

Utilisation d'une ontologie du domaine pour la découverte du contenu de bases de données géographiques

Ammar Mechouche, Nathalie Abadie, Emeric Prouteau, Sébastien Mustière

Institut Géographique National, Laboratoire Cogit, 73 Av. de Paris, 94160 St-Mandé, France

Résumé. L'essor récent des technologies associées à la géomatique a permis la production rapide de nombreuses données géographiques. Or, pour tirer profit de ces données, il convient de pouvoir évaluer leur pertinence et leur complexité vis-à-vis de l'application à laquelle on les destine. Dans cet article, nous présentons une application permettant à un utilisateur de découvrir le contenu de bases de données géographiques, à savoir, quels types d'entités géographiques sont représentés au sein de chaque base et comment. Pour accéder à ces informations, l'utilisateur interroge le système via une ontologie globale du domaine qui décrit les types d'entités topographiques du monde réel. Des ontologies locales (ou d'application) sont utilisées pour formaliser les spécifications de chaque base de données décrite. Elles sont annotées à l'aide de concepts issus de l'ontologie globale. Ce système est implémenté sous la forme d'une interface Web et inclut un affichage cartographique d'échantillons de données.

1 Contexte et objectifs

Au cours des dernières décennies, les Systèmes d'Information Géographique (SIG) n'ont cessé de gagner en importance, suscitant le développement rapide de la production de données géographiques, de leur diffusion, ou encore d'applications de géolocalisation (Craglia et al., 2008). Les nombreuses sources de données géographiques désormais disponibles s'avèrent utiles pour de nombreuses applications telles que l'environnement, l'urbanisme ou l'agriculture. Cependant, d'une base à l'autre, un même type d'entités géographique pourra être représenté différemment selon le point de vue adopté par son producteur lors de la conceptualisation de la base et l'établissement de ses spécifications (Fonseca et al., 2003), phénomène que l'on qualifie d'hétérogénéité sémantique des données (Partridge, 2002), et qui est à l'origine de nombreuses difficultés dans la réutilisation conjointe de bases de données hétérogènes.

En raison de la variété des données géographiques disponibles et de la complexité de leurs spécifications, les utilisateurs peuvent éprouver de grandes difficultés à évaluer et comprendre précisément le contenu de ces bases de données. Le développement de portails Web permettant aux utilisateurs de visualiser les données géographiques disponibles a permis d'améliorer la compréhension de ces données. Cependant, en dépit de ces géoportails, un grand nombre d'informations concernant les données restent inaccessibles à des utilisateurs plus expérimentés en géomatique ; il demeure impossible pour des spécialistes de différents domaines désireux d'apprécier et de comparer les contenus des diverses bases de données disponibles vis-à-vis de leurs besoins spécifiques d'accéder simplement aux informations utiles, et en particulier aux spécifications de ces bases.

Découverte de données géographiques via une ontologie

L'objectif de l'application décrite dans cet article est de fournir à un utilisateur une application lui permettant de découvrir de façon simple les données les plus appropriées pour ses besoins, au travers d'une interface Web conviviale. Celle-ci devra mettre à disposition des utilisateurs des informations issues des spécifications de chaque base de données géographique qui jusqu'alors n'étaient pas accessibles, à moins de lire les volumineuses spécifications textuelles fournies par les producteurs de données. Plus précisément, cette application vise à :

- Aider les utilisateurs à retrouver simplement les types d'entités géographiques qu'ils recherchent, en leur proposant d'utiliser des termes courants issus d'une ontologie du domaine au lieu des termes techniques utilisés dans les schémas conceptuels de bases de données.
- Retrouver automatiquement dans les bases disponibles les données intéressant l'utilisateur.
- Fournir des informations supplémentaires sur les données : quels types d'entités du monde réel sont représentés par ces données (par exemple, s'agit-il de tous les cours d'eau ou bien seulement des cours d'eau permanents ?), comment sont-ils représentés dans ces bases de données (dans quelle classe, avec quels attributs ?), et comment se distinguent-ils des autres types d'entités géographiques (i.e. les informations fournies par la base permettent-elles de distinguer les cours naturels des cours d'eau artificiels et si oui comment ?) ?
- Visualiser les données correspondant aux besoins des utilisateurs à l'aide de techniques de cartographie pour le Web.

Plusieurs systèmes fondés sur des ontologies ont déjà été proposés dans le cadre d'applications de découverte et de recherche automatique de données (Paul et Ghosh, 2006) (Nambiar et al., 2006) (Klien, 2008). Cependant, aucun d'eux ne propose d'approche d'annotation sémantique des bases de données géographiques fondée sur les spécifications de ces bases afin de disposer d'informations plus détaillées sur les données elles-mêmes. A titre d'exemple, si un utilisateur recherche des données concernant les forêts, notre système ne lui indiquera pas seulement que les forêts sont représentées dans la classe « zone arborée » de notre base, mais également que les zones arborées représentées dans cette classe ont nécessairement une superficie supérieure à 5 hectares.

Cet article est structuré de la façon suivante : tout d'abord, nous présentons notre approche et les divers éléments nécessaires à la compréhension du système proposé. Puis, nous détaillons l'architecture du système. Enfin nous décrivons le prototype implémenté et les résultats obtenus avec nos échantillons de données.

2 Approche et méthodes mises en œuvre

2.1 Notions préliminaires

Dans un souci de clarté, nous allons tout d'abord nous attacher à décrire les éléments essentiels de notre système : les spécifications de bases de données géographiques, l'ontologie du domaine de la topographie, et enfin les ontologies d'application décrivant les spécifications de nos bases de données géographiques.

2.1.1 Les spécifications de bases de données géographiques

Comme toutes les bases de données, les bases de données géographiques vectorielles sont en premier lieu décrites par leur schéma. Leurs classes sont nommées à l'aide de termes

désignant généralement les concepts géographiques que nous manipulons couramment. Leurs instances, les objets géographiques, sont décrites par des attributs et une représentation géométrique sous forme de point, de ligne ou de polygone.

Surface d'eau

<p>Définition : Surface d'eau terrestre, naturelle ou artificielle.</p> <p>Géométrie : Surfacique tridimensionnelle</p>	<p>Attributs</p> <ul style="list-style-type: none"> • Identifiant ⁽¹⁾ • Source géométrique des données ⁽¹⁾ • Nature • Régime des eaux • Z_Minimal ⁽¹⁾⁽²⁾ • Z_Maximal ⁽¹⁾⁽²⁾ <p><small>(1) voir les spécifications générales (2) uniquement pour les formats 2D</small></p>
---	---

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

Sélection :
 Toutes les surfaces d'eau de plus de 20 m de long sont incluses, ainsi que les cours d'eau de plus de 7,5 m de large.
 Tous les bassins maçonnés de plus de 10 m sont inclus.
 Les zones inondables périphériques (zone périphérique d'un lac de barrage, d'un étang à niveau variable) de plus de 20 m de large sont incluses (attribut régime des eaux = « intermittent »).

Attribut : Nature

Définition : Attribut permettant de distinguer les bassins des surfaces hydrographiques naturelles

Type : Énuméré

Valeurs : Bassin / Surface d'eau

Nature = « Bassin »

Définition : Construction non couverte destinée à recevoir de l'eau temporairement ou de manière permanente.

Regroupement : Bassin d'élevage piscicole | Bassin d'épuration | Bassin de décantation | Bassin de filtrage | Bassin de lagunage | Bassin de rétention | Bassin ostréicole | Cressonnière | Écrêteur de crues | Marais salant | Réservoir d'eau à ciel ouvert | Retenue collinaire | Saline | Vivier

Sélection : Tous les bassins à ciel ouvert de plus de 10 m de long et 5 m de large. Les bassins de natation des piscines découvertes sont exclus (voir classe <terrain de sport>).

Modélisation géométrique : Rebord extérieur du bassin.

Contrainte de modélisation : Des bassins très proches les uns des autres (séparation < 10 m et très petite devant la largeur des bassins), et qui ne sont pas séparés par un objet linéaire de la base (voie de communication, cours d'eau, etc.) peuvent, dans certains cas, être modélisés par un seul objet englobant la zone de bassins (ex. zone ostréicole, pisciculture).

Fig. 1- Extrait des spécifications de la BDTOPPO® (IGN, 2002).

Cependant, chaque base de données géographique représente le point de vue particulier de son producteur sur le monde réel (Fonseca et al., 2003). Par exemple, si une classe se nomme « route », dans les faits il se peut qu'elle ne représente que les voies carrossables, ou bien qu'elle inclue également les voies non carrossables comme les chemins, et les sentiers forestiers. Par ailleurs, une base de données géographique est associée à un certain niveau de détail, et les entités géographiques du monde réel sont saisies ou non dans la base conformément à ce niveau de détail ; c'est pourquoi seules les plus pertinentes vis-à-vis de la description du paysage sont saisies. Enfin, dans les bases de données géographiques vectorielles la représentation géométrique des objets géographiques peut varier. Ainsi, une route peut être

Découverte de données géographiques via une ontologie

représentée par une ligne saisie le long de son axe ou bien par un polygone couvrant l'ensemble de la chaussée.

Dans la mesure où la saisie des données est réalisée par plusieurs personnes, l'ensemble des critères complexes de sélection et de représentation des entités géographiques au sein d'une base de données est consigné de façon aussi précise et peu ambiguë que possible dans de volumineux documents textuels destinés aux opérateurs de saisie; ce sont les spécifications de saisie de la base de données (voir figure 1). Celles-ci sont garantes de l'homogénéité de la sémantique des données au sein de la base, c'est-à-dire de l'homogénéité des relations entre les objets géographiques de la base et les entités géographiques du monde réel qu'ils représentent (Kavouras et Kokla, 2008).

Ainsi, les spécifications constituent la source de connaissances la plus détaillée dont nous disposons sur le contenu des bases de données géographiques. Elles décrivent la sémantique précise de chaque élément du schéma conceptuel de la base. Cependant, dans la mesure où elles sont destinées en priorité aux opérateurs de saisie, elles sont rédigées en langage naturel et ne peuvent donc pas être mises à profit directement dans une application informatique. Il est donc nécessaire de disposer d'une représentation formelle de ces spécifications afin de tirer parti des connaissances qu'elles contiennent dans le cadre de notre application de découverte du contenu de bases de données géographiques (Mustière et al., 2003).

2.1.2 L'ontologie du domaine de la topographie

Notre approche requiert une ontologie commune des concepts géographiques décrits dans les bases de données traitées (ou bien plusieurs ontologies alignées, cependant dans un souci de clarté et de simplicité une telle approche ne sera pas traitée ici). Nous utilisons ici une ontologie bilingue (français et anglais) créée à partir des textes de plusieurs spécifications de bases de données géographiques à l'aide d'outils de traitement automatique du langage naturel (Abadie et al., 2008). Cette ontologie du domaine de la topographie contient plus de 760 concepts, mais demeure limitée ; à l'avenir nous souhaiterions nous munir d'une ontologie plus riche à la fois en termes de nombre de concepts et en termes de description de ces concepts. La création de cette ontologie fait l'objet de travaux en cours, qui visent à enrichir l'ontologie dont nous disposons déjà à l'aide de traitement automatique du langage naturel appliqué à des spécifications de bases de données géographiques et des récits de voyages, d'outils d'alignement d'ontologies, de dictionnaires, et de bases de données de toponymes.

2.1.3 Les ontologies locales de spécifications

Notre approche repose également sur des ontologies locales de spécifications. Ces ontologies d'application formalisent le contenu des spécifications de chaque base de données considérée. Celles-ci sont construites selon l'approche proposée par (Abadie et al., 2008).

Il s'agit, dans un premier temps, de traduire le schéma de la base de données considérée en OWL. Puis les classes de ce schéma sont annotées à l'aide des concepts issus de l'ontologie du domaine de la topographie. Pour chaque classe du schéma, des connaissances supplémentaires tirées des spécifications de cette dernière, comme par exemple des critères de sélection, sont ajoutées dans l'axiome utilisé pour l'annoter. Ainsi, des règles comme « La classe 'Rivière' comprend tous les cours d'eau permanents de plus de 10 mètres de large » ou bien « La géométrie des rivières est saisie au niveau de l'axe de leur lit » sont formalisées. Ces ontologies locales de spécifications constituent un lien explicite entre les entités du

monde réel décrites par l'ontologie globale du domaine de la topographie et leur représentation dans la base de données.

Afin de garantir un bon niveau d'homogénéité dans la façon dont sont formalisées les spécifications de bases de données géographiques, l'approche suivie s'appuie sur une ontologie commune implémentée en OWL2 et nommée « ontologie des spécifications » (SO). Celle-ci fournit un ensemble de concepts communs propres aux spécifications, comme par exemple les concepts de « photographie aérienne » ou de « contour » d'une entité géographique. Cette ontologie réutilise des ontologies préexistantes comme GeoRSS-Simple¹.

Formalisons la spécification suivante selon cette approche : « Les objets géographiques de la classe de base de données 'Autre_point_d'eau' représentent des 'pertes', des 'lavoirs', des 'puits' ou encore des 'bassins' et possèdent une géométrie de type 'Point'. Les entités géographiques de type 'bassin' doivent mesurer moins de 10 mètres de long pour être saisies. » Les concepts issus de l'ontologie des spécifications sont précédés de 'so:', ceux de l'ontologie locale de spécifications de 'lso:', et ceux de l'ontologie globale du domaine de la topographie de 'topo:'.

```
Class: lso:db_Autre_point_d'eau EquivalentTo: so:hasForGeometry some
gml:Point and so:represents some (topo:Perte or topo:Lavoir or
topo:Puit or (topo:Bassin and topo:length some double[<10.0]))
```

Les ontologies locales de spécifications et l'ontologie globale du domaine de la topographie constituent le cœur de notre système : les premières constituent une interface entre chaque base et le système et la seconde une interface entre le système et les utilisateurs.

2.2 Architecture du système

L'architecture globale de notre système est présentée en figure 2. L'utilisateur exprime sa requête dans les termes de l'ontologie du domaine, qui lui permet d'interroger plusieurs bases à l'aide d'un vocabulaire unifié. Notre système fonctionne selon une architecture client-serveur, et comporte une application de cartographie pour le Web offrant un moyen convivial de visualisation des données coté client. Il est composé de trois modules:

Le module de recherche, tout d'abord, guide l'utilisateur dans la formulation de sa requête en lui proposant, à l'aide d'une application d'auto-complétion, d'interroger le système dans les termes de l'ontologie globale du domaine de la topographie. Ceci présente deux avantages : d'une part, le vocabulaire issu de cette ontologie est supposé commun à l'ensemble de la communauté de l'information géographique et reste indépendant de toute application technique, et d'autre part toutes les ontologies locales de spécifications qui formalisent les spécifications de bases de données reposent sur cette ontologie globale, ce qui fait d'elle un élément central, pivot de notre système.

Par ailleurs, le module d'extraction d'informations recherche, dans les ontologies locales de spécifications, les données disponibles dans les bases de données référencées correspondant aux termes de la requête de l'utilisateur, i.e. les classes qui sont annotées par le concept géographique correspondant au label entré par l'utilisateur dans l'interface de requête. L'ensemble des connaissances disponibles se rapportant à ces classes (définitions, géométrie des objets géographiques, etc.) est renvoyé à l'utilisateur via l'interface de réponse.

¹ <http://mapbureau.com/neogeo/neogeo.owl>

Découverte de données géographiques via une ontologie

Enfin, le module cartographique affiche sous forme de cartes les échantillons de données identifiés par le module d'extraction d'informations afin de permettre à l'utilisateur de visualiser simplement les diverses données disponibles correspondant au thème qui l'intéresse.

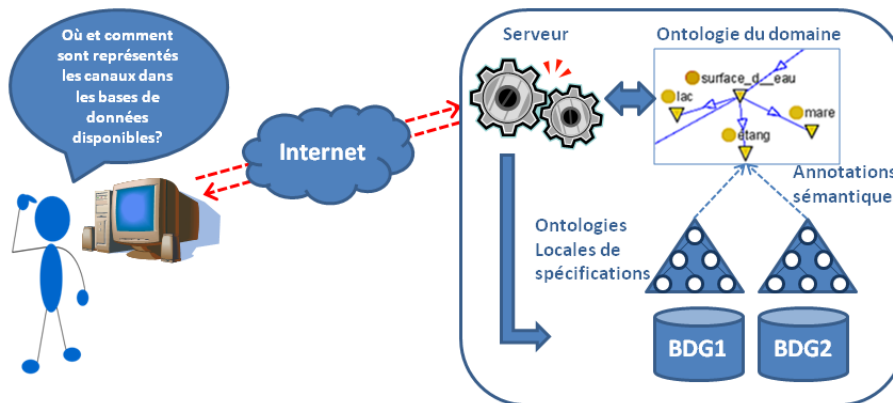


Fig. 2 - Architecture globale du système.

Les informations renvoyées par notre système à l'utilisateur concernant les données sur lesquelles il souhaite se renseigner, ainsi que la visualisation cartographique de ces données lui permettent de comparer plusieurs jeux de données géographiques et l'aident à estimer lequel sera le plus adapté à ses propres besoins.

3 Implémentation du système

Le système proposé a été implémenté sous forme d'une application Web (voir figure 3).

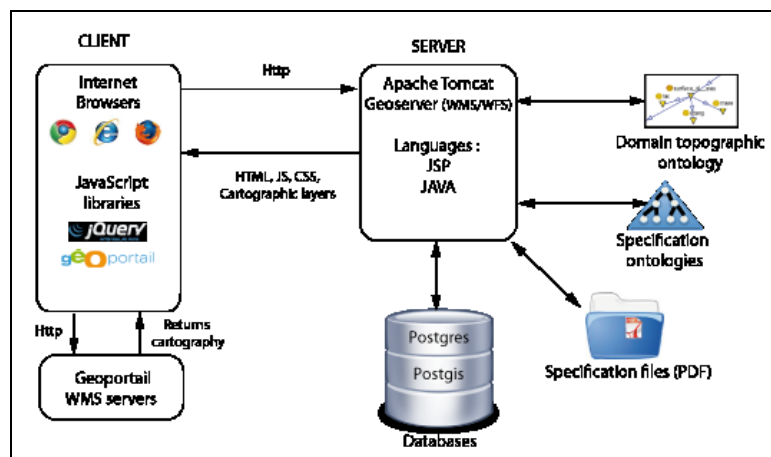


Fig. 3 - Implémentation du système.

Deux bases de données géographiques sont référencées par le système (la BDTPOPO® et la BDCARTO®), avec leurs ontologies locales de spécifications associées. Ces ontologies

sont représentées dans le langage OWL 2, et demeurent limitées au thème de l'hydrographie. L'application fonctionnant côté serveur a été développée en Java et utilise l'API OWL 2² pour manipuler les différentes ontologies nécessaires au système. Le fonctionnement des pages Web repose sur JSP, HTML, Javascript et JQuery. Pour permettre l'interprétation des pages JSP, le serveur choisi est Apache Tomcat. Enfin, le choix de Geoserver³ comme serveur cartographique permet de n'utiliser qu'un serveur pour faire fonctionner le système. Les jeux de données utilisés sont stockés dans deux bases de données gérées par le système de gestion de bases de données spatiales PostgreSQL couplé à son extension spatiale PostGIS. Le système utilise des services WMS pour l'affichage de données et l'API du GéoPortail afin de disposer de fonds de cartes pour le module d'affichage cartographique.

L'interface Web, composée de trois parties, est présentée en figure 4:

- La première partie, en haut de la page, est composée d'un champ texte permettant à l'utilisateur de spécifier le type de données qui l'intéresse à l'aide d'un mot-clé comme dans un moteur de recherche classique : dans notre exemple, canal.
- La deuxième partie, à gauche de la page, est composée d'onglets ; chacun d'eux correspond à une base de données et sert à afficher les informations envoyées par le système sur les données issues de cette base et qui correspondent à la requête de l'utilisateur.
- La troisième partie, à droite de la page, comporte l'affichage cartographique des données qui correspondent à la requête de l'utilisateur. Cet affichage est synchronisé avec les onglets de la deuxième partie: les données affichées sur la carte sont celles appartenant à la base de données correspondant à l'onglet sélectionné.

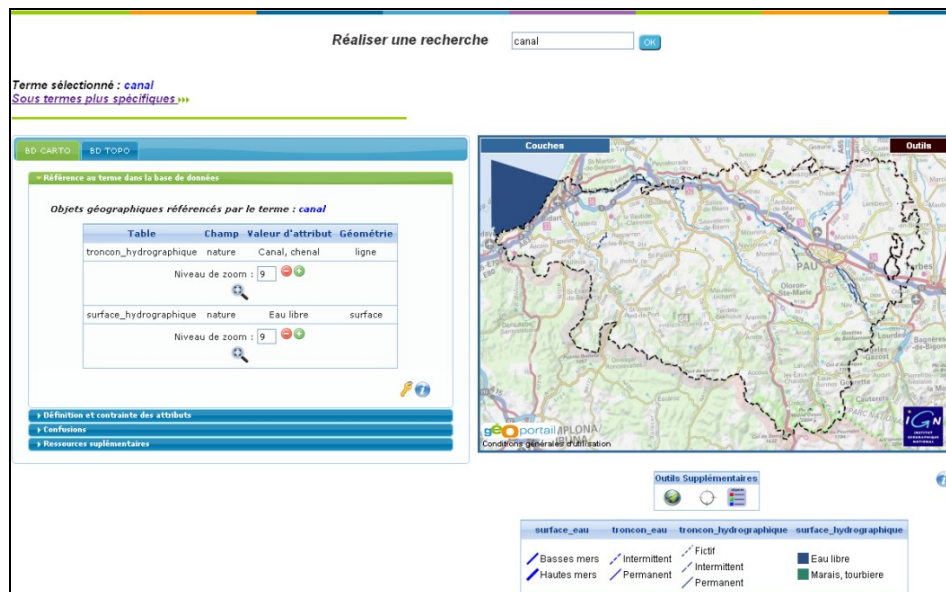


Fig. 4 - Interface Web du prototype implémenté.

Le module de recherche a été implémenté en utilisant le script d'auto-complétion fourni avec la bibliothèque JQuery. Celui-ci affiche les résultats sous la forme d'une liste de termes

² <http://owlapi.sourceforge.net/>

³ <http://geoserver.org/display/GEOS/Welcome>

Découverte de données géographiques via une ontologie

correspondant aux labels associés à chaque classe de l'ontologie du domaine. Afin de compenser d'éventuelles fautes de frappe lors de la saisie des requêtes, le module de recherche s'appuie sur un calcul de distance de Levenshtein⁴ normalisée (Yujian, 2007) entre les caractères entrés par l'utilisateur et les labels des concepts de l'ontologie du domaine. Ainsi les labels de concepts les plus proches orthographiquement du terme entré par l'utilisateur lui seront proposés en priorité. Le module de recherche permet aussi à l'utilisateur d'affiner sa requête en lui proposant une liste de termes correspondants aux concepts de l'ontologie du domaine subsumés par le concept correspondant au terme initialement demandé.

Le module d'extraction d'informations extrait des ontologies locales de spécifications les connaissances disponibles sur les données désignées par la requête de l'utilisateur. Dans le cas de la requête « canal », le système va extraire les classes des différentes ontologies locales de spécifications annotées par le concept de 'canal' ainsi que des informations sur leur mode de représentation (lignes ou polygones), et renvoyer ces résultats à l'utilisateur via l'interface Web (voir figure 5). Un onglet par base est proposé à l'utilisateur.

Table	Champ	Valeur d'attribut	Géométrie
troncon_hydrographique	nature	Canal, chenal	ligne
Niveau de zoom : 9			
surface_hydrographique	nature	Eau libre	surface
Niveau de zoom : 9			

Fig. 5 - Informations sur les données renvoyées par le système à l'utilisateur.

Un onglet se présente sous la forme d'un accordéon composé de quatre sections. La première section (voir figure 6) indique :

1. La(les) classe(s) de la base de données où sont représentées les entités géographiques désignées par le terme choisi par l'utilisateur.
2. Les valeurs d'attributs qui permettent de distinguer les objets géographiques correspondant bien à la requête de l'utilisateur d'éventuels autres objets présents dans la classe. Ici, les canaux sont représentés dans les classes 'surface d'eau' et 'cours d'eau' où ils se distinguent des cours d'eau naturels grâce à l'attribut booléen 'artif'.
3. Le nom de l'attribut pouvant prendre ces valeurs.

⁴ http://fr.wikipedia.org/wiki/Distance_de_Levenshtein

4. Le type de géométrie utilisé pour la représentation des entités géographiques.







Table	Champ	Valeur d'attribut	Géométrie
surface_eau	regime	Permanent	surface
surface_eau	nature	Surface d'eau	surface
Niveau de zoom : <input type="text" value="9"/>   			
troncon_eau	franchisst	Tunnel	ligne
troncon_eau	artif	1	ligne
troncon_eau	franchisst	Sans objet	ligne
Niveau de zoom : <input type="text" value="9"/>   			

Fig. 6 - Informations sur la localisation des données dans la base.

La deuxième section (voir figure 7) présente deux informations. Elle reprend des informations présentées dans la première section et y ajoute les définitions de valeurs d'attributs présentes dans les spécifications. De plus, elle décrit les critères de sélection que doivent vérifier les entités du monde réel pour être représentées dans cette classe de la base de données : ici, les canaux sont représentés comme des instances de 'surface d'eau' à condition d'avoir une largeur supérieure à 7.5 mètres.

Définition des attributs :

Table	Champ	Valeur d'attribut	Définition
surface_eau	regime	Permanent	Objet hydrographique caractérisé par la présence permanente ou quasi-permanente d'eau.
surface_eau	nature	Surface d'eau	Surface d'eau non marine.
troncon_eau	franchisst	Sans objet	Valeur prise par exclusion des cinq autres.
troncon_eau	franchisst	Tunnel	Tronçon de cours d'eau artificiel passant sous un tunnel.
troncon_eau	artif	1	Canal ou cours d'eau naturel dont le tracé a été remanié.

Contrainte(s) sur attribut(s) :

Table	Champ	Valeur d'attribut	Contrainte	Valeur
surface_eau	nature	Surface d'eau	largeur	> 7.5
troncon_eau	franchisst	Tunnel	souterrain	true
surface_eau	regime	Permanent	largeur	> 7.5

Fig. 7 - Définitions et contraintes sur les données de la base.

La troisième section (voir figure 8) dresse la liste de tous les types d'entités géographiques du monde réel qui sont représentés au sein d'une même classe de la base de données

Découverte de données géographiques via une ontologie

avec les mêmes valeurs d'attributs. Par exemple, les portions de canaux, de biefs ou de cours d'eau artificialisés sont représentés comme des instances de la classe 'tronçon de cours d'eau', dont la valeur d'attribut 'artif' vaut 1, sans que rien ne permette de les distinguer.

Table	Champ	Valeur d'attribut	Termes représentés
troncon_eau	artif	1	<u>canal bief cours d'eau</u>
surface_eau	nature	Surface d'eau	<u>mare étang rivière lac canal surface d'eau fleuve</u>
troncon_eau	franchisst	Sans objet	<u>torrent rivière canal fossé bief cours d'eau ruisseau fleuve</u>
troncon_eau	franchisst	Tunnel	<u>canal</u>
surface_eau	regime	Permanent	<u>mare étang rivière lac canal surface d'eau fleuve</u>

Fig. 8 - Listes des types d'entités géographiques regroupés dans la base.

La quatrième section fournit des informations supplémentaires quand elles sont disponibles. L'utilisateur peut télécharger les fichiers de spécifications originaux (en PDF) des classes correspondant à sa requête, et télécharger un échantillon de données au format KML.

Le module cartographique est implémenté à l'aide de l'API du Géoportail, et les différentes couches cartographiques affichées sont contrôlées par le système de façon à correspondre à la requête de l'utilisateur. Afin de garantir l'affichage des seuls objets géographiques intéressant l'utilisateur, le système effectue des requêtes CQL afin de filtrer les couches WMS envoyées par Géoserver. L'affichage cartographique est synchronisé avec les onglets présentant les informations sur chaque base de données. En plus des fonctionnalités offertes par l'API du Géoportail, de nouvelles applications ont été développées, comme les requêtes par adresse, ou l'ajout de couches vectorielles, ou images, de façon à permettre à l'utilisateur de comparer ses propres données avec celles proposées par le système.

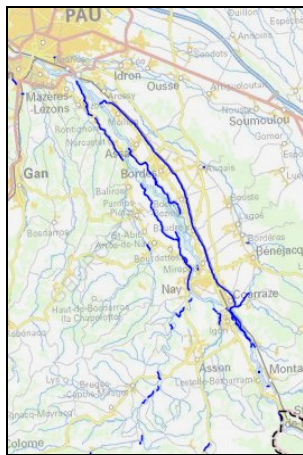


Fig. 9 - Représentation des canaux dans la BDTOPO®.

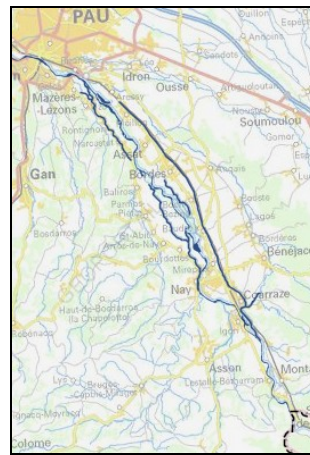


Fig. 10 - Représentation des canaux dans la BDCARTO®.

Grâce à cette application, il est désormais possible de comparer les contenus de plusieurs bases de données géographiques et d'évaluer leur pertinence vis-à-vis d'un besoin spécifique. Ainsi, si un utilisateur recherche quelle base de données représente des canaux de la façon la plus détaillée possible, il peut obtenir rapidement des informations sur la représentation des canaux au sein des différentes bases référencées par le système, et visualiser les données correspondantes (voir figures 9 et 10).

4 Conclusion et perspectives

Dans cet article, nous avons proposé un système fondé sur une ontologie du domaine de la topographie et plusieurs ontologies d'application pour faciliter la découverte du contenu de bases de données géographiques. Le système proposé permet une meilleure compréhension du contenu de bases de données géographiques grâce, d'une part, à la formalisation de leurs spécifications sous forme d'ontologies, et d'autre part à un affichage adéquat de ces connaissances associé à une représentation cartographique des données concernées dans les différentes bases. Ce deuxième aspect constitue une approche simple, efficace et réaliste pour faciliter la compréhension et la comparaison des données.

A l'avenir, nous prévoyons un certain nombre d'améliorations pour notre système. En premier lieu, il serait utile de permettre à l'utilisateur d'interroger le système sur plus d'un type d'entités géographiques à la fois, de façon à pouvoir comparer des données représentant différents types d'entités géographiques. De plus le système d'auto-complétion pourrait être revu de sorte à renvoyer à l'utilisateur les labels de concepts dans un ordre respectant leur hiérarchie taxonomique au sein de l'ontologie, afin de faciliter ses choix.

Les ontologies locales de spécifications formalisent en OWL 2 la plupart des connaissances contenues dans les spécifications des bases de données géographiques. Cependant certaines règles de saisie comme « L'attribut 'largeur' de la classe 'tronçon hydrographique' prend la valeur 'petite' si la largeur du tronçon de cours d'eau mesure entre 0 et 10 mètres » ne peuvent être représentées en OWL 2, car il s'agit de contraintes portant sur les valeurs de plusieurs propriétés. Il reste donc des aspects sur lesquels la formalisation des spécifications pourrait être améliorée, notamment en suivant les évolutions du langage OWL.

Le prototype implémenté est actuellement restreint au thème de l'hydrographie de deux bases de données géographiques et gagnerait à être étendu à d'autres thèmes et d'autres bases.

Remerciements : Cette recherche a été réalisée dans le cadre du projet Geonto, en partie financé par l'Agence Nationale de la Recherche (ANR-O7-MDCO-005).

Références

- IGN (2002). *BD Topo Pays, Version 1.2, Descriptif de Contenu, Edition 1*, Institut Géographique National, 118 p, Paris, France.
- Partridge, C. (2002). *The role of ontology in integrating semantically heterogeneous databases*. Technical Report 05/02 LADSEB-CNR, Padoue, Italie.

Découverte de données géographiques via une ontologie

- Fonseca, F., Clodoveu, D., Camara, G. (2003). *Bridging Ontologies and Conceptual Schemas in Geographic Information Integration*. *Geoinformatica*, 7(4), 355—378.
- Mustière, S. Gesbert, N., Sheeren, D. (2003). *A formal model for the specifications of geographic databases*, GEOPRO Conference: Semantic Processing of Spatial Data, Mexico.
- Nambiar, U., Ludscher Ludäscher, B., Lin K., Baru, C. (2006). *The GEON portal: accelerating knowledge discovery in the geosciences*, 8th ACM International Workshop on Web Information and Data Management, Arlington, Virginia, USA, 83-90.
- Paul, M., Ghosh S.K. (2006). *An Approach for Service Oriented Discovery and Retrieval of Spatial Data*, International Workshop on Service Oriented Software Engineering, Shanghai, Chine, 84-94.
- Yujian,L. (2007). *Bo, A Normalized Levenshtein Distance Metric*, IEEE Trans. Pattern Anal. Mach. Intell., 29(6), 1091-1095.
- Abadie N. and Mustière S. (2008). *Constitution d'une taxonomie géographique à partir des spécifications de bases de données*, Colloque International de Géomatique et d'Analyse Spatiale, Montpellier, France.
- Craglia M, Goodchild M, Annoni A, Camara G, Gould M, Kuhn W, Mark D, Masser I, Mauguire D, Liang S, Parsons E. (2008). *Next-Generation Digital Earth - A Position Paper from the Vespucci Initiative for the Advancement of Geographic Information Science*, International Journal of Spatial Data Infrastructures Research, 3, 146-167. JRC47746 1.5 Article contribution to other periodicals.
- Kavouras M., Kokla M. (2008). *Theories of geographic concepts : Ontological Approaches to Semantic Integration*, CRC Press, Taylor & Francis Group, Boca Raton, FL, USA.
- Klien E.M. (2008). *Semantic Annotation of Geographic Information*. Thèse de doctorat, Institute for Geoinformatics, University of Muenster. Muenster, Germany.
- Abadie N., Mechouche A., Mustière S. (2010). *OWL-based formalisation of geographic databases specifications*, 17th International Conference on Knowledge Engineering and Knowledge Management, Lisbonne, Portugal.

Summary

Nowadays, a huge amount of geodata is available. However, using them efficiently implies being able to evaluate their fitness for use and their complexity. In this paper we propose a system allowing a user to precisely discover which entities of interest are represented in the different available geodatabases and how. The system uses a global 'domain' ontology describing the topographic real world entities, which can be queried by the user to express his/her needs. Local (or application) ontologies are used to formalize the content of each database's specification. These local ontologies are annotated with concepts from the global ontology. The described system is implemented as a Web application and includes a web mapping solution for the visualization of data samples.