seront propagées). Le tableau 2 illustre l'instanciation des constantes sémantiques métiers associées aux attributs Nom et Surface de l'objet Ville. Le schéma de Ville est construit sur une agrégation d'informations hétérogènes. Elles n'ont donc aucune raison d'être interdépendantes. Le choix fait dans cet exemple est de montrer une dépendance entre la surface et le nom (AttacheIntra) mais la présence d'une information comme par exemple la date d'édification initiale de la ville n'aurait aucun lien avec la surface (actuelle). Ceci n'entraînerait donc pas d'introduction de dépendance sémantique (AttacheIntra). Pour simplifier la présentation nous n'introduisons pas de document dérivé.

Attribut	**Domaine**	**Constante sémantique**	
		AttacheIntra	**AttacheInter**
Nom	Chaîne	Surface	Nulle
Surface	Réel	Nom	Nulle

Tableau 2. *Schéma des attributs de l'objet Ville pour les constantes métier*

3.3. *Les constantes sémantiques structurelles*

L'objectif est de permettre de garantir la cohérence de l'étiquette du document produit enrichi. Deux constantes sémantiques Existence et Nature sont définies. Les instances sont définies par le concepteur-BD.

La constante sémantique Existence indique l'indépendance ou la dépendance d'un attribut envers les autres attributs de l'étiquette. Ceci est exprimé par une relation de dépendance existentielle (RDE) « a -> b » qui signifie que l'existence de a dans l'étiquette du document produit final (enrichi) dépend de la présence de b.

La constante sémantique Nature indique la modalité d'obtention de la valeur d'un attribut : *fixe* ou *calculé* à partir de la représentation spatiale d'un objet géographique. Le tableau 3 illustre l'instanciation des constantes sémantiques structurelles des attributs Nom et Surface de l'objet Ville.

Attribut	**Domaine**	**Constante sémantique**	
		Existence	**Nature**
Nom	Chaîne	Indépendant	Fixe
Surface	Réel	Nom	Calculé

Tableau 3. *Instanciation des constantes sémantiques Existence et Nature*

Le tableau 4 propose un récapitulatif pour l'objet géographique « Ville ».

Attribut / Constante	**Nom**	**Population**	**Surface**	**Densité**	**Mur Extérieur**
AttacheInter	Nulle	Nulle	Nulle	Nulle	Nulle
AttacheIntra	Population, Surface	Nom	Nom	Nom, Population, Surface	Nom
Existence	Indépendant	Nom	Nom	Nom, Population	Nom
Nature	Fixe	Fixe	Calculé	Calculé	Fixe
Topologie	Globalité	Intérieur	Intérieur	Intérieur	Limite
Granule	Sous-ensemble	Intégralité	Intégralité	Intégralité	Sous-ensemble

Tableau 4. *Récapitulatif des constantes sémantiques pour l'objet Ville*

4. Approche interprétée

La signature d'un opérateur de manipulation se définit à partir du type des opérandes qu'il utilise (appelé partie gauche) et du type du résultat qu'il produit (appelé partie droite). Dans une approche interprétée, le typage de la partie droite de la signature d'un opérateur n'est déterminé qu'à l'exécution. La structure du document à produire est donc inconnue *a priori* pour cet opérateur. L'étiquette du document produit est générée en utilisant les valeurs des constantes sémantiques *Topologie*/*Granularité*. Cette opération détermine les attributs pertinents pour la représentation spatiale en fonction de la sémantique de l'opérateur (Mainguenaud, 1994). Dans cette section, nous décrivons le principe de cette approche et nous l'illustrons par un exemple.

4.1. *Principe*

La sélection intentionnelle, au niveau schéma, a pour but de constituer la partie intentionnelle de la base. C'est une projection au sens relationnel (Flory *et al*., 1996). Elle utilise les valeurs des Constantes Sémantiques Topologie/Granularité pour déterminer les attributs pertinents. À chaque opérateur est défini un tableau déterminant les caractéristiques des constantes sémantiques à retenir. Les attributs présentant ces caractéristiques formeront l'étiquette finale. Le tableau se présente comme une fonction f de l'ensemble (Topologie x Granularité x Opérateur) vers l'ensemble ((Topologie x Granularité) $\cup$ ¬Pertinent), avec ¬Pertinent représentant le terme spécifiant la non pertinence de la caractéristique de la constante sémantique. Cette approche permet de transférer la sémantique de l'attribut dans un *espace intermédiaire* (Topologie x Granularité). À chaque attribut pris dans l'ensemble fini des attributs du système, correspond un point unique dans cet espace. La définition de ce point correspond aux ajouts sémantiques introduits dans la définition du schéma d'un attribut. Cette définition représente une application injective dont le but est de définir la pertinence d'un attribut suivant la représentation de l'espace de l'objet.

4.2. *Exemple*

Pour illustrer cette approche, nous étendons la base de données composée des objets géographiques {Ville, EspaceVert, Forêt} où Forêt est dérivée de EspaceVert. Nous ne présentons pour Forêt que les attributs supplémentaires dans la mesure où elle contient tous les attributs de EspaceVert. Le tableau 5 expose le schéma de Forêt et EspaceVert (cf. tableau 4) ainsi que les classifications associées (pour simplifier sans les domaines).

Forêt			
Attribut	**Observateur**	**GardeBarrière**	
Topologie	Intérieur	Limite	
Granule	Sous Ensemble	Sous Ensemble	
EspaceVert			
Attribut	**Nom**	**HautMaxAutorisée**	**Périmètre**
Topologie	Globalité	Globalité	Limite
Granule	Sous Ensemble	Sous Ensemble	Intégralité

Tableau 5. *Affectation des constantes sémantiques pour Forêt et EspaceVert*

Pour illustrer le mécanisme de sélection intentionnelle, nous présentons un exemple avec l'opérateur d'intersection spatiale ($\cap$). Le tableau 6 définit les règles de cette sélection pour cet opérateur et expose pour chaque cellule, la pertinence ou non d'une classification. La partie graphique de l'opérateur d'intersection entre deux

objet (ici deux représentations spatiales) retourne par définition un sous-ensemble des représentations spatiales (la partie commune aux deux représentations). Les informations uniquement valides sur l'intégralité de la représentation ne pourront donc pas être pertinentes pour le résultat (quelle que soit la topologie retenue). Les informations valides pour un sous-ensemble le resteront avec leur topologie associée. Le tableau 7 expose le schéma de ∩ (Ville, EspaceVert).

Topologie / Granule	Intégralité	Sous-Ensemble
Limite	¬Pertinent	(Limite, Sous-Ensemble)
Intérieur	¬Pertinent	(Intérieur, Sous-Ensemble)
Global	¬Pertinent	(Global, Sous-Ensemble)

Tableau 6. *Règles de sélection intentionnelle de l'intersection* ∩

Attribut	Topologie	Granularité
Ville.Nom	Global	Sous-Ensemble
Ville.MurExtérieur	Limite	Sous-Ensemble
EspaceVert.Nom	Global	Sous-Ensemble
EspaceVert.HautMaxAutorisée	Intérieur	Sous-Ensemble

Tableau 7. *Schéma de l'étiquette résultant de* ∩ *(Ville, EspaceVert)*

Pour cet opérateur, seuls les attributs qualitatifs comme les noms ont été gardés. Par contre la Population ou la Surface n'ont pas été retenues. L'étiquette du document produit décrivant la nouvelle entité est certes pertinente pour la représentation spatiale mais elle est incomplète dans la mesure où elle ne contient pas d'attributs calculés.

5. Approche compilée

Pour contourner le problème soulevé dans l'approche interprétée, l'approche compilée permet à l'opérateur de connaître au préalable le type du résultat. L'étiquette du document à produire est donc connue. Dans cette section, nous décrivons le principe de cette approche et nous l'illustrons par un exemple.

5.1. *Principe*

L'approche compilée comporte deux phases. La première consiste à produire un document initial (DPI^c) par un opérateur connaissant au préalable la structure de l'étiquette, à partir d'un (ensemble de) documents sources. La deuxième consiste à compléter le schéma associé au document résultat par le transfert des attributs sources pertinents. La figure 2 représente la deuxième phase, le processus du transfert des attributs sources pertinents. Trois étapes sont nécessaires. La première

étape consiste à identifier les attributs sources théoriquement pertinents. La deuxième consiste à vérifier la pertinence réelle de ces attributs. La troisième étape consiste à analyser la relation de dépendance existentielle. Nous détaillons l'utilisation de la Figure 2 à partir de l'exemple développé dans la section suivante.

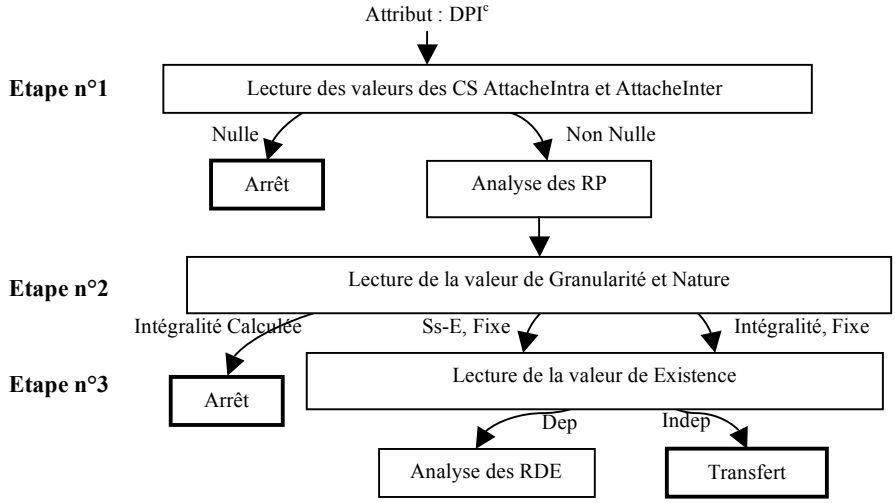

Figure 2. *Règles de transfert – approche compilée*

5.2. *Exemple*

Le tableau 8 complémente les constantes sémantiques pour les attributs de Forêt et EspaceVert (cf. tableau 5). Nous utilisons l'opérateur d'intersection spatiale (∩) pour générer l'étiquette du document produit initial dont la structure est composée des attributs : Ville.Nom, Ville.Surface, EspaceVert.Nom. La signature formelle de cet opérateur (Ville x Forêt → DPI^c) est différente de sa signature effective (Ville x EspaceVert -> DPI^c). Mais, son exécution est possible dans la mesure où Forêt est dérivée de EspaceVert. Grâce à AttacheInter, il est possible de récupérer automatiquement des attributs et leurs instances de Forêt pour les retranscrire dans l'étiquette du document produit final.

L'étape suivante consiste à déterminer pour chacun des attributs de DPI^c, les attributs sources pertinents. Par exemple, pour l'attribut Ville.Nom : AttacheIntra permet d'identifier deux attributs sources Ville.Population et Ville.Surface. La valeur de (Granule, Nature) pour Ville.Population est égale à (Intégralité, Fixe). Celui-ci est réellement pertinent. La classification existentielle de Population marque le lien avec le nom de la ville. : Population.Existence = Dep (Ville.Nom). Le nom de la ville apparaît dans DPI^c. L'attribut Ville.Population est donc transféré. (idem pour l'attribut EspaceVert.Nom : AttacheInter permet d'identifier deux attributs dérivés sources Forêt.Observateur et Forêt.GardeBarrière). Pour ces deux

attributs, la classification est (Granule, Nature) = (Sous-Ensemble, Fixe). La pertinence réelle est vérifiée. De par la caractéristique Existence = Indep ces deux attributs sont transférés. Nous exposons dans le tableau 9 le schéma de ∩ (Ville, EspaceVert) après enrichissement.

Forêt		
Attribut	**Observateur**	**GardeBarrière**
AttacheIntra	Nulle	Nulle
AttacheInter	Nulle	Nulle
Nature	Fixe	Fixe
Existence	Indépendant	Indépendant

EspaceVert			
Attribut	**Nom**	**HautMaxAutorisée**	**Périmètre**
AttacheIntra	Nulle	Nom	Nom
AttacheInter	Forêt : Observateur, GardeBarrière	Forêt : Nulle	Forêt : Nulle
Nature	Fixe	Fixe	Calculé
Existence	Indépendant	Nom	Nom

Tableau 8. *Classifications associées à Forêt et EspaceVert*

Attribut	AttacheIntra	AttacheInter	Granule	Nature	Existence
Ville					
- Nom	Surface, Population	Nulle	Sous-Ensemble	Fixe	Indépendant
- Surface	Nom	Nulle	Intégralité	Calculé	Nom
- Population	Nom	Nulle	Intégralité	Fixe	Nom
EspaceVert					
- Nom	Forêt Observateur Gardebarrière	Nulle	Sous-Ensemble	Fixe	Indépendant
Forêt					
- Observateur	Nulle	Nulle	Sous-Ensemble	Fixe	Indépendant
GardeBarrière	Nulle	Nulle	Sous-Ensemble	Fixe	Indépendant

Tableau 9. *Résultat de l'enrichissement de ∩ (Ville, EspaceVert)*

Pour fournir au décideur un document riche en information, l'opérateur doit puiser, en dehors de Ville et EspaceVert, ces informations à partir des documents dérivés sources (Forêt). Ceci est possible grâce à la constante sémantique AttacheInter. Le problème soulevé ici, concerne la présence d'attributs qui ne sont pas sémantiquement pertinents pour l'opérateur (Ville.Population). En effet, la valeur de Ville.Population est incohérente pour la représentation spatiale de l'objet résultant de l'opération d'intersection entre Ville et EspaceVert. Le transfert des attributs sources pertinents se fait uniquement en fonction du schéma, ce qui peut engendrer une étiquette finale non cohérente. L'approche hybride a pour objectif de pallier ces incohérences potentielles.

6. Approche hybride

L'approche hybride combine les deux approches précédente. Le processus de l'approche hybride comporte deux phases. La première phase correspond à la génération parallèle de l'étiquette issue pour le document produit par l'approche compilée et par l'approche interprétée. La deuxième phase correspond à la génération de l'étiquette du document final, i.e. retranscription des attributs issus des approches compilée et interprétée dans l'étiquette du document produit final. Il y a ensuite transfert des attributs sources pertinents. Ce transfert est réalisé en trois étapes : la première est identique à celle de l'approche compilée. La deuxième étape repose sur la valeur du couple **(**Granularité, Nature**)** pour un attribut jugé théoriquement pertinent. Elle consiste à faire un appariement de la valeur de Granule définie pour cet attribut avec celle définie pour l'opérateur. La troisième est identique à celle de l'approche compilée.

Nous utilisons l'opérateur (∩) pour générer DPI^{I} (cf. Tableau 7) et DPI^{c} dont la structure est composée des attributs : Ville.Nom, Ville.Surface, EspaceVert.Nom. Nous ne décrivons que la 2ème étape de la 2ème phase de l'approche hybride, qui consiste à vérifier la valeur de Granule pour chacun des attributs sources jugés théoriquement pertinents pour les attributs de $DPI^{c} \cup DPI^{I}$.

La valeur de Granule pour *Ville.Population* est égale à Intégralité et celle définie pour l'opérateur (∩) est égale à Sous-Ensemble, l'attribut *Ville.Population* ne peut donc être transféré. Pour les deux autres attributs Forêt.Observateur et Forêt.GardeBarrière, la valeur de Granule est égale à Sous-Ensemble et celle de Existence est égale à Indépendant ce qui autorise le transfert de ces deux attributs. Comme exemple de résultats, nous présentons dans le tableau 10 le schéma de ∩ *(*Ville, EspaceVert) après enrichissement.

Grâce à l'approche hybride, nous avons obtenu une étiquette finale riche en attributs pertinents à la fois pour la représentation spatiale et pour les attributs des documents initiaux. Nous avons réussi à éliminer le risque d'incohérence dans la

mesure où tout attribut qui n'est pas sémantiquement pertinent pour l'opérateur n'est pas transféré.

Attribut	AttacheIntra	AttacheInter	Granule	Nature	Existence
Ville					
- Nom	Surface	Nulle	Sous-Ensemble	Fixe	Indépendant
- Surface	Nom	Nulle	Intégralité	Calculé	Nom
- MurExtérieur	Nulle	Nulle	Sous-Ensemble	Fixe	Nom
EspaceVert					
- Nom	Forêt : Observateur, GardeBarrière	Nulle	Sous-Ensemble	Fixe	Indépendant
- HautMaxAutorisée	Nom	Nulle	Sous-Ensemble	Fixe	Nom
Forêt					
- Observateur	Nulle	Nulle	Sous-Ensemble	Fixe	Indépendant
- GardeBarrière	Nulle	Nulle	Sous-Ensemble	Fixe	Indépendant

Tableau 10. *Résultat de l'enrichissement de* ∩ *(Ville, EspaceVert)*

7. Conclusion

A notre connaissance, peu de travaux étudient l'aspect dynamique et transparent de l'enrichissement de l'information. Nous avons présenté dans cet article trois approches de production de documents sémantiquement riches à partir d'un ensemble de documents sources hétérogènes. La technique utilisée se résume en une sélection de sémantique. Notre contribution s'est orientée sur l'enrichissement dynamique et transparent des documents produits. L'approche interprétée utilise les constantes sémantiques centrées sur la représentation spatiale pour générer une étiquette associée au document pertinente pour la représentation spatiale mais incomplète. L'approche compilée utilise les constantes sémantiques métiers pour transférer des attributs sources pertinents, néanmoins les données générées en sortie risquent d'être incohérentes. L'approche hybride permet de produire un document cohérent, complet et riche.

Comme perspectives, nous pouvons envisager deux axes de développement. Le premier concerne le développement des interfaces de saisie (graphiques) des valeurs

des constantes sémantiques (spatiales et métier). La seconde perspective concerne la pertinence des attributs pour le décideur, nous pouvons envisager ainsi le masquage de certains attributs non pertinents pour un profil donné.

8. Bibliographie

Barrera R., Buchmann A., « Schema definition and query language for geographical database system », *IEEE Transactions on Computer Architecture : Pattern Analysis and image Database Management*, vol. 11, p. 250-256, 1981.

Date C. J., Darwen H., *A Guide to the SQL Standard*, ISBN : 0-201-96426-0, Addison-Wesley, 1997.

Feldman R. D., Campbell N., Larochelle P., Bolli P., Burgess E. D., Carruthers S. G., Floras J. S., Haynes R. B., Honos G., Leenen F. H. H., Leiter L. A., Logan A. G., Myers M. G., Spence J. D., Zarnke K. B., « Recommandations de 1999 pour le traitement de l'hypertension artérielle au Canada », *Canadian Medical Association Journal*, vol. 161, n° 12, p. 1477-1624, 1999.

Flory A., Laforest F., *Les bases de données relationnelles*, Economica, 1996.

Georg G., Séroussi B., Bouaud J., « Extending the GEM model to support knowledge extraction from textual guidelines », *International Journal of Medical Informatics*, vol. 74, n° 2-4, p. 79-87, 2005.

Giraldo G., Reynaud C., « Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine », *13èmes journées francophones d'Ingénierie des Connaissances*, Rouen, 28-30 Mai, 2002.

Mainguenaud M., « Consistency of Spatial Database Query Results », *Computer Environment and Urban Systems*, vol. 18, n° 5, p. 333-342, Pergamon, 1994.

Matillon Y., Durieux P., « L'évaluation médicale, du concept à la pratique », *Medecine-Sciences*, Flammarion, p.43-54, 2000.

Sackett D. L., Rosenberg W. M., Gray J. A., Haynes R. B., Richardson W. S., « Evidence based medicine : what it is and what it isn't », *British Medical Journal*, vol. 312, n° 7023, p. 71-72, 1996.

Shiffman R. N., Karras B. T., Agrawal A., Chen R., Marenco L., Nath S., « GEM : a proposal for a more comprehensive guideline document model using XML ». *J Am Med Inform Assoc*, vol. 7, n° 5, p. 488-498, 2000.

Tierney W. M., Overhage J. M., Takesue B. Y., Harris L. E., Murray M. D., Varco D. L., Mcdonald C. J., « Computerizing guidelines to improve care and patient outcomes : the example of heart failure », *J Am Med Inform Assoc*, vol. 2, n° 5, p. 316-322, 1995.