Atelier DECOR « Découverte de Correspondances », EGC'2007

Appariement de schémas de BD géographiques à l'aide d'ontologies déduites des spécifications

Sébastien Mustière, Nathalie Abadie, Frédéric Laurens

Institut Géographique National
Laboratoire COGIT
2 av. pasteur
94160 SAINT-MANDE
sebastien.mustiere@ign.fr, nathalie-f.abadie@ign.fr
http://recherche.ign.fr/labos/cogit/

Résumé. L'intégration de base de données géographiques peut être facilitée par l'utilisation d'ontologies du domaine. Ces ontologies peuvent être déduites de l'analyse semi-automatique des spécifications textuelles de ces bases de données. Cet article présente des travaux en cours et les principales difficultés qui apparaissent pour réaliser cela.

1 Introduction

De nombreuses bases de données géographiques coexistent pour représenter un même espace. Ces bases ont été réalisées pour répondre à différents besoins (urbanisme, navigation…) et possèdent différents niveaux d'analyse (échelle du pays, de la ville…). Une gestion relativement indépendante de ces bases pose divers problèmes pour le producteur comme pour l'utilisateur des données. Tout d'abord il peut y avoir des incohérences entre les bases. Ensuite les efforts de saisie, de maintenance et de mise à jour sont multipliés. Enfin, il est difficile de réaliser des analyses combinant des données avec différents points de vue. Une solution possible à ces problèmes est de rendre explicites les relations entre les bases de données. Cette intégration soulève deux problèmes, l'appariement des schémas des bases, et l'appariement des données elles-mêmes rendu complexe par l'absence d'identifiant universel sur les données géographiques. Les travaux présentés ici portent sur l'appariement des schémas de bases de données géographiques, l'appariement de données faisant l'objet de recherches parallèles (Olteanu et al. 2005).

2 Des ontologies pour apparier des schémas de BD

Pour apparier deux schémas, une comparaison directe de ceux-ci se révèle souvent difficile, et ceci d'autant plus que le monde représenté dans les bases est complexe, conduisant à des choix de modélisation variés, ce qui est le cas pour les données géographiques. Une approche de plus en plus privilégiée, autant dans le monde des bases de données en général (Partridge 2002), que dans celui des systèmes d'information géographiques (Uitermark 2001, Gesbert 2005), est d'appuyer l'appariement sur une

ontologie du domaine concerné. Chaque schéma de données est alors relié à l'ontologie (voir un exemple en figure 1), et des relations schéma/schéma peuvent être ensuite déduites des deux ensembles de relations schéma/ontologie. C'est la voie que nous privilégions dans nos travaux.

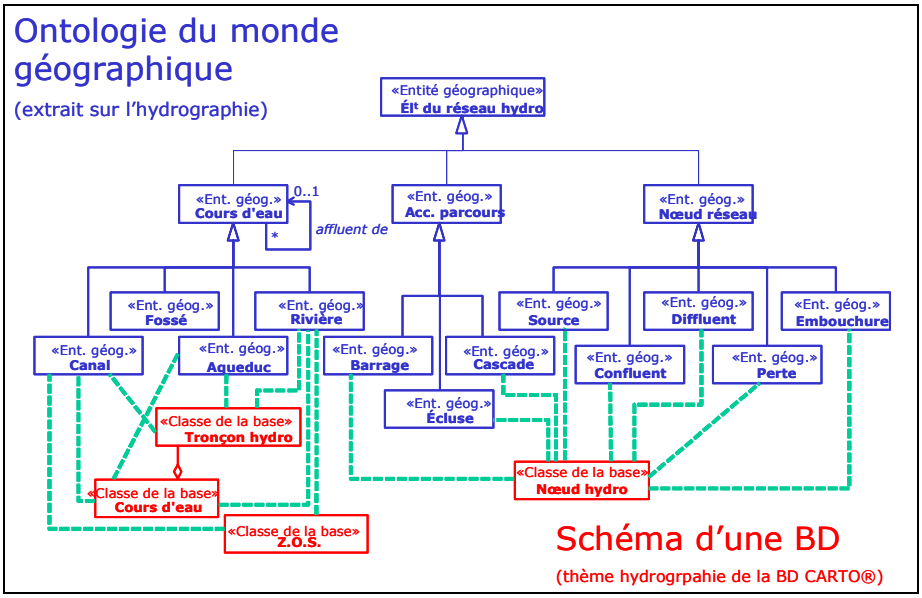
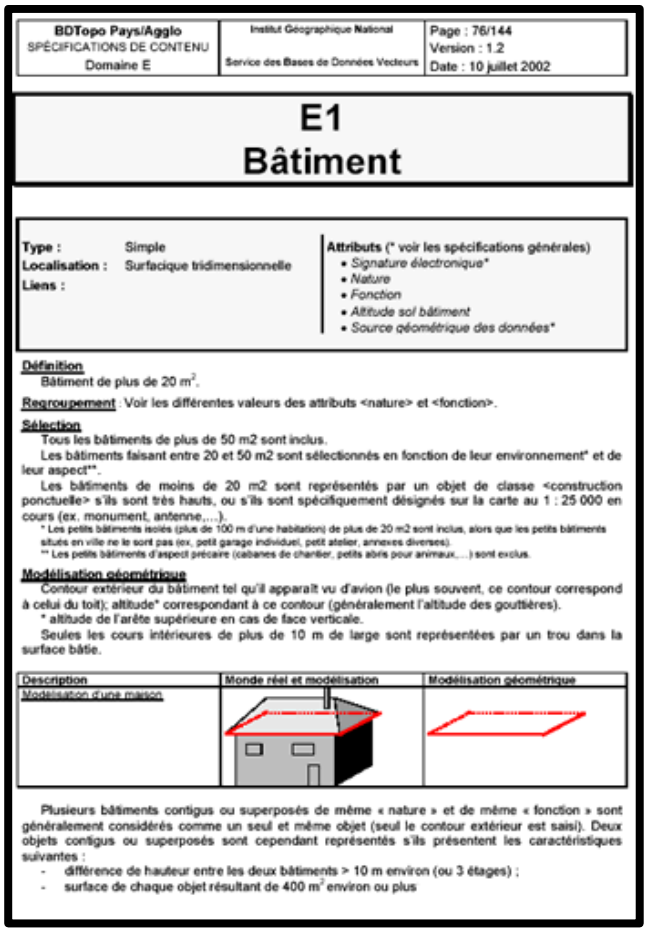


Fig. 1 - Liens entre une ontologie et un schéma de BD, d'après Gesbert (2005)

3 Les spécifications comme sources des ontologies

Deux questions se posent alors : comment constituer une ontologie du domaine, et comment représenter et instancier le lien entre un schéma et une ontologie ? Comme proposé par Gesbert (2005), nos recherches aspirent à répondre à ces deux questions pour les données géographiques à partir d'une même source de connaissances : les spécifications textuelles des bases de données. En raison de la complexité du monde géographique, et surtout en raison de la diversité des choix réalisés pour modéliser et instancier les bases de données géographiques, les producteurs associent généralement à leurs bases de volumineuses spécifications (typiquement d'une centaine de pages, voir un extrait en figure 2). Prenons l'exemple de la classe *Tronçon hydrographique* d'une base de données, pour laquelle les spécifications précisent que « Un *tronçon hydrographique* correspond à l'axe du lit d'une rivière, d'un ruisseau ou d'un canal », mais aussi que la base contient « tous les axes principaux, y compris dans la zone d'estran et dans les zones de marais, à l'exception des "culs-de-sac" d'une longueur inférieure à un kilomètre […] ». Ces spécifications font donc de fait référence à une ontologie du domaine (concepts de *rivière*, *ruisseau*, *canal*, *estran*, *marais*…), et explicitent le lien entre les classes et l'ontologie (la classe *Tronçon hydrographique* représente *toutes les rivières, à l'exception de*….). Ces spécifications sont des documents assez fortement structurés, mais principalement rédigés en langage naturel.



BDTopo Pays/Agglo SPÉCIFICATIONS DE CONTENU Domaine E	Institut Géographique National Service des Bases de Données Vecteurs	Page : 76/144 Version : 1.2 Date : 10 juillet 2002

E1
Bâtiment

Type : Simple
Localisation : Surfacique tridimensionnelle
Liens :

Attributs (* voir les spécifications générales)
- *Signature électronique**
- *Nature*
- *Fonction*
- *Altitude sol bâtiment*
- *Source géométrique des données**

Définition
Bâtiment de plus de 20 m².

Regroupement : Voir les différentes valeurs des attributs <nature> et <fonction>.

Sélection
Tous les bâtiments de plus de 50 m2 sont inclus.
Les bâtiments faisant entre 20 et 50 m2 sont sélectionnés en fonction de leur environnement* et de leur aspect**.
Les bâtiments de moins de 20 m2 sont représentés par un objet de classe <construction ponctuelle> s'ils sont très hauts, ou s'ils sont spécifiquement désignés sur la carte au 1 : 25 000 en cours (ex. monument, antenne,...).
* Les petits bâtiments isolés (plus de 100 m d'une habitation) de plus de 20 m2 sont inclus, alors que les petits bâtiments situés en ville ne le sont pas (ex. petit garage individuel, petit atelier, annexes diverses).
** Les petits bâtiments d'aspect précaire (cabanes de chantier, petits abris pour animaux,...) sont exclus.

Modélisation géométrique
Contour extérieur du bâtiment tel qu'il apparaît vu d'avion (le plus souvent, ce contour correspond à celui du toit); altitude* correspondant à ce contour (généralement l'altitude des gouttières).
* altitude de l'arête supérieure en cas de face verticale.
Seules les cours intérieures de plus de 10 m de large sont représentées par un trou dans la surface bâtie.

Description	Monde réel et modélisation	Modélisation géométrique
Modélisation d'une maison		

Plusieurs bâtiments contigus ou superposés de même « nature » et de même « fonction » sont généralement considérés comme un seul et même objet (seul le contour extérieur est saisi). Deux objets contigus ou superposés sont cependant représentés s'ils présentent les caractéristiques suivantes :
- différence de hauteur entre les deux bâtiments > 10 m environ (ou 3 étages) ;
- surface de chaque objet résultant de 400 m² environ ou plus

Fig. 2 - Extrait de spécifications textuelles d'une base de données géographiques

4 Automatisation de la construction des ontologies

4.1 Traitement semi-automatique du langage naturel

Afin de passer à l'échelle, c'est-à-dire d'apparier l'intégralité des schémas de nombreuses bases, il nous semble nécessaire d'automatiser autant que possible la construction d'une ontologie à partir de ces spécifications en langage naturel, ainsi que la détermination des liens entre le schéma et l'ontologie. Il peut être objecté à notre approche que si à chaque spécification correspond une ontologie sous-jacente, le problème de l'appariement des schémas sera uniquement déplacé et ramené à un problème d'alignement des ontologies sous-jacentes. Cependant, nous constatons que ce nouveau problème, s'il reste entier, est plus facile à résoudre que le problème initial, car les ontologies déduites des spécifications sont beaucoup plus proches les unes des autres que ne le sont les schémas des bases de données correspondantes. Par exemple, les aqueducs sont représentés dans deux bases

distinctes de l'IGN, mais ceux-ci sont représentés dans la classe *Tronçon hydrographique* du thème *Hydrographie* dans une base, et dans la classe *Canalisation* du thème *Transport d'énergie et de fluides* dans l'autre base : il est difficile de faire a priori le lien entre ces deux classes. Mais a contrario, on remarque que le concept d'*Aqueduc* est présent et similaire dans les ontologies sous-jacentes aux deux bases. Seule une analyse des spécifications textuelles met en valeur le lien entre les deux classes mentionnées à travers le concept d'aqueduc présent dans les deux ontologies sous-jacentes.

Nous avons donc réalisé une analyse des spécifications par traitement automatique du langage naturel pour, dans un premier temps, découvrir les ontologies sous-jacentes aux spécifications (voir figure 3, Laurens 2006). Ceci est nécessaire avant de, dans un deuxième temps, détecter et formaliser les relations entre schémas et ontologies.

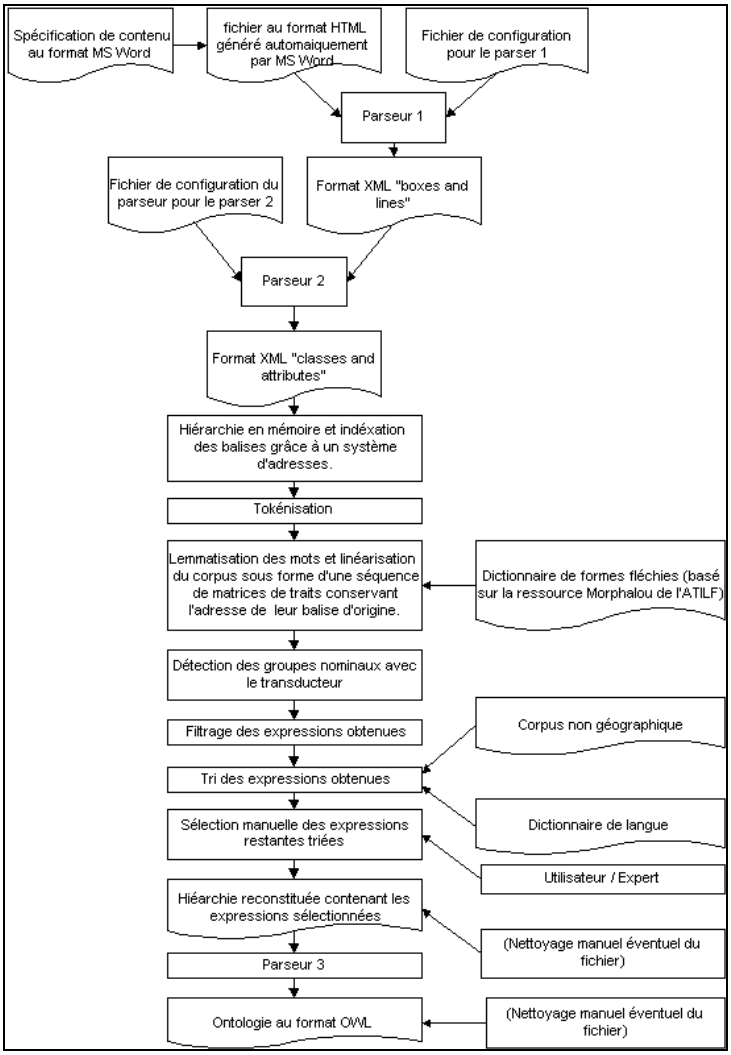


Fig. 3 - Processus d'analyse semi-automatique des spécifications (Laurens 2006)

4.2 Résultats de l'automatisation

Les résultats obtenus (cf. figure 4) montrent en premier lieu qu'une détermination semi-automatique d'ontologie est possible, même si un post-traitement interactif se révèle toujours nécessaire : concrètement, on évalue à une journée de travail le post-traitement interactif nécessaire à la finalisation de l'ontologie obtenue automatiquement.

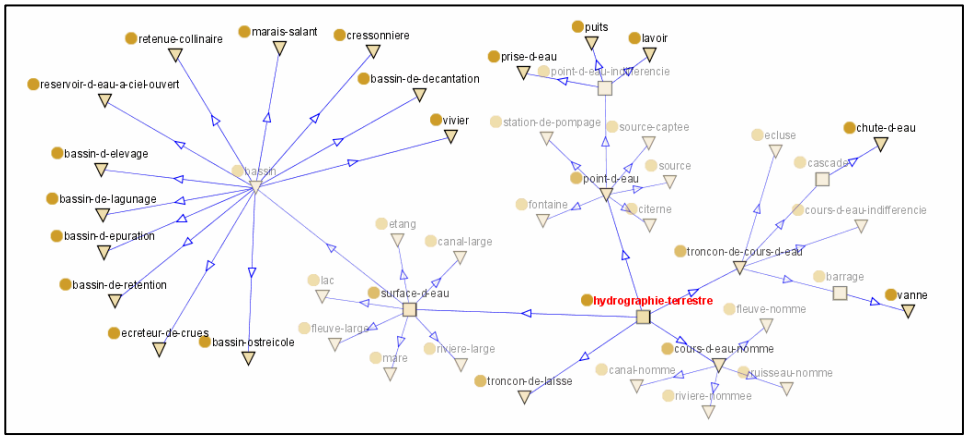


Fig. 4 - Extrait de l'ontologie issue de l'analyse semi-automatique de spécifications (extrait de Laurens 2006, Visualisation : Protégé-OWL)

Nos résultats montrent également que l'analyse automatique du langage naturel des spécifications est plus efficace en suivant une approche par règles qu'une approche statistique, ceci en raison de la faible répétition des concepts de l'ontologie dans les spécifications. En effet, certaines parties caractéristiques des spécifications telles que les noms de classe ou de thèmes qui les regroupent renferment presque toujours des concepts intéressants, et il est donc plus avantageux de se baser sur une analyse de la structure du document et des heuristiques linguistiques que sur une analyse fréquentielle ou distributionnelle des termes utilisés. Nos résultats mettent enfin en valeur les principales difficultés rencontrées pour une analyse automatique des spécifications.

En premier lieu, les concepts pertinents peuvent se situer à divers endroits et il est donc difficile de les détecter tous. Ils peuvent apparaître soit dans le nom même des classes (ex : *Tronçon hydrographique*), soit dans la définition associée (ex : « lit d'une *rivière*, *ruisseau*, *canal*… »), soit dans les règles de sélection associées (ex : « …dans les zones de *marais*… »), soit dans les valeurs possibles d'un attribut (ex : valeur « *aqueduc* » possible pour l'attribut *Nature*), soit dans la définition associée à cette valeur (ex : « *Tuyau* ou *chenal* artificiel… »). Ajoutons à cela une difficulté supplémentaire du au fait que derrière le titre « définition » rencontré dans les spécifications on trouve différentes acceptions du terme : il peut s'agir d'une réelle définition (ex : définition de la valeur d'attribut *cap :* « proéminence dans le contour d'une côte »), mais il peut aussi s'agir en fait d'une précision de ce qui est saisi dans la base (ex : définition de *carrière* : « carrière à ciel ouvert »).

En second lieu, certaines hiérarchies sont difficiles à expliciter. Tout d'abord, de nombreuses hiérarchies naturelles de concepts sont mises à plat dans les spécifications et il est donc impossible de les détecter automatiquement sans source de connaissances extérieure (ex : un attribut peut avoir pour valeurs « sommet », « pic », « vallée » ou « plage », sans que la proximité naturelle entre sommet et pic soit explicitée). Ensuite, il existe des concepts difficiles à nommer autrement que par un des sous-concepts représentatifs qui peuvent être mentionnés (ex : la valeur *dépression* représente « une cuvette, un bassin fermé, une *dépression*, ou une doline » et il est difficile de trouver un terme plus général). Enfin, les sous-concepts mentionnés sont parfois des réelles spécialisations (« Vallée : combe, ravin, val, talweg… »), parfois des termes qui peuvent être plutôt vus comme des synonymes (« Isthme : isthme, cordon littoral »).

5 Conclusion

Nos travaux en cours nous encouragent à poursuivre dans cette voie de l'analyse des spécifications pour apparier les schémas, mais montrent aussi la limite à une complète automatisation de l'extraction d'ontologie pour réaliser cela. D'autres voies sont aussi explorées, comme l'induction de règles de correspondances entre schémas à partir d'un appariement de données réalisé sur des critères géométriques par exemple (Sheeren 2005).

Références

Gesbert N. (2005). *Formalisation des spécifications de bases de données géographiques en vue de leur intégration*, Thèse de doctorat, Université de Marne La Vallée.

Laurens F. (2006). *Création d'une ontologie à partir de textes en langage naturel*. Stage de Master 1 Linguistique-Informatique, Université Paris 7.

Olteanu A.M., Mustière S., Ruas A. (2006). Matching Imperfect Data. Dans les actes de International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science (Accuracy'2006), Lisbon, pp.694-704.

Partridge C. (2002). *The role of ontology in integrating semantically heterogeneous databases*. Technical Report 05/02 LADSEB-CNR, Padoue.

Sheeren D. (2005). *Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales – Une approche combinant l'utilisation de métadonnées et l'apprentissage automatique*. Thèse de doctorat de l'université Paris 6.

Uitermark H. (2001). *Ontology-Based Geographic Data Set Integration*. PhD thesis, Universiteit Twente, Pays-Bas.

Summary

Integration of geographic databases may be facilitated by the use of domain ontologies. These one can be deduced from the processing of textual specifications of the databases. This paper presents ongoing works on this subject and the main difficulties encountered.