Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques : les spécifications au cœur du processus d'intégration

Nathalie Abadie

UNIVERSITÉ PARIS-EST

UNIVERSITÉ PARIS-EST

ÉCOLE DOCTORALE MSTIC

Thèse

Pour obtenir le grade de docteur de l'Université Paris-Est

Spécialité : Sciences et Technologies de l'Information Géographique

ABADIE Nathalie

Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques

Les spécifications au cœur du processus d'intégration

Thèse dirigée par Anne RUAS

Soutenue le 20 novembre 2012

Jury :

LIBOUREL Thérèse, Professeur des universités	Rapporteur
TEISSEIRE Maguelonne, Directrice de recherche	Rapporteur
EUZENAT Jérôme, Directeur de recherche	Président du jury
GENSEL Jérôme, Professeur des universités	Examinateur
CURÉ Olivier, Maître de conférences, HDR	Examinateur
RUAS Anne, Ingénieur des Ponts, des Eaux et des Forêts, HDR	Directrice
MUSTIÈRE Sébastien, Ingénieur IGN, Docteur	Encadrant

Cette thèse a été réalisée au Laboratoire COGIT de l'Institut National de l'Information Géographique et Forestière, sous la direction d'Anne Ruas et l'encadrement de Sébastien Mustière.

Institut National de l'Information Géographique et Forestière

Service de la Recherche, Laboratoire COGIT

73 Avenue de Paris

94165 Saint-Mandé Cedex

Tél. : 01 43 98 80 00

RÉSUMÉ

Cette thèse traite de l'intégration de bases de données topographiques qui consiste à expliciter les relations de correspondance entre bases de données hétérogènes, de sorte à permettre leur utilisation conjointe. L'automatisation de ce processus d'intégration suppose celle de la détection des divers types d'hétérogénéités pouvant intervenir entre les bases de données topographiques à intégrer. Ceci suppose de disposer, pour chacune des bases à intégrer, de connaissances sur leurs contenus respectifs. **Ainsi, l'objectif de cette thèse réside dans la formalisation, l'acquisition et l'exploitation des connaissances nécessaires pour la mise en œuvre d'un processus d'intégration virtuelle de bases de données géographiques vectorielles.**

Une première étape du processus d'intégration de bases de données topographiques consiste à apparier leurs schémas conceptuels. Pour ce faire, nous proposons de nous appuyer sur une source de connaissances particulière : les spécifications des bases de données topographiques. Celles-ci sont tout d'abord mises à profit pour la création d'une ontologie du domaine de la topographie. Cette ontologie est utilisée comme ontologie de support, dans le cadre d'une première approche d'appariement de schémas de bases de données topographiques, fondée sur des techniques d'appariement terminologiques et structurelles. Une seconde approche, inspirée des techniques d'appariement fondées sur la sémantique, met en œuvre cette ontologie pour la représentation des connaissances sur les règles de sélection et de représentation géométrique des entités géographiques issues des spécifications dans le langage OWL 2, et leur exploitation par un système de raisonnement.

Mots-clés : Intégration de bases de données topographiques, appariement de schémas, spécifications, représentation de connaissances, ontologie.

ABSTRACT

Formalisation, Acquisition and Implementation of Specifications Knowledge

for Geographic Databases Integration

This PhD thesis deals with topographic databases integration. This process aims at facilitating the use of several heterogeneous databases by making the relationships between them explicit. To automatically achieve databases integration, several aspects of data heterogeneity must be detected and solved. Identifying heterogeneities between topographic databases implies comparing some knowledge about their respective contents. Therefore, we propose to formalise and acquire this knowledge and to use it for topographic databases integration.

Our work focuses on the specific problem of topographic databases schema matching, as a first step in an integration application. To reach this goal, we propose to use a specific knowledge source, namely the databases specifications, which describe the data implementing rules. Firstly, they are used as the main resource for the knowledge acquisition process in an ontology learning application. As a first approach for schema matching, the domain ontology created from the texts of IGN's databases specifications is used as a background knowledge source in a schema matching application based on terminological and structural matching techniques. In a second approach, this ontology is used to support the representation, in the OWL 2 language, of topographic entities selection and geometry capture rules described in the databases specifications. This knowledge is then used by a reasoner in a semantic-based schema matching application.

Keywords: Topographical databases integration, schema matching, specifications, knowledge representation, ontology.

REMERCIEMENTS

Avant d'encadrer cette thèse, Sébastien Mustière a encadré mes premiers travaux au COGIT. Je le remercie pour la disponibilité et la patience dont il a fait preuve à mon égard au cours de ces sept années, pour les conseils avisés dont il m'a fait bénéficier, et pour la confiance qu'il m'a toujours témoignée. C'est avec plaisir, confiance et optimisme que je poursuis aujourd'hui mes travaux au COGIT sous sa direction.

Je remercie Anne Ruas de m'avoir accueillie au COGIT qu'elle dirigeait alors à mon entrée dans les services de l'Institut National de l'Information Géographique et Forestière. Je la remercie de m'avoir orientée sur le sujet de l'interopérabilité sémantique, et d'avoir dirigé cette thèse.

Ce travail de thèse doit beaucoup à Thérèse Libourel qui, au travers des thèses de Nils Gesbert et Sandrine Balley qu'elle a dirigées au COGIT, a largement contribué à la définition du sujet et à la façon dont je l'ai abordé. Nos conversations sur la représentation de connaissances propres à l'information géographique m'ont toujours été d'une aide précieuse : je l'en remercie, ainsi que d'avoir porté sur mon travail un jugement éclairé et constructif. Je remercie Maguelonne Teisseire pour l'intérêt qu'elle a manifesté à l'égard de mon travail. Les remarques et questions pertinentes émises dans son rapport et lors de ma soutenance m'ont permis de réexaminer mon travail sous un nouveau jour. Ce travail de thèse s'intègre en grande partie dans le domaine de l'alignement d'ontologies : je remercie Jérôme Euzenat d'avoir accepté de l'évaluer et d'y avoir apporté son regard expert, à la fois rigoureux, pragmatique et constructif. Je remercie enfin Jérôme Gensel et Olivier Curé pour leurs nombreuses questions et remarques lors de la soutenance, auxquelles je reviens encore aujourd'hui avec intérêt.

Je tiens également à remercier l'ensemble des membres du projet GéOnto. Si mener à bien une thèse tout en participant à un projet de recherche peut sembler difficile, j'en garde pour ma part le souvenir d'une expérience extrêmement enrichissante et bénéfique, tant sur le plan des nombreuses connaissances que j'ai pu acquérir au contact des membres du projet que sur le plan humain : on ne peut souhaiter meilleure initiation au monde de la recherche, et c'est avec plaisir que je travaillerais de nouveau avec chacun.

J'adresse plus particulièrement mes remerciements à Ammar Mechouche avec qui j'ai eu la chance de travailler durant son année de post-doctorat au COGIT, et dont les conseils et les encouragements m'ont grandement aidée à aborder le domaine du Web sémantique. Je garde un souvenir très positif de nos travaux communs, et j'espère que nos chemins se croiseront de nouveau.

Je remercie également Fayçal Hamdi que j'ai côtoyé grâce au projet GéOnto. Ses travaux de thèse que j'ai suivis tout au long du projet et nos nombreux échanges ont fortement contribué à me faire progresser dans le domaine de l'alignement d'ontologies. C'est avec plaisir que je l'ai retrouvé sur le projet Datalift. J'espère que nous aurons encore longtemps l'occasion de travailler ensemble.

J’adresse mes remerciement aux stagiaires que j’ai eu le plaisir d’encadrer (ou co-encadrer) durant ces années pour leur enthousiasme, leur persévérance et la qualité de leur travail : Frédéric Laurens, Virginie Picard, Thomas Horel, Emeric Prouteau et Léa Massiot.

Je remercie l’ensemble des membres du COGIT pour leur soutien indéfectible et leur bonne humeur quotidienne. J’adresse également mes remerciements à Bénédicte Bucher pour nos conversations fructueuses ainsi que pour la confiance qu’elle m’a témoignée durant ces années de thèse.

Je remercie enfin ma famille pour son soutien de toujours, avec une mention particulière à Valérie Abbadie pour les longues heures passées à relire ma thèse et à corriger mon usage très germanique de la virgule.

TABLE DES MATIÈRES

1 Introduction

Les bases de données topographiques vectorielles visent à représenter la topographie du monde réel. Elles fournissent une vision de l'espace abstraite, partielle, et non unique. Les entités topographiques du monde réel y sont représentées de façon simplifiée, conformément au point de vue adopté par le producteur de la base de données sur le monde réel, et de sorte à répondre aux besoins d'applications bien spécifiques. Les progrès récents en matière d'acquisition de données topographiques alliés à l'essor de plateformes de partage de données géographiques, appelées Infrastructures de Données Géographiques, ont permis la mise à disposition de nombreuses sources de données pouvant être mises à profit dans diverses applications.

Le partage et la réutilisation de données géographiques issues de sources diverses, institutionnalisés par la directive européenne INSPIRE 2007/2/CE, supposent que celles-ci puissent être combinées de sorte à former un ensemble cohérent. Les efforts consentis en matière de normalisation de l'information géographique ont permis de résoudre l'essentiel des problèmes liés à l'hétérogénéité des formats et des modèles conceptuels de données entre bases de données géographiques indépendantes. La question de l'interprétation de la signification des données disponibles, i.e. de leur sémantique, en revanche, n'est à ce jour pas prise en charge par les normes et standards actuels et demeure un enjeu majeur pour l'intégration de bases de données topographiques hétérogènes.

En effet, l'automatisation de ce processus d'intégration de bases de données topographiques vectorielles suppose celle de la détection des divers types d'hétérogénéités pouvant intervenir entre les bases de données à intégrer, qu'il s'agisse d'hétérogénéité sémantique ou bien de types d'hétérogénéités liés à la structuration et à la représentation géométrique des données. L'existence de ces diverses formes d'hétérogénéités entre bases de données géographiques est due à des règles de saisie différentes entre ces bases, qui sont décrites en détail dans de volumineux documents textuels : les spécifications des bases de données géographiques. La détection de ces divers types d'hétérogénéités entre bases suppose donc de disposer, pour chacune des bases à intégrer, de connaissances relatives à leurs spécifications de saisie.

Nous nous intéressons, dans ce mémoire de thèse, à l'exploitation de cette source de connaissances particulière que sont les spécifications de saisie des bases de données topographiques, dans le processus d'intégration de ces bases.

La partie 2 s'attache à situer les enjeux actuels en matière d'intégration de bases de données topographiques, à rappeler les solutions d'ores et déjà adoptées pour parvenir à cet objectif, et à identifier les difficultés devant encore être surmontées. Dans ce cadre, elle s'attache à préciser les objectifs de ce travail de thèse : formalisation, acquisition et mise en œuvre des connaissances issues des spécifications des bases de données topographiques pour l'intégration de ces bases.

La partie 3 dresse un état de l'art des travaux récents en matière d'intégration de bases de données géographiques hétérogènes, dans le domaine des Infrastructures de Données Géographiques où priment actuellement les questions d'intégration relatives à la mise en œuvre de la directive INSPIRE 2007/2/CE. Elle s'intéresse également aux approches proposées pour la représentation et l'acquisition des connaissances nécessaires à la mise en œuvre des approches d'intégration présentées au début du chapitre. Enfin, elle présente les travaux préconisant d'utiliser les spécifications des bases de données topographiques comme sources de connaissances pour l'intégration de ces bases.

La partie 4 présente notre proposition pour la formalisation, l'acquisition et l'exploitation de connaissances pour la mise en œuvre d'un processus d'intégration de bases de données topographiques. Celle-ci place les spécifications de saisie des bases de données topographiques au cœur du processus d'intégration de ces bases.

Enfin, la partie 5 présente les conclusions et perspectives de ce travail.

2 Contexte et objectifs

Cette partie décrit le contexte général dans lequel s'inscrit ce travail de thèse : l'intégration de bases de données géographiques à l'heure de la directive INSPIRE 2007/2/CE.

Elle présente les particularités des bases de données géographiques auxquelles nous nous intéressons ainsi que les diverses approches proposées pour harmoniser ces données : saisie des données conforme à des spécifications très détaillées, normalisation des diverses composantes d'une base de données géographique, ou encore utilisation d'ontologies comme sources de connaissances externes pour détecter des correspondances entre schémas conceptuels de données.

Elle s'attache enfin à définir les objectifs de cette thèse: la formalisation et l'acquisition des connaissances nécessaires à la mise en œuvre d'un processus d'intégration de bases de données géographiques.

2.1 L'intégration d'information au cœur des défis des infrastructures de données géographiques

Il arrive fréquemment qu'une même zone géographique soit couverte par plusieurs bases de données géographiques, conçues et produites indépendamment les unes des autres, à différentes périodes, par des organismes divers, dans des buts et avec des moyens variés (Craglia et al., 2008). La disponibilité de ces nombreuses sources de données géographiques, susceptibles de renfermer des informations utiles et complémentaires, ouvre la voie à de nombreuses applications dans des domaines où la dimension spatiale revêt un aspect important, tels l'environnement, la gestion de crise, l'agriculture, l'aménagement du territoire, etc. Cependant, les diverses sources de données aujourd'hui disponibles demeurent encore généralement indépendantes les unes des autres, ce qui tend à entraver leur maintenance et leur réutilisation. L'effort de résolution de ces problèmes s'est concentré, au cours des années 1990 et 2000, autour du développement d'applications visant à faciliter l'utilisation de données géographiques : les infrastructures de données géographiques.

2.1.1 Le besoin d'intégration des bases de données géographiques

Les applications exploitant des données géographiques réutilisent généralement des données existantes. En effet, les délais, les coûts et la spécificité des compétences requises pour la production et la maintenance des bases de données géographiques alliés à la variété des bases de données géographiques actuellement disponibles encouragent les utilisateurs à se tourner vers des sources d'information existantes et pouvant répondre à leurs besoins, plutôt qu'à se doter de bases de données spécialement créées pour leurs applications. Cette problématique de réutilisation de données se retrouve également chez les producteurs de données soucieux d'assurer la mise à jour et de garantir la cohérence de leurs bases de données à des coûts raisonnables.

Pour les utilisateurs potentiels de données, la première difficulté réside dans le choix du jeu de données correspondant le mieux à leur besoin particulier, parmi l'éventail des données disponibles. Ceci suppose, d'une part, d'obtenir la liste des données disponibles sur un (ou plusieurs) thème(s) particulier(s), et d'autre part d'avoir accès à des métadonnées suffisamment claires et détaillées pour apprécier l'adéquation de ces données à l'application prévue. Cette étape de recherche des données disponibles et d'évaluation de leur pertinence, dite de « découverte » des données géographiques, est cruciale pour garantir une réutilisation optimale des données existantes. Une deuxième difficulté consiste à accéder simplement aux données choisies. Si l'on s'abstrait des considérations de droits d'accès et d'utilisation des données, les utilisateurs se trouvent encore confrontés à différents obstacles d'ordre technique, qu'il s'agisse d'effectuer une extraction spatiale ou thématique des données pertinentes, ou de bénéficier de modalités simples et rapides de transfert des données du producteur à l'utilisateur. Enfin, pour tirer pleinement profit des diverses bases de données disponibles, celles-ci doivent pouvoir être combinées de façon à former un ensemble cohérent, de sorte que leur utilisation conjointe soit transparente pour l'utilisateur.

Pour un producteur de données géographiques, la constitution et la maintenance en parallèle de plusieurs bases de données indépendantes constituent un processus coûteux en temps et en moyens dans la mesure où les efforts de saisie et de mise à jour sont multipliés. En outre, cette gestion

indépendante des données accroît les risques d'incohérences entre les diverses bases de données produites, qu'il s'agisse de différences d'actualité dues à des mises à jour décalées, ou bien d'erreurs de saisie affectant l'une ou l'autre des bases. Ces incohérences entre bases de données indépendantes peuvent conduire à des difficultés d'interprétation des données si l'on considère les différentes bases dans leur ensemble. Pour pallier ces problèmes, une solution consisterait à mettre à jour en priorité la base de données la plus détaillée et à propager ces modifications sur les autres bases. Cependant, ceci suppose de disposer de liens explicites entre les données des différentes bases, ce qui, en l'absence d'identifiants universels pour les objets géographiques, n'est pas le cas actuellement en France.

En outre, la généralisation de l'usage d'Internet et la mise en place de standards pour le Web, dont ceux du consortium OpenGeospatial (Open Geospatial Consortium, OGC[1]), consortium industriel international regroupant plus de quatre cents entreprises, universités et agences cartographiques nationales avec pour objectif la définition et la mise en œuvre de standards ouverts pour l'information géographique, ont permis une large diffusion de grandes quantités de données géographiques. Parallèlement à cet essor, un mode de production de données géographiques nouveau, fondé sur la collaboration d'amateurs bénévoles, est récemment apparu. Cette forme d'acquisition de données, déjà répandue dans d'autres domaines comme la météorologie ou l'environnement, que Goodchild (2007) nomme « Volunteered Geographic Information », a été rendue possible dans le domaine de l'information géographique grâce au développement et à la démocratisation de technologies permettant à tout un chacun d'acquérir et de partager certaines données géoréférencées. OpenStreetMap[2], qui propose à ses contributeurs de constituer une base de données géographique sur les réseaux routiers du monde entier, à partir de leurs propres données GPS ou d'autres sources de données, ou encore Wikimapia[3], qui propose d'annoter des données géographiques issues de Google Maps[4], constituent ainsi deux exemples de cette nouvelle forme de production participative de données géographiques. A ce nouveau mode de production d'information géographique vont correspondre de nouveaux enjeux. Les producteurs de données géographiques traditionnels, d'une part, peuvent mettre à profit ces nouvelles sources d'informations afin de détecter des mises à jour ou enrichissements possibles de leurs bases de données. Ceci nécessite d'établir les correspondances entre leurs bases de données et celles issues du processus de production participative, afin d'étudier les différences et les complémentarités entre ces bases de données. Les utilisateurs, d'autre part, peuvent souhaiter combiner ces données participatives avec les données dont ils disposent déjà. Enfin, l'évaluation de la qualité de ces données participatives et leur éventuelle correction requièrent une comparaison avec des référentiels existants, ce qui suppose leur mise en correspondance préalable avec ces référentiels.

La mise en cohérence de données issues de sources diverses, que l'on désigne par l'appellation plus générale d'*intégration d'information*, constitue donc un des aspects majeurs de l'interopérabilité qui s'avère crucial à la fois pour les utilisateurs et les producteurs de données géographiques.

[1] http://www.opengeospatial.org/
[2] http://www.openstreetmap.fr/
[3] http://wikimapia.org/wiki/Accueil
[4] http://maps.google.fr/

2.1.2 Des infrastructures de données géographiques pour faciliter l'utilisation des données

Une infrastructure de données géographiques est une solution de médiation entre des utilisateurs ayant besoin de données géographiques et des données issues de sources diverses, qui vise à faciliter l'accès, le partage, l'échange et l'utilisation de ces données ainsi qu'à fournir des services liés à ces données (Bucher, 2009). A ces fins, elle s'appuie sur un ensemble de technologies, de politiques, de normes et de ressources nécessaires pour acquérir, manipuler, stocker, distribuer et améliorer l'utilisation de données géographiques (Nebert, 2004). La création, dès 1996, de l'association internationale pour les infrastructures de données géographiques (Global Spatial Data Infrastructure Association[5]) dédiée à la diffusion des bonnes pratiques et au retour d'expériences en matière de développement d'infrastructures de données géographiques a contribué à leur essor et à leur rayonnement dans la communauté de l'information géographique.

En particulier, l'adoption en 2007, par le Parlement et le Conseil de l'Union Européenne, de la directive INSPIRE 2007/2/CE, impose aux États membres de l'Union Européenne de mettre en place une infrastructure d'information géographique européenne au sein de laquelle « il soit possible de combiner de manière cohérente les données géographiques tirées des différentes sources de la Communauté et de les partager entre plusieurs utilisateurs et applications » (Parlement Européen et Conseil de l'Union Européenne, 2007). Voulue par la commission chargée de l'environnement, cette infrastructure doit faciliter l'accès et l'utilisation des données géographiques permettant « la prise de décision concernant les politiques et les activités susceptibles d'avoir une incidence directe ou indirecte sur l'environnement ». Pour ce faire, elle doit s'appuyer sur les infrastructures de données géographiques nationales de chacun des États membres, rendues compatibles par l'adoption de règles communes de mise en œuvre.

En France, il s'agit en particulier du portail de l'information géographique publique, nommé Géoportail, lancé par la Direction Générale de la Modernisation de l'État. Il a pour objectif de « constituer un point d'entrée le plus large possible pour rechercher les principales données géographiques de l'État, de ses établissements publics et des collectivités territoriales, en connaître leurs caractéristiques et les moyens d'y accéder et de les visualiser et les co-visualiser ». Ce portail doit être « ouvert et interopérable, permettant ainsi la fédération des données, en s'appuyant sur les normes en vigueur, sans restreindre les choix techniques » (Charte du portail de l'information géographique publique, 2006). Il intègre à la fois les données de référence de l'Institut National de l'Information Géographique et Forestière (IGN), en particulier le Référentiel à Grande Echelle, et celles de divers partenaires producteurs de données géographiques. Les services de visualisation de données, sous maîtrise d'œuvre de l'IGN, et ceux de découverte des données, sous maîtrise d'œuvre du Bureau de Recherches Géologiques et Minières (BRGM), sont donc implémentés conformément aux règles de mise en œuvre de la directive INSPIRE 2007/2/CE.

Dans le cadre de la directive INSPIRE 2007/2/CE, un état des lieux des infrastructures de données géographiques nationales européennes est régulièrement mené afin de dresser le bilan de leurs niveaux de développement respectifs (Vandenbroucke et al., 2010). Le projet européen eSDI-NET+[6] a

[5] http://www.gsdi.org/

[6] http://www.esdinetplus.eu/

également conduit une évaluation des infrastructures de données géographiques européennes de niveau local, en termes de bonnes pratiques (eSDI Net+ Consortium, 2010). De même, l'Association Française pour l'Information Géographique (AFIGEO[7]) a dressé un catalogue des infrastructures de données géographiques françaises (Dewynter et Ladurelle-Tikry, 2010), recensant quelques 54 applications de ce type.

Ces nombreuses initiatives témoignent du besoin de pouvoir accéder aisément à des données géographiques cohérentes. Les bonnes pratiques en vigueur dans le domaine des infrastructures de données géographiques s'attachent à apporter des solutions concernant les différents aspects de l'interopérabilité des données géographiques, notamment au travers de l'usage de normes.

2.1.3 L'intégration de données dans les infrastructures de données géographiques

Une infrastructure de données géographiques comporte à la fois des données géographiques, de la documentation sur ces données, ou métadonnées, des outils pour découvrir, visualiser et évaluer les données disponibles, appelés catalogues, et des moyens pour accéder aux données. Des services supplémentaires exploitant les données disponibles peuvent être proposés. Les premiers travaux dédiés à la mise en place d'infrastructures de données géographiques se sont donc concentrés sur la définition de modèles standards et de normes s'attachant à permettre, d'un point de vue technique, la réalisation de cet objectif de connexion des usages aux données. En particulier, les efforts du comité technique TC 211 de l'organisation internationale de normalisation (International Standardisation Organisation - ISO), spécialement dédié à l'information géographique, ainsi que ceux du consortium OpenGeospatial, ont conduit à la mise en place d'un ensemble de standards et de normes dédiés à cet objectif.

Pour les données vectorielles, auxquelles nous nous intéressons plus particulièrement dans le cadre de cette thèse, la norme « ISO 19101 - Reference model » constitue le socle d'un ensemble de normes visant à permettre la représentation de l'information géographique numérique. La norme « ISO 19107 - Spatial schema » fournit un schéma conceptuel de données permettant de décrire et manipuler les caractéristiques spatiales d'un phénomène géographique. Elle s'articule autour de la notion de ***Feature***, entité représentant une abstraction d'un phénomène du monde réel associée à une localisation sur la Terre. Ainsi un ensemble de *Features* peut être vu comme une représentation de l'espace géographique. A un *Feature* peuvent être rattachées une ou plusieurs primitives géométriques et topologiques permettant la représentation de sa forme et de sa localisation, également décrites par cette norme, ainsi qu'un ensemble d'attributs décrivant sa sémantique. La norme « ISO 19109 - Rules for application schema » propose un schéma conceptuel permettant de décrire des schémas conceptuels de jeux de données géographiques. Elle s'appuie sur la notion de ***Feature Type*** ou « type d'entité » qui regroupe un ensemble de *Features* du même type, ayant des propriétés et des relations communes. Le Tableau 1 présenté ci-dessous dresse un aperçu des principales normes dont l'usage est recommandé en tant que bonne pratique dans la mise en place d'une infrastructure de données géographiques.

[7] http://www.afigeo.asso.fr/

Norme ou standard	Objectif
ISO 19107 - Spatial schema	Fournit un schéma conceptuel de données permettant de décrire et manipuler les caractéristiques spatiales d'un phénomène géographique.
ISO 19108 - Temporal schema	Fournit un schéma conceptuel de données permettant de décrire et manipuler les caractéristiques temporelles d'un phénomène géographique.
ISO 19109 - Rules for application schema	Fournit un schéma conceptuel permettant de décrire des schémas conceptuels de jeux de données géographiques.
ISO 19110 - Methodology for feature cataloguing	Fournit la méthodologie à adopter pour cataloguer l'ensemble des *Feature Types* présents dans un jeu de données.
ISO 19111 - Spatial referencing by coordinates	Fournit un modèle pour le géoréférencement par coordonnées géographiques se rapportant à un système de coordonnées dont la description est également prise en charge par la norme.
ISO 19112 - Spatial referencing by geographic identifiers	Fournit un modèle pour le géoréférencement par identifiants géographiques appartenant à un index géographique dont la description est également prise en charge par la norme.
ISO 19115 - Metadata	Fournit une spécification abstraite de métadonnées pour la découverte et l'exploration de données géographiques.
ISO 19117 - Portrayal	Fournit un schéma conceptuel pour la représentation graphique des Features.
ISO 19119 - Services	Fournit les métadonnées d'identification d'un service permettant d'accéder à des données ou de les traiter.
ISO 19131 - Data product specification	Fournit un schéma conceptuel de données pour décrire les spécifications d'un jeu de données géographiques.
ISO 19136 - Geography Markup Language	Fournit un ensemble de schémas XML pour l'encodage de données géographiques. Les concepts fondamentaux sur lesquels repose ce format de données sont ceux définis dans la série de normes ISO 191XX et dans les spécifications abstraites de l'OGC. L'ISO 19136 correspond au standard OCG GML 3.2.1 (OGC, 2007).
ISO 19139 - Metadata implementation specification	Fournit des règles d'encodage XML pour des métadonnées conformes à la norme ISO 19115.
OGC Web Feature Service (WFS)(OGC, 2010)	Spécifie l'interface d'un service d'accès aux données.
OGC Catalogue Services for the Web (CSW) (OGC, 2007)	Spécifie l'interface d'un service de catalogage de données.

Tableau 1: Normes recommandées dans les infrastructures de données géographiques

Les infrastructures de données géographiques s'appuient sur ces normes pour décrire et manipuler des données géographiques (Reed, 2009). Elles les mettent également à profit au sein d'applications de catalogage ou d'intégration de données. Ainsi, dans le cadre de la mise en place de la directive INSPIRE 2007/2/CE, les normes ISO 19109 et 19131 ont été utilisées pour définir le schéma global de l'infrastructure de données géographiques européenne et ses spécifications. Les données elles-mêmes, conformes au schéma INSPIRE 2007/2/CE, devront être fournies par les États membres au format GML. Des métadonnées conformes à la norme ISO 19115 et à ses spécifications d'implémentation ISO 19139 devront venir documenter les données entrant dans le cadre de la directive et serviront de source d'information à des applications de catalogage visant à permettre la découverte et l'exploration des données.

La définition et la mise en œuvre de ces normes a donc permis de résoudre l'essentiel des problèmes liés à l'hétérogénéité des formats et des modèles conceptuels de données entre bases de données géographiques indépendantes. En effet, la majorité des logiciels de Système d'Information Géographique (SIG) actuels sont capables de lire les formats de données standards et donc d'afficher les attributs et la géométrie d'objets modélisés en tant que Feature. De même, la plupart des outils de catalogage peuvent facilement intégrer des métadonnées structurées selon les normes ISO en vigueur. En revanche, aucun de ces logiciels n'est capable d'interpréter les données qu'il lit et d'en appréhender la signification exacte. En effet, l'interprétation de la signification des données, ou sémantique, qui est conçue par (Kavouras et Kokla, 2008) comme la relation des données aux phénomènes du monde réel qu'elles représentent, n'est toujours pas prise en charge par les standards actuels. Ainsi un logiciel de SIG disposant de données ayant pour Feature Type, *Tronçon Hydrographique*[8] et *Cours d'eau* ne pourra pas faire de rapprochement sémantique entre ces deux jeux de données représentant pourtant un même type d'entités géographiques du monde réel. En outre, il ne sera pas plus capable de déterminer quels Features de type *Tronçon Hydrographique* et *Cours d'eau* représentent la même entité géographique du monde réel. C'est pourquoi l'intégration de données issues de sources hétérogènes au sein d'une même infrastructure demeure une tâche complexe et constitue encore l'un des grands défis que doivent relever les infrastructures de données géographiques (Craglia et al., 2008). L'un des principaux obstacles à cette intégration demeure l'hétérogénéité des bases de données géographiques vectorielles.

2.2 Les difficultés d'interprétation des données géographiques vectorielles liées à leur hétérogénéité

Dans cette thèse, nous nous intéressons à un type particulier de bases de données géographiques vectorielles : les bases de données topographiques. Ces bases de données visent à décrire la topographie du terrain. Elles fournissent une représentation de l'espace géographique en un instant donné, partielle et non unique. Elles résultent d'une abstraction du monde réel qui dépend du point de vue adopté par le producteur de la base (Fonseca et al., 2003). Leur schéma conceptuel est donc composé de classes dont les noms désignent le plus souvent des concepts géographiques que l'on manipule couramment (les forêts, les unités administratives, les rivières, les routes, les bâtiments,

[8] Pour des raisons de lisibilité, tout au long de ce mémoire, les noms des éléments de schémas de bases de données seront notés en italiques. Les labels d'éléments d'ontologies seront notés en petites majuscules.

etc.). Les entités géographiques du monde réel y sont représentées de façon simplifiée, sous la forme d'objets élémentaires et identifiables. Leur forme et leur localisation sont représentées par une ou plusieurs primitives géométriques (point, ligne ou polygone), choisies en fonction de la nature de l'entité géographique à représenter et du niveau de détail géométrique de la base, tandis qu'un ensemble d'attributs alphanumériques vient compléter leur description.

Conçues et produites de façon indépendante, avec des objectifs et des moyens différents, les diverses bases de données géographiques existantes peuvent s'avérer hétérogènes à différents points de vue (Sheth et Larson, 1990). Si ce phénomène d'hétérogénéité entre bases de données indépendantes n'est pas propre aux données géographiques, et est largement décrit dans la littérature dédiée aux bases de données (Batini et al., 1986, Sheth et Larson, 1990, Kashyap et Sheth, 1996, Wache, 2003), celles-ci y sont probablement plus sujettes en raison des nombreuses difficultés liées à la catégorisation des entités géographiques, à leur modélisation au sein d'une base de données, ainsi qu'à leur représentation géométrique. La représentation de l'espace géographique dépend donc fortement du point de vue adopté.

2.2.1 L'absence de catégorisation universelle des entités géographiques

Smith et Mark (1998) définissent les entités géographiques comme des entités qui non seulement ont une localisation dans l'espace, mais sont intrinsèquement liées à cette localisation qui conditionne la plupart de leurs propriétés physiques, qu'elles soient géométriques, topologiques ou méréologiques. Ils soulignent l'importance de ces propriétés dans le processus de catégorisation des entités géographiques.

Leur taille ou leur forme, tout d'abord, constituent des propriétés discriminantes, conditionnant la classification d'entités géographiques, pourtant semblables à de nombreux points de vues, au sein de catégories différentes. C'est le cas, par exemple, des lacs, des étangs et des mares, qui représentent tous trois des étendues d'eau terrestres que l'on distingue essentiellement en raison de leur taille. Cependant, il n'existe pas pour autant de seuil précis et universellement admis permettant de déterminer à laquelle de ces catégories appartient une étendue d'eau terrestre. Ainsi dans la plupart des cas, le critère de taille est observé de façon plus qualitative que quantitative, par rapport à un contexte géographique et culturel, ce qui d'une communauté à l'autre peut induire des recouvrements entre catégories d'entités géographiques.

Le rôle de la topologie, par ailleurs, et en particulier la détermination de la localisation des limites des entités géographiques, sont particulièrement mis en avant par Smith et Mark (1998). Ces derniers insistent sur l'importance qu'il y a à établir une distinction entre les limites qui relèvent de discontinuités visibles de l'espace, qu'ils nomment « bona fide boundaries » – les frontières authentiques, comme le contour d'une île par exemple, et celles résultant d'un processus de délimitation d'origine humaine, ou « fiat boundaries » – les frontières établies par décret. Ces dernières peuvent découler d'un choix arbitraire, comme pour des frontières entre États, ou d'un processus cognitif fondé sur la vérification de critères de décision subjectifs, établie sur la base d'un ensemble d'observations, comme c'est le cas pour la détermination de limites entre types de sols. Pour autant ces deux principaux types de limites ne sont pas incompatibles. En effet, une limite de type « fiat » peut être composée d'éléments de type « bona fide » ; c'est le cas d'une limite entre

deux communes qui suit le tracé d'un cours d'eau. Par ailleurs, si certaines limites d'entités géographiques sont précisément identifiables, d'autres en revanche sont vagues par nature. Les vallées, les montagnes, ou encore les baies, dont les contours peuvent seulement être estimés par une région de l'espace, constituent des exemples de ces entités géographiques dont les limites exactes ne peuvent être déterminées dans la mesure où elles n'existent pas. En effet, celles-ci sont des éléments constitutifs d'un même phénomène continu, le relief, que l'on cherche à discrétiser en le partitionnant en catégories composées de formes de relief semblables. Ce processus cognitif dépend largement de considérations linguistiques et culturelles, et conduit nécessairement à la définition de catégories différentes, ainsi qu'à des critères de délimitation des entités géographiques appartenant à chacune de ces catégories différents d'une communauté à l'autre. Ainsi, la définition des catégories d'entités géographiques dotées de limites de type « fiat », et la localisation des limites de chacune de leurs entités, seront plus soumises à variation d'une communauté à l'autre que dans le cas des catégories d'entités dotées de limites de type « bona fide ».

Ainsi, la complexité de l'espace géographique peut conduire à de nombreuses interprétations selon le point de vue adopté. Or, chaque communauté d'acteurs de l'information géographique possède un point de vue particulier sur l'espace géographique. Aussi, des communautés différentes peuvent-elles avoir des points de vue divergents concernant une même catégorie d'entités du monde réel, ou à l'inverse utiliser des termes différents pour désigner une même catégorie d'entités géographiques. Une même entité géographique du monde réel pourra donc avoir différentes descriptions et interprétations selon la base de données au sein de laquelle elle est représentée, phénomène que Bishr (1998) qualifie d'**hétérogénéité sémantique**. Une classe de base de données nommée *Point d'eau* pourra donc avoir diverses interprétations : s'agit-il ici de représenter uniquement les points d'eau naturels, comme les sources, ou bien s'agit-il des points d'eau aménagés, comme les sources captées, les citernes, les abreuvoirs et les fontaines, ou bien souhaite-t-on disposer de l'ensemble des points d'eau potable auxquels un randonneur pourra se ravitailler? De plus, au sein d'une autre base de données, des choix terminologiques différents pourront être faits, de sorte que les points d'eau représentés ici dans la classe *Point d'eau* seront instanciés au sein de l'autre base dans une classe nommée *Source*, par exemple.

2.2.2 Différents schémas conceptuels pour structurer les données

Les schémas conceptuels de bases de données géographiques ne visent pas seulement à catégoriser les entités géographiques mais à structurer l'information qui sera représentée dans la base (Partridge, 2002). Une même catégorie d'entités géographiques du monde réel pourra donc être modélisée de façons différentes d'un schéma conceptuel de données à un autre. Une instance d'une base de données pourra par exemple être traitée comme un attribut dans une autre base. On parle alors d'**hétérogénéité schématique** (Bishr, 1998) entre bases de données. Les schémas conceptuels de données ne reflètent donc pas directement la conceptualisation du terrain de leur producteur. La nécessaire prise en compte de l'application à laquelle la base est destinée, ainsi que celle de certaines contraintes techniques de modélisation introduisent un biais dans le processus de conceptualisation du terrain : il ne s'agit plus ici de définir les catégories d'entités géographiques existant sur le terrain, mais de déterminer quelle sera la façon la plus appropriée de les représenter au sein d'une base de données, ce qui peut également conduire à des cas d'hétérogénéité sémantique entre bases.

En effet, une même classe de base de données pourra représenter des entités géographiques appartenant à des catégories différentes. Ce type de regroupement vise généralement à simplifier la structure de la base en réduisant son nombre de classes. Dans ce cas, le regroupement est effectué en raison de l'existence d'une relation de généralisation-spécialisation entre les différentes catégories regroupées. Considérons l'exemple de l'une des bases de données topographiques vectorielles produites par l'IGN, la BDTOPO© 2.0 (IGN, 2011a) ; la classe *Point d'eau*, renferme à la fois des citernes, des fontaines, des sources, des sources captées et des stations de pompage. Ces diverses entités géographiques sont représentées dans une même classe nommée à l'aide d'un terme faisant référence à une catégorie d'entités géographiques plus générique. La catégorie exacte à laquelle appartient chaque instance de la classe est fournie à travers les valeurs possibles d'un attribut appelé *Nature*. En outre, ces regroupements résultent généralement de l'adoption d'un point de vue particulier sur une catégorie d'entités géographiques pouvant être considérée comme une spécialisation de différentes catégories plus génériques. Le choix effectué est alors guidé par l'application à laquelle la base est destinée. C'est le cas, par exemple, pour les aqueducs dans les bases de données topographiques BDTOPO© Pays 1.2 (IGN, 2002) et BDCARTO© 3.0 de l'IGN (IGN, 2006). Les aqueducs sont représentés, dans la première, au sein de la classe *Canalisation*, alors qu'ils sont représentés, dans la seconde, dans la classe *Tronçon hydrographique*.

De plus, la représentation au sein d'une base de données géographique de certaines catégories d'entités géographiques nécessite une modélisation particulière du phénomène représenté. La prise en compte de la topologie des réseaux, comme les réseaux hydrographiques par exemple, impose un découpage des entités linéaires à représenter. Ainsi, la catégorie d'entités géographiques ROUTE est représentée au sein du schéma conceptuel de la BDTOPO© 2.0 produite par l'IGN par une classe nommée *Route* dont les instances sont en fait des tronçons de route. La définition de la classe précise qu'un tronçon de route est une « Portion de voie de communication destinée aux automobiles, aux piétons, aux cycles ou aux animaux, homogène pour l'ensemble des attributs et des relations qui la concerne […] ». Notons que la définition des tronçons de route dépend des valeurs d'attributs que prendront les instances de la classe *Route*, au lieu de dépendre des propriétés des entités géographiques de type « Route » elles-mêmes. Cette classe ne se rapporte donc pas directement à une catégorie d'entités géographiques spontanément identifiables et ses instances dépendront avant tout du schéma conceptuel adopté. A ce niveau peuvent donc intervenir des **conflits de fragmentation** (Devogele, 1997). Ceux-ci concernent le découpage de certaines entités géographiques complexes en plusieurs instances de bases de données en fonction de la valeur qui sera affectée à l'un des attributs de chacune de ces instances. Parmi ces conflits, les **conflits de segmentation** interviennent lorsque le critère de fragmentation porte sur des attributs différents des deux classes homologues. A titre d'exemple, des routes peuvent être découpées en tronçons de route selon leur nombre de voies dans une base et selon leur type de revêtement dans une autre. La représentation d'une même route fournie par deux bases dotées de critères de segmentation différents variera donc d'une base à l'autre.

Ainsi, au-delà des cas d'hétérogénéité schématique (Bishr, 1998) que l'on rencontre aussi bien entre bases de données géographiques qu'entre bases de données classiques, la définition des schémas de bases de données géographiques peut également conduire à des cas d'hétérogénéité sémantique entre ces bases. En effet, des choix de regroupement ou de classification différents peuvent conduire à des cas d'hétérogénéité sémantique entre bases. De plus, la modélisation de certaines catégories d'entités géographiques du monde réel conduit à la définition d'entités géographiques dites

« modélisées », qui correspondent généralement à des portions ou des agrégations d'entités géographiques. Elles ne correspondent donc pas directement à des entités géographiques spontanément identifiables sur le terrain, mais seront néanmoins instanciées dans la base. Dans ce cas, des choix de modélisation différents conduiront nécessairement à la définition d' « entités modélisées » différentes, et donc à la saisie d'instances éventuellement très différentes d'une base à l'autre pour représenter un même phénomène du monde réel. Ainsi les schémas conceptuels de données adoptés peuvent s'avérer sources de nombreuses hétérogénéités entre bases de données, à la fois au niveau de schémas eux-mêmes, mais également au niveau des données.

2.2.3 L'hétérogénéité des niveaux de détail des bases de données géographiques

La notion de point de vue recoupe celle de niveau de détail de la base de données qui est le pendant, dans le domaine des bases de données géographiques, de celle d'échelle dans celui de la cartographie. Elle englobe à la fois les notions de résolution – ordre de grandeur géométrique des phénomènes présents dans la base, de précision géométrique – écart entre la position des objets sur le terrain et leur position dans la base, et de granularité – taille des plus petites formes représentées dans la base (Ruas et Bianchin, 2002). Ainsi, une base de données dotée d'une échelle caractéristique de l'ordre du 1 :25 000 est une base de données contenant des données à un niveau de détail équivalent à celui d'une carte au 1 :25 000. Le choix des types d'entités géographiques devant figurer au sein d'une base de données géographiques et celui de leur représentation, dépendent donc, non seulement du domaine modélisé et de la communauté d'acteurs visée, mais également du niveau de détail géométrique de celle-ci. Ceci va conduire à des types d'hétérogénéités propres aux données géographiques. Devogele (1997) propose une taxonomie de ces conflits dont nous avons directement extrait les cas présentés ci-dessous et illustrés en figure 1.

Certains concepts géographiques ne peuvent être représentés à des échelles trop grandes ou trop petites. A titre d'exemple, une forêt ne pourra pas être représentée à très grande échelle dans la mesure où il ne sera ni possible, ni pertinent de chercher à déterminer la position de sa lisière. Il s'agit, en effet, d'une entité mésoscopique qui regroupe un ensemble d'entités microscopiques, les arbres, pouvant être identifiées individuellement (Ruas, 1999). A petite échelle, en revanche, dans la mesure où ses dimensions le permettent, elle pourra être représentée comme un objet unique dont le contour est cohérent avec la précision perceptible. L'échelle caractéristique de la base de données va donc contraindre les possibilités de représentation de certains concepts géographiques au sein de la base.

En fonction du niveau de détail, une sélection sera opérée parmi les types d'entités géographiques devant figurer dans la base. En effet, seules les entités géographiques dont les dimensions sont cohérentes avec le niveau de détail de la base seront retenues. De ce fait, entre deux bases de données géographiques ayant des niveaux de détail différents éclateront nécessairement des **conflits de critères de sélection** (Devogele, 1997). Ainsi, on peut observer un conflit de critère de sélection entre la classe *Massif boisé* de la BDCARTO© 3.1 (IGN, 2011b) et la classe *Zone de végétation* de la BDTOPO© 2.0. Les critères de sélection de la première stipulent que doivent être sélectionnés pour figurer dans cette classe « […] les bois et forêts d'une superficie supérieure à 500 ha […] ». Ceux de la seconde désignent les « […] bois de plus de 500 m^2 […] ». Le critère de taille que doivent vérifier les zones arborées pour figurer dans l'une ou l'autre des deux bases porte ici sur des seuils différents, ce

qui impliquera des différences d'exhaustivité importantes entre les deux classes, dont on pourrait croire, à en juger par leurs dénominations respectives, qu'elles représentent des phénomènes homologues du monde réel.

La représentation de certaines entités géographiques complexes au sein d'une base de données impose parfois leur décomposition en plusieurs instances. Cette décomposition intervient lorsqu'une (ou plusieurs) propriété(s) géométrique(s) des entités du monde réel concernées atteint (atteignent) un certain seuil. Un **conflit de critère de décomposition** (Devogele, 1997) peut donc intervenir entre deux classes homologues lorsque les seuils de décomposition imposés à chacune des classes sont différents.

Les choix de modélisation géométrique des entités géographiques sont également soumis à celui de l'échelle caractéristique de la base. En effet, à grande échelle la surface occupée par un cours d'eau pourra être représentée sous la forme d'un polygone, tandis que ce même cours d'eau se résumera à l'axe de son lit principal, représenté par une simple ligne, à plus petite échelle. Des **conflits de description géométrique** (Devogele, 1997) peuvent donc survenir entre bases de données dotées de niveaux de détail différents et au sein desquelles un même phénomène est représenté par des primitives géométriques différentes.

La granularité d'une base de données géographique vectorielle désigne la taille du plus petit objet représentable dans cette base. Par exemple, les spécifications de la classe *Bâti indifférencié* de la BDTOPO© 2.0 précisent que "Seules les cours intérieures de plus de 10 m de large sont représentées par un trou dans la surface bâtie.". Une base de données dotée d'un critère de saisie des cours intérieures différent présenterait alors ce que Devogele (1997) appelle un **conflit de résolution** – « conflit de granularité » semblerait plus approprié ici - avec la BDTOPO© 2.0. Dans le même ordre d'idées, Devogele (1997) nomme **conflits de granularité**, expression à laquelle on pourrait préférer « conflit de seuil de segmentation », les conflits qui se produisent lorsque deux classes homologues présentent un même critère de fragmentation mais des contraintes différentes sur la taille minimale des instances pouvant être créées. Ainsi des tronçons de route peuvent être découpés dans deux bases en fonction de leur nombre de voies, mais une contrainte supplémentaire peut également préciser que ce découpage n'interviendra qu'à la condition que chaque tronçon saisi mesure plus de 200 mètres dans une base et plus de 1000 mètres dans l'autre.

Enfin, de nombreux **conflits de données** (Devogele, 1997) peuvent résulter du processus de saisie d'une base de données géographique vectorielle, en particulier en raison de la saisie manuelle de la géométrie qui peut conduire à des restitutions différentes d'une base à l'autre. Les conflits de données engendrés par les opérations de généralisation effectuées sur les données, par exemple, sont particulièrement complexes à identifier et à traiter. Cette dénomination de **conflits de données** demeure très générale, aussi semblerait-il préférable de lui substituer des expressions dédiées à chacun des conflits de ce type, comme « conflit d'agrégation » pour l'exemple présenté en figure 1.

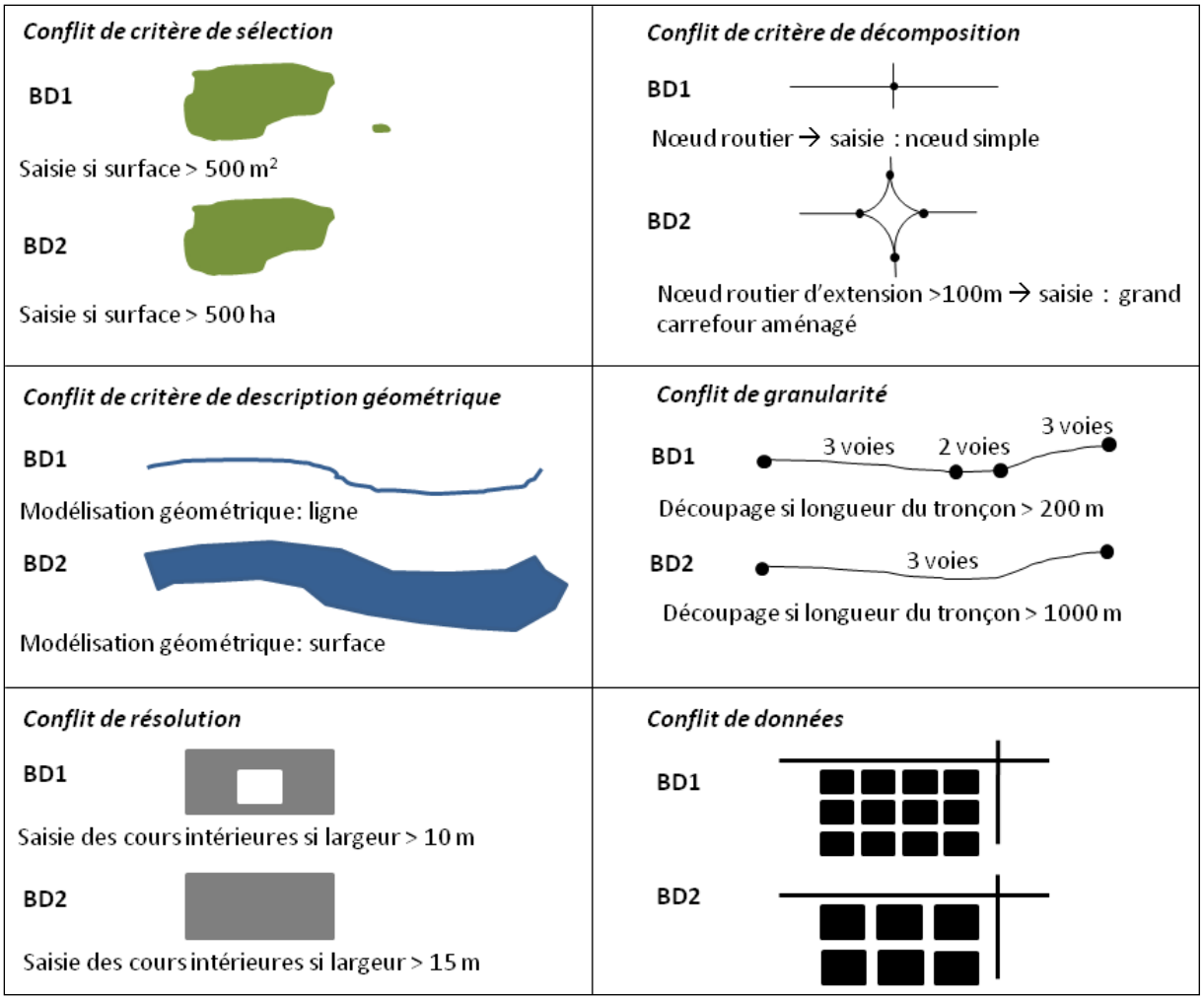


Figure 1: Exemples d'hétérogénéités propres aux données géographiques

Deux bases de données géographiques pourvues de niveaux de détail différents ne fourniront donc pas la même description du territoire, ceux-ci influant sur leurs contenus, à la fois en termes de concepts représentés, de modélisation géométrique de ces concepts, de sélection des entités géographiques qui figureront dans chacune des bases, et de représentation géométrique des entités. L'ensemble des choix effectués obéit généralement à des critères stricts et le plus souvent quantitatifs. Cependant, à cette étape intervient la notion subjective de « trait caractéristique » du terrain qui permet d'inclure dans la base des entités géographiques ne satisfaisant pas aux critères de sélection géométriques, mais dont la représentation au sein de la base constitue une information pertinente vis-à-vis de la description du paysage. Une certaine liberté concernant l'interprétation du monde réel est donc laissée aux opérateurs de saisie. Cette liberté permet de garantir la qualité de la base de données dans la mesure où elle fait intervenir les connaissances et l'expérience cartographique des opérateurs chargés de la saisie sur les points les plus délicats. En revanche, cela implique des variations possibles d'un opérateur à l'autre.

2.2.4 Des spécifications pour assurer l'homogénéité des données au sein d'une même base de données géographique

Afin de garantir un niveau d'homogénéité satisfaisant au sein d'une base de données géographique, les producteurs de données consignent généralement l'ensemble des règles définissant le contenu des classes de leurs bases dans des documents textuels destinés aux opérateurs de saisie de la base : les spécifications des bases de données géographiques.

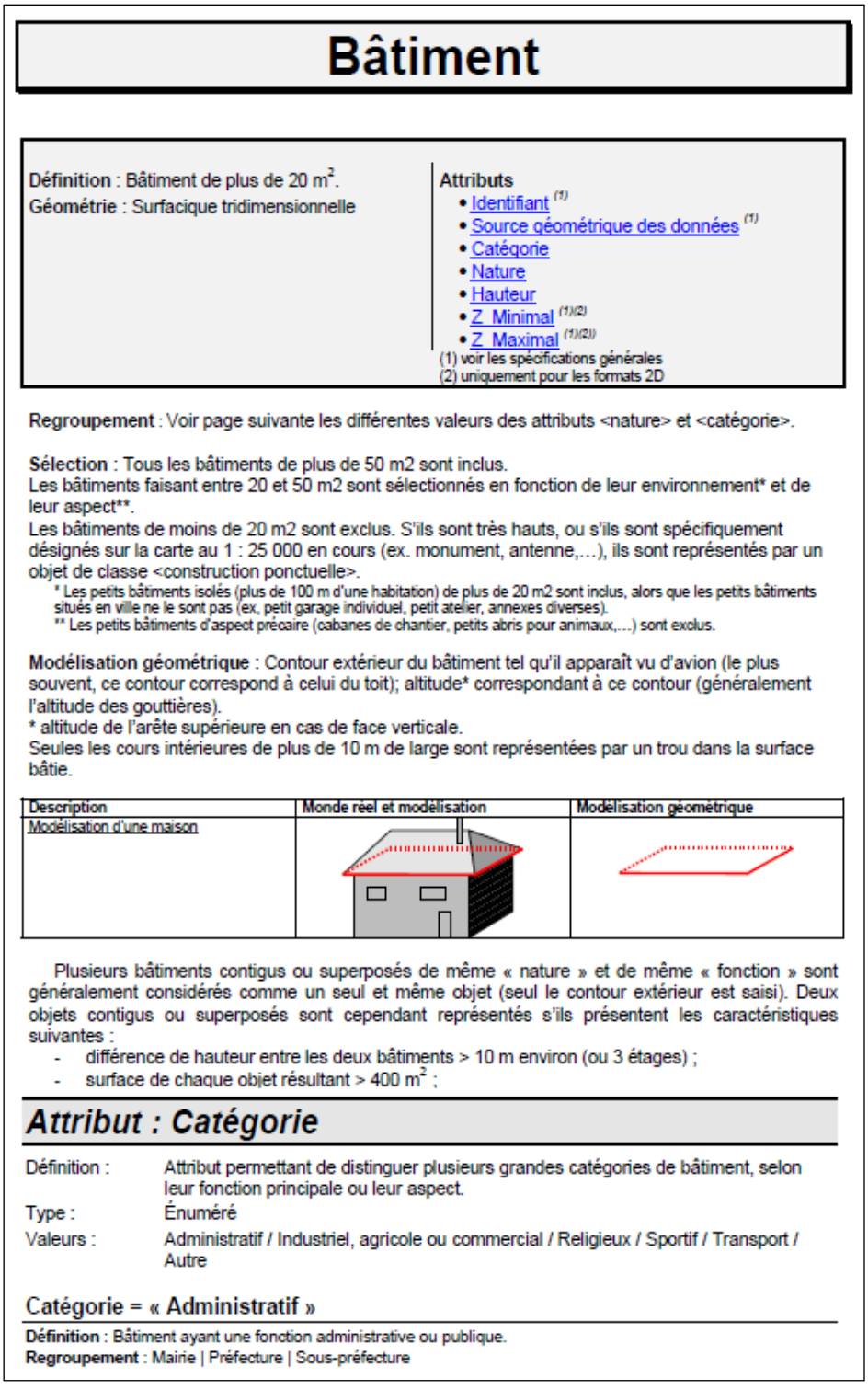
Dans les spécifications des bases de données de l'IGN (voir extrait figure 2), le contenu de chaque classe d'une base est décrit au sein d'une fiche de spécifications particulière. Celle-ci, dont le titre correspond au nom de la classe qu'elle décrit, fournit une définition du (ou des) type(s) d'entités géographiques représenté(s) au sein de la classe. Elle précise également le type de primitive géométrique utilisé pour représenter la forme et la localisation des entités géographiques qui seront saisies dans cette classe.

Puis, l'ensemble des règles de sélection que doivent vérifier les entités du monde réel correspondant au type d'entités à représenter dans cette classe afin d'être représentées au sein de cette classe est fourni. Il s'agit généralement de critères fondés sur les dimensions des entités géographiques du monde réel, seules les entités disposant de dimensions conformes au niveau de détail de la base étant retenues. Ainsi, sur la fiche de spécifications présentée en figure 2, qui décrit la classe *Bâtiment* de la BDTOPO©, on peut lire le critère de sélection suivant : « Tous les bâtiments de plus de 50 m^2 sont inclus ». Une concession à cette règle est tout de même ajoutée, afin de permettre la saisie de bâtiments de superficie plus faible mais dont la présence au sein de la base paraît pertinente au regard de leur rôle structurant dans la description du paysage : « Les bâtiments faisant entre 20 m^2 et 50 m^2 sont sélectionnés en fonction de leur environnement et de leur aspect ».

D'autres règles de représentation, concernant la manière dont les instances de la classe doivent être saisies et représentées sont ajoutées dans la partie « modélisation géométrique ». Ainsi, chaque bâtiment du monde réel saisi comme instance de la classe « Bâtiment » de la BDTOPO© est représenté par un polygone dont les contours correspondent à ceux de son toit: « Contour extérieur du bâtiment tel qu'il apparaît vu d'avion (le plus souvent, ce contour correspond à celui du toit) ; ». A ces règles de modélisation géométrique générales peuvent s'ajouter des règles plus complexes faisant intervenir des contraintes de généralisation cartographique, procédé qui consiste à adapter la représentation de l'information géographique à un niveau d'analyse donné. Ainsi, des bâtiments spatialement très proches et ayant des propriétés similaires seront saisis dans la base comme un seul et même bâtiment : « 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) ».

Enfin, la liste des attributs descriptifs des instances de la classe est détaillée. En particulier, on trouvera pour chaque attribut, sa définition, son type et ses valeurs possibles. Chacune des valeurs possibles sera à son tour décrite, lorsque son affectation aux instances de la classe est sujette à conditions. La figure 2 présente en partie la description de l'attribut *Catégorie* de la classe *Bâtiment*. Il s'agit d'un attribut énuméré de type « chaîne de caractères », dont les valeurs possibles sont : *Administratif*, *Industriel, Agricole ou commercial*, *Religieux*, *Sportif*, *Transport* et *Autre*. Pour qu'une instance de la classe *Bâtiment* se voie affecter, pour l'attribut *Catégorie*, la valeur *Administratif*, il

faut que cette instance représente un bâtiment du monde réel ayant comme fonction d'être une mairie, une préfecture ou une sous-préfecture.



Bâtiment

Définition : Bâtiment de plus de 20 m^2.
Géométrie : Surfacique tridimensionnelle

Attributs
- Identifiant (1)
- Source géométrique des données (1)
- Catégorie
- Nature
- Hauteur
- Z_Minimal (1)(2)
- Z_Maximal (1)(2)

(1) voir les spécifications générales
(2) uniquement pour les formats 2D

Regroupement : Voir page suivante les différentes valeurs des attributs <nature> et <catégorie>.

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 exclus. S'ils sont très hauts, ou s'ils sont spécifiquement désignés sur la carte au 1 : 25 000 en cours (ex. monument, antenne,...), ils sont représentés par un objet de classe <construction ponctuelle>.

* 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 > 400 m^2 ;

Attribut : Catégorie

Définition : Attribut permettant de distinguer plusieurs grandes catégories de bâtiment, selon leur fonction principale ou leur aspect.
Type : Énuméré
Valeurs : Administratif / Industriel, agricole ou commercial / Religieux / Sportif / Transport / Autre

Catégorie = « Administratif »

Définition : Bâtiment ayant une fonction administrative ou publique.
Regroupement : Mairie | Préfecture | Sous-préfecture

Figure 2: Extrait des spécifications de la BDTOPO® 1.2: la classe "Bâtiment"

L'intérêt des spécifications des bases de données géographiques vectorielles est double. D'une part, elles permettent de guider la saisie d'une base de données géographique, garantissant ainsi son homogénéité. D'autre part, une fois la base saisie, elles constituent une source de connaissances

extrêmement riche concernant le contenu exact de la base et permettent à leurs lecteurs de comprendre précisément quelle représentation du territoire leur est proposée à travers cette base de données.

2.3 Approches proposées pour l'intégration d'informations dans le domaine de l'informatique

La nécessité de résoudre les conflits d'intégration inhérents à l'hétérogénéité des bases de données afin de disposer, à partir d'un ensemble de bases de données hétérogènes, d'un ensemble intégré d'informations constitue, depuis les années 1980, un enjeu majeur dans le domaine de la recherche en informatique (Rahm et Berstein, 2001). Cette intégration peut se faire via différentes approches, dépendant le plus souvent de l'application finale visée, et de façon plus ou moins poussée. Cependant, dans tous les cas elle nécessitera de résoudre l'hétérogénéité sémantique des bases de données à intégrer.

2.3.1 Typologie des approches d'intégration d'informations

Il existe plusieurs classifications des systèmes intégrés, fondées sur différents critères. Une classification des approches d'intégration repose sur l'objectif final du processus d'intégration et sur le type des entrées et sorties de ce processus (Hacid et Reynaud, 2003). Parmi les systèmes fortement couplés, les auteurs distinguent trois approches majeures.

L'**intégration de schémas** pour laquelle l'entrée est constituée d'un ensemble de schémas sources, et la sortie est un schéma correspondant à la représentation intentionnelle réconciliée des schémas en entrée. L'entrée comporte aussi la spécification de la représentation des liens entre schéma global et schémas locaux. C'est sur ces correspondances que s'appuiera l'étape suivante qui consiste, à l'aide de requêtes ou d'un programme spécifique, à traduire les données de leur schéma initial vers le schéma intégré. Une variation de ce scénario d'intégration de schémas consiste à intégrer un schéma conceptuel préexistant avec un autre schéma conceptuel cible, développé indépendamment (Rahm et Bernstein, 2001). Le cas des bases de données fédérées fortement couplées, qui contrairement aux bases de données fédérées faiblement couplées, conduit à la création d'un schéma intégré, dit schéma fédéré, peut être considéré comme un scénario d'intégration de schémas. L'architecture de référence en matière de bases de données fédérées est décrite par Sheth et Larson (1990). Parent et Spaccapietra (2000) décrivent le processus d'intégration de schémas dans le cadre de la réalisation d'un système de bases de données fédérées et proposent des solutions pour la résolution des divers types de conflits d'intégration rencontrés.

L'**intégration de données virtuelle** pour laquelle l'entrée est constituée d'un ensemble de sources de données hétérogènes, et la sortie d'une spécification décrivant comment fournir un accès global et unifié aux sources de données afin de bénéficier des informations qui y sont contenues sans toutefois interférer avec l'autonomie de ces sources. Un exemple typique d'intégration de données virtuelle est l'approche médiateur, proposée par Wiederhold (1992), qui consiste à définir une interface destinée à fournir un accès unifié à différentes sources de données qui ne sont pas nécessairement des bases de données, et peuvent par conséquent être peu structurées et

documentées. L'objectif est de donner l'impression d'interroger un système centralisé et unifié, alors que les différentes sources de données sont réparties, autonomes et hétérogènes. Un médiateur comprend une base de connaissances qui modélise le domaine d'application, fournit un vocabulaire commun pour l'expression des requêtes, et fait le lien entre les différentes sources de données accessibles. Un médiateur est donc une couche logicielle permettant à l'utilisateur d'interroger de façon transparente plusieurs sources de données réparties et hétérogènes.

L'**intégration de données matérialisée** pour laquelle l'entrée est également constituée d'un ensemble de sources de données, mais dont la sortie consiste en un ensemble de données représentant une vue réconciliée des sources, à la fois au niveau des schémas et au niveau des données. C'est le cas de l'approche entrepôt de données (Doucet et Gançarski, 2001) ; les données sont organisées, coordonnées, intégrées et stockées pour donner à l'utilisateur une vue globale et unifiée des données. Les entrepôts de données sont conçus dans le but de rassembler les données d'une entreprise au sein d'une base unique afin de faciliter l'analyse et la prise de décisions. Il y a duplication des données et il n'est pas nécessaire d'accéder aux sources pour répondre à une requête. Le schéma global n'est pas figé : il est amené à évoluer, à chaque ajout ou réorganisation de données.

2.3.2 La prise en compte de la sémantique pour l'intégration d'informations

Les différentes approches décrites ci-dessus présentent des points communs. En particulier, on devra, dans tous les cas, procéder à un appariement de schémas (Rahm et Bernstein, 2001). Cette tâche, qui consiste à déterminer quelles sont les classes de chacune des bases à intégrer qui représentent un même type d'entités du monde réel, nécessite, pour dépasser les limitations inhérentes aux approches terminologiques, structurelles, et en extension, une parfaite compréhension du contenu des diverses bases à intégrer (Uschold et Gruninger, 1996).

Pour y parvenir, un consensus s'est formé autour de l'utilisation d'ontologies comme sources de connaissances externes permettant de spécifier sans ambiguïté à quels types d'entités du monde réel se rapporte chaque classe des bases à intégrer (Charlet et al., 2005). Gruber (1995) définit une **ontologie** comme « la spécification explicite et formelle d'une conceptualisation partagée ». Uschold et Gruninger (1996) précisent qu' « une ontologie implique ou comprend une certaine vue du monde vis-à-vis d'un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts (par exemple, des entités, des attributs et des processus), leurs définitions et les relations qui existent entre eux. On appelle cela une conceptualisation. [...] Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification (i.e. des définitions). ».

Différentes typologies des ontologies sont proposées dans la littérature. Celle de Uschold et Gruninger (1996) est fondée sur le niveau d'expressivité des ontologies considérées et va de simples glossaires à des ontologies formelles décrites à l'aide d'axiomes logiques, en passant par des taxonomies et des schémas conceptuels de bases de données. Guarino (1998) propose une classification considérant le caractère plus ou moins générique des concepts décrits par les ontologies. Ainsi les **ontologies de haut niveau** décrivent des concepts très généraux, comme l'espace, le temps, ou la matière, indépendamment de tout domaine d'application. Les **ontologies du**

domaine et les **ontologies de tâche** décrivent respectivement le vocabulaire propre à un domaine générique et à une tâche générique, en spécialisant les concepts d'une ontologie de haut-niveau. Enfin, les **ontologies d'application** décrivent des concepts relevant à la fois d'un domaine particulier et d'une tâche particulière. Ces concepts correspondent donc le plus souvent à des rôles joués par des entités du domaine dans le cadre d'une tâche précise.

Il s'agit donc de se doter d'une modélisation du domaine d'application des bases de données à intégrer qui recense l'ensemble des types d'entités du monde réel en jeu, les définisse de façon non ambiguë et fasse consensus au sein de la communauté visée. Annoter les éléments de schémas des bases de données à intégrer via cette ontologie permet alors de désambiguïser leur signification et de déduire automatiquement les relations de correspondance entre eux. Cette opération d'annotation consiste à définir un lien entre un élément du schéma conceptuel de données ou de métadonnée et la description formelle d'une catégorie d'entités géographiques du monde réel ou de l'une de ses propriétés, lorsque les instances de cet élément représentent des entités géographiques appartenant à cette catégorie ou des valeurs de cette propriété, et permet donc de définir chaque élément de schéma ou de métadonnée à l'aide d'un vocabulaire formel commun.

2.4 Objectif de cette thèse

De nombreuses applications nécessitent d'intégrer des bases de données géographiques : constitution d'un référentiel cohérent à partir de bases de données hétérogènes, transformation de schéma, recalage de données, enrichissement d'une base de données à partir de données issues de sources externes, géocodage, mutualisation de mises à jour, contrôle qualité, suivi de versions, requêtes sur des bases de données hétérogènes et distribuées, etc. Ces différentes applications requièrent toutes, dans un premier temps, de déterminer les correspondances entre les éléments de schémas des différentes bases de données à intégrer. L'automatisation de ce processus d'appariement de schémas suppose celle de la détection des divers types d'hétérogénéité pouvant intervenir entre les bases de données à intégrer (Fichtinger et al., 2009), qu'il s'agisse d'hétérogénéité sémantique ou bien de types d'hétérogénéité liés à la structuration et à la représentation géométrique des données dont on retrouve la justification dans les spécifications des bases de données. La détection de ces divers types d'hétérogénéités permettra d'établir des correspondances précises entre éléments de schémas de bases de données qui pourront être exploitées différemment par la suite en fonction de l'application visée. En effet, si une application de découverte de bases de données hétérogènes nécessite essentiellement de résoudre des questions d'hétérogénéité sémantique, une application de recalage exigera de déterminer quelle base de données possède le niveau de détail géométrique le plus fin et doit donc être utilisée comme référentiel. En outre, le processus d'appariement des données pourra nécessiter de disposer de connaissances sur d'éventuels conflits de description géométrique des données dans les bases à intégrer.

Les approches proposées pour l'intégration d'informations dans le domaine informatique sont peu à peu reprises dans le domaine de l'information géographique pour résoudre les difficultés liées à l'hétérogénéité sémantique entre bases de données. Celles-ci s'appuient sur des ontologies utilisées comme sources de connaissances externes permettant d'expliciter la sémantique exacte des éléments des schémas des bases de données à apparier. Un préalable indispensable à la mise en

œuvre de ces approches consiste donc à se doter d'une ontologie couvrant le domaine sur lequel portent les bases de données à intégrer.

Par ailleurs, l'appariement des schémas de bases de données vectorielles requiert de déterminer, outre les relations de correspondance entre éléments de schémas, les restrictions dues à l'hétérogénéité des spécifications des différentes bases qui s'appliquent à ces correspondances. Considérons le conflit de critère de sélection entre la classe *Massif boisé* de la BDCARTO© 3.1 et la classe *Zone de végétation* de la BDTOPO© 2.0 présenté au paragraphe 2.2.3. La relation de correspondance établie entre ces deux classes devra donc comporter une restriction sur la superficie des instances de ces classes et préciser, dans le cas présent, que seules les instances de la classe *Zone de végétation* de plus de 500 hectares correspondent potentiellement aux instances de la classe *Massif boisé*. Détecter automatiquement ce type de restriction suppose d'inclure dans le processus d'appariement de schémas des connaissances issues des spécifications de chacune des bases à intégrer. Il est donc nécessaire de formaliser ces connaissances afin de pouvoir les traiter automatiquement. Or cette étape de formalisation s'avère complexe dans la mesure où il s'agit de représenter des connaissances très spécifiques faisant intervenir à la fois des notions de modélisation de bases de données géographiques et des notions de représentation géométrique.

L'objectif de cette thèse réside donc dans la formalisation et l'acquisition des connaissances nécessaires pour la mise en œuvre d'un processus d'intégration virtuelle de bases de données géographiques, en accord avec les normes et standards actuels du domaine. Il s'agit des connaissances nécessaires à la description des domaines de la topographie et de la saisie de données géographiques. Ainsi, une première étape consistera à se doter d'une ontologie du domaine de la topographie, et à l'exploiter pour annoter et apparier des schémas de bases de données géographiques hétérogènes. Une seconde étape sera d'étendre l'approche adoptée initialement pour intégrer à l'ensemble du processus d'appariement de schémas des connaissances issues des spécifications des bases de données et exploiter ces connaissances dans le cadre de deux applications classiques dans le domaine des infrastructures de données géographiques : l'intégration de données "métier" sur un référentiel et la découverte de bases de données géographiques. Ce travail de thèse s'inscrit dans le cadre du projet ANR GéOnto (ANR-O7-MDCO-005) pour la création, la comparaison et l'exploitation d'ontologies géographiques.

3 État de l'art

Cette partie dresse un état de l'art des travaux récents en matière d'intégration d'informations dans le domaine des infrastructures de données géographiques. Les problématiques de découverte et d'accès aux données ainsi que de transformation de schémas de bases de données géographiques sont au cœur de la mise en œuvre de la directive INSPIRE 2007/2/CE, et constituent l'essentiel des efforts de recherche actuels dans le domaine de l'intégration de bases de données géographiques. C'est pourquoi notre état de l'art s'attache, dans un premier temps, à décrire les travaux réalisés dans ces deux domaines. La troisième partie est consacrée à la principale ressource sur laquelle reposent les travaux précédents : les ontologies du domaine de la topographie. Nous y dressons un bilan des ontologies existantes, et décrivons brièvement les principales méthodologies proposées pour la création d'ontologies du domaine. Enfin, la quatrième partie est consacrée à la prise en compte de connaissances issues des spécifications de bases de données géographiques dans le processus d'intégration de ces bases.

3.1 L'intégration dans les infrastructures de données géographiques

La directive INSPIRE 2007/2/CE constitue un exemple typique des usages actuels en matière d'intégration d'informations pour les infrastructures de données géographiques. Elle fournit un cadre légal sous la forme d'un ensemble de spécifications[9] définissant la forme sous laquelle les États membres de l'Union Européenne devront fournir leurs données géographiques entrant dans le cadre de la directive. Afin de garantir la cohérence des spécifications de chaque thème identifié, une équipe d'experts, dite « Data Specification Drafting Team », a élaboré un ensemble de quatre documents techniques visant à encadrer leur définition. En particulier, ces documents fournissent un modèle conceptuel de données reposant sur la norme ISO 19107, une méthodologie pour le développement de spécifications et des recommandations pour l'encodage des données, encourageant l'usage du format GML. Les spécifications sont élaborées thème par thème, par des groupes d'experts dédiés et sont rédigées selon la norme ISO 19131 « Data product specifications ». La découverte et l'accès aux données sont assurés à l'aide d'outils de catalogage. Ceux-ci s'appuient sur des métadonnées comme source d'information sur les données, permettent aux utilisateurs d'effectuer des requêtes par mots-clés pour retrouver les jeux de données correspondant à leurs besoins, et fournissent éventuellement des services Web reposant sur les standards de l'OGC pour faciliter l'accès à ces jeux de données. Cependant, les outils de recherche par mots-clés implémentés dans les applications de catalogage standardisées actuellement préconisées souffrent de faiblesses liées aux ambiguïtés du langage naturel. Pour pallier ces manques, des approches d'intégration de données virtuelles (médiateur) sont également proposées dans la littérature dédiée aux infrastructures de données géographiques et sont présentées en 3.1.1. Par ailleurs, les producteurs de données géographiques des États membres de l'Union Européenne chargés d'alimenter l'infrastructure de données géographiques européenne se voient dans l'obligation de traduire leurs données pour les rendre conformes aux spécifications INSPIRE 2007/2/CE. Cette traduction peut être réalisée de façon plus ou moins automatique et dynamique. Cette intégration s'apparente à une approche de bases de données fédérées dans la mesure où le schéma global a été défini par analyse des schémas locaux et où l'intégration demeure virtuelle. En effet, les différentes bases conservent leur autonomie de départ et sont hébergées et maintenues par leur producteur d'origine. Les approches proposées pour réaliser ces opérations de transformation de schémas sont présentées en 3.2.2.

3.1.1 Découverte et accès aux données

La découverte et l'accès à des données géographiques réparties et hétérogènes sont généralement assurés, au sein d'une infrastructure de données géographiques, à l'aide de catalogues de données. C'est le cas pour l'infrastructure européenne définie par la directive INSPIRE 2007/2/CE, qui dispose d'un catalogue de données en ligne[10]. Un catalogue vise à indexer et décrire les jeux de données disponibles afin de fournir à leurs utilisateurs potentiels un accès efficace et aisé à diverses informations les concernant. Seront fournies, en particulier, des informations sur la thématique

[9] http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2

[10] http://www.inspire-geoportal.eu/index.cfm

couverte par chaque jeu de données, la zone couverte, le système de coordonnées utilisé, la généalogie des données, ou encore leur qualité. Des indications permettant l'accès aux données pourront également être fournies par le catalogue lorsque cela est possible, au travers d'un service Web OGC par exemple. Enfin, un catalogue pourra être doté d'un outil de visualisation afin d'afficher un échantillon pour chaque jeu de données référencé. Toutes ces informations sont structurées et stockées sous la forme de métadonnées. La plupart des outils de catalogage de données géographiques s'appuient, comme le préconisent les spécifications de la directive INSPIRE 2007/2/CE, sur les normes ISO 19115 et ISO 19139 pour la structuration et l'encodage des métadonnées. Celles-ci sont interrogées par mots-clés à l'aide d'un moteur de recherche fourni par le catalogue. Les réponses aux requêtes sont ensuite renvoyées à l'utilisateur via l'interface du catalogue, afin de l'aider à découvrir les données disponibles et à évaluer leur pertinence vis-à-vis de ses besoins (cf. figure 3).

Figure 3: Résultats de la recherche "sentiers de randonnée" dans le Géocatalogue

En dépit de ces efforts de normalisation, les catalogues de données présentent des lacunes concernant la gestion de la sémantique des données. En effet, la découverte des données est réalisée à l'aide de requêtes par mots-clés sur les métadonnées. La chaîne de caractères entrée par l'utilisateur est comparée à un ensemble de mots-clés fournis par les métadonnées. Au cours de ce processus, de simples fautes de frappe, ou des cas de synonymie ou d'homonymie peuvent conduire à des réponses inexactes ou incomplètes, voire à l'absence de réponse.

Les travaux récents dans le domaine de la découverte et de l'accès aux données géographiques se sont donc particulièrement concentrés sur la résolution de l'hétérogénéité sémantique dans les catalogues de données géographiques. Ces travaux s'appuient généralement sur des approches d'intégration de données virtuelle proposées dans le domaine de l'informatique. Les différentes sources de données référencées par un catalogue sont annotées à l'aide d'une (ou plusieurs) ontologie(s) du domaine qui sert(servent) de point d'entrée aux requêtes des utilisateurs. En établissant une relation de correspondance explicite entre les éléments de l'ontologie du domaine et ceux des diverses ressources descriptives des sources de données, qu'il s'agisse de leurs schémas conceptuels, d'ontologies d'application ou de métadonnées, le processus d'annotation permet la détection des sources de données pertinentes vis-à-vis des requêtes des utilisateurs, ainsi que la réécriture de ces requêtes et de leurs réponses par le système. Il s'agit donc d'une étape critique du processus d'intégration qui suppose au préalable de déterminer les relations de correspondance entre ressources descriptives des données et ontologie du domaine, ce qui revient à apparier des éléments de ces ressources avec l'ontologie. Selon les cas, cette tâche d'appariement pourra être réalisée de façon plus ou moins automatique et plus ou moins détaillée. C'est à cette étape de mise en correspondance que nous allons nous intéresser en particulier. Nous présentons ci-dessous les divers travaux relevant de cette approche d'intégration de données virtuelle, en nous attachant tout particulièrement à comparer les architectures proposées, les ressources mises en jeu et les approches préconisées pour la mise en correspondance de ces ressources.

Découverte et accès aux données : approche proposée par (Paul et Ghosh, 2006)

L'application proposée par Paul et Ghosh (2006) vise à permettre la découverte et l'accès à des sources de données géographiques hétérogènes et réparties. Elle repose sur une architecture orientée services qui utilise une ontologie du domaine pour interpréter et rediriger les requêtes des utilisateurs vers les services concernés. L'accès aux données est réalisé à l'aide de services Web OGC de type WFS (voir tableau 1 p. 18). La sémantique de chaque source de données est décrite, non seulement à l'aide de métadonnées comme c'est traditionnellement le cas dans les catalogues de données, mais également via une ontologie d'application construite à partir de l'ontologie du domaine. Chaque service est référencé par le système via un fichier de métadonnées appelé « Comprehensive Source Description ». Celui-ci est décrit à l'aide d'un langage dérivé d'XML, et permet le stockage de métadonnées sur les données disponibles, ainsi que la description de l'ontologie d'application de chaque source de données. Lors de la phase de découverte des données, l'utilisateur fournit une requête dans les termes de l'ontologie du domaine et ceux-ci sont alignés avec ceux des ontologies d'application à l'aide du *système de raisonnement* RACER qui recherche des relations d'équivalence ou de subsomption entre concepts de la requête et concepts des ontologies d'application. La figure 4 dresse la synthèse des différentes ressources exploitées par le système, et identifie les étapes de mise en correspondance automatique ou manuelle de ces ressources.

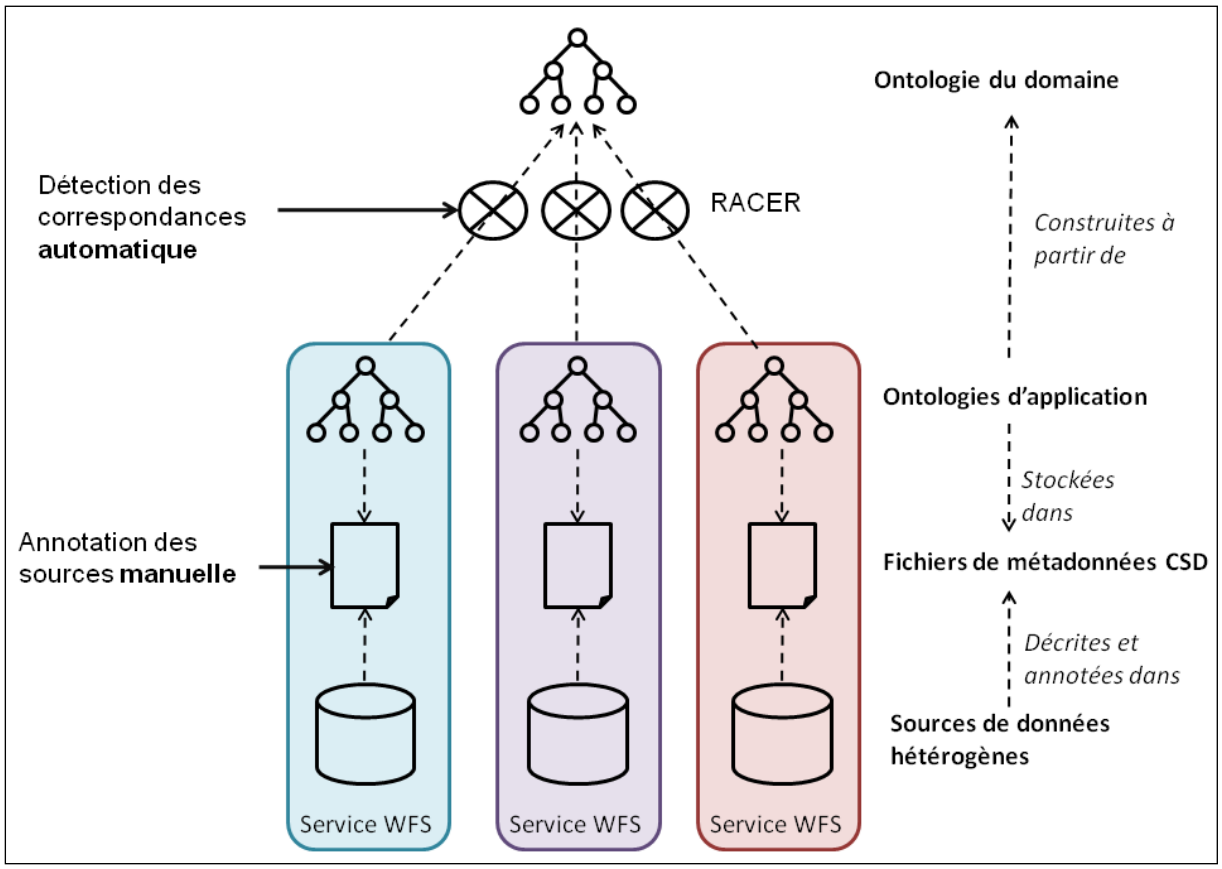


Figure 4: Approche proposée par (Paul et Ghosh, 2006) pour la découverte et l'accès aux données: identification des étapes d'annotation automatiques et manuelles

Une fois les sources de données pertinentes découvertes, les services Web permettant d'y accéder peuvent être à leur tour découverts à l'aide de requêtes getCapabilities. A cette requête, chaque service va renvoyer à l'utilisateur la description des données disponibles. Les requêtes d'accès aux données sont interprétées par un module d'exécution de requêtes. Celui-ci fait office de médiateur: il utilise l'ontologie globale, les ontologies d'application et leurs relations de correspondance pour décomposer les requêtes en sous-requêtes lorsque cela est nécessaire afin d'aller interroger les différents services pertinents.

Découverte et accès aux données : approche proposée par (Nambiar et al., 2006)

Le réseau GéoScience (GeoScience Network, GEON) propose un portail[11] dédié à la gestion de données géographiques hétérogènes distribuées. L'objectif de ce portail est de faciliter le partage de données géographiques entre communautés de recherche, en fournissant des outils pour la découverte et l'accès aux sources de données. L'application développée (Nambiar et al., 2006) offre différentes possibilités aux fournisseurs de données. Selon le cas, ils peuvent soit héberger eux-mêmes les données mises à disposition, soit les dupliquer au sein d'un entrepôt centralisé fourni par le portail. Puis ils doivent référencer leurs jeux de données auprès du système. Ce processus comporte deux aspects. D'une part, afin de permettre au système de résoudre d'éventuels

[11] http://www.geongrid.org/index.php

problèmes d'hétérogénéité syntaxique des données, les fournisseurs doivent fournir des métadonnées décrivant les formats de données utilisés. D'autre part, afin de résoudre les problèmes d'hétérogénéité sémantique des données, celles-ci doivent être annotées à l'aide d'une (ou plusieurs) ontologie(s) du domaine. Cette annotation, peut être plus ou moins détaillée selon les cas. Les fournisseurs peuvent en effet annoter globalement un jeu de données soit à l'aide de métadonnées, soit à l'aide d'un ou plusieurs concepts issus des ontologies disponibles. Afin de permettre un plus haut niveau d'intégration, ils peuvent également annoter chaque élément du schéma d'un jeu de données à l'aide de concepts d'ontologies. Pour faciliter ce processus d'annotation pour les fournisseurs de données, une interface conviviale a été développée. Les annotations, ou mappings, entre éléments de schémas et concepts d'ontologies sont décrits et stockés dans un langage dédié, doté d'une syntaxe proche de celle du langage pour ontologies recommandé par le W3C, OWL - Ontology Web Language (W3C, 2004), nommé Ontological Database Annotation Language (ODAL). Ces mappings peuvent être de différents types; les éléments de schémas peuvent être associés aux concepts des ontologies via différentes relations, comme « possède des instances du type », « mentionne », ou « utilise ». Afin de permettre l'ajout de données géolocalisées issues de diverses communautés, les utilisateurs ont la possibilité d'ajouter leurs propres ontologies au système, et de les rendre disponibles pour l'annotation d'autres sources de données. La figure 5 présente les différentes ressources exploitées par le système, ainsi que les étapes de mise en correspondance automatique ou manuelle de ces ressources.

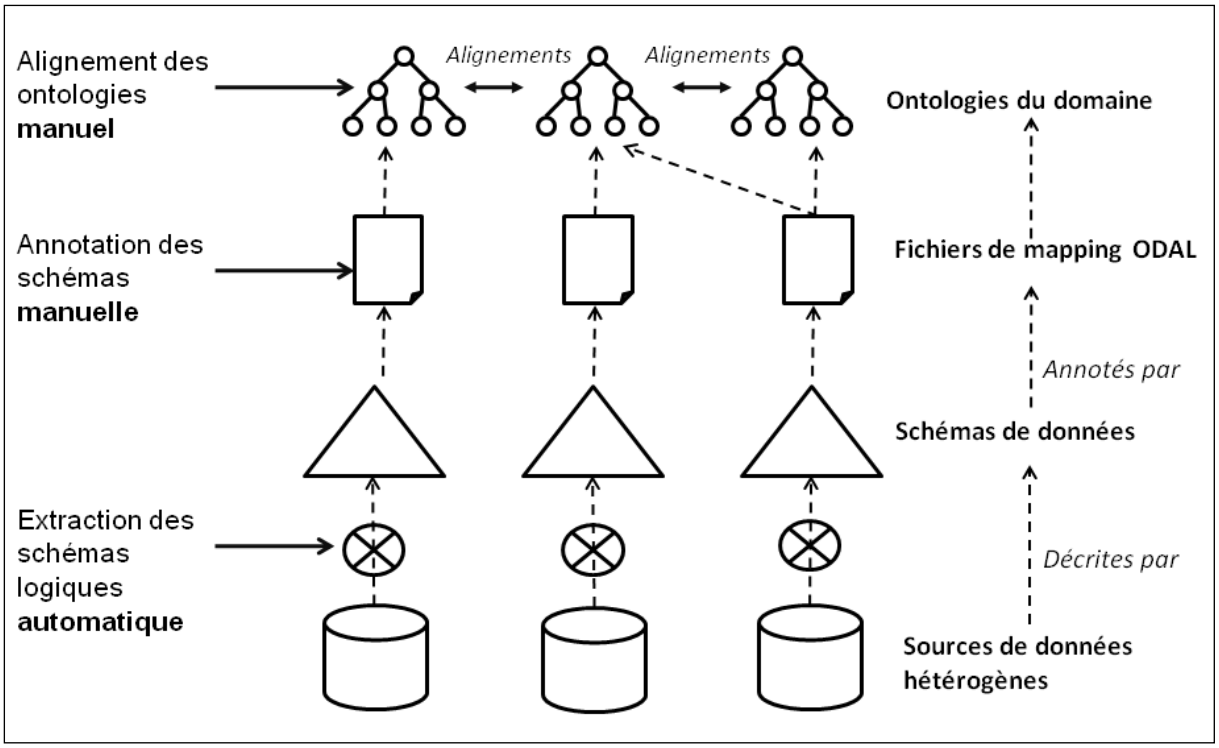


Figure 5: Approche proposée par (Nambiar et al., 2006) pour la découverte et l'accès aux données: identification des étapes d'annotation automatiques et manuelles

Dans le portail GEON, la découverte des données est assurée à l'aide d'un moteur de recherche. L'utilisateur a le choix entre deux interfaces de recherche de données. La première, la plus simple, permet d'effectuer des requêtes par mots-clés sur les métadonnées associées aux sources de données référencées par le système. La seconde intègre les annotations sémantiques associées aux

sources de données. L'utilisateur peut donc découvrir les sources de données disponibles en sélectionnant un concept parmi les ontologies du domaine proposées et en spécifiant le type de mapping recherché entre les sources de données et les ontologies. Les requêtes de ce type peuvent également être étendues par le système en recherchant les sources de données annotées via des concepts sémantiquement proches de celui recherché par l'utilisateur, afin de lui proposer d'autres sources de données potentiellement proches de ses centres d'intérêt. Des métadonnées de découverte sont affichées afin de permettre à l'utilisateur d'évaluer la pertinence des sources de données détectées vis-à-vis de ses besoins, et d'accéder aux sources de données s'il le souhaite. Enfin, une interface de requêtes permet aux utilisateurs d'interroger les différentes sources de données. Ceci peut être réalisé de deux façons. D'une part, l'utilisateur peut sélectionner les sources de données qu'il souhaite interroger, visualiser leurs schémas respectifs et écrire lui-même des requêtes SQL adaptées aux schémas des différentes sources de données. D'autre part, il peut interroger les sources de données au travers des ontologies du domaine fournies par le système. Dans ce second cas, le système de médiation implémenté va, grâce aux annotations sémantiques dont il dispose, déterminer les sources de données pertinentes et écrire des requêtes adaptées pour chacune de ces sources.

Découverte et accès aux données : approche proposée par (Desconnets et al., 2007)

MDweb (Desconnets et al., 2007) est un outil de catalogage de ressources générique offrant des possibilités de création de catalogues dédiés aux données géographiques. A ce titre, il intègre l'ensemble des normes relatives aux catalogages de données propres aux infrastructures de données géographiques. En particulier, les métadonnées sont structurées conformément à la norme ISO 19115 et les catalogues créés implémentent la spécification OGC « Catalogue Services for the Web » (CSW) (OGC, 2007) permettant à tout client compatible de les interroger. L'indexation des ressources référencées par le système est réalisée à l'aide de deux bases de référence. D'une part, une base de données géographique permet la description de l'emprise géographique des ressources. D'autre part, une base de référence thématique assure le stockage de divers thesauri ou ontologies aux formats SKOS (Simple Knowledge Organization System[12]) ou RDF (Resource Description Framework[13]) destinés à l'annotation sémantique des ressources, comme Agrovoc[14] ou Gemet[15]. Les mots-clés utilisés dans les métadonnées correspondent à des termes issus de la base de référence thématique. Ces deux bases de référence sont exploitées, à la fois pour faciliter la création des métadonnées et pour la recherche de ressources pertinentes pour un utilisateur. La figure 6 passe en revue les relations de correspondance entre les ressources mises en œuvre par le système.

[12] http://www.w3.org/2004/02/skos/
[13] http://www.w3.org/RDF/
[14] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[15] http://www.eionet.europa.eu/gemet

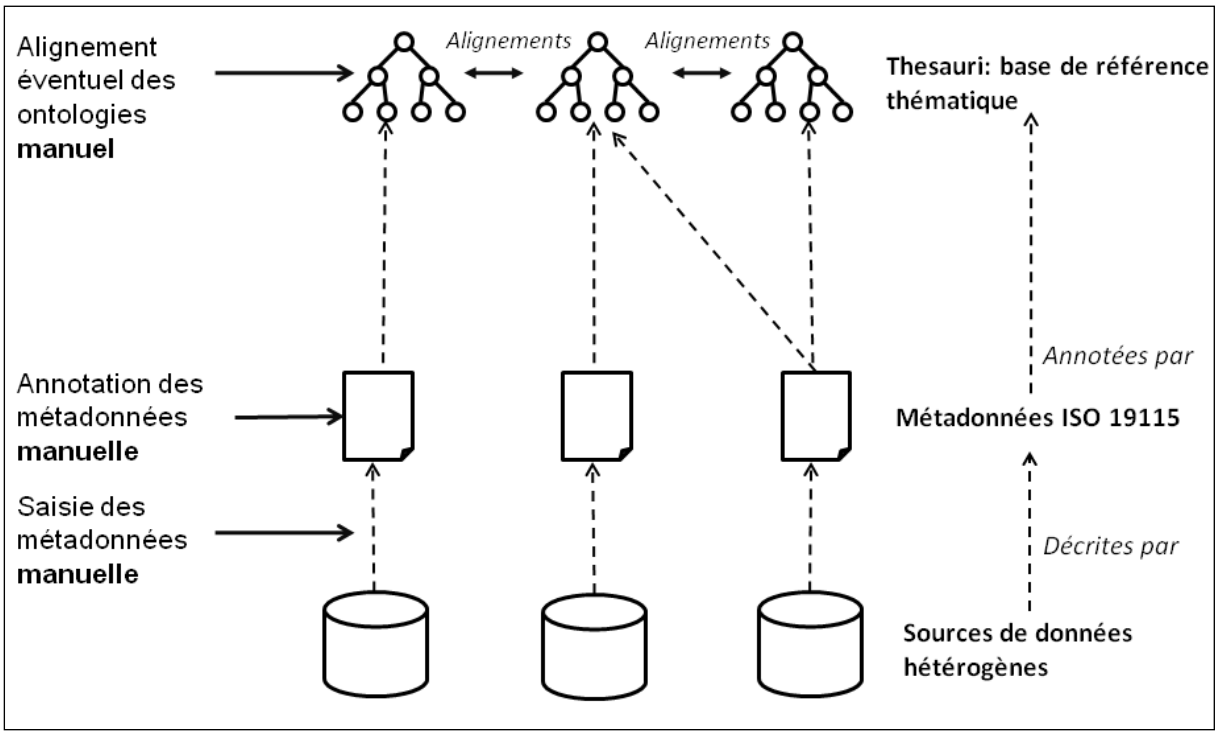


Figure 6: Approche proposée par (Desconnets et al., 2007) pour la découverte des données: identification des étapes d'annotation automatiques et manuelles

Découverte et accès aux données : approche proposée par (Lassoued et al., 2008)

Lassoued et al. (2008) présentent une approche de médiation fondée sur des ontologies pour la découverte de données géographiques répertoriées par différents catalogues conformes au standard OGC « Catalogue Services for the Web » (CSW) (OGC, 2007). Des ontologies d'application au format OWL sont utilisées pour définir de façon non ambiguë les mots-clés employés dans les métadonnées fournies par les différents producteurs de données pour alimenter leurs catalogues de données. Celles-ci sont alignées manuellement avec une ontologie globale, également décrite en OWL, qui couvre l'ensemble des concepts partagés du domaine d'intérêt et les relations de correspondance ou « mappings » sont stockés au sein d'une ontologie dédiée, dite « ontologie de mappings », au format OWL. La figure 7 présente deux extraits des ontologies de mapping utilisées. L'ontologie globale y est préfixée par le terme « global » et les mappings entre concepts des ontologies locales et concepts de l'ontologie globale sont des relations de subsomption présentées en gras. Les requêtes CSW *getRecords* permettant de découvrir les données disponibles sont écrites dans les termes de l'ontologie globale. Un médiateur exploite les mappings entre ontologies d'application et ontologie du domaine pour identifier les ontologies d'application concernées par une requête donnée, la réécrire dans les termes des ontologies d'application pertinentes, envoyer ces différentes requêtes, collecter les réponses et les traduire dans les termes de l'ontologie globale afin de fournir une réponse unifiée à l'utilisateur. La figure 8 présente l'architecture globale du système et identifie les étapes de mise en correspondance des diverses ressources mises en jeu.

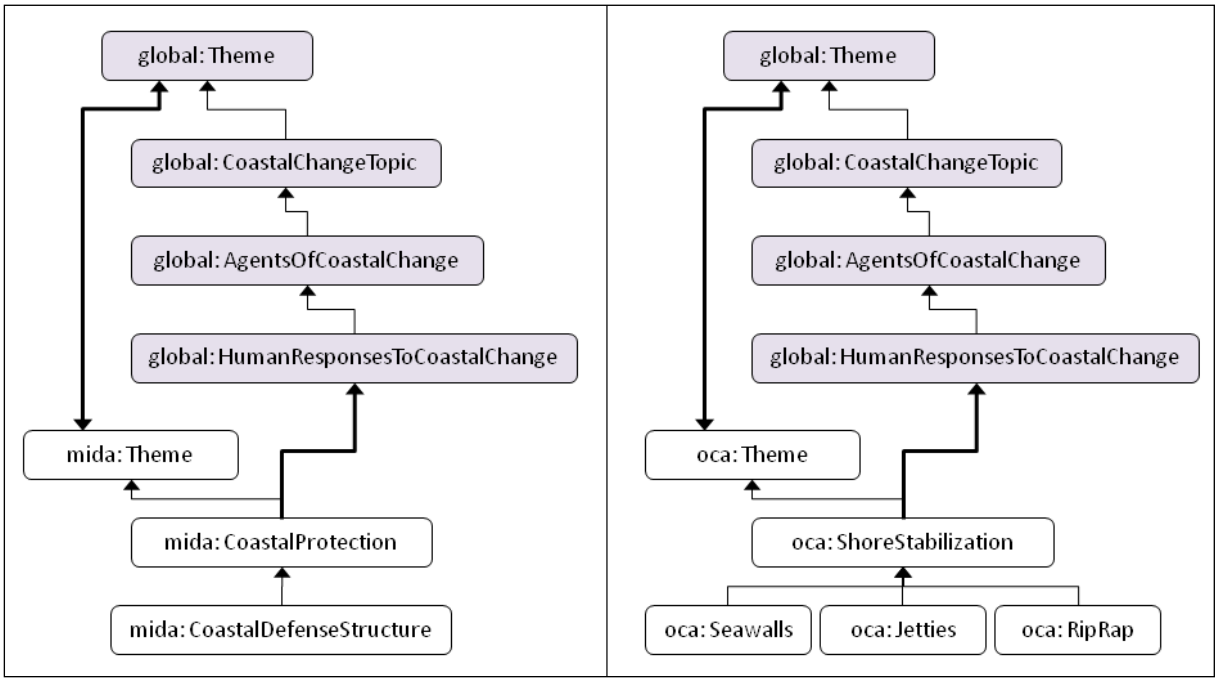


Figure 7: Extraits de deux ontologies de mappings d'après (Lassoued et al., 2008). A gauche celle reliant l'ontologie Marine Irish Digital Atlas à l'ontologie globale, à droite celle reliant l'ontologie Oregon Coastal Atlas à l'ontologie globale.

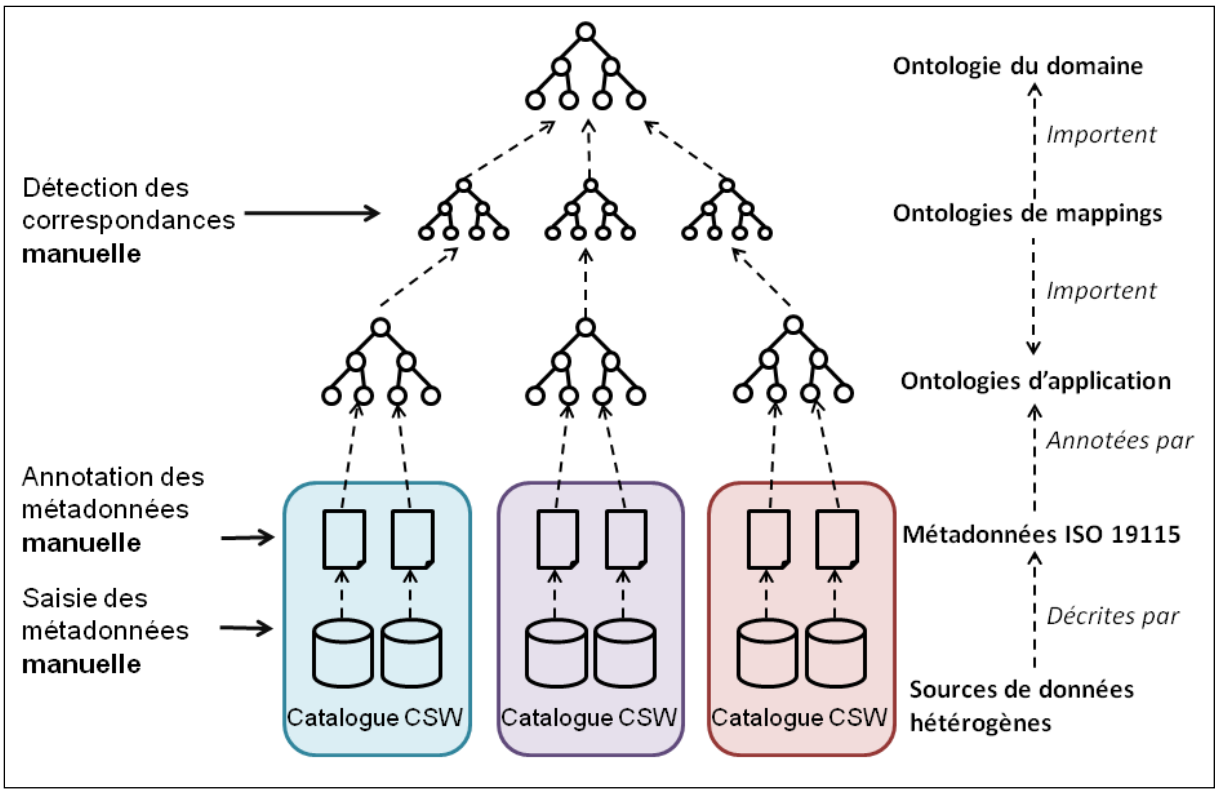


Figure 8: Approche proposée par (Lassoued et al., 2008) pour la découverte des données: identification des étapes automatiques et manuelles

Découverte et accès aux données : approche proposée par (Zhao et al., 2008)

Zhao et al. (2008) présentent également une approche pour la découverte et l'accès à des sources de données géographiques hétérogènes et distribuées. Il s'agit de fournir une interface reposant sur une ontologie pour interroger différentes sources de données de façon uniforme. Une première étape consiste donc à créer une ontologie du domaine qui servira de point d'entrée au système. Celle-ci est décrite dans le langage RDF Schema[16], et se restreint aux concepts spatiaux et à leurs propriétés les plus génériques: Feature, SpatialFeature, hasID et geometry. Elle est étendue par une ontologie d'application sur le thème des réseaux de transport routier. Notons que dans la typologie des ontologies proposées par (Guarino, 1998), les ontologies présentées ici correspondent plutôt à deux ontologies du domaine, l'une étant plus spécifique que l'autre et s'appuyant sur cette dernière. La figure 9 présente ces ontologies.

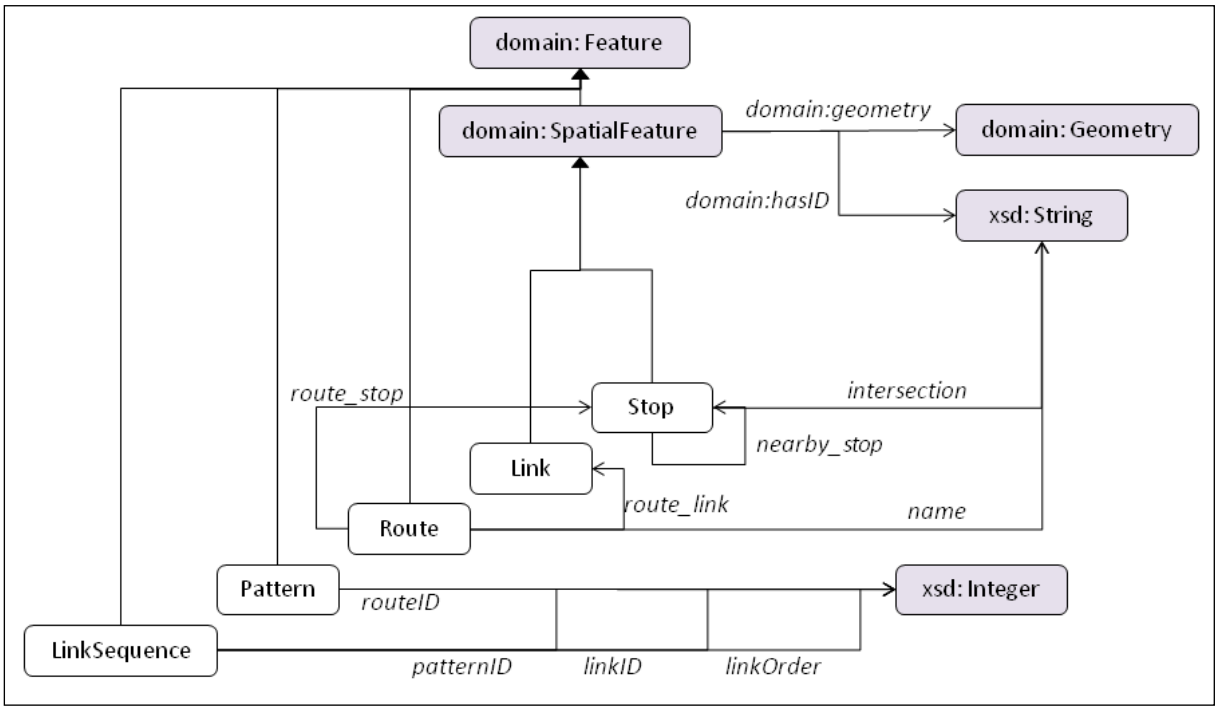


Figure 9: Ontologies utilisées par (Zhao et al, 2008)

Puis des règles de correspondance entre schémas des bases de données et ontologie sont établies. Les auteurs supposent ici que les données sont accessibles soit directement via un système de gestion de bases de données, soit par un service Web de type WFS. Ainsi, les correspondances entre *WFS features* ou tables de base de données et ontologie sont définies par des vues RDF, décrites dans un langage proche de Datalog, langage de requêtes et de règles pour les bases de données déductives, permettant la réécriture de requêtes RDF en requêtes getFeature ou SQL selon le cas. Les vues RDF sont définies en deux étapes. La première consiste à créer des règles de correspondance entre chaque *WFS feature* ou une table relationnelle et l'ontologie RDF. Ces correspondances s'expriment au niveau des attributs des tables et des propriétés de l'ontologie. La seconde réside dans la définition de règles d'inférences permettant de dériver des relations entre objets des bases de données à partir de leurs valeurs d'attributs. Celles-ci s'expriment dans les termes de l'ontologie

[16] http://www.w3.org/TR/rdf-schema/

RDF, ce qui revient à définir des relations entre concepts à partir de propriétés de ces concepts. Les requêtes d'accès aux données sont exprimées en SPARQL[17], le langage de requêtes dédié aux données RDF. Les vues RDF définies précédemment offrent la possibilité de traduire les données de leur format source vers RDF, afin de pouvoir exécuter les requêtes SPARQL directement sur les données RDF. Ici, pour des raisons de performance, et afin de pouvoir bénéficier des opérateurs spatiaux disponibles au niveau des données sources, les auteurs privilégient la solution consistant à conserver les données dans leur formats d'origine et à réécrire les requêtes SPARQL en requêtes *getFeature* WFS et SQL. Pour ce faire, les règles d'inférences sont appliquées en premier sur la requête SPARQL afin de remplacer les clauses comportant des relations entre classes de l'ontologie par leur équivalent en termes de propriétés de ces classes. Puis les règles de correspondance entre schémas de données sources et ontologies sont exploitées de sorte à remplacer les termes de l'ontologie par ceux des schémas appropriés au sein de la requête qui peut alors être envoyée directement aux sources de données pertinentes. La figure 10 présente l'architecture globale de l'application et identifie les étapes de mise en correspondance des diverses ressources mises en jeu.

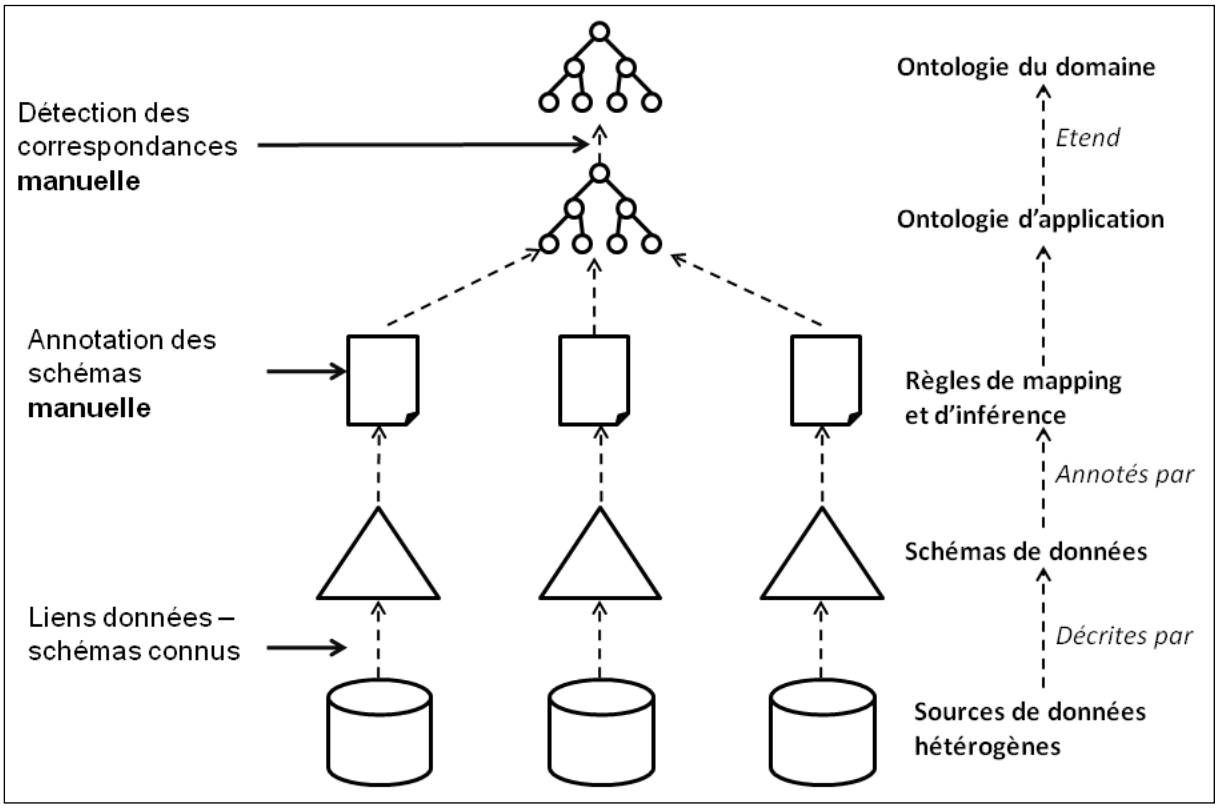


Figure 10: Approche proposée par (Zhao et al., 2008) pour la découverte des données: identification des étapes automatiques et manuelles

[17] http://www.w3.org/TR/rdf-sparql-query/

Découverte et accès aux données : approche proposée par (Lutz et al., 2006)

Lutz et al. (2006) proposent une approche fondée sur des ontologies pour la découverte, l'accès, et l'intégration de données géographiques. Cette approche est relativement proche de celle proposée par (Paul et Gosh, 2006) pour les phases de découverte et d'accès aux données. Les différentes sources de données sont annotées par des ontologies d'application, elles-mêmes construites à l'aide de concepts issus d'ontologies du domaine. Ainsi, les termes utilisés au sein des ontologies du domaine et des ontologies d'application demeurent les mêmes et sont donc facilement comparables. Ces ontologies sont décrites à l'aide d'une logique de description. Les requêtes des utilisateurs s'appuient sur les concepts des ontologies du domaine. Ainsi, le système exploite les capacités d'inférences offertes par les logiques de description pour détecter des relations d'équivalence ou de subsomption entre concepts exprimés dans la requête et concepts décrivant les sources de données. Lorsqu'une relation d'équivalence ou de subsomption est identifiée par le *système de raisonnement*, le concept issu de l'ontologie d'application est considéré comme une réponse pertinente à la requête. Le lien entre chaque source de données et son ontologie d'application est assuré par un fichier de mappings fourni par le fournisseur de données. Lors de l'envoi d'une requête, l'utilisateur reçoit en réponse la liste des éléments de schémas de données annotés par le concept recherché. La figure 11 présente l'ensemble des ressources mises en jeu par le système, ainsi que les différentes étapes d'établissement des correspondances entre ces ressources.

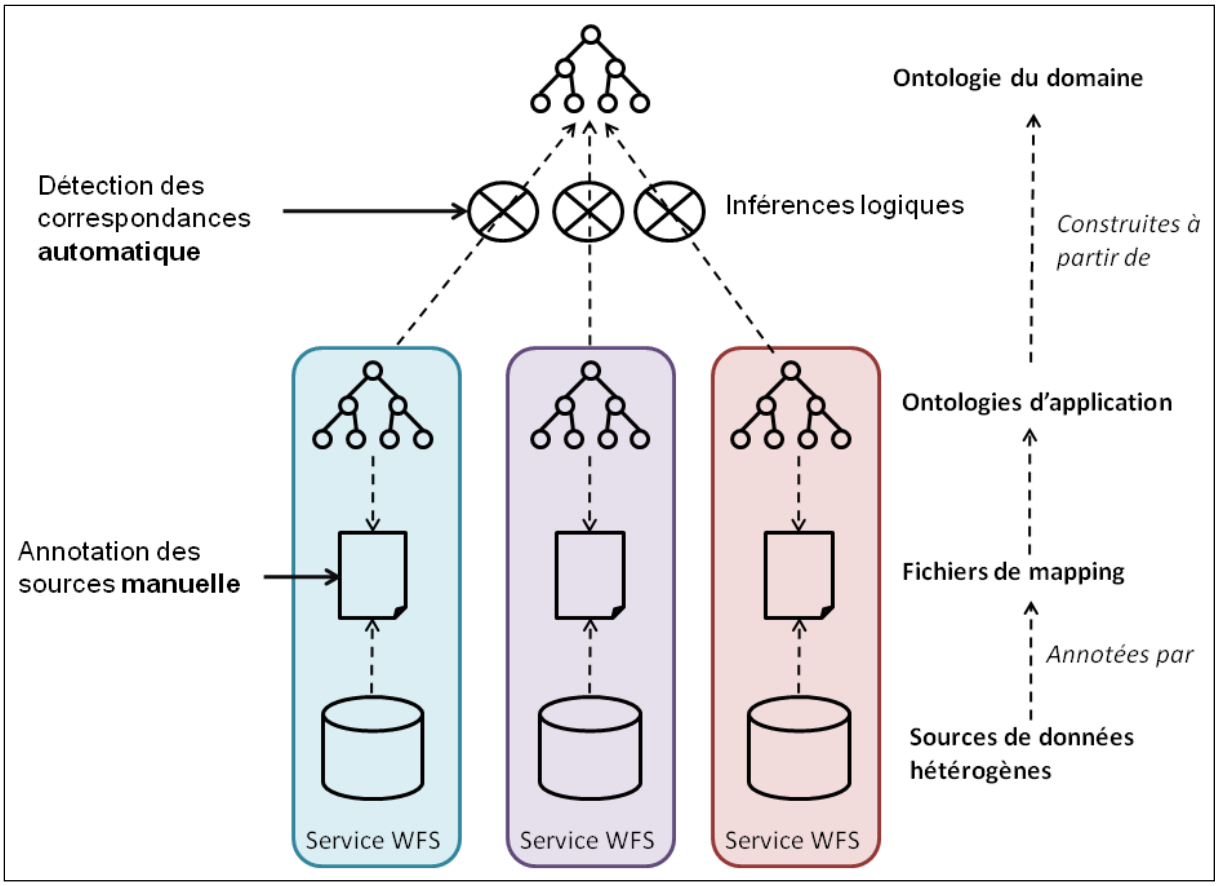


Figure 11: Approche proposée par (Lutz et al., 2006) pour la découverte des données: identification des étapes automatiques et manuelles

L'approche proposée s'intègre dans le cadre des architectures traditionnelles des infrastructures de données géographiques. Les sources de données sont donc accessibles par services Web OGC de type WFS. Pour accéder aux données, l'utilisateur écrit sa requête en SQL. Celle-ci est traduite par le système en concept « requête » dans le langage de logique de description adopté, à l'aide des relations inférées entre concepts de la requête issus de l'ontologie du domaine et ceux des ontologies d'application. Ce concept « requête » est ensuite exploité dans un processus de raisonnement pour identifier les sources de données pertinentes. Enfin une requête de type *getFeature* est générée par le système à l'aide des concepts identifiés comme pertinents parmi les ontologies d'application et des mappings entre sources de données et ontologies d'application disponibles. Celle-ci est envoyée aux sources de données identifiées comme pertinentes afin de renvoyer les données correspondantes à l'utilisateur.

Les auteurs envisagent en outre le cas de la composition de services Web qui nécessite de transformer les données obtenues par ce biais dans le schéma d'un autre service Web destiné à effectuer de nouvelles opérations sur celles-ci. La difficulté majeure d'une telle opération consiste à spécifier les transformations nécessaires. Pour ce faire, les schémas de données en entrée des services doivent également être décrits à l'aide d'ontologies d'application selon la même méthode que pour les sources de données. Des relations d'équivalence et de subsomption entre l'ontologie d'application du service invoqué et celles des sources de données sont inférées par le système afin de détecter les données pertinentes pour le service que l'on souhaite exécuter. Une fois les données pertinentes identifiées, des règles de transformation sont générées. Celles-ci peuvent être de deux types. D'une part, les règles de décomposition de requête visent à convertir les opérations de sélection prévues par le service sur son propre schéma de données en requêtes destinées aux schémas de données des sources identifiées comme pertinentes et fournissant des résultats équivalents. Elles sont créées à l'aide des relations d'équivalence détectées entre ontologies d'application des sources de données et du service ainsi que des mappings entre schémas de données des sources et du service invoqué et leurs ontologies d'application respectives. Les règles de transformation de contexte spécifient les opérations à effectuer pour adapter les données de leur propre contexte vers celui du service invoqué. Celles-ci sont générées à partir des relations détectées entre ontologies d'application des sources de données et du service invoqué. Dans le cadre de ce travail ces règles se cantonnent à spécifier des conversions d'unités de mesure de longueur. Ces différentes règles de transformation sont générées automatiquement et exprimées en logique de Horn.

Découverte et accès aux données : approche proposée par (Bügel et al., 2007)

Le projet ORCHESTRA (Open Architecture and Spatial Data Infrastructure for Risk Management) est un projet intégré du sixième programme cadre de la Commission Européenne. Il vise à favoriser l'interopérabilité entre les différents acteurs européens impliqués dans la gestion des risques environnementaux. A ce titre, l'une des tâches principales du projet consiste à définir et à implémenter les spécifications d'une infrastructure de données géographiques orientée services, en s'appuyant sur les technologies, normes et standards actuels, et en prenant en compte les besoins des divers acteurs du domaine. Un modèle de référence pour l'architecture globale de l'infrastructure (Usländer, 2007) a donc été élaboré et publié dans le cadre du groupe de travail

« Gestion des risques et des crises » de l'OGC. Celui-ci garantit l'interopérabilité syntaxique des données et des services mis en œuvre dans le cadre de l'infrastructure. L'interopérabilité sémantique des données et des services est quant à elle assurée à l'aide d'ontologies utilisées comme sources de connaissances de support permettant de décrire la signification exacte des données disponibles et des entrées et sorties des services implémentés. L'architecture proposée (Bügel et al., 2007) distingue deux principaux types de services. Les services architecturaux visent à fournir des fonctionnalités génériques et indépendantes de toute application ou domaine spécifiques. Ceux-ci sont exploités par les services thématiques qui offrent des fonctionnalités plus spécifiques, comme c'est le cas du service de catalogue sémantique. Parmi les premiers, trois services dits « sémantiques » ont pour but d'assurer l'interopérabilité sémantique des données et des services thématiques qui s'appuient sur eux. Un service d'annotation sémantique, tout d'abord, exploite des outils de traitement automatique du langage naturel pour annoter automatiquement diverses ressources via les ontologies disponibles. Les entités nommées détectées sont validées par comparaison avec une ressource de support, comme une base de connaissances gérée par le service du même nom. Le service d'accès aux ontologies assure le stockage, la gestion et l'interrogation des ontologies disponibles. Le service « base de connaissances », enfin, assure le transfert de requêtes envoyées par une application cliente à une base de connaissances ORCHESTRA et renvoie les réponses du moteur d'inférence à l'application cliente. Le service de catalogue sémantique élaboré dans le cadre du projet s'appuie sur les services d'accès aux ontologies et d'annotation sémantique afin de permettre l'exécution de requêtes uniformes sur des catalogues indépendants mais répondant aux principaux standards en vigueur (OGC CSW-ISO, OGC CSW-ebRIM, UDDI, etc.). La figure 12 présente l'ensemble des ressources mises en jeu par le système, et identifie les étapes d'établissement des correspondances entre ces ressources automatiques et manuelles.

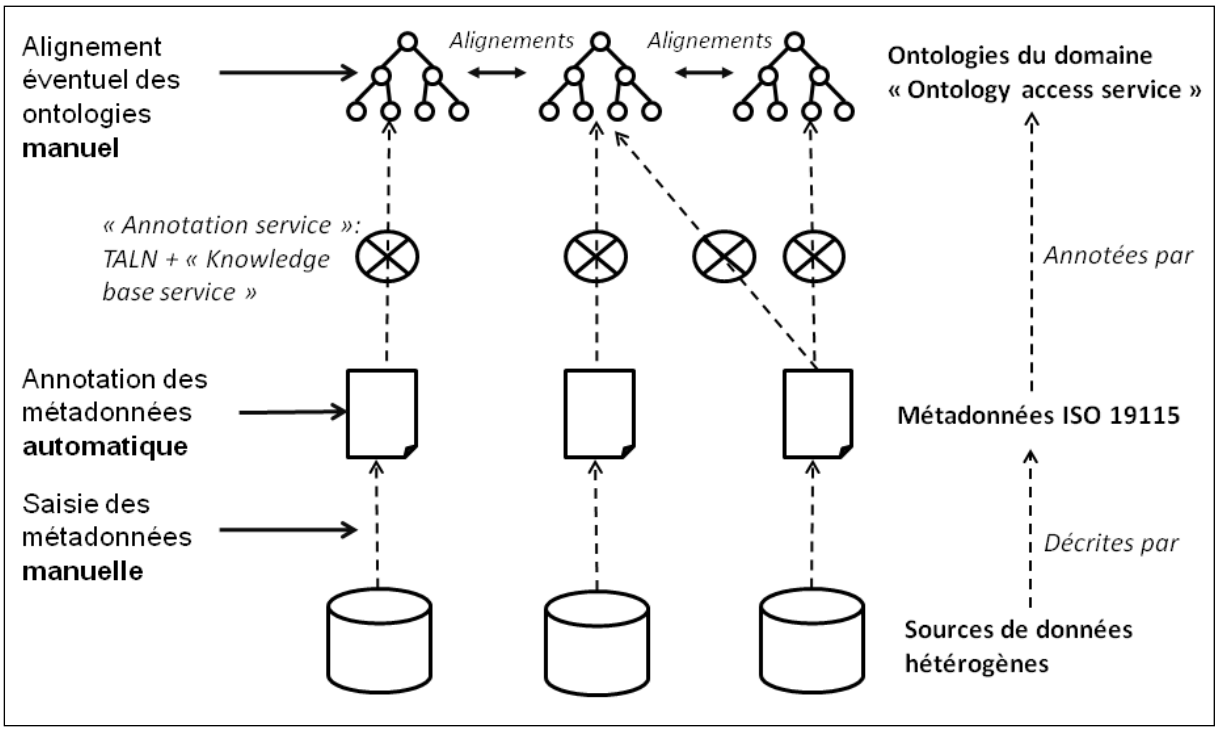


Figure 12: Approche proposée par (Bügel et al., 2007) pour la découverte des données: identification des étapes automatiques et manuelles

Découverte et accès aux données : approche proposée par (Klien, 2008)

Les travaux de thèse de (Klien, 2008) portent sur la découverte et l'accès à des données géographiques hétérogènes dans le cadre des infrastructures de données géographiques. Elle propose un modèle conceptuel pour l'annotation sémantique de données géographiques hétérogènes à l'aide d'ontologies du domaine de la topographie. Ainsi, elle s'intéresse aux divers composants à mettre en œuvre dans le processus d'annotation, à leurs relations et à la forme que doivent prendre les annotations elles-mêmes afin de faciliter, voire automatiser en partie leur définition, l'évaluation de leur cohérence et leur extension à d'autres points de vue. Le modèle conceptuel proposé s'appuie sur le modèle de référence de l'Open Geospatial Consortium (OGC, 2003).

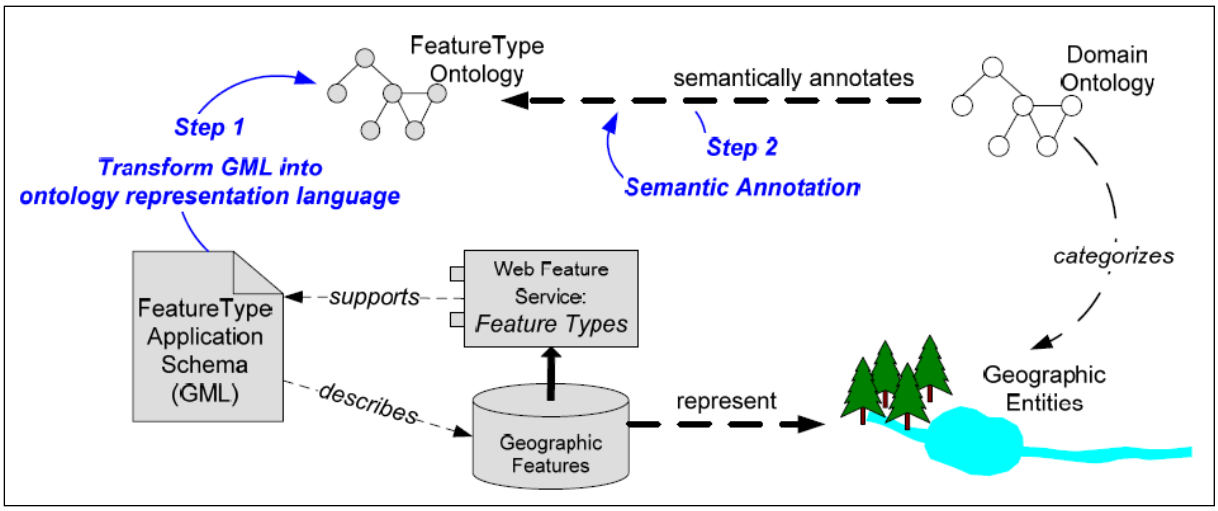


Figure 13: Modèle global pour l'annotation de bases de données géographiques extrait de (Klien, 2008)

De ce fait, la modélisation de l'information géographique repose sur la notion de *Feature* qui représente une abstraction d'une entité du monde réel dotée d'une localisation. Un ensemble de *Features* du même type est décrit par un *Feature Type*. Un jeu de données géographique est décrit par son *Feature Type Schema*. De façon similaire, la description formelle des catégories d'entités géographiques du monde réel est réalisée à l'aide d'une ontologie du domaine. L'auteur insiste sur la nature des liens d'annotation à établir entre *Feature Types* et catégories: il ne doit en aucun cas s'agir de relations taxonomiques dans la mesure où un schéma conceptuel de données et une ontologie du domaine représentent des entités de natures différentes. A la place de relations de type *is-a*, l'auteur propose d'introduire une relation nommée *annotate* afin de relier chaque *Feature Type* d'un schéma conceptuel donné à la(les) catégorie(s) correspondante(s) de l'ontologie du domaine. L'approche proposée pour l'annotation semi-automatique des données géographiques comprend deux étapes.

La première consiste à traduire le schéma conceptuel de données dans la syntaxe utilisée pour décrire l'ontologie du domaine. L'intérêt de cette étape est double. Elle permet, d'une part de formaliser les liens d'annotations dans un langage commun aux schémas et à l'ontologie du domaine, sous la forme d'axiomes et d'autre part d'effectuer des raisonnements à la fois sur les schémas conceptuels de données et sur l'ontologie du domaine. L'ontologie d'application résultante est appelée *Feature Type Ontology* (FTO). Cette étape de traduction syntaxique peut être relativement facilement automatisée, ce qui simplifie d'autant le processus d'annotation sémantique.

La seconde est l'étape d'annotation sémantique proprement dite. Il s'agit d'une étape manuelle qui consiste à déterminer à quelle catégorie - ou concept - de l'ontologie appartient chaque *Feature* du jeu de données puis à annoter le *Feature Type* correspondant via cette catégorie. Cette étape peut être validée de façon automatique en s'appuyant sur les capacités de raisonnement propres aux ontologies. En effet, en supposant que l'on dispose d'une ontologie du domaine présentant des catégories décrites de façon formelle et non ambiguë, un moteur de raisonnement peut déduire quelles sont les instances de l'ontologie FTO remplissant les critères d'appartenance à une catégorie donnée de l'ontologie du domaine. Le *Feature Type* auquel correspondent ces instances peut alors être annoté via la catégorie à laquelle il a été inféré qu'elles appartiennent. La mise en place d'un tel processus présente un double intérêt. D'une part, cela permet de vérifier automatiquement l'annotation sémantique de données géographiques, opération qui, lorsqu'elle est réalisée

manuellement nécessite des connaissances approfondies sur les données elles-mêmes et sur l'ontologie du domaine utilisée ainsi qu'un investissement en temps important de nature à dissuader les producteurs de données. Elle peut, en outre, être source d'erreurs si l'interprétation qui est faite par l'opérateur de la sémantique des données ou des concepts de l'ontologie s'avère erronée. D'autre part, cela permet de détecter de nouvelles annotations possibles via des ontologies issues de domaines connexes.

Cependant, cela nécessite de disposer d'une ontologie du domaine axiomatisée et, dans le cas où l'on dispose de plusieurs ontologies du domaine, de pouvoir les comparer facilement afin de détecter d'éventuelles correspondances (ou alignements) entre elles. Ainsi la structure utilisée pour les ontologies du domaine doit permettre de représenter et comparer différentes conceptualisations d'un même domaine. Ceci suppose que ces ontologies se réfèrent à un dénominateur commun (Masolo et al., 2003; Probst, 2007), c'est-à-dire à une même ontologie de haut niveau, suffisamment générique pour permettre la représentation de connaissances reflétant divers points de vues sur les concepts géographiques, tout en conservant une cohérence dans les descriptions ainsi formalisées. Klien (2008) propose donc d'aligner les catégories les plus génériques de l'ontologie du domaine utilisée avec les catégories fournies par l'ontologie de haut niveau DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Masolo et al., 2003). En cela, elle rejoint (Kuhn, 2003) en proposant un cadre de référence sémantique permettant le référencement sémantique des schémas conceptuels de données des jeux de données mis à disposition. Deux implémentations successives sont proposées dans la thèse: une première est réalisée en OWL avec des annotations en SWRL[18], le langage de règles associé, et la seconde à l'aide du Web Service Modeling Language (WSML[19]), choisi pour sa plus grande expressivité. La figure 14 présente une implémentation en OWL de l'ontologie du domaine utilisée[20].

[18] http://www.w3.org/Submission/SWRL/
[19] http://www.wsmo.org/TR/d16/d16.1/v0.2/
[20] http://ifgi.uni-muenster.de/~klien/tgis/LandscapeClassification.owl

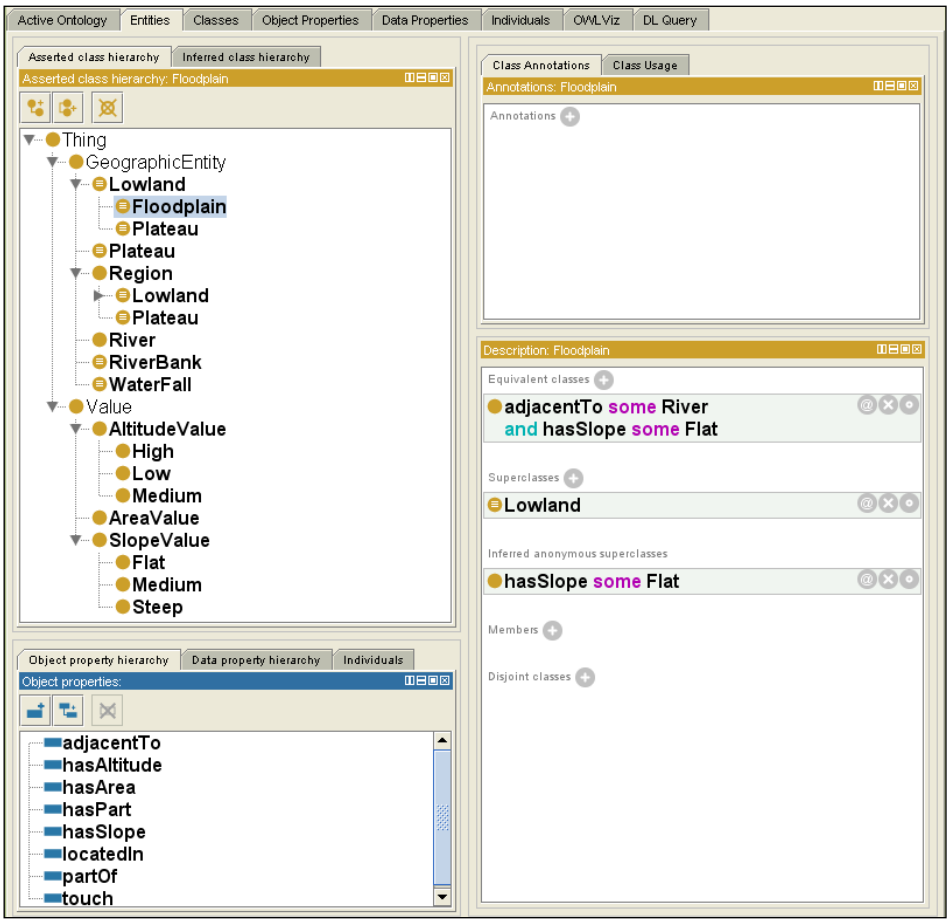


Figure 14: Ontologie du domaine OWL proposée par (Klien, 2008), visualisation sous Protégé[21]

L'évaluation de la qualité des annotations produites manuellement est réalisée en testant si les instances du *Feature* annoté via une catégorie de l'ontologie du domaine remplissent les critères d'appartenance à cette catégorie et représentent donc des membres de cette catégorie. Ces critères reposent essentiellement sur les propriétés physiques que doivent posséder les entités géographiques pour appartenir à une catégorie et sur les relations spatiales qu'elles entretiennent entre elles. Les *Features* possédant une géométrie, il est possible de tester s'ils possèdent ou non les propriétés physiques et s'ils vérifient ou non les relations spatiales caractéristiques d'une catégorie donnée. Ceci est réalisé en deux temps. En effet, les moteurs d'inférence pour ontologies ne disposent pas d'opérateurs spatiaux. Les analyses spatiales utiles sont donc effectuées par un logiciel de Système d'Information Géographique (SIG). Pour ce faire, les règles d'appartenance à une catégorie sont traduites en opérations d'analyse spatiale et stockées dans un fichier XML. Celui-ci est utilisé en entrée de l'application d'évaluation de l'appartenance à une catégorie, ainsi que des annotations complémentaires fournies par l'utilisateur, afin de définir les tests à effectuer sur les données pour déterminer si elles représentent ou non une certaine catégorie d'entités géographiques. Lorsque ces critères d'appartenance à une catégorie de l'ontologie du domaine sont vérifiés par un *Feature*, on peut alors en déduire qu'il représente très probablement une entité géographique membre de cette catégorie. L'annotation qui relie son *Feature Type* à la catégorie testée peut alors être validée. La figure 15 présente l'ensemble des ressources mises en jeu par

[21] http://protege.stanford.edu/

l'application et les étapes automatiques et manuelles permettant l'établissement des relations de correspondance entre elles.

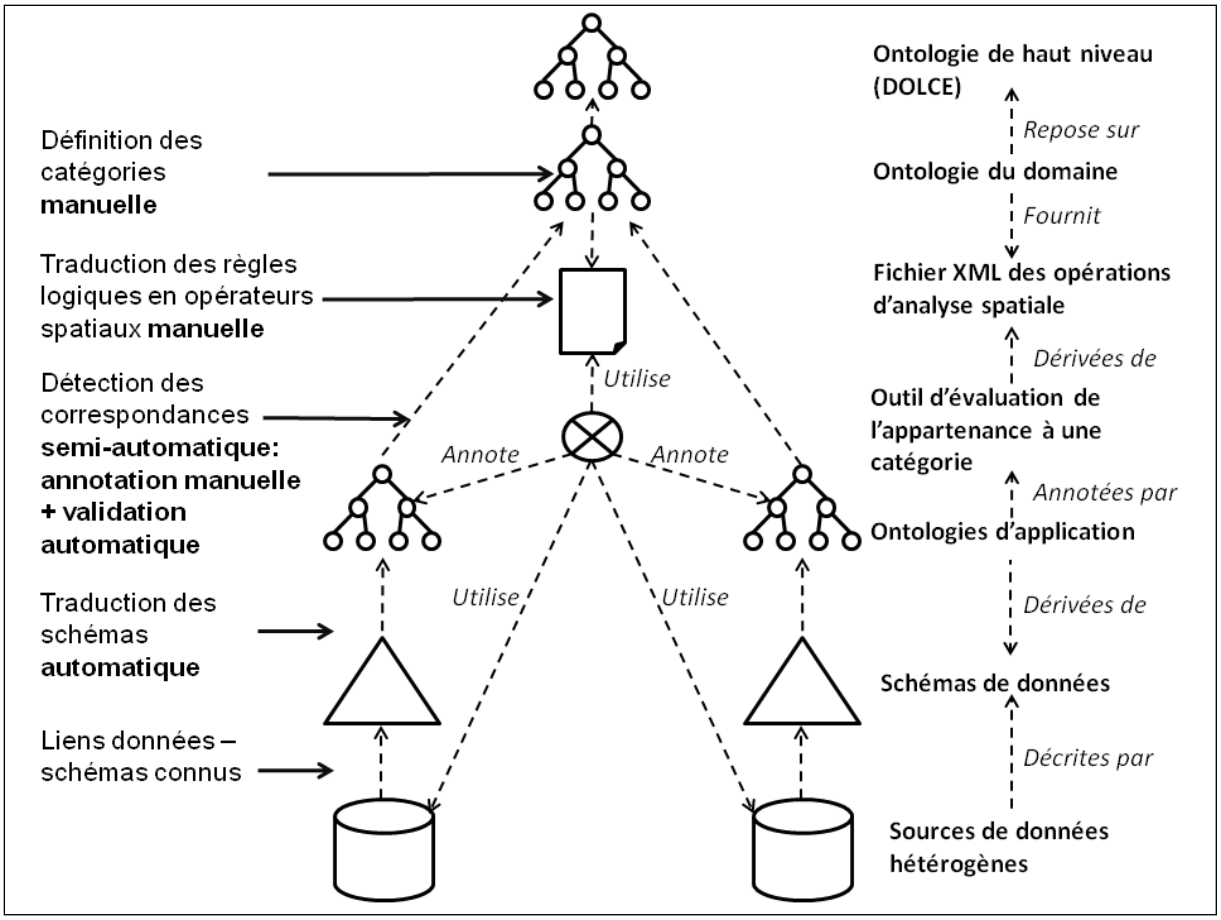


Figure 15: Approche proposée par (Klien, 2008) pour l'annotation semi-automatique de données géographiques: identification des étapes automatiques et manuelles

Ce modèle a été appliqué à l'annotation de *Feature Types* par le concept de zone inondable. L'essentiel du travail de constitution d'une ontologie du domaine se concentre sur la définition du cadre de référence sémantique sur lequel s'appuient les définitions des concepts du domaine. Seuls quelques concepts géographiques sont définis dans l'ontologie proposée. De plus, les définitions utilisées pour les concepts de l'ontologie du domaine demeurent très spécifiques et subjectives, s'appuyant sur des seuils plus propres à définir des *Feature Types* que des concepts génériques. Par exemple, le concept de zone inondable est défini comme l'ensemble des zones adjacentes à une rivière et dont la différence d'altitude locale par rapport à cette rivière est inférieure à quatre mètres.

Conclusion

Les applications pour la découverte et l'accès aux données géographiques détaillées ci-dessus présentent, à quelques variations près, des architectures semblables. Les approches adoptées s'apparentent en effet à une architecture de type médiateur ; les sources de données hétérogènes sont décrites via leur schéma, une ontologie d'application ou encore des métadonnées qui sont

annotés par une ontologie du domaine qui fournit une vue unifiée des données pour les utilisateurs (voir figure 16). Les principales différences d'une application à l'autre résident dans les choix des langages de description des ressources (OWL, RDF, WSML, etc.) et d'implémentation des requêtes d'accès aux sources de données.

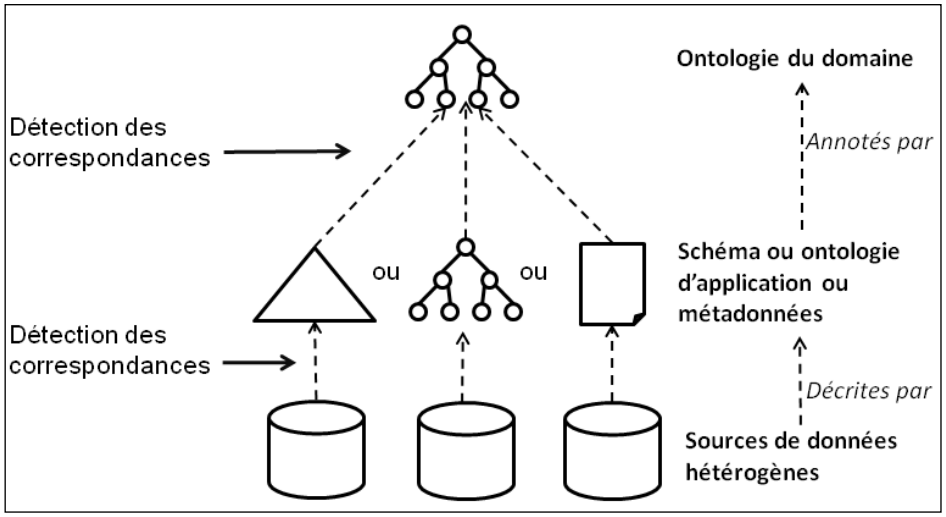


Figure 16: Architecture globale des applications décrites dans ce chapitre

Une autre différence fondamentale concerne la détection des correspondances entre les ressources descriptives des données (schémas, ontologies d'application ou métadonnées) et l'ontologie du domaine. Cette étape d'appariement est indispensable pour l'annotation sémantique des sources de données sur laquelle repose le système. Les relations de correspondance établies entre ressources descriptives et ontologie du domaine sont en effet exploitées pour détecter les sources de données pertinentes pour un utilisateur et pour traduire les requêtes de l'utilisateur exprimées dans les termes de l'ontologie du domaine, dans les termes des sources de données disponibles.

Selon les cas, l'annotation sémantique est réalisée de façon manuelle ou automatique. Dans la plupart des cas, le travail complexe d'appariement des ressources descriptives des données avec l'ontologie du domaine est laissé à la charge du fournisseur de données. Dans les rares cas où la mise en correspondance des ressources et de l'ontologie du domaine est réalisée de façon automatique, c'est par simple comparaison de chaînes de caractères, les auteurs ayant, au préalable, pris soin de décrire leurs ressources dans les termes exacts de l'ontologie du domaine. Les travaux de (Klien, 2008) proposent une approche différente, prenant en compte la géométrie des données géographiques, pour valider des annotations existantes et tenter d'en détecter de nouvelles en testant l'appartenance des instances des diverses sources de données à un concept de l'ontologie du domaine. Plus généralement, le modèle d'annotations sémantiques proposé présente un intérêt particulier dans la mesure où il permet la représentation et l'exploitation de connaissances détaillées sur les éléments de schémas hétérogènes et les concepts géographiques qu'ils représentent.

De plus, seule l'application proposée par (Lutz et al., 2006) exploite les annotations disponibles pour apparier les schémas de données décrits via l'ontologie globale – ici dans le cadre d'une application de transformation de schémas pour la composition de services Web. Dans les autres applications, aucune mention n'est faite d'une éventuelle tentative de réconciliation de références lors des réponses aux requêtes. Les annotations ne sont donc pas utilisées pour apparier directement les

schémas des différentes sources, mais pour rediriger et réécrire les requêtes à destination des sources pertinentes.

Enfin, les ontologies du domaine mises en œuvre sont très peu documentées et semblent généralement restreintes en termes de couverture d'un domaine particulier. Si les travaux de (Desconnets et al., 2007) réutilisent des thesauri existants, les autres approches exploitent des ontologies du domaine créées spécialement pour les besoins de l'application. Celles-ci sont présentées de façon succincte : les auteur décrivent généralement assez peu, voire pas du tout, le processus de création de leurs ontologies, et leur contenu, généralement limité à quelques concepts, propriétés et relations, n'est que rarement présenté en détails. Seul (Klien, 2008) présente le processus de construction de l'ontologie utilisée, mais cette description se concentre pour l'essentiel sur la définition du cadre de référence sémantique sur lequel repose sa conceptualisation du domaine.

3.1.2 Intégration de schémas pour les infrastructures de données géographiques

La mise en œuvre pratique de la directive INSPIRE 2007/2/CE fait l'objet de plusieurs projets au sein de l'association Eurogeographics[22] qui regroupe des agences nationales de cartographie et de cadastre européennes. En particulier, le projet European Spatial Data Infrastructure Network (ESDIN[23]), qui vise à améliorer l'automatisation du processus de création du jeu de données fédéré constitué des thèmes de l'annexe 1 de la directive, à partir des données des États membres, illustre l'importance de ce processus. La transformation automatique de schémas, ou encore le recalage de données transfrontalières constituent donc deux tâches d'intégration auxquelles le projet s'intéresse particulièrement. Parmi les projets partenaires d'ESDIN on compte le projet européen Humboldt[24], qui s'attache à fournir des outils open source pour l'harmonisation de données géographiques dans le cadre de la mise en place de l'infrastructure de données géographiques européenne. En particulier, des outils dédiés à l'appariement de schémas et à la transformation de schémas sont développés par ce projet. La transformation de schéma constitue une variante de l'intégration de schémas. Elle consiste à intégrer un schéma développé indépendamment avec un schéma cible donné (Rahm et Bernstein, 2001). Il s'agit donc, partant d'un schéma source, de permettre sa transformation, ainsi que celle des données de la base correspondante, vers un schéma cible. Transformer des données existantes afin de les rendre conformes à un nouveau schéma comporte plusieurs étapes (De Vries et Reitz, 2009). Une fois le schéma cible spécifié, et le schéma source connu, il convient de déterminer les relations de correspondance entre eux, c'est-à-dire de procéder à leur appariement. Puis, ces relations de correspondance sont exploitées afin de traduire les données sources conformément au schéma cible. La transformation des données peut être réalisée à la volée, ou a priori, par le producteur de données lui-même. Cette étape de transformation requiert la mise au point et l'exécution d'algorithmes permettant la traduction des données. Ceci implique des opérations diverses comme renommer des entités de schémas, filtrer les données à l'aide de requêtes attributaires et (ou) spatiales, réaliser des opérations complexes de concaténation, d'agrégation, de découpage ou de transformation de géométries, recoder des valeurs d'attributs ou

[22] http://www.eurogeographics.org/

[23] http://www.esdin.eu/fr/about

[24] http://www.esdi-humboldt.eu/home.html

encore affecter des valeurs par défaut aux attributs. A ceci vient s'ajouter la création de métadonnées décrivant le jeu de données cible ainsi produit. Une étape clé de ce processus de transformation de schéma est l'appariement du (ou des) schéma(s) sources avec le schéma cible. Celui-ci doit, en effet, être réalisé très finement afin de permettre une définition cohérente et précise des opérations de transformation à effectuer. C'est à cette étape d'appariement de schémas que nous allons nous intéresser plus particulièrement dans la mesure où celle-ci se retrouve dans toutes les applications d'intégration de données géographiques, quelle que soit leur finalité.

3.1.2.1 Approches manuelles pour l'appariement de schémas de bases de données géographiques

Dans le cadre de la mise en œuvre de la directive INSPIRE 2007/2/CE, l'étape d'appariement des schémas est généralement réalisée manuellement par les producteurs de données, à l'aide d'outils dédiés visant à simplifier la saisie des relations de correspondance entre éléments de schémas à apparier. C'est le cas notamment des logiciels commerciaux GoPublisher (Snowflake Software[25]), et Feature Manipulation Engine (FME) (Safe Software[26]) ou encore du logiciel open source Spatial Data Integrator (SDI) (Camp To Camp et NeoGeo Technologies [27]), qui fournissent une interface intuitive aux utilisateurs leur permettant de définir les correspondances entre les schémas source et cible. La transformation des données est réalisée à partir des relations de correspondances fournies par l'utilisateur. Cependant, les relations de correspondance et les opérations de transformation proposées ici sont définies au niveau des schémas logiques et de ce fait demeurent encore limitées vis-à-vis de la complexité des données à traiter.

Une suite d'outils logiciels open source dédiée à la transformation de schémas pour les infrastructures de données géographiques a également été développée dans le cadre du projet Humboldt[28]. Celle-ci se compose tout d'abord du Humboldt GeoModel Editor, un logiciel d'édition de schémas conceptuels de données spécialement dédié aux données géographiques et fondé sur UML, qui permet de créer et d'exporter, dans divers formats (notamment XMI ou GML) des schémas conceptuels de données géographiques. Le Humboldt Alignment Editor (HALE) est un logiciel doté d'une interface conviviale permettant de spécifier manuellement les relations de correspondance entre un schéma source et un schéma cible, éventuellement décrits à l'aide du GeoModel Editor. Enfin le Conceptual Schema Transformer (CST) est une bibliothèque permettant d'exécuter des transformations de données d'un schéma à un autre, en s'appuyant sur des correspondances décrites à l'aide de HALE. Cette dernière application est mise à disposition sous la forme d'un *Web Processing Service*. Lors de la conception de HALE, l'accent a été mis sur deux aspects particuliers. D'une part un effort important a été consenti concernant l'ergonomie de l'interface (Reitz et Kuijper, 2009), présentée en figure 17.

[25] http://www.snowflakesoftware.com/products/gopublisher/
[26] http://www.safe.com/fme/fme-technology/
[27] http://www.talendforge.org/wiki/doku.php?id=sdi:mainpage
[28] http://www.esdi-humboldt.eu/home.html

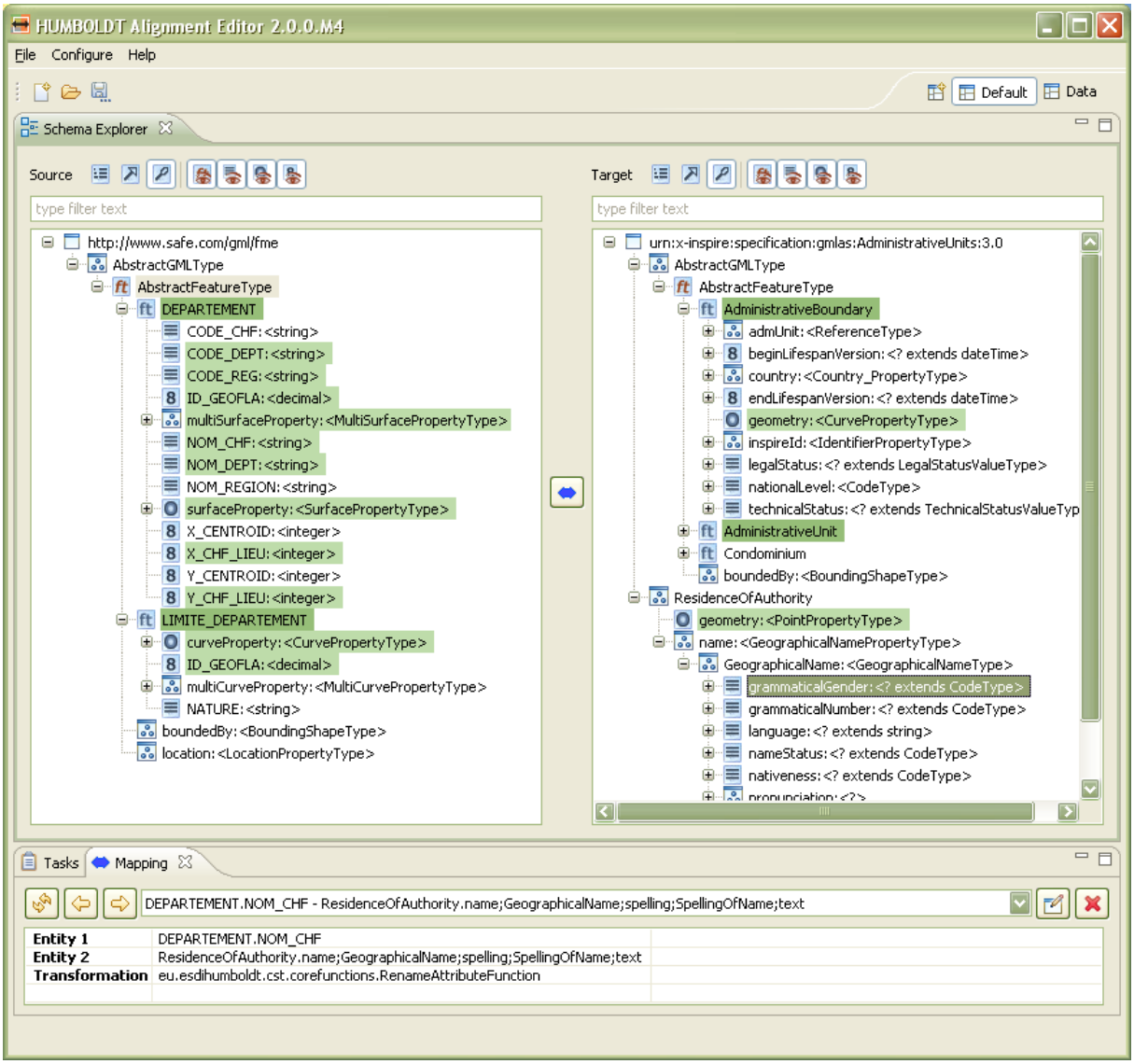


Figure 17: Interface d'appariement manuel de schémas fournie par HALE : appariement du schéma conceptuel de données GEOFLA© avec le schéma INSPIRE « Administrative Units »

Celle-ci a en effet été conçue afin de faciliter la déclaration des relations de correspondance entre éléments de schémas, ainsi que des opérations de transformation à effectuer pour passer du schéma source au schéma cible. L'interface proposée résulte d'une étude réalisée auprès d'utilisateurs concernant l'ergonomie des interfaces de deux outils d'alignement d'ontologies, PROMPT et COMA++, afin de pouvoir prendre en compte leurs avantages et inconvénients dans le cas d'une application d'appariement de schémas de bases de données géographiques. Ainsi, contrairement aux interfaces des outils d'alignement d'ontologies étudiés, l'interface de visualisation des schémas permet leur affichage selon divers modes d'organisation (relations d'héritage, relations d'agrégation, autres relations sémantiques). De plus, au fur et à mesure de leur définition par l'utilisateur, les diverses tâches de transformation sont ordonnées et affichées par le système selon des critères de complexité et d'influence sur le résultat final. Ceci permet à l'utilisateur d'identifier plus facilement les tâches prioritaires. Enfin les opérations de transformation sont appliquées à la volée sur un jeu de données source et les données résultantes sont affichées au sein d'une fenêtre de visualisation cartographique dédiée permettant l'affichage simultané des données sources et du résultat de leur

transformation pour deux portions contiguës de la zone géographique couverte. Cet outil de visualisation vise à fournir à l'utilisateur un outil intuitif pour évaluer rapidement la cohérence des relations de correspondance et des opérations de transformation qu'il définit entre les deux schémas.

D'autre part la conception de HALE a fait l'objet de travaux approfondis sur la définition d'un langage dédié à l'appariement de schémas de bases de données géographiques. Un appariement entre schéma source et schéma cible est défini à l'aide du langage geographic Ontology Mapping Language (gOML) (Reitz et al., 2009) dérivé du langage d'alignement d'ontologies proposé par (Euzenat et al., 2007), choisi pour son expressivité. Dans ce langage, l'ensemble des relations de correspondance entre deux schémas est appelé *alignement*. Une relation de correspondance entre éléments de schémas ou entre instances est, quant à elle, appelée *mapping*. Un alignement est donc constitué d'un ensemble de mappings. gOML s'appuie ainsi sur ce langage pour décrire de façon déclarative des relations entre ontologies, ou entre schémas dans ce cas de figure, à la fois au niveau des classes, des attributs, des associations et des instances, et permet l'expression de restrictions sur ces correspondances. Ces restrictions peuvent être définies directement à l'aide d'OML, mais gOML intègre également la possibilité d'utiliser le langage OGC Common Query Language (CQL) (OGC, 2007) pour exprimer des restrictions sur un mapping entre éléments de schémas ou entre instances. De plus, un langage de description des incohérences entre mappings (Mismatch Description Language) (Reitz, 2010) a été intégré à gOML afin de documenter les mappings éventuellement imparfaits. Par exemple, deux classes peuvent être mises en correspondance en raison d'un certain nombre d'instances communes, mais posséder chacune des instances n'appartenant pas à leur classe homologue. La déclaration de ces incohérences lors de l'appariement des schémas permet d'anticiper d'éventuelles erreurs lors de l'utilisation future des mappings. En effet, ceux-ci sont destinés à être exploités par le système pour déduire les opérations de traduction nécessaires qui devront être exécutées sur les données source. De plus, dans l'éventualité de leur réutilisation pour une application d'intégration de schémas avec une base de données nouvelle, il convient de pouvoir tenir compte de ces incohérences afin d'éviter de générer par transitivité des mappings erronés entre ce nouveau schéma et ceux déjà appariés. L'exemple présenté ci-dessous montre un mapping simple en gOML entre l'attribut *NOM_CHF* de la classe *DEPARTEMENT* de GEOFLA© et l'attribut *name* de la classe *ResidenceOfAuthority* du schéma de données INSPIRE.

```
<align:map>
  <align:Cell>
    <omwg:entity1>
    <omwg:Property df:about=
"http://www.safe.com/gml/fme/DEPARTEMENT/NOM_CHF">
      <omwg:transf rdf:resource =
"eu.esdihumboldt.cst.corefunctions.RenameAttributeFunction"/>
      </omwg:Property>
    </omwg:entity1>
    <omwg:entity2>
      <omwg:Property rdf:about= "urn:x-
inspire:specification:gmlas:AdministrativeUnits:3.0/ResidenceOfAuthority/na
me;GeographicalName;spelling;SpellingOfName;text"/>
    </omwg:entity2>
    <align:relation>Equivalence</align:relation>
  </align:Cell>
</align:map>
```

3.1.2.2 Approches automatiques pour l'appariement de schémas

L'appariement manuel de schémas conceptuels hétérogènes constitue une tâche longue et fastidieuse, qui se limite à des applications statiques. On rencontre donc, dans la littérature, de nombreuses approches développées afin d'automatiser, autant que faire se peut, cette étape d'appariement de schémas (Rahm et Bernstein, 2001). De façon générale, la problématique de l'intégration de schémas rejoint, dans le domaine de l'intelligence artificielle, celle de l'intégration d'ontologies (Rahm et Bernstein, 2001). Si schémas conceptuels et ontologies diffèrent en ceci que les schémas conceptuels ne fournissent pas de sémantique explicite sur les données qu'ils décrivent, là où les ontologies permettent de définir les concepts qu'elles décrivent sous forme d'axiomes, ces deux notions présentent néanmoins des similitudes. Les schémas conceptuels de bases de données et les ontologies ont en effet en commun le fait de fournir un vocabulaire de termes décrivant un domaine d'application, et de contraindre la sémantique de ces termes (Shvaiko et Euzenat, 2005). Il est donc courant que des approches similaires soient employées pour l'appariement de schémas et pour l'alignement d'ontologies. Celles-ci consistent généralement en diverses combinaisons de techniques d'appariement élémentaires. Ces techniques d'appariement élémentaires peuvent être classées de diverses façons. Une classification à double entrée de ces approches a été proposée par (Euzenat et Shvaiko, 2007). Celle-ci repose, d'une part, sur le niveau de granularité des entités en entrée de l'algorithme d'appariement (éléments de schémas seuls ou considérés au travers leurs relations) et sur la façon dont ce dernier interprète ces informations (en s'appuyant sur les seules données en entrée, à l'aide de ressources externes ou via une sémantique formelle) et d'autre part sur le type des données en entrée de l'algorithme d'appariement (chaînes de caractères, structures, instances, etc.). Nous reprenons cette classification dans le tableau 2 pour présenter succinctement les diverses techniques élémentaires d'appariement couramment utilisées. Ces techniques élémentaires sont également présentées en détail dans (Euzenat et Shvaiko, 2007). Le sens de lecture de la classification suivi est celui fondé sur la distinction granularité/interprétation des entrées. Pour une vision complémentaire de l'état de l'art des techniques élémentaires d'appariement de schémas on pourra également se référer à (Kavouras et Kokla, 2008).

Techniques considérant les éléments de schémas seuls	
Techniques terminologiques	Techniques fondées sur la comparaison des termes (i.e. des chaînes de caractères) utilisés pour désigner les éléments de schémas.
Techniques linguistiques	Techniques fondées sur la comparaison des termes (i.e. des chaînes de caractères) utilisés pour désigner les éléments de schémas, et mettant en œuvre des techniques de traitement automatique du langage naturel permettant la prise en compte des particularités grammaticales de ces termes.
Techniques utilisant des ressources linguistiques externes	Techniques fondées sur la comparaison des termes (i.e. des chaînes de caractères) utilisés pour désigner les éléments de schémas et exploitant des ressources externes telles des lexiques ou des thesauri, qui apportent des connaissances au système sur l'existence de relations de synonymie ou d'hyperonymie entre termes, afin de permettre la prise en compte des relations

	linguistiques entre termes utilisés pour nommer les éléments de schémas.
Techniques fondées sur les contraintes internes des schémas	Techniques fondées sur la comparaison des contraintes internes de définition des éléments de schémas - types de données imposés (chaînes de caractères, entiers, flottants, booléens, etc.), cardinalité des attributs ou encore clés primaires et étrangères - afin de détecter des relations de correspondance entre eux.
Réutilisation d'alignements	Techniques fondées sur la comparaison des résultats d'un alignement a_1 (i.e. d'un ensemble de relations de correspondance) entre un schéma s_1 et un autre schéma du domaine s_3 et ceux d'un alignement a_2 entre un schéma s_2 et s_3, afin de déduire des relations de correspondance entre s_1 et s_2.
Utilisation d'ontologies de haut-niveau et d'ontologies du domaine formelles	Techniques fondées sur la définition formelle des différents éléments de schémas à apparier à l'aide de concepts fournis par une ontologie de haut niveau ou une ontologie du domaine formelle (cf. réutilisation d'alignements), et de divers opérateurs logiques. Ces définitions faisant référence à des concepts communs, elles deviennent comparables pour des applications de raisonnement pouvant traiter les opérateurs logiques utilisés.
Techniques considérant les relations entre éléments de schémas	
Techniques fondées sur les graphes	Techniques fondées sur l'assimilation des schémas en entrée à des graphes étiquetés dont les nœuds et les arcs possèdent des labels. L'évaluation de la similarité entre deux éléments de schémas (i.e. entre deux nœuds des graphes à comparer) est réalisée par comparaison de leurs voisinages au sein de chacun des schémas (i.e. au sein de chacun des graphes).
Techniques fondées sur les relations taxonomiques	Ces techniques constituent une spécialisation des techniques d'appariement fondées sur les graphes. On ne considère ici que les relations de généralisation/spécialisation dans l'analyse des voisinages des nœuds à comparer.
Référentiels de structures	Techniques fondées sur l'évaluation préalable de la similarité globale entre (portions d'/de) ontologies ou schémas, désignés ici par le terme commun de « structure », afin de déterminer si les deux structures à apparier sont suffisamment proches l'une de l'autre pour justifier l'utilisation d'algorithmes d'appariements élaborés et souvent lourds à mettre en œuvre, ou de détecter des alignements réutilisables.
Techniques fondées sur la	Techniques fondées sur la comparaison de l'interprétation sémantique des éléments d'ontologies à apparier. Les techniques

sémantique	de raisonnement fondées sur les logiques de description qui permettent de détecter des relations de subsomption et d'équivalence entre concepts d'ontologies décrits sous forme d'axiomes entrent dans le cadre de ces techniques.
Analyse de données et statistiques	Techniques d'appariement dites en extension. Elles consistent à rechercher des relations de correspondance au niveau des instances, en s'appuyant sur divers critères, afin d'en déduire des relations de correspondance au niveau des schémas. Les techniques fondées sur la classification automatique ou l'analyse formelle de concepts entrent dans le cadre de ces techniques d'appariement.

Tableau 2: Classification des techniques d'appariement de schémas élémentaires proposée par (Euzenat et Schvaiko, 2007)

Enfin un état de l'art plus récent dressant le bilan des avancées en matière d'alignement d'ontologies, et proposant des pistes de recherche dans ce domaine est également proposé par (Shvaiko et Euzenat, 2011). Les auteurs y présentent plusieurs systèmes d'alignement d'ontologies choisis en raison de leurs participations régulières aux campagnes 2004 à 2010 de l'OAEI (Ontology Alignment Evaluation Initiative[29]), initiative internationale organisant des campagnes d'évaluation annuelles des systèmes d'alignement d'ontologies au travers de la comparaison des résultats obtenus par chaque système sur des cas d'alignement imposés. Un tableau récapitulatif dresse l'inventaire des formats d'ontologies acceptés, des types de relations de correspondance produites, et des techniques d'appariement élémentaires disponibles au sein de chaque système. Au final, il s'avère que tous ces systèmes prennent en entrée des ontologies au format OWL. Certain acceptent, en outre, des formats comme SKOS[30]. La plupart de ces systèmes, à quelques exceptions près, produisent des relations de correspondance de type 1 :1. Celles-ci sont détectées à l'aide de diverses combinaisons de techniques d'appariement, pour l'essentiel terminologiques et structurelles, le recours à des techniques en extension ou à des techniques fondées sur la sémantique étant plus rare. Les auteurs analysent, en outre, les progrès réalisés au fil du temps par chaque système et s'interrogent sur les pistes d'amélioration pour les années à venir. Ils identifient ainsi huit défis que les systèmes d'alignement d'ontologies devront relever. Le premier concerne l'évaluation de résultats d'alignement sur des ontologies de très grande taille, qu'il s'agisse d'évaluer le degré de difficulté d'une tâche d'alignement ou de produire un alignement – i.e. un ensemble de relations de correspondance - de référence. Le deuxième porte sur l'optimisation des systèmes d'alignement en termes de ressources nécessaires et de temps d'exécution. Le troisième se rapporte à l'identification, la sélection et l'exploitation pertinentes de sources de connaissances de support pour l'alignement. Le quatrième concerne la sélection, la combinaison, le paramétrage automatique et l'adaptation d'outils de mise en correspondance afin de tirer le meilleur parti des qualités de chacun. Le cinquième consiste à proposer des solutions pour aider les utilisateurs des systèmes d'alignement à parcourir et contrôler les résultats, éventuellement très volumineux, obtenus. Dans le même souci d'implication des utilisateurs, le sixième point s'intéresse à la clarté des résultats d'alignement qui

[29] http://oaei.ontologymatching.org/

[30] http://www.w3.org/2004/02/skos/

leur sont présentés ; les relations de correspondance détectées par le système doivent être facilement compréhensibles si l'on souhaite permettre leur validation par les utilisateurs. Le septième point porte sur l'élaboration de systèmes d'alignement collaboratifs. Enfin le dernier vise la création d'outils permettant le stockage et le partage des alignements produits.

3.1.2.3 Mise en œuvre de ces techniques dans des approches (semi-)automatiques pour l'appariement de schémas de bases de données géographiques

La plupart des outils d'appariement de schémas ou d'alignement d'ontologies (semi-)automatiques combinent ces différentes techniques élémentaires afin de tirer profit des avantages de chacune. Nous présentons ici quelques approches d'appariement de schémas proposées dans le domaine des bases de données géographiques.

Appariement automatique de schémas de bases de données géographiques: approche proposée par (Volz, 2005)

Une approche en extension pour l'appariement automatique de schémas a été proposée par (Volz, 2005). En l'absence d'identifiants universels pour les objets géographiques, les processus d'appariement de données géographiques doivent s'appuyer sur divers critères afin d'établir les correspondances entre objets de bases de données géographiques hétérogènes qui représentent une même entité géographique du monde réel. Aussi, en première approche pourra-t-on se référer à la localisation de ces objets. En effet, deux instances de deux bases de données géographiques situées au même endroit sont très susceptibles de représenter un même objet du monde réel. Cependant, en raison de niveaux de détail et de précisions géométriques différents d'une base à l'autre, ce critère de localisation seul ne suffit pas toujours à déterminer quels sont les objets correspondants. La prise en compte d'autres critères comme les catégories d'entités géographiques auxquelles les données se rapportent, leur forme, la toponymie, les valeurs d'attributs, ou encore leur topologie, ainsi que de connaissances d'experts sur les données elles-mêmes, comme leur précision géométrique, doit alors être envisagée afin de pallier ce problème. Dans cette optique, (Olteanu, 2008) propose une approche pour l'appariement de données géographiques qui s'appuie sur la théorie des fonctions de croyance afin d'inclure dans la processus d'appariement des connaissances d'experts concernant les données, éventuellement imparfaites mais néanmoins utiles, et de combiner divers critères d'appariement (géométrique, sémantique, topologique, etc.). L'appariement de données géographiques constitue à lui seul un domaine de recherche complexe. Les algorithmes mis en œuvre dépendent fortement des données à apparier et de l'application à laquelle est voué le processus d'appariement de données (intégration de bases de données, contrôle qualité, recalage de données, etc.). Nous ne nous attacherons donc pas ici à fournir un état de l'art complet de ce domaine.

Appariement semi-automatique de schémas de bases de données géographiques: approche proposée par (Cruz et al., 2007)

L'application proposée par (Cruz et al., 2007), l' « AgreementMaker », associe des approches lexicales, structurelles, et manuelles pour l'alignement d'ontologies géographiques. L'application développée permet en effet d'aligner automatiquement deux ontologies géographiques en

comparant les termes utilisés pour désigner les concepts de chacune des ontologies ainsi que leurs propriétés. Les relations de correspondance ainsi détectées sont ensuite vérifiées par comparaison des hiérarchies de subsomption auxquelles appartiennent les concepts mis en relation sur la seule base de leurs labels. Une interface conviviale pour l'alignement manuel des ontologies a également été développée. Les ontologies y sont visualisables sous la forme d'arbres et l'utilisateur peut saisir manuellement les relations de correspondance qui lui paraissent pertinentes. Divers types de relations sont proposées: équivalence entre classes (de type 1:1), généralisation entre classes (de type 1:1 ou 1:n), spécialisation entre classes (de type 1:1 ou 1:n). Les mappings détectés par ces deux biais ainsi que la structure de chacune des ontologies sont ensuite exploités par un troisième algorithme d'alignement afin de déduire de nouveaux mappings. Un quatrième algorithme, dit de consolidation, permet de détecter d'éventuelles incohérences entre les mappings établis par les trois algorithmes précédents. Le résultat du processus d'alignement est ensuite stocké dans un fichier listant tous les concepts de l'ontologie cible et leur associant les concepts de l'ontologie source qui leur correspondent, selon chacune des méthodes utilisées.

Transformation automatique de schémas de bases de données géographiques: approche proposée par (Schade, 2010)

Schade (2010) propose une approche fondée sur des ontologies permettant d'automatiser la phase d'appariement de schémas d'un processus de transformation. Le but de cet appariement est de permettre l'exécution des opérations de transformation de données nécessaires et de générer des métadonnées concernant la qualité des données obtenues en sortie du processus, à l'instar de (Balley, 2007).

Pour ce faire, les schémas source et cible sont représentés sous la forme d'ontologies exprimées à l'aide du langage Web Service Modeling Language (WSML[20]) et annotés via une ontologie du domaine (voir figure 18). En cela, il s'appuie sur les travaux de (Klien, 2008) sur l'annotation sémantique de schémas conceptuels de bases de données géographiques, présentés dans la partie 3.1.1, dont il reprend et adapte le modèle global. L'ontologie *Feature Type Ontology* (FTO) constitue, comme chez (Klien, 2008), l'élément central du modèle. Elle vise à représenter de façon formelle les éléments constitutifs des schémas conceptuels de données à apparier, ainsi que leurs instances. Deux autres types d'ontologies entrent en jeu: des ontologies d'application concernant les modèles ISO spatial schema et OGC GML sur l'encodage des données géographiques d'une part, et les ontologies du domaine et de haut niveau composant le cadre de référence sémantique d'autre part. Les premières sont générées de façon automatique à partir des spécifications des normes correspondantes. L'ontologie FTO est également dérivée automatiquement à partir des schémas de *Feature Types* GML correspondants. Cette traduction s'appuie sur les ontologies d'encodage de données qui fournissent les types de données de base utiles à la description des schémas conceptuels. Les liens entre les éléments de schémas de *Feature Types* GML et les éléments correspondants de l'ontologie FTO sont explicités au niveau des schémas GML sous la forme d'une annotation nommée « model reference » empruntée au standard SA-WSDL. La relation inverse est implémentée au sein de l'ontologie FTO sous la forme d'une propriété WSML. La construction du cadre de référence s'appuie sur les travaux de (Probst, 2007) pour la partie concernant les propriétés des entités géographiques et leurs mesures sur ceux de (Klien, 2008) pour la classification des types d'entités géographiques. Tous deux utilisent l'ontologie de haut niveau DOLCE qu'ils étendent en spécialisant les concepts pertinents pour la description des entités géographiques. (Schade, 2010)

reprend et adapte ces extensions afin de disposer des concepts de base nécessaires à la description formelle détaillée des éléments de schémas qu'il souhaite apparier, ainsi qu'à la description des entrées et sorties des diverses opérations de transformation pouvant être effectuées sur les éléments d'un schéma source.

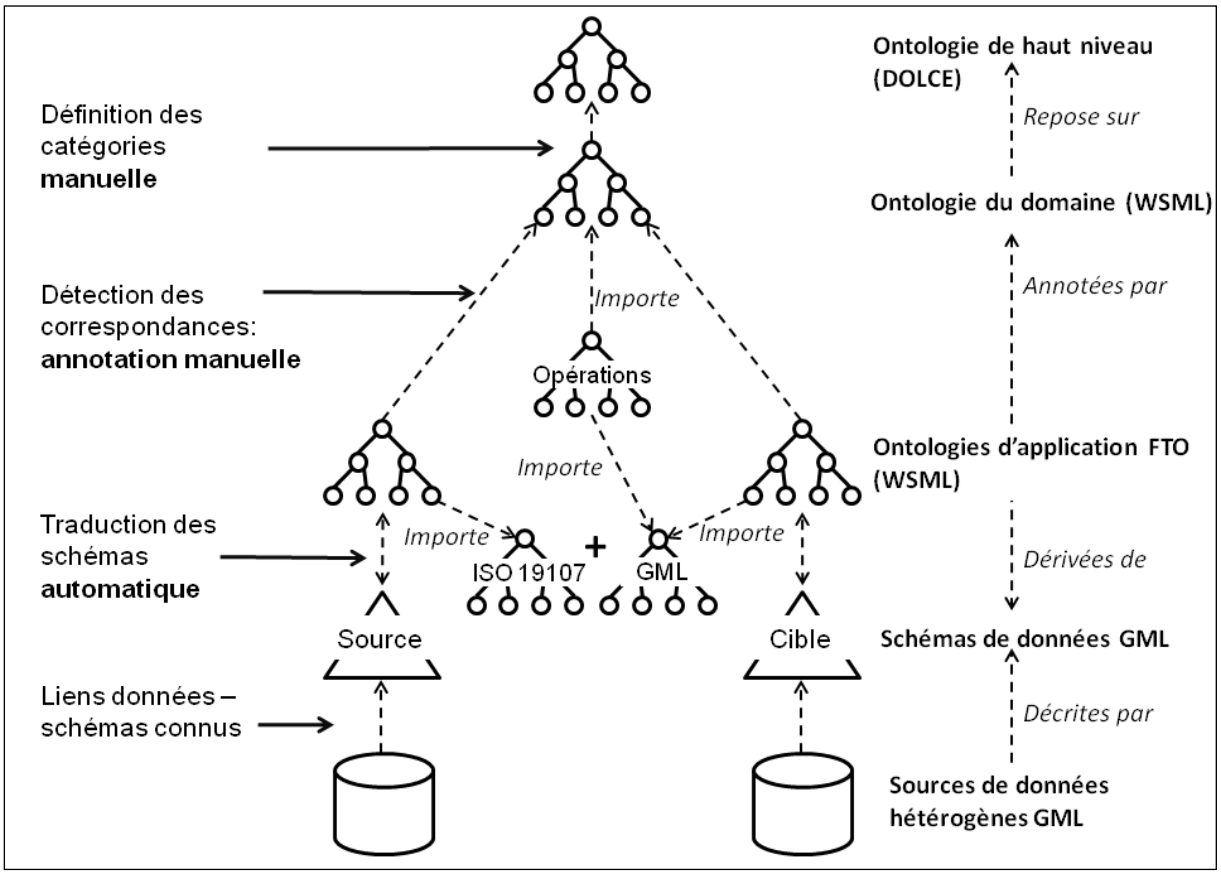


Figure 18: Approche proposée par Schade (2010): identification des étapes d'annotation automatiques et manuelles

L'annotation sémantique des schémas source et cible est réalisée conformément aux travaux de (Klien, 2008), seul le nom de la relation d'annotation est modifié pour plus de clarté: au lieu d'une relation nommée « annotate », l'auteur utilise deux relations distinctes « domain reference » pour les classes et attributs des schémas, et « measure reference » pour les valeurs pouvant être prises par les attributs des schémas. Les opérations de transformation disponibles sont décrites au sein d'une ontologie d'application particulière nommée ontologie des opérations. Les caractéristiques de chaque opération disponible y sont décrites sous la forme d'axiomes. Dans le cas de ces travaux, cette approche est testée sur les schémas de données dédiés aux réseaux routiers ATKIS[31], le référentiel de données géographiques allemand, et INSPIRE. C'est pourquoi seul le cas des calculs de longueur pour les objets linéaires est traité. Les types de données en entrée sont explicités en pointant vers l'ontologie GML, les conditions que doivent vérifier les entités du monde réel pour que cette opération soit applicable sont précisées à l'aide d'un axiome réutilisant des concepts du cadre

[31] http://www.atkis.de/

de référence et les données en sortie sont annotées via le cadre de référence. Ce processus est présenté en figure 19.

Une fois les schémas source et cible et les opérations disponibles annotés, il devient possible à l'aide d'un *système de raisonnement* de déterminer, non seulement les relations de correspondance entre schémas, mais également les opérations pouvant être appliquées sur des éléments du schéma source afin de dériver des données en entrée les éléments attendus par le schéma cible. Ces résultats sont stockés au sein d'un script de traduction qui pourra être utilisé afin de lancer la traduction des données du schéma source vers le schéma cible à l'aide d'un outil adapté.

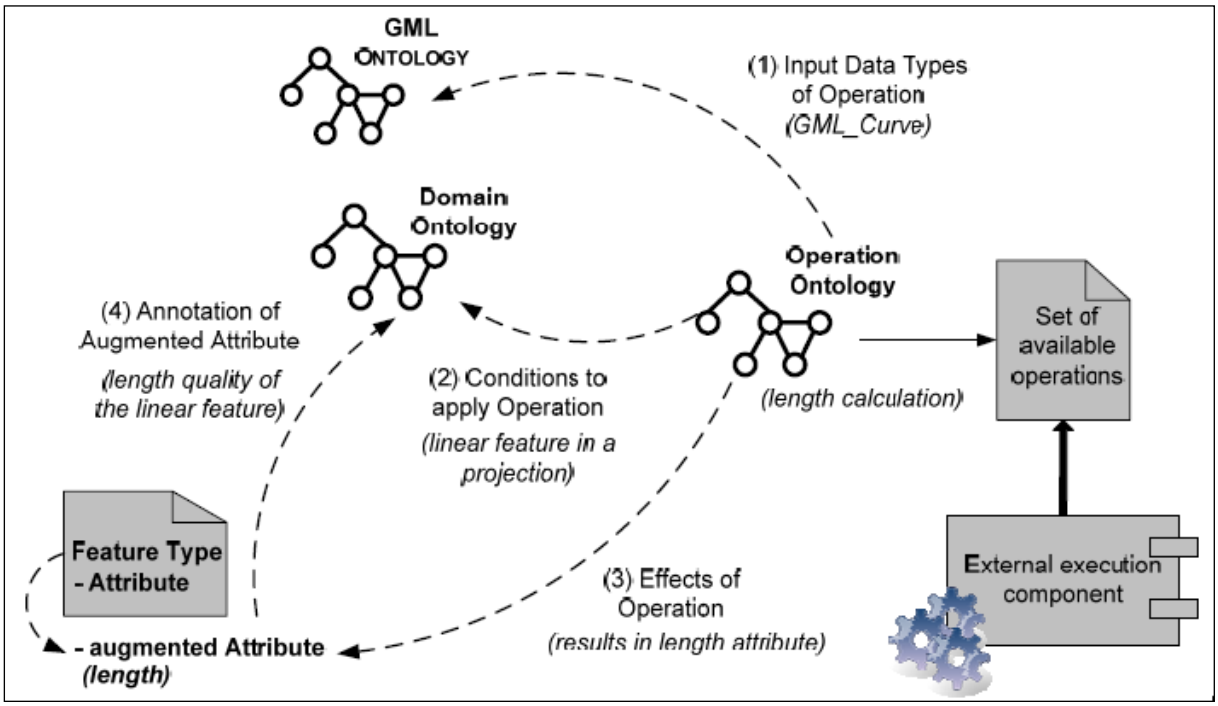


Figure 19: Description des opérations de transformation de données disponibles via l'ontologie des opérations extrait de (Schade, 2010)

Conclusion

L'appariement de schémas conceptuels de données représente une étape essentielle dans un processus d'intégration de schémas de bases de données géographiques. Une des principales difficultés de cet appariement réside dans la détection et la définition de relations de correspondance entre schémas suffisamment détaillées pour permettre la mise en œuvre automatique des diverses opérations de transformation, d'appariement de données, ou encore de généralisation de données qui suivent cet appariement dans le processus d'intégration.

Ainsi, il ne s'agit pas seulement de détecter les classes, attributs et valeurs d'attributs des différents schémas à intégrer qui se correspondent, mais également de mettre en évidence les diverses restrictions qui s'appliquent à ces relations de correspondance, et qui résultent du processus de catégorisation et de saisie des données géographiques. C'est pourquoi, en dépit de l'existence d'un

grand nombre d'approches automatiques pour l'appariement de schémas, ce processus est encore largement réalisé de façon manuelle dans le cas des bases de données géographiques.

Des approches semi-automatiques (Cruz et al., 2007) ou automatiques (Volz, 2005, Schade, 2010) ont néanmoins été proposées, mais les relations de correspondance détectées ne couvrent pas l'ensemble des hétérogénéités existant entre bases de données géographiques. L'approche fondée sur des ontologies proposée par (Schade, 2010) qui permet la représentation et l'exploitation de connaissances à la fois sur les éléments de schémas de données, les concepts géographiques qu'ils représentent et sur les opérations applicables aux données, permet non pas la détection d'hétérogénéités entre bases de données mais celle de relations entre éléments de schémas moyennant l'application d'opérations de transformation sur les données. Elle est donc bien adaptée à un processus de transformation de schémas mais pas nécessairement à d'autres applications nécessitant l'intégration de données géographiques. De plus, sa mise en œuvre effective suppose de compléter l'ontologie des opérations de transformation, qui à ce jour se cantonne au calcul de longueur d'objets linéaires. Enfin, comme chez (Klien, 2008) la description de l'ontologie du domaine utilisée demeure restreinte au cas d'application présenté.

3.2 Ontologies du domaine de la topographie: existant et approches de création

Les travaux présentés dans la partie précédente nécessitent de résoudre les conflits d'intégration inhérents à l'hétérogénéité sémantique des bases de données géographiques, qu'il s'agisse de permettre la découverte et l'accès à des bases de données hétérogènes et distribuées ou d'effectuer des transformations de schémas de données. Pour ce faire, les approches proposées s'inspirent des travaux réalisés dans le domaine de l'intégration d'informations qui préconisent le recours à des ontologies du domaine pour décrire la sémantique des sources de données à intégrer. Un préalable indispensable à la mise en œuvre de ces approches consiste donc à se doter d'une telle ontologie, qu'il s'agisse d'une ressource déjà existante, ou qu'elle soit construite par les auteurs pour les besoins de leur application.

3.2.1 Ontologies du domaine de la topographie : bilan de l'existant

Une approche largement encouragée dans la communauté de l'intégration d'informations pour se doter d'une ontologie du domaine consiste à réutiliser une ontologie existante (Suárez-Figueroa et Gómez-Pérez, 2009). La création d'une ontologie dédiée à un domaine constitue en effet une tâche longue et complexe. C'est pourquoi, il est préférable, lorsqu'une application nécessite ce type de ressource, d'établir au préalable un recensement des ontologies existantes. Outre un gain de temps et de ressources, la réutilisation d'ontologies existantes favorise l'interopérabilité : annoter de nombreuses sources de données via la même ontologie permet en effet leur alignement et contribue à la diffusion de bonnes pratiques si l'ontologie choisie est de bonne qualité.

Accéder aux ontologies existantes sur un domaine pour évaluer leur pertinence vis-à-vis d'une application peut se faire de différentes manières (d'Aquin et Lewen, 2009). En effet, il n'existe pas de dépôt centralisant l'ensemble des ontologies existantes. En revanche, de nombreux dépôts

regroupent des ontologies et peuvent être consultés à ces fins : TONES Ontology Repository, le wiki Ontology Design Patterns, LIRMM Ontologies publishing platform, SWEET Ontologies, etc. De plus, des moteurs de recherche pour ontologies, comme SWOOGLE ou Watson peuvent être mis à contribution.

En outre, il convient ensuite de déterminer si l'une des ontologies identifiées peut convenir à l'application d'intégration visée. Ce choix d'une ontologie s'opère selon divers critères : couverture du domaine, langue(s) utilisée(s) pour décrire les concepts, leurs propriétés et leurs relations, langage de description, niveau de complexité (ontologies légères dites « Lightweight », qui sont généralement des taxonomies, ou ontologies lourdes dites « Heavyweight » qui proposent une modélisation plus fine, incluant des restrictions sur la sémantique des concepts du domaine), profondeur, cohérence, documentation, modularité, etc. Il peut arriver que seule une partie d'une ontologie corresponde aux besoins de l'application. Dans ce cas, il faudra que cette ontologie soit suffisamment modulaire pour permettre l'utilisation de cette partie indépendamment du reste. A l'inverse, plusieurs ontologies peuvent s'avérer complémentaires pour assurer la couverture de l'ensemble du domaine. Il faudra dans ce cas envisager leur utilisation conjointe, et donc leur alignement, voire leur fusion.

Dans le cadre du projet GéOnto (ANR-O7-MDCO-005), dans lequel s'inscrit cette thèse, un état de l'art des ontologies du domaine de la topographie a été réalisé (Minard, 2008). Celui-ci a permis de recenser une cinquantaine d'ontologies traitant de concepts topographiques (Towntology, Fodomust, GIEA, WalkOnWeb, etc.). Celles-ci ont été examinées selon divers critères : type (ontologie du domaine, ontologie de tâche, ontologie de haut niveau, ou thesaurus), domaine couvert, niveau de complexité, profondeur, langage d'implémentation et langue employée pour décrire les concepts et leurs propriétés.

Il ressort de cette étude que la majorité des ontologies découvertes sont des ontologies du domaine, mais ne couvrant que partiellement le domaine de la topographie ; certaines portent sur l'hydrographie, d'autres sur les toponymes, ou encore l'occupation du sol. Par ailleurs, il s'agit pour la plupart d'ontologies légères, de profondeurs très variables. Pour la majorité, ces ontologies sont décrites en OWL, le langage pour ontologies recommandé par le W3C. Enfin, la langue la plus répandue est l'anglais, seules quatre d'entre elles, dont un thésaurus, employant le français, et six, dont trois thesauri, étant multilingues, mais pas nécessairement francophones. Ainsi, très peu de ressources ontologiques, couvrant même très partiellement le domaine de la topographie, sont disponibles en langue française. Or, traduire des ontologies en langue anglaise vers le français demeure délicat, voire inapproprié et celles analysées lors de cette étude ne présentent ni une couverture globale du domaine, ni un niveau de détail très fin dans la description du domaine. Si la réutilisation d'ontologies existantes doit être privilégiée, le bilan de l'existant en matière d'ontologies francophones du domaine de la topographie plaide plutôt en faveur de la création d'une nouvelle ontologie. En effet, réutiliser les ontologies existantes nécessiterait un travail important en termes de traduction de labels pour les ontologies non francophones, d'alignement de ces ontologies, et éventuellement d'ajout de concepts afin d'affiner la description des thèmes existants ou de compléter les thèmes du domaine non couverts par les ontologies disponibles. En comparaison, l'effort à fournir pour la création d'une ressource francophone, adaptée à notre objectif d'intégration de données topographiques et offrant une couverture aussi complète que possible du domaine semble raisonnable.

3.2.2 Approches proposées pour la création d'ontologies du domaine de la topographie

(Gómez-Pérez et al., 2003) définissent l'ingénierie ontologique comme l'ensemble des activités concernant le processus de développement d'ontologies, le cycle de vie des ontologies, et les méthodes, outils et langages dédiés à la création d'ontologies. Les travaux réalisés dans ce domaine au cours des deux dernières décennies ont conduit à l'élaboration de nombreuses méthodologies de création d'ontologies du domaine. Dans la communauté de l'ingénierie des connaissances et dans une moindre mesure dans celle de l'information géographique, les différentes méthodologies proposées semblent s'accorder sur les aspects principaux du processus de développement d'ontologies à adopter. Pour autant, à ce jour, aucune ne semble emporter l'adhésion générale.

3.2.2.1 Approches proposées dans le domaine de l'ingénierie des connaissances

En décrivant précisément le cycle de vie d'une ontologie, les méthodologies de création d'ontologies guident experts du domaine et spécialistes en ingénierie des connaissances tout au long du processus de création d'une ontologie du domaine, en décomposant ce dernier sous la forme d'une succession d'activités, elles-mêmes composées de tâches élémentaires, à réaliser pour parvenir à un résultat opérationnel.

(Gómez-Pérez et al., 2003) présentent le processus global de développement d'ontologies défini dans le cadre de la méthodologie de création d'ontologies METHONTOLOGY (Gómez-Pérez et al., 1996). Celui-ci se compose de trois types d'activités. Les activités de **gestion de l'ontologie**, tout d'abord, concernent l'organisation de l'ensemble du processus de création de l'ontologie, c'est-à-dire la définition des tâches devant être réalisées, ainsi que celle de leur chronologie et des délais et ressources attribués à chacune. Elles s'attachent également à la vérification du bon déroulement de chaque tâche devant être réalisée, ainsi qu'au contrôle de la qualité de l'ontologie produite. Les activités de **développement de l'ontologie** se rapportent à l'ensemble des tâches visant directement la création de l'ontologie, ainsi qu'à celles situées en amont et en aval de cette étape. Les activités de pré-développement comprennent ainsi une étude de l'environnement dans lequel l'ontologie sera utilisée, qu'il s'agisse de l'application visée ou de la plateforme utilisée. Une étude de faisabilité est également préconisée afin de déterminer si le recours à une ontologie est bien la bonne approche pour l'application visée et si la création de cette ontologie constitue une tâche réalisable. La phase de développement proprement dite débute par une activité de spécification visant à définir les objectifs de l'ontologie. Cette dernière est suivie d'une phase de conceptualisation destinée à définir et structurer les connaissances du domaine. Le schéma conceptuel ainsi défini est ensuite transformé en modèle formel, avant d'être implémenté dans un langage de représentation de connaissances. Les activités de post-développement englobent, quant à elles, la maintenance – mise à jour et amélioration – et la réutilisation de l'ontologie par d'autres applications. Le dernier type d'activités, dites de **support**, regroupe l'ensemble des activités indispensables à la réussite des activités de développement de l'ontologie. C'est le cas de l'activité d'acquisition de connaissances qui consiste à extraire des connaissances sur le domaine couvert par l'ontologie, soit en s'adressant directement à des experts du domaine, soit à l'aide de techniques semi-automatiques. De nombreuses techniques ont été mises au point afin de permettre l'extraction des connaissances nécessaires à la création

d'une ontologie du domaine à partir de divers types de ressources, qu'il s'agisse de corpus textuels (Aussenac-Gilles et al., 2008), de bases de données (Baglioni et al., 2007), de taxonomies, ou encore de folksonomies (Hotho et Jäschke, 2010). Des approches collaboratives, comme les VoCamps qui rassemblent les acteurs d'une communauté donnée lors de sessions de travail destinées à faire émerger un vocabulaire partagé au sein de la communauté, ont également été mises en œuvre. Des guides méthodologiques, tels celui proposé par (Noy et McGuinness, 2001), prodiguent un ensemble de conseils et de bonnes pratiques en matière d'acquisition et de structuration de connaissances pour la création d'ontologies. L'usage de patrons de conception, largement répandu en génie logiciel, a été étendu au domaine de la création d'ontologies du domaine pour résoudre des problèmes de modélisation récurrents (Gangemi et Presutti, 2009), pouvant survenir lors de l'étape de structuration des connaissances. Les patrons de conception d'ontologies (Ontology Design Patterns, ODPs) sont des modèles de schémas apportant des solutions types à des questions de modélisation bien identifiées. Ils consistent en un ensemble d'éléments d'ontologie, de métadonnées sur leurs cas d'utilisations, leur provenance, leurs avantages et inconvénients, etc. L'activité d'évaluation s'attache à estimer, d'un point de vue technique, la qualité de l'ontologie créée, ainsi que celles des logiciels et documentations qui l'accompagnent. Les approches proposées, en matière d'évaluation d'ontologies, s'intéressent principalement à l'identification et la formalisation des erreurs de modélisation les plus courantes et proposent des actions correctives pour remédier à ces imperfections. Ainsi, la méthodologie OntoClean (Guarino et Welty, 2002) définit un ensemble de propriétés génériques portant sur les classes, propriétés et relations constituant une ontologie, et dont la vérification permet l'évaluation, et éventuellement la correction, des choix de modélisation effectués. En se fondant sur cette méthodologie d'évaluation, ainsi que sur l'analyse d'ontologies existantes, Poveda-Villalón et al. (2010) dressent une liste de vingt-quatre erreurs de modélisation courantes, illustrées par des exemples typiques. Cette activité d'évaluation présente un double intérêt. Elle permet, d'une part, lors de la création d'une ontologie, de s'assurer de la qualité du travail effectué et éventuellement de remédier aux diverses imperfections détectées. D'autre part, dans un contexte de réutilisation d'ontologies, elle peut contribuer à faciliter le choix d'une ontologie pour une application donnée. C'est le cas, par exemple, de l'application d'évaluation d'ontologies proposée par (Duque-Ramos et al., 2010), qui s'inspire des standards d'évaluation de la qualité d'applications logicielles, pour définir un ensemble de métriques permettant de comparer plusieurs ontologies selon divers critères – structure, adéquation aux objectifs, fiabilité, maintenance, etc. L'activité d'intégration désigne la réutilisation d'ontologies existantes pour la création de l'ontologie du domaine. Cette démarche de réutilisation se retrouve dans les activités d'alignement et de fusion. La première désigne l'établissement de relations de correspondance explicites entre les éléments constitutifs de plusieurs ontologies portant, au moins en partie sur le domaine d'intérêt, et fait l'objet de très nombreux travaux (Euzenat et Schvaiko, 2007). La seconde consiste à unifier les concepts, les vocabulaires et les définitions issus de plusieurs ontologies portant sur le domaine d'intérêt afin de produire une ontologie harmonisée. L'activité de documentation décrit chaque étape du processus de création de l'ontologie et est complétée par une activité de gestion des versions successives des différents documents produits ainsi que de celles de l'ontologie créée.

Ce processus global de développement d'ontologie se retrouve en grande partie dans les différentes méthodologies recensées par (Gómez-Pérez et al., 2003) : Cyc, Uschold et King, Grüninger et Fox, KACTUS, METHONTOLOGY, SENSUS et On-To-Knowledge. (Gómez-Pérez et al., 2003) proposent des tableaux comparatifs synthétisant les principales caractéristiques de chacune des méthodologies

présentées, sans se prononcer quant au meilleur choix possible parmi les différentes méthodologies. Les principales différences d'une méthodologie à l'autre résident dans le cycle de vie de l'ontologie proposé, le degré d'indépendance de l'ontologie créée vis-à-vis de son application d'origine, ou encore dans le nombre et la nature des activités préconisées et le niveau de détail de description de ces activités. Des états de l'art plus récents (Simperl et al., 2010)(Sure et al., 2009) confirment cette absence de consensus au sein de la communauté de l'ingénierie des connaissances concernant la méthodologie à adopter.

3.2.2.2 Approches proposées dans le domaine de l'information géographique

L'intérêt de la communauté de l'information géographique pour les ontologies s'est manifesté de façon plus tardive (Frank, 1997)(Smith et Mark, 1998)(Rodriguez et al., 1999). Cet essor s'est accompagné d'un ensemble d'interrogations sur les différents aspects que recouvre la notion d'ontologie géographique, la place des ontologies dans l'interopérabilité des données et des applications géographiques, ainsi que la façon la plus appropriée d'envisager leur création en vue de leur mise en œuvre (Winter, 2001). En effet, différentes visions de la notion d'ontologie se côtoient au sein de la communauté de l'information géographique (Winter, 2001). Kavouras et Kokla (2008) relèvent cette coexistence d'approches philosophiques et informatiques dans les ontologies géographiques. Les premières s'attachent à définir les concepts, processus, et relations géographiques fondamentaux, tandis que les secondes se concentrent sur la description et la formalisation des catégories d'entités géographiques et de leurs relations dans un but d'intégration d'information. Agarwal (2005) souligne l'ironie de cette ambivalence des ontologies géographiques, pourtant destinées à fournir une représentation partagée et non ambigüe des concepts du domaine, et s'interroge sur la probabilité de parvenir à une vision unifiée de la notion d'ontologie géographique dans le contexte fondamentalement pluridisciplinaire qu'est l'information géographique.

Une solution couramment proposée pour parvenir à une conceptualisation partagée du domaine consiste à recourir à une ontologie de haut niveau afin de disposer d'une définition commune des concepts les plus généraux du domaine, comme l'espace, le temps, les qualités, les relations, les événements, etc. Ces concepts possèdent, en effet, la particularité de demeurer communs aux diverses ontologies du domaine de la géographie pouvant être développées, quel que soit leur contexte applicatif. Ils constituent donc un socle de connaissances partagées sur lequel ces différentes ontologies peuvent faire reposer leurs conceptualisations du domaine respectives, et constituent un point d'ancrage commun des connaissances qu'elles modélisent, favorisant ainsi leur alignement (Guarino, 1998, Mark et al., 2004). Kuhn (2003) reprend et étend cette proposition en introduisant la notion de *système de référence sémantique*. Celle-ci est développée par analogie avec les systèmes de références spatiaux, qui formalisent la définition des coordonnées géographiques de façon à permettre de localiser sans ambiguïté un lieu donné à la surface de la Terre. Les données géographiques présentent généralement trois composantes principales: spatiale, temporelle, et sémantique. Les deux premières s'appuient sur des systèmes de référence standardisés (ISO 19107 et 19108) permettant d'effectuer des transformations et des projections d'un système à l'autre. La troisième, en revanche, repose sur le seul schéma conceptuel de la base de données, qui à lui seul ne permet pas d'interpréter sans ambiguïté la signification des données. Partant de ce constat, Kuhn (2003) prône la mise en place d'un système de référence sémantique permettant d'expliciter la

signification exacte des données géographiques, de traduire les données dans les termes d'une autre communauté, et d'intégrer des données sémantiquement hétérogènes. Filant la métaphore des systèmes de référence spatiaux, il introduit les notions de *datum sémantique* et de *cadre de référence sémantique* comme parties constituantes du système de référence sémantique ainsi que les processus associés à un tel système: *référencement*, *projection* et *transformation*. Kuhn et Raubal (2003) détaillent ces notions et proposent une implémentation de système de référence sémantique. Ainsi un *datum sémantique* correspond à un mécanisme permettant d'ancrer la sémantique des termes du système de référence sémantique hors du système lui-même. Il s'agit, en particulier, de définir les primitives du système de référence sémantique. Ceci peut être réalisé en utilisant les concepts d'une ontologie extérieure au système ou bien en s'appuyant sur la sémantique propre du langage utilisé pour implémenter le système de référence sémantique. Un *cadre de référence sémantique* consiste en une structure conceptuelle indépendante de toute application et définie de façon formelle. Aussi cette notion est-elle assimilée ici à une ontologie de haut niveau (Guarino, 1998). Pour Schade (2010), un cadre de référence sémantique peut également comporter une ontologie du domaine dans la mesure où celle-ci repose sur une ontologie de haut niveau. Le cadre de référence sémantique, qui permet le référencement sémantique de termes issus de divers domaines, constitue le cœur du système de référence sémantique. Cette opération consiste, pour un jeu de données, à annoter son schéma conceptuel à l'aide des primitives fournies par le cadre de référence sémantique. L'interprétation des termes utilisés dans le schéma conceptuel du jeu de données est ainsi décrite sans ambiguïté à l'aide des concepts définis dans le cadre de référence. Cette étape de référencement sémantique permet d'accéder aux deux fonctionnalités restantes: projeter un schéma conceptuel de données vers un schéma simplifié ou encore transformer des données fournies dans un schéma particulier vers un autre schéma, éventuellement défini à l'aide d'un autre système de référence sémantique. Ainsi, un système de référence sémantique vise à favoriser l'interopérabilité en fournissant des méthodes pour expliquer les interprétations de vocabulaires et traduire les expressions utilisées d'une communauté à l'autre (Kuhn 2003, Kuhn 2005). Il est constitué d'ontologies dont les classes, relations et axiomes servent de structure conceptuelle formelle et indépendante de toute application sur laquelle appuyer la description des termes employés au sein d'un schéma conceptuel de données.

Ainsi, dans le domaine des ontologies géographiques, une attention particulière a été portée à la définition des concepts de haut niveau. Casati et al. (1998) identifient quatre types de questions ontologiques propres aux entités géographiques. En premier lieu viennent les questions relatives à la conceptualisation de l'espace géographique. Les auteurs s'interrogent sur l'opportunité de définir des primitives pour les concepts de « région de l'espace » et d' « entité géographique », ainsi que pour les relations existant entre ces derniers. Cette question est étroitement liée à la conceptualisation de l'espace adoptée. Celui-ci peut en effet être considéré comme un conteneur existant indépendamment des entités qu'il renferme. Dans cette perspective, l'espace est exclusivement constitué d'entités géographiques et sa conceptualisation nécessite une fonction de localisation permettant d'attribuer à chaque entité sa position au sein de l'espace. Vieu (1997) parle alors d'*espace absolu*. A l'opposé, la notion d'*espace relatif* (Vieu, 1997) considère l'espace géographique comme constitué d'un ensemble d'entités non nécessairement géographiques associées les unes aux autres par un ensemble de relations spatiales. Une question complémentaire posée par Vieu (1997) concerne le choix de la définition de la localisation des entités géographiques. Dans un *espace global*, une localisation définie dans un système de référence est attribuée à chaque

entité géographique de sorte que sa position relative par rapport aux autres entités est connue. Dans un *espace local*, en revanche, chaque entité géographique est localisée via l'explicitation de ses relations spatiales par rapport aux autres entités. Un troisième point abordé par Casati et al. (1998) concerne le type des entités géographiques considérées, qui conditionne leur relation à l'espace. Ils distinguent les entités matérielles qui occupent l'espace où elles sont situées de façon exclusive vis-à-vis des autres entités matérielles, des entités immatérielles, comme les phénomènes géographiques qui peuvent partager leur localisation avec d'autres entités. Enfin, la dernière question traitée par les auteurs porte sur les difficultés particulières relatives à la définition des catégories d'entités géographiques en comparaison de celles posées par la définition des catégories d'entités que l'on côtoie quotidiennement comme les animaux ou les objets manufacturés. Les entités géographiques, et par extension les catégories d'entités géographiques auxquelles elles appartiennent, sont le plus souvent définies via leurs limites qui décrivent à la fois leur localisation et leur forme. Cette approche visant à décrire les entités géographiques via leurs contours – ou autres éléments caractéristiques saillants – se retrouve dans le modèle traditionnel de données géographiques dit vectoriel, où chaque entité géographique est perçue comme un *objet* bien identifié et représentée par une primitive géométrique – point, ligne ou polygone – destinée à décrire sa localisation et sa forme. Or, les frontières délimitant les entités géographiques n'étant pas nécessairement bien déterminées, la distinction entre entités géographiques adjacentes ou bien entre catégories d'entités géographiques peut s'avérer délicate. Les auteurs soulignent donc la nécessité de discerner les catégories d'entités dont les limites relèvent de discontinuités visibles de l'espace - *bona fide entities* - de celles dont les frontières sont issues d'un processus de délimitation d'origine humaine - *fiat entities* - (cf. 2.2.1). De plus, certaines catégories d'entités géographiques, comme les montagnes ou les vallées, ne possèdent pas de limites aisément distinguables et localisables. Peuquet et al. (1998) s'intéressent tout particulièrement à ces catégories d'entités géographiques qui résultent des variations des valeurs d'une mesure donnée dans l'espace – on identifie les montagnes par leur altitude élevée, et s'interrogent sur leur représentation au sein d'une ontologie géographique. En effet, ces entités sont traditionnellement traitées, dans le domaine des sciences géographiques, comme des entités de type *champ*, et représentées sous la forme de données raster – matrice à pas constant dont les valeurs correspondent aux valeurs d'une mesure donnée pour chaque région de l'espace concernée. Cependant, dans la mesure où elles occupent une place importante dans notre conceptualisation courante de l'espace topographique, il est essentiel de pouvoir les représenter également selon les formalismes orientés objets adoptés pour la représentation de connaissances. Faute de seuils universellement admis permettant de déterminer des limites linéaires strictes pour ces catégories d'entités géographiques, Peuquet et al. (1998) introduisent donc la notion de *limites graduelles*, qui permettent de délimiter les entités géographiques de type champ à l'aide de régions surfaciques de l'espace. Ceci permet d'envisager la représentation de leur localisation et de leur forme selon une approche orientée objet, telle celle pratiquée pour les données géographiques vectorielles. Cependant, les auteurs s'interrogent sur la pertinence d'une représentation des entités de type champ selon un formalisme orienté objet et identifient un ensemble de questions relatives à ce sujet: développement de formalismes pour ontologies dédiés aux entités de type champ, établissement d'une typologie des entités de type champ, opérations de raisonnement adaptées aux entités de type champ, topologie spécifique aux entités de type champ, représentation et raisonnement conjoints avec les formalismes orientés objets classiques, etc.

Grenon et Smith (2004) proposent une double ontologie cadre destinée à permettre la représentation conjointe d'entités dotées de modes d'existence différents au cours du temps. En cela, ils s'appuient sur la distinction établie au sein de l'ontologie de haut niveau Basic Formal Ontology (BFO[32]). Dans cette ontologie, les *continuants* encore appelés *endurants*, d'une part, désignent des entités dont l'existence est continue et qui conservent leur identité tout au long de cette existence, en dépit des évolutions qui peuvent survenir. C'est le cas, par exemple, des humains, ou des villes. Les *occurents* ou *perdurants*, d'autre part, font référence aux événements, aux processus, aux changements et aux activités, comme les phénomènes météorologiques, par exemple. Si ces différents types d'entités existent de différentes façons dans le temps, elles n'en sont pas moins interdépendantes les unes des autres. En effet, les *continuants* subissent des changements et constituent le support des *occurents*. Les auteurs proposent donc deux ontologies, nommées *SNAP* et *SPAN* respectivement destinées à la représentation des *continuants* et des *occurents*, interconnectées par un ensemble de relations permettant une représentation conjointe cohérente de ces deux types d'entités et de leurs propriétés. La première propose une conceptualisation de la réalité sous la forme d'instantanés, tandis que la seconde conçoit la réalité sous la forme d'un ensemble de processus se déroulant dans le temps. Les auteurs spécialisent cette double conceptualisation en l'appliquant au cas des concepts spatio-temporels. Il en résulte trois ontologies, regroupées sous le nom de *GeO*, et reposant sur les notions de *région de l'espace,* de *région du temps* et de *région de l'espace-temps*, respectivement vues comme des portions de l'espace, du temps, et de l'espace-temps. La première ontologie définit les objets géographiques de type *SNAP*, comme les *entités géographiques* – i.e. les montagnes, les lacs, les rivières, les États, etc. les *artéfacts géographiques* – i.e. les bâtiments, les routes, les villes, etc., les *agents géographiques* – i.e. les instances agissant sur une portion du territoire, *les lieux géographiques* – i.e. les portions du territoire occupées par des objets géographiques et éventuellement désignées par un toponyme, les *frontières* – de type *fiat* ou *bona fide*, et les *qualités géographiques* – i.e. l'altitude d'une montagne, la superficie d'une ville, etc. La deuxième distingue deux principaux types d'objets géographiques *SPAN* selon les entités géographiques qu'ils impliquent. Ainsi, les *processus physiques*, comme les inondations ou les feux de forêts, concernent uniquement des objets physiques, tandis que les *processus sociaux*, comme les guerres et les épidémies, se rapportent à des groupes humains. A ces concepts s'ajoutent les *actions*, qui résultent d'une intention, et les *événements*, qui délimitent les processus. Les auteurs insistent également sur la nécessité de pouvoir représenter les changements qui peuvent intervenir au cours du déroulement d'un processus, qu'il s'agisse de changements concernant l'intensité du processus observé ou de changements concernant l'emprise géographique de ce dernier. La troisième ontologie, enfin, concerne les objets géographiques *SNAP* de type *champ*. Elle comporte les concepts de *champ* et d'*attribut de champ*, ainsi que les relations permettant d'associer un *attribut de champ* à la localisation de cet objet *champ*.

La conceptualisation que l'on peut avoir de l'espace géographique s'appuie donc sur différentes caractéristiques de ce dernier et des entités géographiques qui le composent. Or, selon le point de vue adopté, les caractéristiques prises en compte peuvent varier, et avec elles, notre vision de l'espace géographique. Partant de constat, Fonseca et al. (2002) proposent un modèle composé de cinq niveaux d'abstraction, chacun de ces niveaux appréhendant l'espace géographique selon un point de vue particulier. Dans ce paradigme, l'*univers physique* désigne l'espace géographique,

[32] http://www.ifomis.org/bfo

considéré indépendamment de la perception que nous en avons. L'*univers cognitif* est le siège du processus de conceptualisation de l'espace géographique. Ce dernier consiste à classer les entités géographiques en différentes catégories définies par consensus au sein d'une communauté donnée. Le lien entre l'univers physique et l'univers cognitif réside dans le consensus établissant les critères de classification des entités géographiques du monde réel au sein des catégories géographiques structurant l'univers cognitif. La définition et l'organisation de ces catégories -ou concepts- suit un cadre logique. L'explicitation formelle, à l'aide de méthodes logiques, de la définition et de l'organisation de ces catégories produit des ontologies, qui constituent l'*univers logique*. La représentation des éléments de l'univers logique conformément à divers modèles conceptuels – en particulier les modèle *objets* et *champs* - permettant de leur appliquer diverses opérations s'effectue dans l'*univers de la représentation*. C'est également au sein de cet univers qu'interviennent les considérations relatives à la mesure et à la saisie des entités géographiques du monde réel. Le lien entre les niveaux *logique* et *représentation* est assurée par des médiateurs. Enfin, les ontologies produites aux niveaux *logique* et *représentation* sont implémentées dans l'*univers des implémentations*, composé des structures de données de base – vecteur et raster - et des algorithmes permettant de les manipuler. Pour parvenir à une description de l'espace géographique rendant compte des divers aspects des entités géographiques, Frank (2001, 2005) propose de structurer celle-ci au sein d'une ontologie en cinq niveaux de description reliés les uns aux autres. L'objectif premier de cette approche est de permettre l'intégration de différentes représentations de l'espace au sein d'un système unifié, et en particulier d'apporter une solution à la question de la représentation et de la gestion conjointes d'entités de type *champ* et d'entités de type *objet* soulevée par (Peuquet et al., 1998). Le premier niveau est dédié à la réalité physique perçue comme un espace à quatre dimensions régi par les lois de la physique. Le deuxième niveau représente la réalité telle que nous l'observons en un instant donné. Ceci implique la prise en compte des limitations, des incomplétudes et des imprécisions induites par les moyens d'observation et de mesure mis en œuvre. Le troisième niveau est consacré à la définition des objets et des catégories d'objets à l'aide de règles s'appuyant sur leurs propriétés caractéristiques. Le quatrième niveau s'attache à la représentation de la réalité sociale, c'est-à-dire à la définition de la sémantique des catégories d'objets dans un contexte donné. Le cinquième niveau décrit la réalité telle qu'elle est perçue, de façon incomplète et subjective, par différents agents cognitifs qui utilisent ces connaissances pour raisonner et prendre des décisions.

Parallèlement à ces approches visant à établir un cadre conceptuel de haut niveau pour la définition des concepts géographiques, l'Ordnance Survey, agence cartographique du Royaume-Uni, a publié une série de documents méthodologiques sur la construction d'ontologies du domaine. Ces documents s'appuient directement sur l'expérience acquise par les membres de l'équipe « GeoSemantics » lors de la création des ontologies publiées par cet organisme sur les thèmes de l'hydrologie, des zones bâties, et des limites administratives. L'approche proposée s'inspire largement des méthodologies de création d'ontologies élaborées dans le domaine de l'ingénierie des connaissances, en particulier celles proposées par (Uschold et King, 1995) et (Gómez-Pérez et al., 1996). Ces documents s'adressent à des experts du domaine désireux de créer une ontologie et proposent un guide méthodologique complet allant de la création d'ontologies du domaine informelles (Kovacs et al., 2006), à l'implémentation des ontologies ainsi définies dans le langage OWL(Goodwin, 2007), en passant par des recommandations de bonnes pratiques de modélisation d'ontologies du domaine (Hart et Goodwin, 2007).

La première étape de la méthodologie proposée consiste donc en la définition d'une ontologie conceptuelle, décrite en langage naturel. Kovacs et al. (2006) proposent un ensemble de tâches et de recommandations systématiques, illustrées par des exemples tirés de leurs propres travaux de création d'ontologies, visant à guider l'expert du domaine dans son travail de conceptualisation du domaine et d'organisation des connaissances nécessaires à sa représentation. Cinq tâches successives sont identifiées par les auteurs. Tout d'abord, il s'agit de définir les objectifs motivant la création de l'ontologie, ainsi que son domaine de couverture et son niveau de détail. Les auteurs insistent particulièrement sur ces deux derniers points; un domaine de couverture et un niveau de détail précisément définis constituent un repère précieux pour les experts du domaine lors de la définition des concepts devant être représentés dans l'ontologie. Ils permettent en effet de distinguer les concepts principaux du domaine, indispensables à une description complète du domaine, des concepts secondaires, ne concernant pas directement le domaine d'intérêt, mais nécessaires à la bonne représentation des connaissances utiles à sa description, et d'éliminer les concepts non pertinents. Puis il convient de rassembler des sources de connaissances sur le domaine modélisé. Il peut s'agir de connaissances issues directement de l'expérience des experts du domaine interrogés, ou bien de documents décrivant le domaine d'intérêt. Au sein de ces documents, les experts doivent identifier les concepts principaux et secondaires du domaine, généralement désignés par des noms en langage naturel. Les relations entre concepts sont plus généralement désignées par des verbes, indiquant une appartenance, une filiation, ou encore une action. L'expression de nuances au sein des phrases décrivant le domaine peut être interprété comme des contraintes s'ajoutant à la description des concepts du domaine. Ce travail d'extraction de connaissances concernant le domaine de couverture de l'ontologie, réalisé à partir des connaissances propres d'experts du domaine et de documents décrivant ce dernier, se concrétise par la rédaction d'un glossaire de connaissances. Celui-ci est composé de deux tables, l'une listant les concepts et l'autre les relations identifiés, ainsi qu'un ensemble d'informations utiles à la réalisation de l'ontologie – définitions, références, synonymes, contraintes, etc. L'étape suivante consiste à structurer les connaissances ainsi listées sous la forme de phrases rédigées en langage naturel, en respectant une structure grammaticale précise. Cette structure vise à permettre une formalisation des connaissances relativement proche de celle atteinte grâce à l'usage des langages pour ontologies, tout en conservant une formulation lisible pour des experts du domaine non spécialistes de ces langages dédiés. Il s'agit donc d'élaborer des phrases permettant de définir les concepts principaux de l'ontologie, puis des phrases permettant de définir les concepts principaux via leurs relations avec d'autres concepts principaux tout d'abord, puis avec les concepts secondaires. Les auteurs mettent en garde contre un usage abusif des relations de généralisation-spécialisation lors de cette étape de définition des concepts. En effet, celui-ci induit généralement l'introduction de concepts abstraits non nécessaires à la description du domaine, et potentiellement contraignants lors de la réutilisation de l'ontologie pour d'autres applications. Les phrases à élaborer possèdent une structure simple, du type « sujet-verbe-complément » où le sujet est un concept issu de la table des concepts du glossaire de connaissances, le verbe, une relation issue de la table des relations de ce même glossaire, et le complément, un second concept issu du glossaire. L'usage des termes « *only* » et « *some* » au sein de ces phrases, traductions en langage naturel des quantificateurs universels et existentiels disponibles dans les langages de représentation d'ontologies comme OWL, est préconisé, afin d'introduire des restrictions dans les définitions proposées. Les auteurs limitent l'usage de « *only* » aux seuls cas où la restriction pourra être exprimée entièrement en un nombre fini de phrases. Des phrases plus complexes, incluant des conjonctions de coordination, comme « *et* » ou « *ou* », ou encore des

pronoms relatifs peuvent être formulées. Un recours parcimonieux à des adjectifs ou des adverbes pour nuancer les phrases est également autorisé. Lorsqu'un terme issu de la table des relations est utilisé au sein d'une phrase, les auteurs recommandent de fournir une définition et un exemple afin de clarifier l'usage de ce terme. Les auteurs préconisent également d'identifier, parmi les définitions ainsi formalisées, les concepts définis à l'aide de conditions nécessaires et suffisantes. Enfin, afin de faciliter la réutilisation de l'ontologie, les auteurs recommandent d'identifier des modules, c'est-à-dire des parties d'ontologie pouvant être pertinentes pour la description d'autres domaines, et qui seront implémentées à part de l'ontologie principale et réutilisées par cette dernière. La dernière tâche concerne l'évaluation et la validation de l'ontologie conceptuelle. Les auteurs fournissent un ensemble de critères à vérifier afin de déterminer si l'ontologie couvre bien le domaine d'intérêt et si elle est cohérente.

Une fois l'ontologie conceptuelle définie et validée, elle peut être implémentée dans un langage dédié à la représentation d'ontologies. Les auteurs préconisent ici l'usage du langage OWL, et fournissent un ensemble de règles de traduction de l'ontologie conceptuelle définie en langage naturel vers le langage OWL. Les auteurs ajoutent à ces règles de traduction des conseils relatifs aux bonnes pratiques d'implémentation d'ontologies préconisées dans le domaine des données liées (Vatant et al., 2011).

L'intérêt de la communauté de l'information géographique pour les ontologies s'est donc essentiellement traduit par la définition de concepts de haut niveau propres au domaine. Cette tâche est, en effet, très majoritairement perçue, au sein de la communauté, comme un préalable indispensable à la création d'ontologies du domaine. Ce postulat a donc conduit la communauté à produire une littérature très riche sur les notions d'espace, de temps, de localisation, d'entité géographique, de frontières, etc. et à formaliser ces concepts en vue de la création d'un cadre conceptuel permettant la représentation des connaissances du domaine de la géographie. Cependant, cet effort a été peu suivi d'effets en termes d'implémentation et de publication des ontologies proposées, même si quelques concepts relatifs au domaine, parmi les plus génériques, figurent dans des ontologies de haut niveau telles BFO ou Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE[33]). A l'inverse, peu de méthodologies alternatives pour le développement d'ontologies du domaine ont été élaborées. Cependant, face au besoin croissant de ressources de ce type, quelques initiatives s'inspirant des méthodologies proposées dans le domaine de l'ingénierie des connaissances émergent désormais. C'est le cas de la méthodologie proposée par l'Ordnance Survey, ou encore du GeoVoCamp[34] récemment organisé.

Conclusion

La mise en œuvre des approches d'intégration de données géographiques présentées au chapitre 3.1 nécessite de disposer d'une ontologie du domaine permettant la description de la sémantique des diverses ressources décrivant les sources de données à intégrer. Notre objectif d'intégration virtuelle de bases de données topographiques s'inscrit dans la lignée de ces travaux. De ce fait, il requiert également de se doter d'une telle ontologie pour notre domaine d'application. Une bonne pratique en la matière consiste généralement à réutiliser, autant que faire se peut, une ou plusieurs

[33] http://www.loa.istc.cnr.it/DOLCE.html

[34] http://vocamp.org/wiki/GeoVoCampSB2012

ontologies existantes. Or, le bilan de l'existant en matière d'ontologies francophones du domaine de la topographie, réalisé dans le cadre du projet GéOnto (ANR-07-MDCO-005), fait état de trop peu de ressources de ce type pour couvrir l'ensemble du domaine représenté au sein des bases de données que nous cherchons à intégrer. La création d'une ontologie francophone du domaine de la topographie s'impose donc comme un préalable à la mise en œuvre de l'approche que nous proposons. Aussi, nous sommes nous intéressée aux méthodologies de création d'ontologies proposées par les communautés de l'ingénierie des connaissances et de l'information géographique. Si aucune des méthodologies proposées ne semble faire consensus, certaines activités du processus global de développement d'ontologies (Gómez-Pérez et al., 2003), défini dans le cadre de la méthodologie de création d'ontologies METHONTOLOGY, se rencontrent dans la plupart des méthodologies proposées par les deux communautés. On retrouve en particulier, les activités de définition de l'application visée, d'évaluation de la faisabilité du projet de création d'ontologie, de spécification de l'ontologie à créer, de conceptualisation – et donc d'acquisition des connaissances du domaine – de formalisation, d'implémentation et enfin d'évaluation de l'ontologie créée. Notre démarche de création d'ontologie du domaine sera donc essentiellement composée de ces activités. En cela, elle s'apparente plus aux méthodologies proposées dans le domaine de l'ingénierie des connaissances qu'aux approches préconisées par la communauté de l'information géographique. En effet, la majorité des approches proposées pour la création d'ontologies géographiques ont ceci de remarquable qu'elles mettent essentiellement l'accent sur les concepts de haut niveau. Un effort important a été consenti dans la définition et la formalisation de ces concepts au cours des deux dernières décennies ce qui se traduit par une littérature extrêmement abondante sur le sujet. Ces travaux seront mis à profit dans la seconde étape de notre proposition consistant à permettre la représentation de connaissances issues des spécifications de bases de données topographiques.

3.3 Prise en compte des spécifications pour l'intégration de données géographiques

L'automatisation du processus d'intégration de données géographiques suppose celle de la détection des divers types d'hétérogénéité pouvant intervenir entre les bases de données à intégrer (Fichtinger et al., 2009). Un consensus s'est formé autour de l'utilisation d'ontologies du domaine pour la détection et la résolution de l'hétérogénéité sémantique entre bases de données géographiques hétérogènes, comme en attestent les travaux récents en matière de découverte et d'accès aux données géographiques ou bien de transformation de schéma (cf. chapitre 3.1). En revanche, peu de travaux se sont penchés sur la détection automatique des types d'hétérogénéités liés à la structuration et à la représentation géométrique des données. Ceci suppose en effet de disposer de connaissances formelles sur les spécifications de saisie des bases de données géographiques que l'on souhaite intégrer. Or, ces spécifications, destinées avant tout à guider les opérateurs de saisie, se présentent sous la forme de documents textuels rédigés en langage naturel.

Un effort de normalisation de ces spécifications de saisie des bases de données géographiques a été entrepris par le comité technique TC 211 de l'organisation internationale de normalisation, et a abouti à la mise en place de la norme ISO 19131 « Data product specifications ». Celle-ci vise à permettre la création de documents de spécifications de saisie de bases de données géographiques harmonisés d'une base de données à l'autre, en décrivant et en structurant les divers éléments de

spécifications de saisie devant être définis au sein de chaque document de spécifications. Elle se compose de plusieurs parties, couvrant chacune un aspect des spécifications : identification de la base de données, définition et structure du contenu, système de référence, spécifications de qualité, livraison des données, métadonnées, etc. Cependant, la partie dédiée à la description des critères de sélection des entités du monde réel devant figurer dans la base ainsi qu'à la modélisation géométrique des instances destinées à les représenter, dite « Data capture », se cantonne à un paragraphe en texte libre. L'exploitation automatique des connaissances qu'elle renferme à des fins d'intégration demeure donc, comme dans le cas des spécifications de saisie des bases de données de l'IGN, extrêmement limitée.

Approche proposée par Uitermark (2001) pour l'intégration de bases de données géographiques à partir de la formalisation de leurs spécifications

Uitermark (2001) propose une approche d'intégration virtuelle de données géographiques s'appuyant à la fois sur une ontologie du domaine et les spécifications des bases de données géographiques à intégrer. L'architecture globale du modèle proposé est présentée en figure 20. L'ontologie du domaine utilisée ici est constituée de six concepts : BÂTIMENT, ROUTE, VOIE FERRÉE, EAU, TERRAIN, AUTRE. L'auteur propose de mettre à profit les connaissances issues des spécifications pour élaborer un schéma fédéré, appelé ici « modèle de référence ». Les classes du modèle de référence sont définies par spécialisation des classes de l'ontologie du domaine, et visent à couvrir l'ensemble des concepts les plus généraux du domaine des bases à intégrer. Ainsi, la classe *Bâtiment Principal* du modèle de référence est définie comme une sous-classe du concept de BÂTIMENT, désignant l'ensemble des bâtiments possédant au moins une adresse. A l'inverse, la classe *Annexe Adjacente* désigne l'ensemble des bâtiments ne possédant pas d'adresse propre et connectés à une instance de *Bâtiment Principal*. Enfin la classe *Annexe Indépendante* est définie comme comprenant les bâtiments isolés ne disposant pas d'adresse et situés en zone urbaine, ou bien les bâtiments isolés ne disposant pas d'adresse, situés en zone rurale et ayant une emprise au sol supérieure à 20 m^2. Les spécifications sont également analysées afin de déceler d'éventuelles relations sémantiques entre classes du modèle de référence. Les relations d'équivalence, de généralisation-spécialisation et de composition sont explicitées en priorité. Enfin, une dernière analyse des spécifications va permettre de mettre en correspondance les classes des schémas des bases à intégrer avec celles du schéma fédéré. Ce sont ces relations qui seront mises à profit afin de dériver automatiquement des relations de correspondance entre classes de schémas hétérogènes. Ces relations peuvent être de différents types - équivalence, généralisation-spécialisation et composition – mais ne permettent pas d'exprimer des restrictions sur des valeurs d'attributs, ou sur des propriétés géométriques des instances de ces classes. L'implémentation du modèle de référence ainsi que des relations entre schémas et modèle est réalisée dans le langage Prolog.

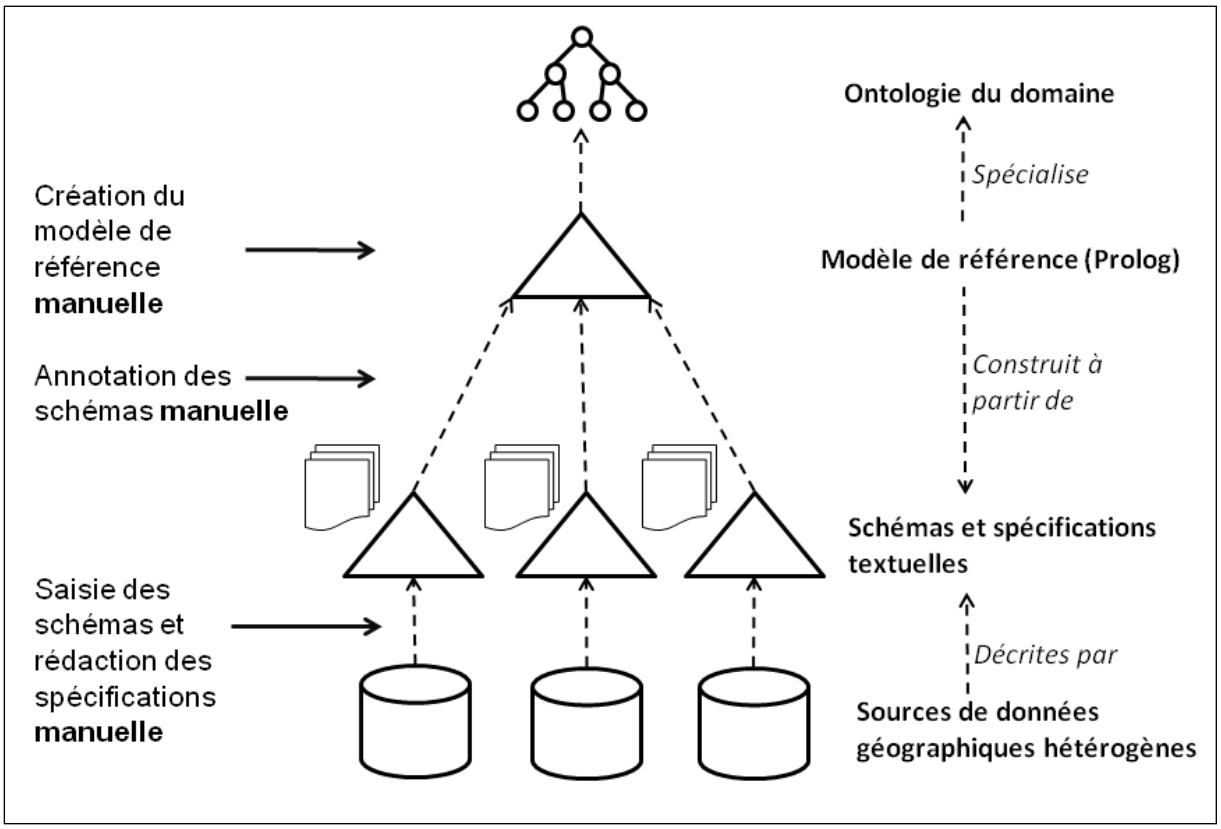


Figure 20: Approche proposée par Uitermark (2001): identification des étapes d'annotation automatiques et manuelles

Enfin, une fois les classes homologues identifiées, leurs instances sont appariées. Cet appariement est réalisé en deux temps. Tout d'abord les géométries des instances des classes homologues sont comparées afin de détecter des paires d'instances candidates à l'appariement. Puis chaque paire d'instances est testée afin de déterminer si chacune des instances impliquées est bien cohérente avec les spécifications des deux classes homologues. En effet, les critères de sélection des entités géographiques du monde réel pouvant être différents d'une classe de base de données à l'autre, deux classes homologues peuvent présenter des niveaux d'exhaustivité différents. Tester la cohérence des instances candidates vis-à-vis des définitions de chacune des classes homologues permet d'éliminer des résultats de l'appariement de données des paires d'instances candidates dont l'une des instances se révèlerait non conforme à ces définitions. Il s'agit donc ici d'effectuer un filtrage a posteriori des instances candidates à l'appariement au sein d'une paire de classes homologues. Par ailleurs, ceci permet également d'éliminer de ces résultats des instances résultant d'erreurs de saisies dans les bases de données.

Approche proposée par Gesbert (2005) pour l'intégration de bases de données géographiques à partir de la formalisation de leurs spécifications

S'inspirant de l'approche de (Uitermark, 2001), (Gesbert, 2005) propose un modèle formel permettant la description des spécifications de bases de données topographiques en vue d'intégrer leurs schémas. Celui-ci s'apparente à une architecture de bases de données fédérées (voir figure 21). D'une part, les schémas conceptuels des bases de données topographiques à intégrer décrivent la structure des bases de données. D'autre part, une ontologie du domaine représente l'ensemble des

concepts topographiques du monde réel. Les relations de correspondance entre éléments de schémas et éléments de l'ontologie reposent sur les connaissances fournies par les spécifications concernant les règles de représentation des entités topographiques du monde réel au sein de la base. Il s'agit ici de relations complexes, incluant des restrictions sur les propriétés que doivent vérifier les entités géographiques pour être représentées au sein d'une classe de base de données, ainsi que des règles de modélisation de ces entités géographiques. Un langage formel pour la description des règles de représentation des entités géographiques est proposé afin de permettre la représentation de ces relations complexes entre ontologie et schémas.

Dans ce langage, l'ensemble des règles décrivant comment un type d'entités géographiques est représenté au sein de la base constitue une *procédure de représentation* (voir figures 21 et 22). Une *procédure de représentation* est composée de diverses sections, correspondant aux divers types de règles de représentation que l'on rencontre dans les spécifications de bases de données topographiques : sélection, découpage, agrégation, attributs et instanciation.

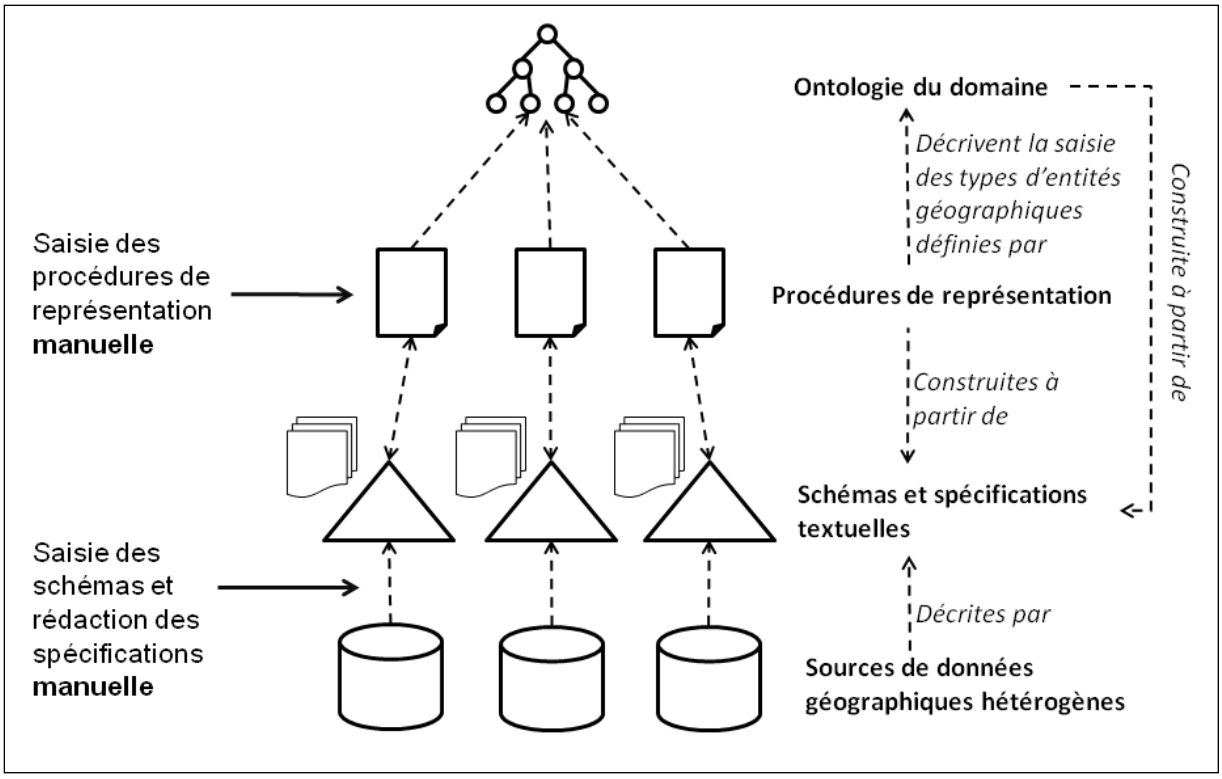


Figure 21: Approche proposée par Gesbert (2005): identification des étapes d'annotation automatiques et manuelles

Ainsi, une *procédure de représentation* comporte toujours au moins une règle de *sélection* des entités topographiques du monde réel devant figurer dans une classe donnée. Cette sélection s'exprime sous la forme de conditions plus ou moins complexes que doivent vérifier ces entités topographiques. Ces conditions peuvent porter sur divers critères, comme la catégorie d'entités topographiques à laquelle appartiennent les entités, des valeurs de propriétés, le plus souvent géométriques, de ces entités, ou encore leurs relations spatiales, essentiellement topologiques ou métriques, avec d'autre entités topographiques. D'autres critères, plus subjectifs, peuvent intervenir, comme la mention d'une entité topographique sur la carte au 1/25 000 en service. Enfin des critères

flous peuvent entrer en ligne de compte. C'est le cas lorsque les règles de sélection font référence à l'importance des entités géographiques dans le paysage. Un ensemble de contraintes est donc défini dans le langage formel de représentation des spécifications proposé afin de permettre l'expression de ces diverses conditions de sélection. Enfin, outre les cas de sélection simple, l'auteur prévoit des cas de critères de sélection particuliers, dédiés à la description des règles de sélection des entités topographiques impliquées dans des réseaux lorsque celles-ci correspondent à des culs-de-sacs ou des axes secondaires.

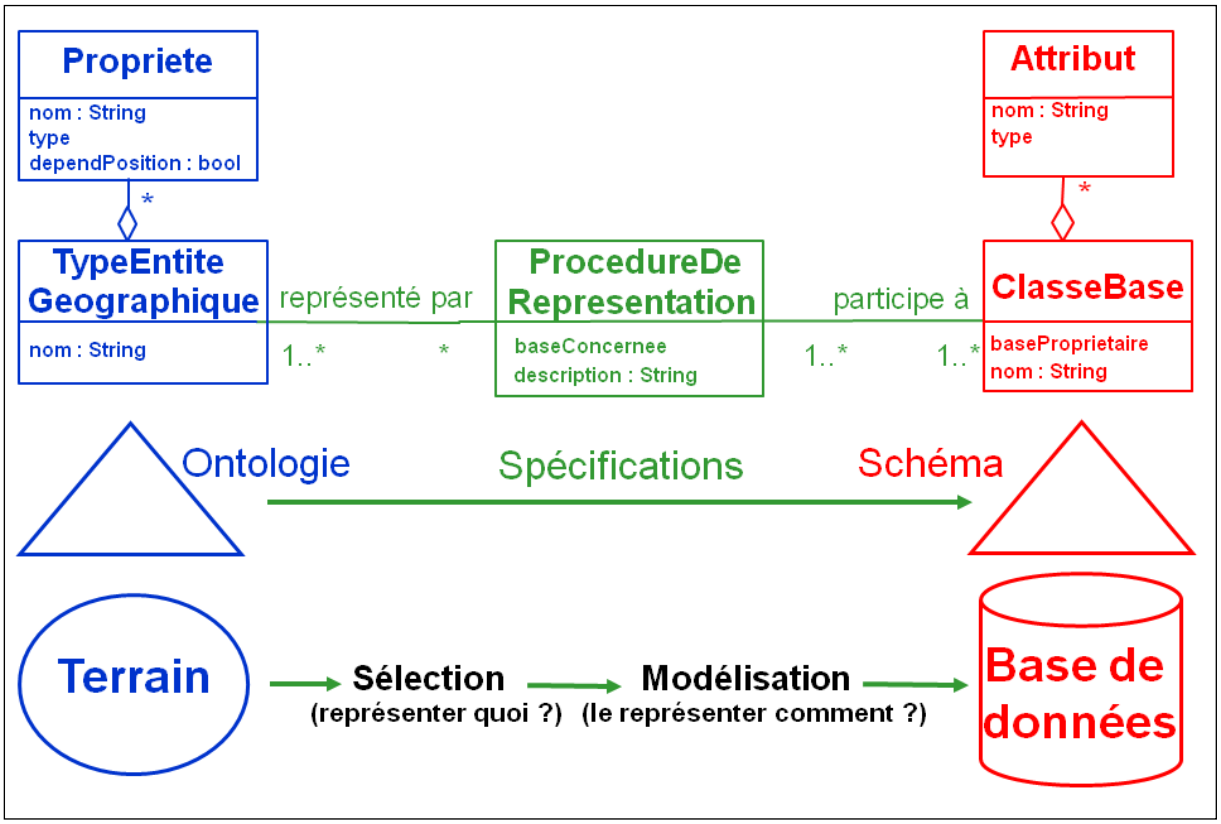


Figure 22: Modèle global pour la représentation des spécifications de bases de données géographiques, extrait de (Gesbert, 2005)

La représentation d'entités topographiques impliquées dans des réseaux au sein de bases de données topographiques implique une modélisation particulière des données saisies, sous la forme d'un graphe, et donc des règles de saisie conformes à cette modélisation. La section « découpage » définie par Gesbert (2005) vise à permettre la représentation de ces règles particulières. En particulier, elle permet de décrire les divers types de critères de découpage couramment utilisés pour la saisie de réseaux. Ainsi, la partie *intersections* introduit une règle de découpage des entités géographiques saisies au niveau des intersections du réseau. La partie *rencontre* désigne une règle de découpage des entités géographiques en cas de présence d'un type d'obstacle particulier sur le réseau. Enfin, la partie *changement* correspond aux règles de découpage fondées sur le changement d'une propriété des entités géographiques sur le terrain et ayant pour conséquence la modification de la valeur prise par l'attribut visant à refléter cette propriété au niveau des objets géographiques instanciés dans la base. Afin de permettre la description formelle des règles de représentation des arcs et nœuds résultant du processus de découpage des entités géographiques, Gesbert (2005)

introduit des règles de représentation particulières identifiées par les mots-clés *sections* et *limites de section*.

La représentation des entités géographiques du monde réel au sein d'une base de données géographique nécessite parfois de recourir à des opérations de généralisation, également décrites dans les spécifications de saisie de ces bases. Il s'agit pour l'essentiel d'opérations d'agrégation d'entités géographiques proches et dotés de propriétés similaires pour former un seul objet au sein de la base de données. Gesbert (2005) propose de formaliser les règles d'application de ces opérations au sein d'une section *agrégation*. Celle-ci comporte des parties *chacun* et *ensemble* permettant de préciser si les conditions que doivent vérifier les entités géographiques du monde réel pour être généralisées doivent l'être par chaque entité indépendamment des autres, ou bien par l'ensemble des entités. Ces conditions sont représentées soit par des contraintes, du même type que celles utilisées dans les sections *sélection* et *découpage*, soit par des relations spatiales, soit enfin par des critères de similarité ou de dissimilarité introduits par les mots-clés *même* et *différence*.

La section *attribut* vise à permettre la description d'attributs locaux au sein de la *procédure de représentation*. Il s'agit d'attributs permettant de représenter certaines propriétés des entités géographiques pour les besoins de l'expression des règles de saisie de ces dernières.

Les *procédures de représentation* définies par Gesbert (2005) sont orientées. Elles partent des entités géographiques définies au sein de l'ontologie du domaine pour aller vers les schémas de bases de données. Une *procédure de représentation* ne décrit donc pas les spécifications d'une classe de base de données particulière, mais l'ensemble des règles de représentation d'une catégorie d'entités géographiques au sein d'une (ou plusieurs) classe(s) de base de données. La partie *instanciation* permet donc de formaliser les conditions sous lesquelles les entités d'une catégorie donnée seront représentées au sein de l'une ou l'autre des classes d'une base de données particulière.

Afin de permettre l'exploitation automatique des *procédures de représentation* décrites dans ce langage, Gesbert (2005) propose un modèle orienté objet, implémenté en Java, permettant de saisir et manipuler les divers éléments de schémas et d'ontologie, ainsi que les principaux éléments du langage formel. L'instanciation du modèle est réalisée de façon manuelle. Pour ce faire, Gesbert (2005) préconise d'extraire les connaissances nécessaires à la création de l'ontologie du domaine et des *procédures de représentation* par analyse des textes des spécifications.

Conclusion

Peu de travaux se sont penchés sur la formalisation des spécifications de bases de données géographiques en vue de mettre à profit les connaissances qu'elles renferment dans un processus d'intégration. (Uitermark, 2001) propose de formaliser, en Prolog, et d'exploiter certaines règles issues des spécifications de bases de données géographiques dans un processus complet d'intégration allant de l'appariement des schémas à celui des données. Cependant, il ne s'intéresse ici qu'aux règles de sélection des entités géographiques, délaissant toutes les règles de représentation géométrique de celles-ci. Ceci est possible dans le cas traité dans ses travaux : les bases de données à intégrer possèdent ici des niveaux de détail très proches et donc des représentations géométriques très semblables. L'appariement de leurs instances ne nécessite donc pas la mise en œuvre d'algorithmes prenant en compte des différences de géométries importantes entre les données des différentes bases. Dans le cas contraire, disposer des connaissances fournies

par ces règles de représentation géométrique s'avère crucial pour permettre un paramétrage fin des algorithmes d'appariement géométrique des données. C'est pourquoi Gesbert (2005) propose un langage formel permettant de représenter l'ensemble des règles décrites dans les spécifications de saisie des bases de données géographiques. Cependant, le modèle proposé n'intègre aucun des standards et normes actuellement préconisés en matière de représentation de schémas de bases de données géographiques, de représentation d'ontologies ou encore de règles. Sa mise en œuvre pratique nécessiterait l'implémentation d'algorithmes ad hoc permettant la comparaison automatique des *procédures de représentation*, ce qui, compte tenu de l'expressivité du langage proposé, pourrait s'avérer extrêmement complexe et conduire à des résultats éventuellement peu exploitables.

3.4 Bilan

Les travaux récents en matière d'intégration d'informations dans le domaine des infrastructures de données géographiques portent prioritairement sur les problématiques de découverte et d'accès aux données ainsi que de transformation de schémas de bases de données géographiques.

Dans les deux cas, ils s'attachent à résoudre les problèmes posés par l'hétérogénéité sémantique des données géographiques. Les solutions proposées s'appuient sur les architectures proposées dans le domaine informatique pour l'intégration d'informations. Ainsi, les applications dédiées à la découverte et à l'accès aux données géographiques reposent généralement sur des architectures de type « médiateurs » tandis que les applications de transformation de schémas s'apparentent à des approches de type « bases de données fédérées ». Dans tous les cas, les sources de données hétérogènes sont décrites par leur schéma, une ontologie d'application ou bien des métadonnées qui sont annotés à l'aide d'une ontologie du domaine décrivant de façon formelle les concepts topographiques représentés pas les données. Cette annotation, qui consiste à mettre en correspondance les éléments constitutifs des diverses ressources décrivant les données avec les éléments d'une ontologie du domaine, constitue une tâche essentielle du processus d'intégration. En effet, les applications de découverte et d'accès aux données ou d'appariement de schémas exploitent directement les relations de correspondance définies à cette étape afin de déterminer les correspondances entre sources de données hétérogènes. En raison de sa complexité, cette annotation est réalisée le plus souvent de façon manuelle, même si on recense quelques tentatives d'automatisation.

En pratique, la plupart des approches proposées se cantonnent à des exemples applicatifs relativement restreints, ne nécessitant pas de recourir à des ontologies du domaine particulièrement étendues. Les auteurs ne présentent d'ailleurs que rarement le contenu des ontologies mises en œuvre. Lorsqu'il s'agit d'ontologies créées spécifiquement pour leur application, les auteurs ne mentionnent généralement pas la méthodologie adoptée pour leur construction et n'indiquent que très exceptionnellement les URLs permettant d'accéder à ces ressources. De façon générale, il existe peu d'ontologies du domaine de la topographie, et celles disponibles n'offrent qu'une couverture partielle du domaine et se présentent généralement en langue anglaise. Les premiers travaux engagés par la communauté de l'information géographique concernant l'élaboration d'ontologies topographiques se sont principalement attachés à définir des concepts de haut niveau pour la description de l'espace topographique et des propriétés des catégories d'entités topographiques.

Plus récemment ont émergé des approches de création d'ontologies du domaine de la topographie s'inspirant des méthodologies proposées dans le domaine de l'ingénierie des connaissances, qui ont permis d'aboutir à la construction d'ontologies sur les thèmes de l'hydrographie ou du découpage administratif du territoire britannique.

Enfin, si l'utilisation d'ontologies du domaine s'avère pertinente pour la prise en compte de l'hétérogénéité sémantique des données, peu d'applications envisagent de traiter d'autres types d'hétérogénéités, liées à la structuration ou à la représentation géométrique des données. Ceci nécessite de disposer de connaissances extrêmement détaillées sur les données, dont les spécifications de saisie des bases de données géographiques sont la principale source. Peu de travaux se sont penchés sur la question de la formalisation de ces spécifications afin de permettre l'exploitation automatique des connaissances qu'elles renferment. L'approche la plus aboutie demeure celle proposée par (Gesbert, 2005). Cependant, si le modèle proposé s'apparente aux architectures d'intégration de données virtuelle désormais de plus en plus répandues pour l'intégration d'informations dans les infrastructures de données géographiques, il n'intègre aucun des standards ou normes actuellement recommandés en matière de gestion de l'information géographique, ou de représentation de connaissances ; sa mise en œuvre, en l'état, au sein d'un processus d'intégration de données géographiques constituerait une solution totalement ad hoc.

4 Proposition

Cette partie présente notre proposition pour la formalisation et l'acquisition de connaissances portant sur la description de données géographiques pour la mise en œuvre d'un processus d'intégration virtuelle de bases de données géographiques.

Elle décrit, dans un premier temps, l'architecture globale de notre modèle. Celle-ci s'inspire des architectures de médiation couramment préconisées pour les applications de découverte et d'accès aux données géographiques présentées au chapitre 3.1.1, ainsi que des approches pour l'appariement de schémas présentées au chapitre 3.1.2, auxquelles elle intègre un modèle pour la représentation des connaissances issues des spécifications de bases de données géographiques.

Une première mise en œuvre du modèle proposé y est présentée en partie 4.2. Celle-ci vise à automatiser, autant que faire se peut, l'étape d'annotation sémantique des ressources décrivant les bases de données à intégrer, sans pour autant contraindre les fournisseurs de données sur le vocabulaire utilisé dans les ressources décrivant leurs bases de données. L'approche adoptée repose sur une combinaison de techniques d'appariement automatique terminologiques, de techniques fondées sur les relations taxonomiques, et de techniques fondées sur une ontologie de support.

Une seconde mise en œuvre est présentée en partie 4.3. Celle-ci vise à mettre à profit les possibilités de représentation de connaissances et de raisonnement offertes par les formalismes de représentation de connaissances fondés sur les logiques de description afin de permettre la prise en compte de connaissances issues des spécifications des bases de données à intégrer dans le processus d'appariement de leurs schémas. Le modèle proposé est également mis en œuvre dans le cadre d'une application de découverte de bases de données topographiques permettant à l'utilisateur de retrouver aisément dans quelle(s) base(s) de données sont représentées les entités topographiques appartenant à une catégorie de son choix, et sous quelle(s) condition(s), et avec quelle(s) représentation(s) géométrique(s) celles-ci y sont représentées.

4.1 Modèle global pour l'intégration de bases de données géographiques

Le modèle que nous proposons vise l'intégration virtuelle de bases de données géographiques. Une étape fondamentale d'un tel processus d'intégration, quelle que soit sa finalité, réside dans la détermination des relations de correspondance entre les éléments des différents schémas conceptuels des bases de données que l'on souhaite intégrer. **La réalisation de cet objectif d'appariement de schémas constitue donc l'enjeu majeur de notre modèle : nous souhaitons permettre la détection automatique des différents types d'hétérogénéités pouvant intervenir entre bases de données topographiques vectorielles afin d'établir des relations de correspondances fines entre éléments de schémas**. Les approches proposées pour la découverte et l'accès aux données, ainsi que celles dédiées à la transformation de schémas, présentées au chapitre 3.1, s'attachent principalement à permettre la détection et la résolution de l'hétérogénéité sémantique entre bases de données géographiques. A ces fins, elles s'appuient sur des ontologies du domaine, destinées à décrire les concepts partagés du domaine d'intérêt des diverses sources de données à intégrer et utilisées comme sources de connaissances externes pour l'annotation sémantique des éléments des diverses ressources descriptives – schémas conceptuels, ontologies d'applications ou métadonnées - des bases de données considérées. Cependant, si les annotations sémantiques proposées permettent de décrire, pour chacune des classes de bases de données annotées, les catégories d'entités géographiques que celles-ci visent à représenter, elles ne rendent pas compte de la complexité de la relation entre chacune de ces diverses représentations du territoire et le terrain réel, relation que Kavouras et Kokla (2008) définissent comme la sémantique des données. Aussi, la modélisation fine de cette sémantique sera-t-elle un élément important de notre proposition.

Les applications de découverte et d'accès aux données présentées au chapitre 3.1 requièrent toutes, dans un premier temps, de déterminer les relations de correspondance entre les éléments des ressources décrivant les bases de données considérées, et les éléments d'une ontologie du domaine. Ces relations de correspondance, sont, dans la plupart des cas, utilisées pour rediriger et réécrire les requêtes des utilisateurs en quête de données vers les sources pertinentes, bien qu'elles puissent également être mises à profit afin de déterminer des correspondances entre éléments de ces ressources décrivant les bases de données à intégrer, comme c'est le cas dans les approches proposées par (Lutz et al., 2006) ou encore (Schade, 2010). Les relations de correspondance entre ressources descriptives et ontologie du domaine pouvant être définies au sein de ces applications peuvent être de différents types. On rencontre principalement des relations d'équivalence ou de subsomption (Paul et Gosh, 2006) (Lassoued et al., 2008) (Lutz et al., 2006), bien que d'autres types de relations puissent parfois être utilisés, comme c'est le cas pour l'approche proposée par (Nambiar et al., 2006) qui offre la possibilité de définir des relations du type « possède des instances du type », « mentionne » ou encore « utilise ». Klien (2008) insiste sur la nécessité d'introduire une relation spécifique pour l'annotation d'ontologies d'applications décrivant des bases de données géographiques à l'aide de concepts issus d'une ontologie du domaine. En effet, le recours à des relations d'équivalence ou de subsomption pour associer des concepts d'ontologies d'applications décrivant des classes de schémas de bases de données géographiques à des concepts d'ontologies du domaine décrivant des catégories d'entités géographiques suggère implicitement que les instances des premières sont également des instances des secondes. Considérant qu'une instance de base de données géographique constitue une représentation d'une entité géographique du monde réel, et ne

peut, de ce fait, en aucun cas être assimilée à une telle entité géographique, Klien (2008) introduit la relation « annotate » afin de définir les relations de correspondances entre ontologies d'applications et ontologie du domaine. Adoptant son point de vue, Schade (2010), reprend le processus d'annotation proposé par Klien (2008) pour l'adapter à une application de transformation de schémas (chapitre 3.1.2), mais préfère néanmoins à la dénomination « annotate » de cette relation, trop vague selon lui, celle de « domain reference ». Dans notre approche, nous adhérons à cette thèse concernant l'existence d'une sémantique particulière de la relation associant les diverses ressources décrivant des bases de données topographiques à une ontologie du domaine.

Les relations définies pour l'annotation sémantique des éléments de ressources descriptives de bases de données géographiques permettent donc de décrire pour chacune des classes de bases de données annotées les catégories d'entités géographiques que celles-ci visent à représenter. Or, un utilisateur recherchant, par exemple, des données sur les forêts pourrait souhaiter connaître, outre les sources de données lui permettant de se procurer des données sur les forêts, les éventuels conflits de critères de sélection (Devogele, 1997) existant entre ces sources. En effet, deux classes de bases de données représentant, l'une « […] les bois et forêts d'une superficie supérieure à 500 ha […] » (BDCARTO© 3.1), et l'autre les « […] bois de plus de 500 m^2 […] » (BDTOPO© 2.0), ne présenteront pas nécessairement le même intérêt pour notre utilisateur, en raison de la différence de granularité des données entre ces deux classes. De plus, il semble également utile de l'informer d'éventuels conflits de description géométrique des données (Devogele, 1997) entre les diverses bases disponibles. Reprenons l'exemple des classes *Massif_Boisé* de la BDCARTO© 3.1 et *Zone_Végétation* de la BDTOPO© 2.0 présenté ci-dessus. En se fondant sur les seuls noms de ces deux classes, on pourrait supposer que celles-ci fournissent deux représentations géométriques semblables des espaces arborés. Or, la classe *Zone_Végétation* décrit les forêts sous la forme de polygones dont les côtés correspondent aux contours extérieurs des forêts représentées, tandis que la classe *Massif_Boisé* correspond en fait à une classe de toponymes propres aux espaces arborés dont les instances possèdent une géométrie de type ponctuel, saisie au centre de chaque forêt représentée. Ces connaissances, portant sur les critères de sélection des entités géographiques devant être représentées au sein d'une classe de base de données ou leur modélisation géométrique, issues des spécifications de saisie des bases de données, constituent la sémantique exacte des données géographiques. A ce titre, elles sont indispensables aux utilisateurs souhaitant découvrir et évaluer des données en vue de leur éventuelle réutilisation. En outre, la dérivation automatique du contenu de la classe *Massif_Boisé* à partir de celui de la classe *Zone_Végétation* à l'aide d'une application de transformation de schémas nécessite, pour produire des données cohérentes avec les spécifications de la classe *Massif_Boisé*, de prendre en compte les différences de critères de sélection et de représentation géométrique entre ces deux classes. De la même façon, l'appariement automatique des données de ces deux classes requiert des algorithmes spécifiques, permettant, en premier lieu, le filtrage des instances de la classe *Zone_Végétation* candidates à l'appariement avec des instances de la classe *Massif_Boisé*, et prenant en compte les différences de représentation géométrique entre les données de ces deux classes. C'est pourquoi nous proposons dans notre approche d'introduire, dans les annotations sémantiques des ressources décrivant les données, des connaissances précisant les critères de sélection des entités topographiques ou leur modélisation géométrique, issues des spécifications de saisie des bases de données topographiques concernées.

Cette nécessité de déterminer, lors de l'étape d'appariement des schémas, outre les relations de correspondance entre éléments de schémas, les restrictions dues à l'hétérogénéité des spécifications des différentes bases qui s'appliquent à ces correspondances a été abordée par Uitermark (2001) et Gesbert (2005). Pour ce faire, Uitermark (2001) propose de formaliser les règles de sélection des entités géographiques devant figurer dans chaque classe des diverses bases de données à intégrer. Celles-ci sont exploitées en aval de l'appariement géométrique des données afin de tester si chaque instance impliquée dans une paire d'instances candidates à l'appariement vérifie bien les critères de sélection des deux classes homologues auxquelles elles appartiennent. Gesbert (2005) s'inspire de cette approche et propose un modèle formel permettant de représenter l'ensemble des règles de sélection et de modélisation géométrique des entités géographiques dans un langage dédié. Cependant, à l'inverse de Uitermark (2001), Gesbert (2005) préconise d'exploiter les spécifications formelles en amont du processus d'appariement géométrique des données afin d'opérer une présélection des instances candidates à l'appariement et de paramétrer les algorithmes d'appariement géométrique utilisés.

L'architecture globale du modèle proposé par Gesbert (2005) (cf. figure 20) s'apparente aux diverses architectures d'intégration virtuelle présentées au chapitre 3.1. Les *procédures de représentation*, qui définissent l'ensemble des règles de représentation d'un type d'entités géographiques au sein de l'une des bases de données à intégrer peuvent, en effet, être vues comme des annotations sémantiques complexes, décrivant les relations entre les éléments d'une ontologie du domaine et ceux du schéma de la base de données en question. C'est pourquoi nous nous proposons de nous inspirer de ces travaux et de les adapter aux standards et normes actuellement recommandés en matière de gestion de l'information géographique et de représentation de connaissances. En effet, les schémas de bases de données et l'ontologie du domaine y sont représentés selon des schémas conceptuels de données définis par Gesbert (2005). Dans un souci d'interopérabilité, nous proposons donc de leur substituer des standards dédiés ; les normes ISO pour la description de l'information géographique, et OWL pour l'ontologie du domaine. De plus, l'instanciation, la vérification de la cohérence et la comparaison automatique des *procédures de représentation* nécessitent le développement d'outils totalement ad hoc. A l'inverse, l'approche proposée par Klien (2008) et reprise par Schade (2010) consiste à décrire, dans un langage de représentation de connaissances (WSML), les schémas des bases de données à intégrer sous la forme d'ontologies d'applications inspirées des schémas ISO pour la représentation de schémas de bases de données géographiques (*Feature Types Ontologies*). Ces ontologies d'applications sont annotées via une ontologie du domaine, intégrée à un cadre de référence sémantique et décrite dans le même langage. Ainsi, les annotations sont directement exploitables par les systèmes de raisonnement existants et permettant d'exploiter des connaissances décrites dans le langage choisi. Cette approche, fondée sur la mise en œuvre de standards adéquats et de leurs outils associés, et en particulier ce dernier aspect de description des schémas conceptuels des bases de données à l'aide d'un formalisme permettant d'effectuer des raisonnements sur les descriptions fournies, nous semblent plus adaptés à notre objectif final d'interopérabilité. C'est pourquoi nous proposons d'en reprendre certains aspects et de les intégrer au modèle proposé par Gesbert (2005).

L'architecture globale du modèle que nous proposons est présentée en figure 23. Elle correspond à celle proposée par Gesbert (2005), exception faite des *procédures de représentation* que nous souhaitons remplacer, à l'instar de Klien (2008) et Schade (2010), par des annotations sémantiques portées par les ontologies d'applications au format OWL décrivant la structure des données des

diverses bases à intégrer. Le choix de ce langage de représentation de connaissances, fondé sur les logiques de description, est motivé par les possibilités de raisonnement associées à un tel formalisme, que nous souhaitons intégrer à nos travaux. Par ailleurs, nous proposons que ces ontologies soient générées de façon semi-automatique à partir des schémas des bases de données décrits selon le schéma conceptuel de données « ISO 19109 - Rules for application schema » et de l'analyse des spécifications textuelles de ces bases. Leur structure s'inspire des normes ISO pour la représentation de l'information géographique et varie selon le type d'annotations sémantiques que l'on souhaite mettre en œuvre. Nous proposons, en effet, deux approches. La première, présentée en détail au chapitre 4.2, consiste à augmenter le niveau de granularité des schémas conceptuels des bases de données à intégrer en mettant en évidence, au sein des ontologies d'applications générées, des concepts topographiques qui, en raison de contraintes de modélisation de ces bases, n'apparaissent dans les schémas que sous forme de valeurs d'attributs. Dans cette approche, l'établissement des annotations sémantiques s'inspire de techniques utilisées dans le domaine de l'alignement d'ontologies. L'application développée détermine, à l'aide de techniques lexicales et structurelles, des relations d'équivalence, de subsomption et de proximité sémantique entre les concepts des ontologies d'applications et ceux de l'ontologie du domaine. Ces relations de correspondance sont ensuite exploitées afin de calculer des relations d'appariement entre les schémas conceptuels des bases de données concernées. Cette approche permet une prise en compte partielle des spécifications dans la mesure où les connaissances permettant de déterminer quelles valeurs d'attributs doivent êtres mises en évidence lors de la création des ontologies d'applications en sont directement issues. Cependant, elle ne permet pas la détermination de relations de correspondance fines entre éléments de schémas hétérogènes. En revanche, elle présente l'avantage de fournir des résultats rapides, en réduisant considérablement le coût lié à une annotation manuelle des ontologies d'applications. La seconde approche proposée, que nous détaillerons au chapitre 4.3, consiste à générer des ontologies d'applications au sein desquelles les classes, géométries, attributs et valeurs d'attributs des schémas des bases de données à intégrer sont réifiés, afin de permettre leur annotation à l'aide d'axiomes décrivant précisément à quels éléments de l'ontologie du domaine ils se rapportent et sous quelles conditions. Ceci suppose de disposer, au sein de l'ontologie du domaine, d'un certain nombre de concepts se rapportant au vocabulaire propre aux spécifications de bases de données géographiques, afin de pouvoir exprimer l'ensemble des règles de représentation des spécifications identifiées par Gesbert (2005). C'est pourquoi nous proposons d'adopter l'approche proposée par Kuhn (2003) et reprise par Klien (2008) et Schade (2010), en intégrant notre ontologie du domaine de la topographie à un cadre de référence sémantique permettant la définition de ces concepts.

Enfin, nous proposons d'utiliser une ontologie du domaine réalisée de façon semi-automatique, à l'aide d'outils de traitement automatique du langage naturel (TALN), appliqués aux textes des spécifications de bases de données de l'Institut National de l'Information Géographique et Forestière (cf. partie 4.2.2.2). Cette première ontologie a par ailleurs été enrichie, par alignement automatique, de termes également extraits de récits de voyages à l'aide d'outils de TALN. Ce travail d'enrichissement de l'ontologie du domaine a été réalisé par l'ensemble des partenaires du projet GéOnto (ANR-O7-MDCO-005) dans lequel s'inscrit cette thèse.

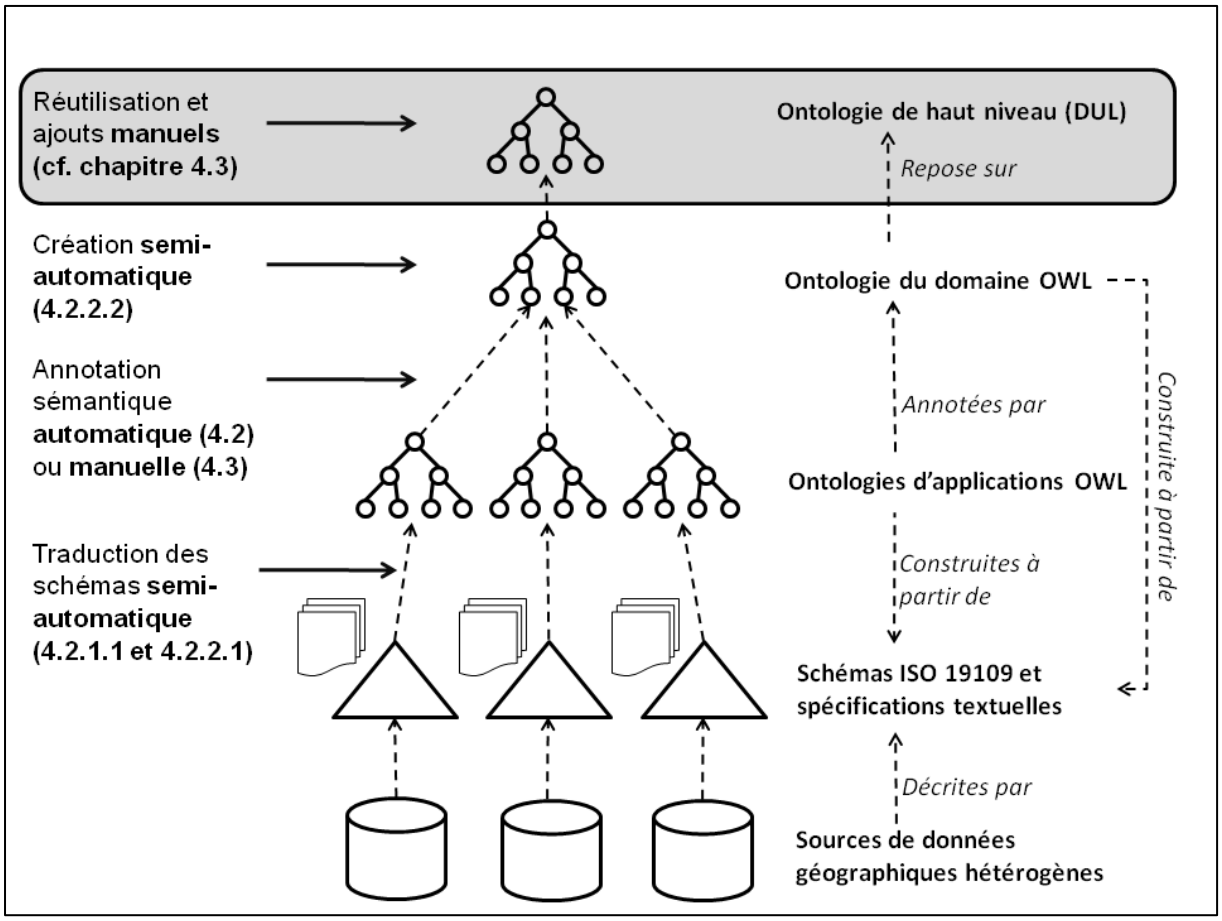


Figure 23: Architecture globale du modèle proposé: identification des étapes d'annotation automatiques et manuelles

4.2 Appariement de schémas fondé sur des valeurs d'attributs et une ontologie de support

L'architecture globale du modèle que nous proposons pour l'appariement de schémas de bases de données géographiques s'apparente à celles des approches présentées au chapitre 3.1 pour l'intégration virtuelle de données géographiques. Au sein de ces modèles, les différentes bases de données à intégrer sont décrites via divers types de ressources (leurs schémas, des ontologies d'applications ou bien des métadonnées), annotées à l'aide d'une ontologie du domaine décrivant l'ensemble des concepts représentés au sein de ces bases de données. Ces annotations, qui mettent en correspondance des éléments de ressources décrivant les bases de données à intégrer et des éléments de l'ontologie du domaine, peuvent être mises à profit afin de déduire des relations d'appariement entre les ressources décrivant les bases de données concernées, et par conséquent entre leurs schémas conceptuels. C'est le cas notamment dans les approches proposées par (Lutz et al., 2006) et (Schade, 2010). Dans la plupart des cas, le travail complexe d'appariement de ces ressources décrivant les données avec l'ontologie du domaine, indispensable à la réalisation de leur annotation sémantique, est laissé à la charge du fournisseur de données, et effectué manuellement par ce dernier. Conscients du coût potentiellement dissuasif que représente cette tâche pour les fournisseurs de données, quelques travaux envisagent donc son automatisation. Or, ces approches

visant l'automatisation de la mise en correspondance d'éléments de schémas (ou d'ontologies d'applications) avec ceux d'une ontologie du domaine s'appuient sur des techniques d'appariement lexicales, comparant les chaînes de caractères utilisées pour désigner éléments de schémas et éléments de l'ontologie (Paul et Ghosh, 2006) (Lutz et al., 2006). Afin de s'assurer d'un recouvrement lexical suffisant entre schémas et ontologie du domaine, ces approches imposent que les schémas soient préalablement traduits par le fournisseur de données dans les termes de cette ontologie, ce qui revient finalement à lui déléguer la tâche de mise en correspondance de son (ou ses) schéma(s) avec l'ontologie du domaine.

En première approche, nous nous sommes donc attachée à proposer une mise en œuvre de notre modèle automatisant, autant que faire se peut, les différentes étapes d'annotation sur lesquelles repose l'appariement des schémas. Nous souhaitons pouvoir apparier au mieux des schémas de bases de données géographiques sans imposer aux fournisseurs de données d'étapes d'annotation manuelle, ni de contraintes concernant le vocabulaire utilisé au sein de leurs schémas conceptuels de données. Pour atteindre cet objectif de mise en correspondance automatique des éléments des schémas conceptuels des bases de données à intégrer avec ceux d'une ontologie du domaine, nous nous sommes inspirée des techniques d'appariement automatique de schémas ou d'alignement d'ontologies rencontrées dans la littérature et présentées au chapitre 3.1.2. L'étape d'appariement des schémas est réalisée à l'aide des relations de correspondance établies entre les schémas et l'ontologie du domaine, utilisée ici comme ressource externe visant à pallier d'éventuels manques de recouvrement lexical et structurel entre schémas, via des ontologies d'applications. Nous présentons donc dans la partie 4.2.1 notre approche globale pour l'appariement automatique de schémas de bases de données topographiques. La partie 4.2.2 décrit les approches adoptées pour l'instanciation des principales composantes de notre modèle : les ontologies d'applications et l'ontologie du domaine. Cette approche est également décrite dans (Abadie, 2009). Enfin, la partir 4.2.3 présente une mise en œuvre de notre approche sur des schémas de bases de données de l'IGN, et discute des perspectives qu'offrent les résultats obtenus pour l'appariement de données topographiques.

4.2.1 Approche globale

L'architecture et le processus global mis en œuvre dans cette première approche d'appariement de schémas de bases de données géographiques sont présentés en figure 24. L'architecture correspond à celle présentée en figure 23, exception faite de l'ontologie de haut-niveau que nous n'utilisons pas ici. Une première étape (cf. figure 24, numéro 1) de notre approche d'appariement automatique de schémas de bases de données géographiques va consister à générer, de façon semi-automatique, les ontologies d'applications destinées à représenter la structure des différentes bases de données à intégrer, à partir des schémas de ces bases représentés selon la norme « ISO 19109-Rules for application schema ». Cette étape de création de ressources descriptives intermédiaires, intervenant entre les schémas des bases de données et l'ontologie du domaine, va nous permettre de faire émerger des connaissances importantes sur le contenu de ces bases. En effet, il arrive fréquemment que des contraintes de modélisation conduisent à préciser la nature exacte des entités géographiques représentées au sein d'une classe à l'aide de valeurs d'attributs. Suivant l'approche proposée par (Manoah et al., 2004), nous souhaitons mettre à profit ces valeurs d'attributs afin d'augmenter le niveau de granularité des schémas lors de la création des ontologies d'applications. Cette opération vise à accroître nos chances de disposer d'un recouvrement lexical et structurel

suffisant lors de l'étape de mise en correspondance des ontologies d'applications et de l'ontologie du domaine (cf. figure 24, numéro 2). Cette dernière est réalisée de façon automatique à l'aide d'une combinaison de techniques d'alignement d'ontologies lexicales et structurelles, élaborée dans le cadre du projet GéOnto (ANR-O7-MDCO-005) (Hamdi et al. 2008). Cette étape est réalisée de préférence à un alignement direct des ontologies d'applications, afin de pallier un éventuel manque de recouvrement lexical et structurel entre elles ; l'ontologie du domaine est utilisée ici comme une ontologie de support, selon l'approche proposée par (Aleksovski et al., 2006). Les relations de correspondance entre ontologies d'applications et ontologie du domaine ainsi détectées sont ensuite mises à profit afin d'aligner automatiquement les ontologies d'applications (cf. figure 24, numéro 3). Enfin, les relations de correspondances entre ontologies d'applications, et celles entre schémas et ontologies d'applications, préalablement conservées lors de l'étape de création de ces dernières, sont exploitées afin de déduire automatiquement les relations de correspondances entre schémas hétérogènes (cf. figure 24, numéro 4). L'ensemble de ce processus est détaillé dans la suite de ce chapitre, et une mise en œuvre pratique est proposée.

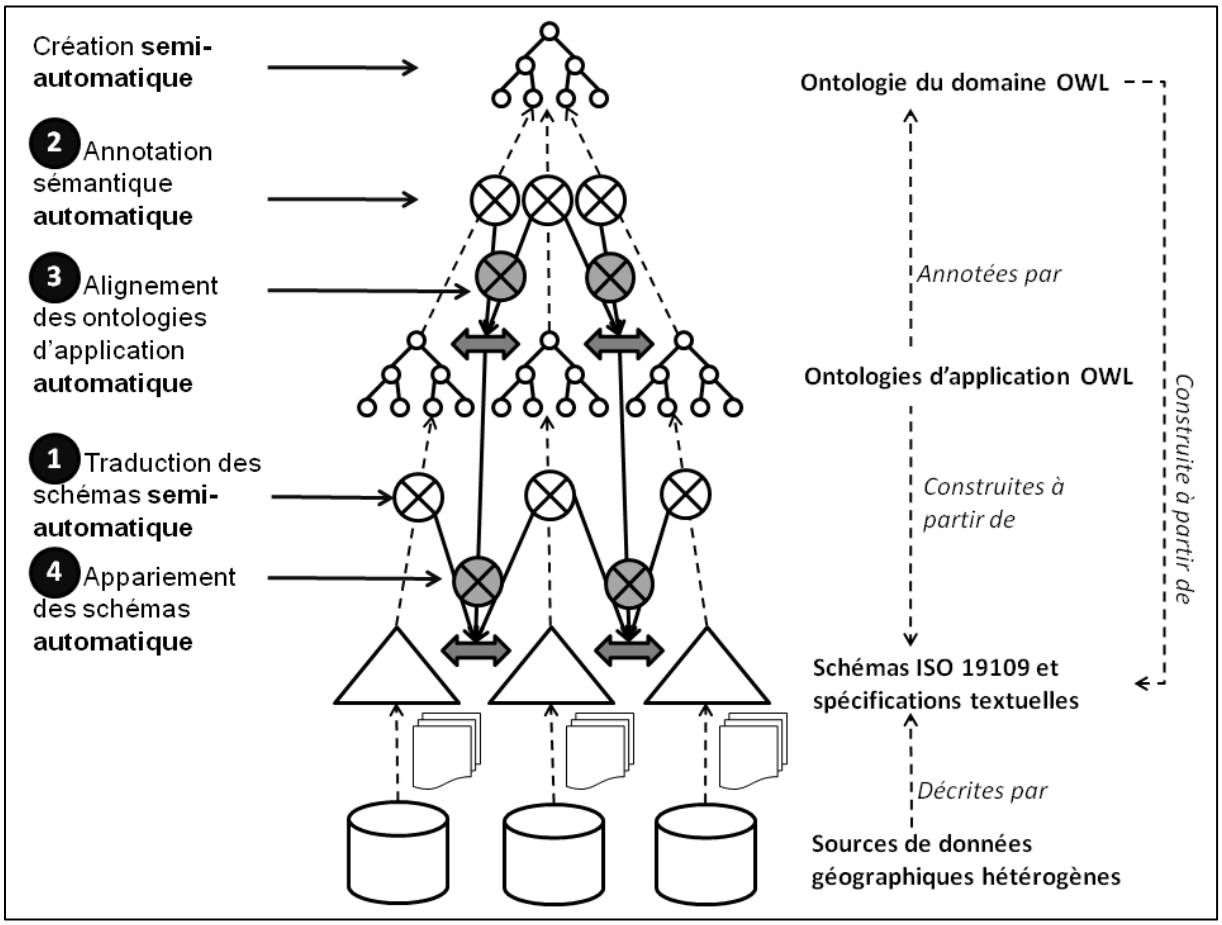


Figure 24: Processus global d'appariement de schémas de bases de données géographiques fondé sur des valeurs d'attributs et une ontologie de support

4.2.1.1 Création des ontologies d'applications et explicitation de concepts topographiques cachés

Les schémas conceptuels de bases de données géographiques résultent d'une conceptualisation du monde réel selon un modèle orienté objet (Fonseca et al., 2003). Les classes qui les composent sont nommées à l'aide de termes désignant le plus souvent les catégories d'entités géographiques représentées au sein de chacune d'entre elles, et leurs attributs, à l'aide de termes désignant des propriétés de ces catégories. Or, comme nous l'avons vu au chapitre 2.2.2, les schémas de bases de données géographiques ne visent pas seulement à catégoriser les entités géographiques mais à structurer l'information qui sera représentée dans la base (Partridge, 2002). En particulier, il arrive que des catégories d'entités topographiques n'apparaissent dans le schéma conceptuel de ces bases que sous la forme de valeurs d'attributs. Nous proposons ici de mettre de telles connaissances concernant le contenu précis de la base en évidence lors de l'étape de création des ontologies d'applications (cf. figure 24, numéro 1).

Une pratique de modélisation de bases de données géographiques couramment rencontrée consiste, en effet à représenter, au sein d'une même classe, des entités géographiques appartenant à des catégories différentes, mais néanmoins associées entre elles par des relations de généralisation-spécialisation. L'objectif d'une telle opération est de réduire le nombre de classes de la base de données afin de simplifier sa structure, en particulier dans les cas de catégories d'entités géographiques possédant peu d'entités représentatives sur le territoire couvert par la base. Par exemple, la classe *Zone_végétation* de la BDTOPO© 2.1 renferme à la fois des zones arborées, des forêts fermées de feuillus, des forêts fermées mixtes, des forêts fermées de conifères, des forêts ouvertes, des peupleraies, des haies, des landes ligneuses, des vergers, des vignes, des bois, des bananeraies, des mangroves, ou des plantations de canne à sucre. Ces diverses entités géographiques sont représentées dans une même classe nommée à l'aide d'un terme faisant référence à une catégorie d'entités géographiques plus générique. La catégorie exacte à laquelle appartient chaque instance de la classe est précisée à l'aide d'un attribut généralement appelé *Nature*, dans les bases de données de l'IGN, dont les valeurs possibles sont précisées dans les spécifications de la base. Nous proposons donc ici de mettre à profit ces valeurs d'attributs afin d'expliciter la nature exacte du contenu des classes de schémas que nous souhaitons mettre en correspondance avec l'ontologie du domaine. Pour y parvenir, nous proposons de générer automatiquement à partir des schémas, des ontologies d'applications utilisées ici comme ressources intermédiaires pour décrire en priorité les catégories d'entités géographiques représentées dans les bases.

L'approche que nous proposons pour la création des ontologies d'applications est relativement intuitive. Chaque classe de schéma conceptuel de base de données géographique représente une catégorie d'entités géographiques du monde réel, et porte généralement un nom désignant cette catégorie d'entités. A ce titre, elle peut être directement traduite, au sein de l'ontologie d'application décrivant sa base de données, sous la forme d'un concept, doté d'un label correspondant à son nom. Les propriétés et les relations de ce concept sont directement dérivées des attributs et des associations de cette classe. Les ontologies d'applications que nous proposons de créer sont structurées selon une hiérarchie de subsomption; les concepts les plus spécifiques sont reliés aux concepts plus généraux via des relations de type ISA. Ainsi, si le schéma conceptuel de notre base de données comporte des relations d'héritage entre classes, celles-ci seront traduites en relations de

type ISA entre les concepts dérivés de ces classes. De plus, afin mettre en évidence l'ensemble des catégories d'entités géographiques plus spécifiques représentées au sein de chaque classe, nous allons mettre à profit les attributs destinés à définir leur nature exacte que Manoah et al. (2004) qualifient de « propriétés déterminantes ». Les valeurs énumérées de ces attributs, qui font référence à des labels de concepts topographiques, constituent en effet une source de connaissances importante concernant le contenu exact de chaque classe de la base de données, qui demeure généralement cachée dans la structure de la base, et donc moins facilement exploitable pour les techniques d'appariement automatique de schémas. Afin de rendre ces connaissances plus accessibles au système, nous proposons d'augmenter le niveau de granularité du schéma en traduisant chaque valeur possible de ces attributs sous la forme d'un concept. Ces concepts sont intégrés à la structure globale de l'ontologie d'application en tant que concepts plus spécifiques, raccordés par une relation de type ISA au concept directement dérivé de la classe à laquelle appartient l'attribut dont ils sont issus. Cette étape de mise en évidence de concepts cachés nécessite de connaître les attributs susceptibles de renfermer des valeurs désignant des catégories d'entités géographiques. En première approche, une analyse manuelle des spécifications des bases de données concernées permettra de dresser la liste de ces attributs. Enfin, il convient de conserver, tout au long de ce processus, une trace des relations entre éléments de schémas et éléments d'ontologies dérivés, en vue de l'étape d'appariement des schémas.

Cette approche permet donc de générer de façon semi-automatique, pour chaque schéma, une ontologie d'application visant à décrire en priorité, non pas la structure des données, mais les catégories d'entités géographiques qu'elles représentent.

4.2.1.2 Alignement des ontologies d'applications

L'alignement des ontologies d'application est réalisé en deux temps. Nous suivons, en effet, l'approche proposée par (Aleksovki et al., 2006), qui consiste à compenser d'éventuelles carences des techniques traditionnelles d'alignement lexical et structurel, dues à des différences de terminologie et de structure entre les ontologies à aligner, en s'appuyant sur une ontologie dite de support, comme source de connaissances externe. Celle-ci est supposée offrir une couverture plus large et plus détaillée du domaine que celles fournies par les diverses ontologies d'applications. Ainsi, l'application de techniques d'alignement lexicales et structurelles entre les ontologies d'applications et cette ontologie de support a plus de chances de conduire à la détection d'un grand nombre de relations de correspondances entre les éléments des premières et de la seconde. Ces relations de correspondances – aussi appelées alignements – sont ensuite exploitées pour calculer les alignements entre ontologies d'applications. Si l'on se réfère à la taxonomie des techniques d'appariement automatiques présentée dans le tableau 2, cette approche revient donc à combiner des techniques terminologiques et éventuellement linguistiques, des techniques fondées sur les graphes ou plus simplement sur les relations taxonomiques, à des techniques de réutilisation d'alignements. Dans le cas de notre processus global d'appariement de schémas présenté en figure 24, c'est l'ontologie du domaine qui joue le rôle d'ontologie de support.

La première phase du processus d'alignement, dite phase d'**ancrage**, consiste donc à aligner chaque ontologie d'application avec l'ontologie de support (cf. figure 24, numéro 2). Cette étape peut être réalisée à l'aide de diverses applications d'alignement automatique, utilisant des techniques

d'alignement classiques : techniques terminologiques, techniques linguistiques, techniques fondées sur les relations taxonomiques, techniques fondées sur les graphes, etc. Notre seule contrainte est de pouvoir obtenir en sortie du processus, une liste de relations de correspondances précisant, pour chaque paire de concepts mis en relation, le type de relation qui les lie et une valeur indiquant le degré de fiabilité de la relation de correspondance. Pour les besoins de notre approche, nous avons choisi d'utiliser un processus d'alignement fondé sur une combinaison de techniques d'alignement d'ontologies terminologiques et de techniques fondées sur les relations taxonomiques, élaborée dans le cadre du projet GéOnto (ANR-O7-MDCO-005). Ce processus d'alignement est orienté ; il vise à détecter des relations de correspondance entre les concepts d'une ontologie source (O_{source}) et ceux d'une ontologie cible (O_{cible}). Pour chaque paire de concepts sources et cibles des relations d'équivalence (*isEq*), de généralisation-spécialisation (*isGeneral* ou *isA*) ou de proximité sémantique (*isClose*) peuvent être détectées. En outre, à chaque paire de concepts alignés est attribué un score de similarité permettant d'évaluer la fiabilité de la relation de correspondance établie par le système. Ce processus est décrit en détail dans (Hamdi et al. 2008), et présenté en annexe 1.

La seconde phase du processus d'alignement, dite phase de **dérivation** (cf. figure 24, numéro 3), vise à détecter des relations de correspondance entre les concepts des ontologies d'applications. Cette tâche est réalisée par analyse conjointe des relations de correspondance établies entre les concepts des ontologies d'applications et ceux de l'ontologie de support lors de l'étape d'ancrage, et les relations structurelles existant entre les concepts de l'ontologie de support. Ainsi, disposant d'alignements typés entre ontologies d'applications et une ontologie du domaine de la topographie, les relations entre concepts des ontologies d'applications peuvent être déduites de la façon suivante :

- Si deux concepts, C_1 et C_2, issus de deux ontologies d'applications O_1 et O_2, possèdent chacun une relation de correspondance avec un même concept C_s de l'ontologie de support et que la relation de correspondance entre C_1 et C_s est du type (C_1 *isEq* C_s), alors une relation de correspondance est générée entre C_1 et C_2. Cette relation est du même type que celle créée lors de la phase d'ancrage entre C_2 et C_s.

- Si deux concepts, C_1 et C_2, issus de deux ontologies d'applications O_1 et O_2, possèdent chacun une relation de correspondance avec un même concept C_s de l'ontologie de support et que les relations de correspondance entre C_1 et C_s et C_2 et C_s sont du même type, mais ne sont pas des relations d'équivalence, alors une relation de type (C_1 *isClose* C_2) est générée.

- Si deux concepts, C_1 et C_2, issus de deux ontologies d'applications O_1 et O_2, possèdent chacun une relation de correspondance de type *isEq* avec des concepts de l'ontologie de support, respectivement C_{s1} et C_{s2}, alors les relations structurelles entre C_{s1} et C_{s2} sont analysées. Si C_{s2} subsume C_{s1}, alors une relation de type (C_1 *isA* C_2) est générée. Si C_{s1} subsume C_{s2}, alors une relation de type (C_1 *isGeneral* C_2) est générée.

A chacune des relations de correspondance ainsi déduites est associé un score égal à la moyenne des scores associés aux relations de correspondance issues de la phase d'ancrage à l'aide desquelles celles-ci ont pu être calculées. En première approche, celui-ci permet de disposer d'une estimation de la fiabilité des relations de correspondance calculées. Cependant, il serait probablement préférable de pondérer cette moyenne en fonction des types de relations de correspondance comparées, afin d'obtenir une estimation plus fine de la similarité des concepts mis en relation.

Ainsi, ce processus permet de générer des alignements entre ontologies d'applications, deux à deux, qui vont ensuite être exploités lors de la phase d'appariement des schémas conceptuels des bases de données hétérogènes.

4.2.1.3 Appariement des schémas conceptuels

L'appariement des schémas hétérogènes de bases de données géographiques est enfin réalisé, à l'aide des alignements calculés entre ontologies d'applications et des relations entre éléments de schémas et éléments d'ontologies d'applications dérivées de ces schémas conservées lors de l'étape de création de ces ontologies (cf. figure 24, numéro 4). Les relations de correspondance déduites entre éléments de schémas sont considérées comme du même type que les relations entre concepts d'ontologies d'applications dont elles sont issues, et se voient attribuer le même score de similarité. Ces types et ces scores pourront être utilisés comme indicateur de la fiabilité des relations d'appariement déduites entre éléments de schémas. En effet, une relation d'appariement sera logiquement considérée comme plus valide dans la mesure où elle aura été déterminée sur la base d'une relation d'équivalence plutôt que sur celle d'une relation de proximité sémantique.

Pour chaque relation de correspondance entre concepts d'ontologies d'applications, les éléments de schémas dont proviennent ces concepts sont appariés de la façon suivante :

- Si les deux éléments de schémas dont proviennent les deux concepts mis en relation lors de l'alignement des ontologies d'applications sont des classes, alors une relation d'appariement entre ces classes est générée. Ceci implique donc que ces deux classes de schémas hétérogènes sont considérées comme représentant des catégories d'entités géographiques associées par une relation du même type que celle attribuée à la relation d'appariement.

- Si les deux éléments de schémas dont proviennent les deux concepts mis en relation lors de l'alignement des ontologies d'applications sont des valeurs d'attributs désignant des concepts topographiques, alors une relation d'appariement entre ces valeurs d'attributs est générée. Ceci implique donc que les instances des deux classes de schémas, dont les valeurs d'attributs sont ainsi appariées, et qui possèdent ces valeurs d'attributs spécifiques, sont considérées comme représentant des catégories d'entités géographiques associées par une relation du même type que celle attribuée à la relation d'appariement.

- Si l'un des éléments de schémas dont proviennent les deux concepts mis en relation lors de l'alignement des ontologies d'applications est une classe, et l'autre une valeur d'attribut, alors une relation d'appariement entre cette classe et cette valeur d'attribut est générée. Ceci implique donc que la classe du premier schéma, et les instances de la classe du deuxième schéma, dont la valeur d'attribut a été appariée, et qui possèdent cette valeur d'attribut spécifique, sont considérées comme représentant des catégories d'entités géographiques associées par une relation du même type que celle attribuée à la relation d'appariement.

4.2.2 Instanciation du modèle

La mise en œuvre du processus d'appariement automatique de schémas de bases de données géographiques présenté ci-dessus nécessite l'instanciation des composantes de base du modèle présenté en figure 24. Celui-ci comprend, en effet, pour chaque schéma de base de données à intégrer, une ontologie d'application destinée à décrire les catégories d'entités géographiques représentées au sein de la base. Nous avons proposé dans la partie 4.2.1.1 une approche pour l'extraction des connaissances nécessaires à la création d'une ontologie à partir du schéma et des spécifications de la base de données concernée. Nous présentons ici une mise en œuvre de cette approche pour la génération semi-automatique des ontologies d'applications associées à deux bases de données topographiques vectorielles produites par l'IGN, la BDTOPO© Pays 1.2 d'une part, et la BDCARTO© 3.0 d'autre part. De plus, l'approche d'alignement par ontologie de support adoptée implique de disposer d'une ontologie du domaine, offrant une couverture plus large et plus détaillée du domaine que celles fournies par les diverses ontologies d'applications. L'approche d'alignement par ontologie de support que nous souhaitons mettre en œuvre ici ne s'attachant à établir de relations de correspondance qu'au niveau des concepts des ontologies à aligner, sans se préoccuper de traiter leurs propriétés ni leurs relations, cette ontologie de support pourra se cantonner à une simple taxonomie, pour peu que cette dernière présente une couverture suffisamment riche du domaine. Afin de disposer rapidement d'une ressource de support nous permettant de mener à bien notre processus d'appariement de schémas, nous nous sommes donc résolue à utiliser une taxonomie, dont le processus de création est présenté dans la partie 4.2.2.2.

4.2.2.1 Création des ontologies d'applications

L'étape de création des ontologies d'applications mises en œuvre dans cette première approche d'appariement automatique de schémas de bases de données géographiques est effectuée de façon semi-automatique. A ces fins, un traducteur générique, prenant en entrée un schéma de base de données géographique, ainsi que la liste des attributs susceptibles de renfermer des labels de concepts topographiques, et produisant en sortie l'ontologie d'application associée à ce schéma, a été développé. Dans un souci de cohérence avec notre objectif général d'interopérabilité, ce traducteur utilise les normes et standards en vigueur pour la description des schémas de bases de données géographiques et pour la représentation d'ontologies. La traduction des schémas conformes au modèle « ISO 19109 - Rules for application schema » en ontologies au format OWL est effectuée conformément à l'approche présentée dans la partie 4.2.1.1.

Le traducteur chargé de la conversion des schémas vers des ontologies d'applications est implémenté dans le langage Java, au sein de la plateforme de système d'information géographique développée au laboratoire COGIT, GéOxygène. Celle-ci implémente de nombreuses normes dédiées à la représentation de l'information géographique; c'est le cas notamment de la norme « ISO 19109 - Rules for application schema » que nous utilisons ici. La représentation et la manipulation en Java des ontologies au format OWL est réalisée à l'aide de l'API protégé-owl[35]. A un schéma ISO 19109 correspondra une ontologie d'application au format OWL.

[35] http://protege.stanford.edu/plugins/owl/api/

Suivant l'approche proposée dans la partie 4.2.1.1, les instances de la classe FeatureType, qui représentent des classes du schéma ISO 19109 en entrée, sont donc converties en instances de la classe OWLNamedClass définie dans l'API protégé-owl. De la même façon, les instances de la classe AttributeType sont directement traduites en instances de la classe OWLDatatypeProperty. Celles des classes AssociationType et AssociationRole donnent naissance à des instances d'OWLObjectProperty, dotées de domaines et de rangs correspondant aux instances de la classe OWLNamedClass issues des instances de FeatureType définies respectivement comme domaines et rangs de ces instances d'AssociationType au sein du schéma d'origine. Enfin, les instances de la classe InheritanceRelation sont utilisées pour définir des relations de type subClassOf entre les instances de la classe OWLNamedClass issues des instances de FeatureType définies comme possédant une relation d'héritage au sein du schéma.

De plus, conformément à notre proposition pour augmenter le niveau de granularité des schémas, les instances de la classe AttributeType sont testées afin de déterminer si leur label « memberName » figure dans la liste des attributs susceptibles de renfermer des labels de concepts topographiques fournie en entrée. Le cas échéant, ces instances ne sont pas converties en instances de la classe OWLDatatypeProperty ; les instances de la classe FeatureAttributeValue associées à l'une de ces instances particulières d'AttributeType sont converties en instances de la classe OWLNamedClass dotées d'une relation de type subClassOf avec l'instance de la classe OWLNamedClass issue de l'instance de FeatureType associée à cette instance d'AttributeType.

La figure 25 présente, sous forme graphique, un extrait d'un schéma en entrée du processus de traduction, l'ontologie résultante et les relations entre éléments de schéma et éléments d'ontologie dérivés de ces éléments de schéma. Le diagramme d'objets situé en bas de la figure correspond à un extrait du schéma de la BDCARTO© 3.0 décrit conformément à la norme ISO 19109. Le haut de la figure correspond à l'ontologie créée à partir de ce schéma et visualisée à l'aide de l'éditeur d'ontologies Protégé. Les flèches matérialisent les opérations de traduction effectuées. Ainsi, les instances de FeatureType, *Zone d'habitat* et *Commune,* sont transformées en classes de mêmes labels au sein de l'ontologie OWL produite. L'instance de la classe AssociationType, *Chef-lieu*, ainsi que les instances d'AssociationRole qui lui sont associées, *Est chef lieu de* et *A pour chef lieu*, qui relient les instances de FeatureType *Commune* et *Zone d'habitat*, sont transformées en OWL en ObjectProperty (et sa propriété inverse) reliant les classes OWL COMMUNE et ZONE_D__HABITAT. L'instance de la classe AttributeType *Importance*, associée à l'instance de FeatureType *Zone d'habitat*, est définie comme possédant trois valeurs possibles, représentées ici par les instances de FeatureAttributeValue *Hameau*, *Quartier de ville* et *Chef-lieu de commune*. Ces valeurs faisant référence à des labels de concepts topographiques, elles sont traduites sous la forme de classes OWL de ces mêmes labels. Ces classes OWL sont ensuite associées à la classe OWL, ZONE_D__HABITAT, via une relation de type subClassOf. Les autres instances de la classe AttributeType, ne faisant pas référence à des labels de concepts topographiques, sont traduites en OWL par des DatatypeProperties (cas non représenté sur cette figure).

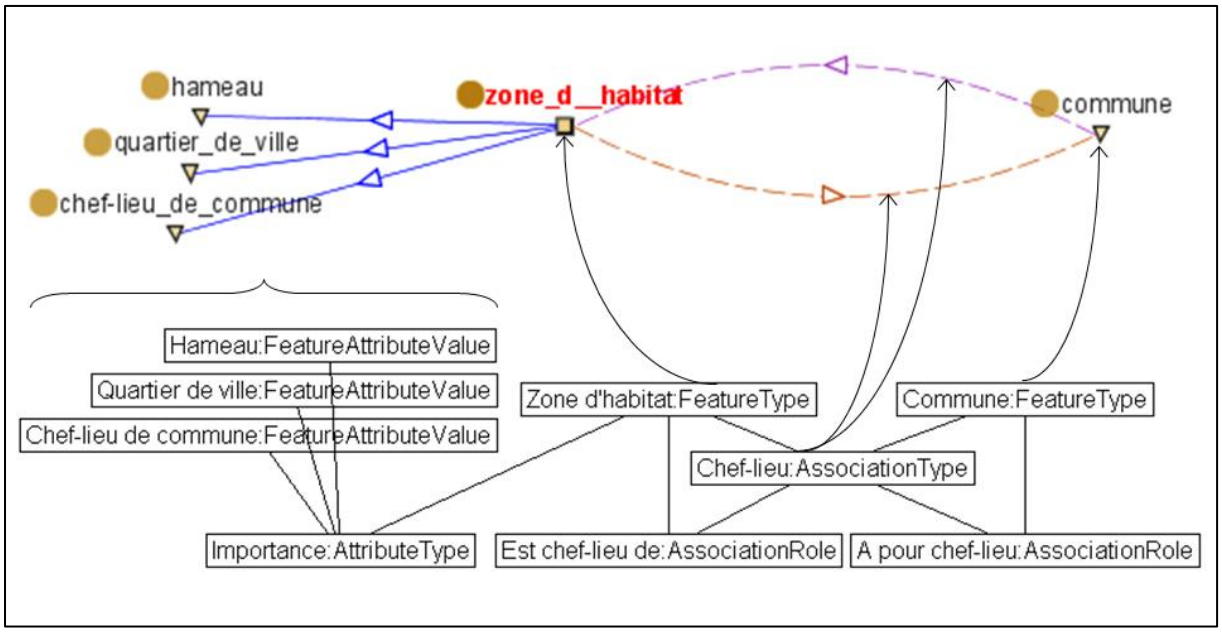


Figure 25: Traduction d'un extrait du schéma de la BDTOPO Pays 1.2 représenté selon le schéma ISO 19109 (en bas de l'image), en ontologie au format OWL (en haut de l'image, visualisation sous Protégé)

4.2.2.2 Création de l'ontologie de support

L'approche d'alignement par ontologie de support que nous souhaitons mettre en œuvre implique de disposer d'une ontologie du domaine de la topographie. Pour remplir son office de source de connaissances de support, celle-ci doit offrir une couverture plus large et plus détaillée du domaine que celles fournies par les diverses ontologies d'applications. Par ailleurs, dans la mesure où nous travaillons sur des bases de données françaises, les schémas décrivant la structure des données que nous cherchons à intégrer, et donc les ontologies d'applications qui en sont dérivées, sont en langue française ; l'ontologie du domaine à mettre en œuvre devra donc disposer de labels en français. Or, le bilan de l'existant, en matière d'ontologies francophones du domaine de la topographie, présenté au chapitre 3.2.1, fait état de trop peu de ressources existantes de ce type pour couvrir les besoins de notre application. C'est pourquoi, nous proposons ici une approche semi-automatique pour la création d'une ontologie francophone du domaine de la topographie, mettant à profit les textes des spécifications des bases de données de l'IGN. Aucune méthodologie de création d'ontologies particulière ne semblant faire consensus (chapitre 3.2.2), nous avons adopté une approche spécialement adaptée à nos besoins et aux ressources dont nous disposons. Celle-ci se compose d'activités, que l'on retrouve dans le processus global de développement d'ontologies défini dans le cadre de la méthodologie de création d'ontologies METHONTOLOGY (Gómez-Pérez et al., 1996), essentiellement parmi les activités de développement de l'ontologie et les activités de support. Notre approche a été mise en œuvre sur les spécifications de deux bases de données produites par l'IGN, la BDTOPO© et la BDCARTO©, et est également décrite dans (Abadie et Mustière, 2010).

Approche pour la création d'une ontologie francophone du domaine de la topographie

Nous souhaitons créer une ontologie francophone du domaine de la topographie, permettant la mise en œuvre de notre approche d'intégration virtuelle de bases de données topographiques

vectorielles. Ainsi, cette ontologie a pour objectif de représenter les catégories d'entités topographiques existant sur le territoire français, leurs propriétés et leurs relations, afin de permettre la description de la sémantique du contenu de bases de données topographiques vectorielles. Cependant, l'approche d'alignement par ontologie de support que nous souhaitons mettre en œuvre ne nécessitant pas de disposer d'une ontologie dotée de propriétés ou de relations sémantiques non hiérarchiques, en première approche, nous pourrons nous cantonner à créer une taxonomie. Nous nous intéressons en premier lieu aux bases de données topographiques produites par l'IGN. Cependant, l'intégration de bases de données topographiques d'autres producteurs, avec les bases de données de l'IGN est également envisagée, dans les limites de la compatibilité de leurs niveaux de détail respectifs. En termes de domaine à couvrir, cela signifie que notre ontologie doit comporter les catégories d'entités topographiques utiles à la description du contenu de bases de données topographiques, dotées d'une couverture thématique du domaine et d'une échelle caractéristique proches de celles de la BDTOPO©, la base de données topographique de référence dont nous disposons. Elle devra donc couvrir les thèmes du routier, de l'hydrographie, du bâti, de l'occupation du sol, etc. sans pour autant rechercher l'exhaustivité au sein de chacun de ces thèmes: compte tenu de l'échelle caractéristique des bases de données que nous traitons, les catégories auxquelles appartiennent des entités topographiques dont la taille est inférieure à la résolution de ces bases ne sont pas indispensables ici. Ainsi, les concepts désignant des éléments de mobilier urbain, comme ceux de PLAQUE D'ÉGOUT ou d'ABRIBUS pourront ne pas figurer dans cette ontologie. En revanche, dans la mesure où nous envisageons également son utilisation dans le cadre de notre approche d'intégration de données fondée sur des connaissances issues des spécifications des diverses bases à intégrer, elle devra comporter des catégories d'entités topographiques dont ces bases ne font pas explicitement mention, mais qui sont néanmoins citées dans leurs spécifications. C'est le cas par exemple des PUITS, dans la BDTOPO© 2.1. Ce terme n'apparaît jamais dans le schéma de la base de données alors que les spécifications précisent que ceux-ci sont représentés au sein de la base comme instances de la classe *Point_Eau* avec comme valeur d'attribut *Nature* « *Autre point d'eau* ».

La création d'une telle ontologie semble raisonnement réalisable dans la mesure où les spécifications des bases de données de l'IGN nous fournissent une importante source de connaissances, déjà partiellement structurées, sur le domaine. Rédigées par un producteur de données de référence, elles se veulent le fruit d'un consensus sur le monde topographique. Destinées, en premier lieu, à garantir un niveau d'homogénéité satisfaisant au sein des bases de données, elles visent à réduire autant que faire se peut les ambiguïtés concernant l'interprétation du contenu des bases. Aussi le vocabulaire utilisé pour y décrire les catégories d'entités du monde réel est-il relativement générique. De plus, les spécifications des bases de données font référence à un très grand nombre de concepts topographiques, à travers les noms de classes, les valeurs d'attributs, les parties de « regroupement » énumérant les catégories d'entités géographiques devant être représentées dans la base, ou de manière plus informelle au milieu de phrases en langage naturel. Disposer de l'ensemble de ces concepts au sein d'une ontologie du domaine de la topographie nous garantirait donc un bon niveau d'exhaustivité, vis-à-vis de notre application. C'est pourquoi, Gesbert (2005) préconise déjà de parcourir manuellement les textes des spécifications afin d'y identifier les concepts topographiques permettant de constituer l'ontologie du domaine nécessaire à la mise en œuvre de son modèle. Enfin, la structure même de ces spécifications est source d'informations concernant notamment les relations entre les concepts topographiques rencontrés au fil du texte. En effet,

celles-ci sont rédigées conformément au schéma de la base de données qu'elles décrivent : à chaque classe du schéma correspond une fiche de spécifications qui définit en premier lieu la classe, puis ses attributs, et enfin, au sein de chaque section décrivant un attribut, les valeurs énumérées que peut prendre ce dernier, et les regroupements de catégories d'entités topographiques effectués. Ces différentes parties du texte sont délimitées par des titres dont la hiérarchie peut être mise à profit, selon une approche proche de celle adoptée pour l'explicitation des concepts cachés dans le processus de création des ontologies d'applications, pour définir des relations généralisation-spécialisation entre concepts dont les labels figurent dans l'une ou l'autre de ces parties de texte. La figure 26 illustre cette approche. Elle présente la fiche de spécifications de la classe *Point d'eau* de la BDTOPO© Pays 1.2. Sur cette fiche sont identifiés, en jaune, des labels de concepts topographiques pouvant être retenus afin de définir des concepts topographiques au sein de notre ontologie du domaine. La hiérarchie de subsomption entre ces concepts est directement dérivée de celle des titres des parties de texte dont sont issus leurs labels respectifs.

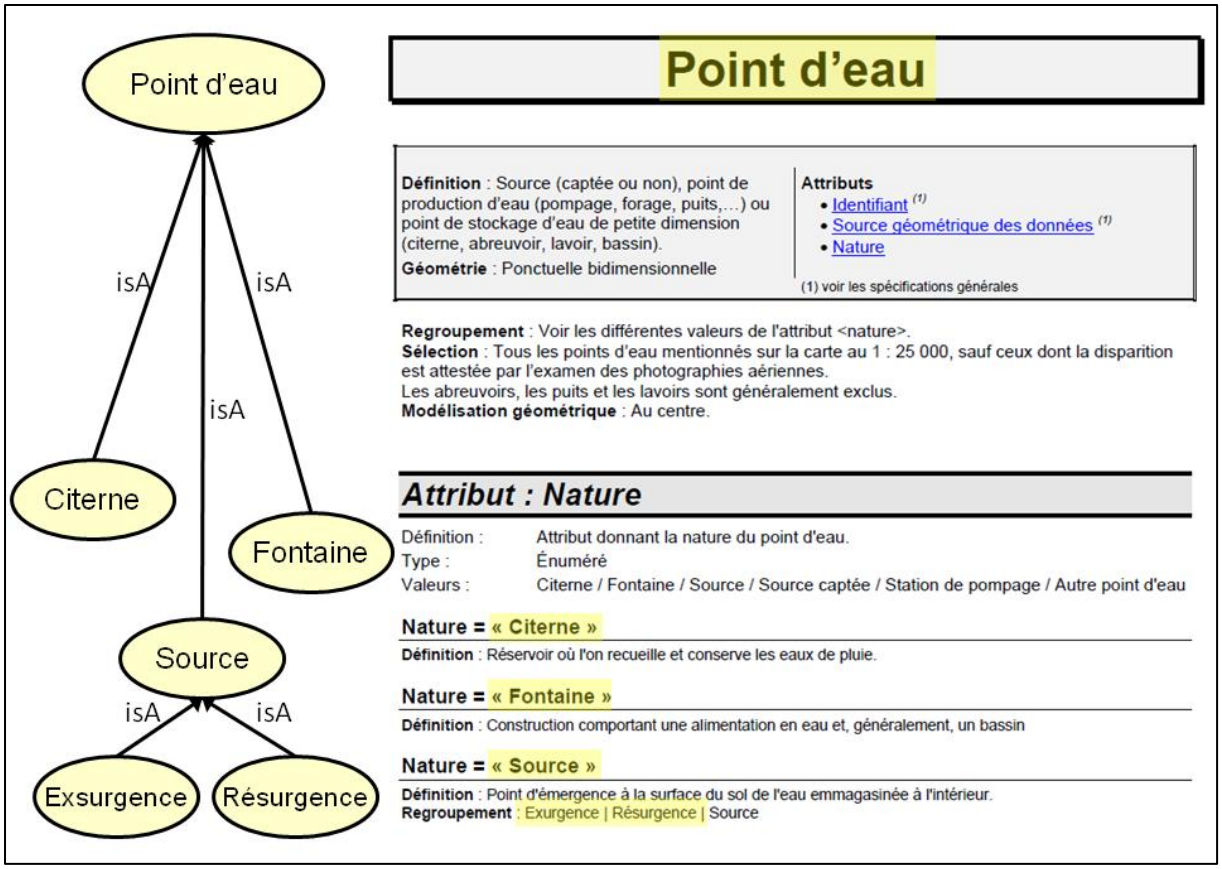


Figure 26: Approche proposée pour la définition et la structuration des connaissances au sein de l'ontologie du domaine de la topographie

La tâche de définition et de structuration des connaissances du domaine peut donc être considérée comme déjà partiellement réalisée au travers des spécifications dont nous disposons. Cependant, l'acquisition, la formalisation et la représentation manuelles de ces connaissances dans un langage de représentation de connaissances s'avérant particulièrement longues, nous proposons de recourir à des techniques de traitement automatique du langage naturel (TALN) afin de les automatiser. Compte tenu de l'organisation particulière des textes des spécifications, une telle approche relève de l'acquisition de connaissances à la fois à partir de corpus textuels et à partir de schémas de bases de

données. Il s'agit donc, dans un premier temps, de retrouver dans les textes les concepts topographiques, puis de reconstituer automatiquement la hiérarchie de ces concepts à partir de leur localisation au sein des textes. Pour ce faire, nous privilégions le recours à des méthodes par règles. En effet, les spécifications de ces deux bases sont des textes relativement courts en termes de TALN, de l'ordre d'une centaine de pages pour chaque base, au sein desquels certains concepts topographiques n'apparaissent que rarement. Elles s'avèrent donc peu propices à l'utilisation de méthodes statistiques et probabilistes de TALN. En revanche, il s'agit de documents extrêmement organisés. C'est pourquoi le choix a été fait de mettre à profit la structure même des documents, et d'employer des heuristiques linguistiques, afin de détecter les concepts topographiques.

Les spécifications des deux bases de données sont traitées dans un premier temps indépendamment, avant de fusionner les ontologies ainsi produites. Cette tâche de fusion des résultats est réalisée à l'aide d'outils d'alignement et de fusion d'ontologies.

Mise en œuvre de l'approche proposée

La mise en œuvre de notre approche pour la création d'une ontologie de concepts topographiques à partir des textes des spécifications de bases de données de l'IGN suppose celle de deux tâches de support essentielles. Les outils d'acquisition semi-automatique de connaissances à partir des textes des spécifications, d'une part, ont été implémentés et appliqués tour à tour pour deux bases de données topographiques de l'IGN, la BDTOPO© Pays 1.2 et la BDCARTO© 3.0 dans le cadre d'un stage (Laurens, 2006). L'analyse des deux taxonomies obtenues, leur post-traitement manuel et leur fusion semi-automatique ont été réalisés ultérieurement.

Acquisition semi-automatique de connaissances par traitement automatique du langage naturel

La figure 27 présente l'enchaînement des traitements appliqués aux textes des spécifications pour l'acquisition semi-automatique de connaissances. Une première étape consiste à structurer le corpus de base. Ainsi, dans un premier temps, le fichier au format *.doc des spécifications est converti en fichier *.html. En effet, ce format présente l'avantage de fournir explicitement les paramètres de mise en forme du texte des spécifications. Ainsi, un premier parseur qui s'appuie sur les styles définis au sein des balises de ce document html a été développé afin d'identifier les titres, sous-titres, parties en gras ou soulignées ou encore simples parties de texte des spécifications, en fonction de leur rendu visuel au sein des spécifications (voir figure 27). On obtient ainsi une liste de lignes typées selon la classification suivante : « nom de champ », « nom de classe », « nom d'attribut », « nom de valeur d'attribut », etc. Les lignes de même type qui se suivent sont ensuite encapsulées dans un même conteneur. En outre, des types différents mais jouant des rôles similaires dans le texte des spécifications sont renommés et fusionnés. Enfin, la hiérarchie du texte est reconstituée à l'aide de ces types. Ce processus se fait de façon ascendante : les conteneurs de type « enfant » qui suivent directement un conteneur de type « parent » sont directement ajoutés dans ce dernier. Dans un deuxième temps, un second parseur opère une sélection dans le fichier *.xml issu du premier parseur afin de ne conserver que les conteneurs utiles pour l'analyse. En effet, nous cherchons ici à extraire des labels de concepts géographiques. Une lecture rapide des spécifications montre que certaines portions de texte bien définies n'en comportent aucun. Celles-ci sont donc supprimées. Enfin, une nouvelle hiérarchie, comportant moins de balises, mais typée plus précisément est créée (voir figure 28).

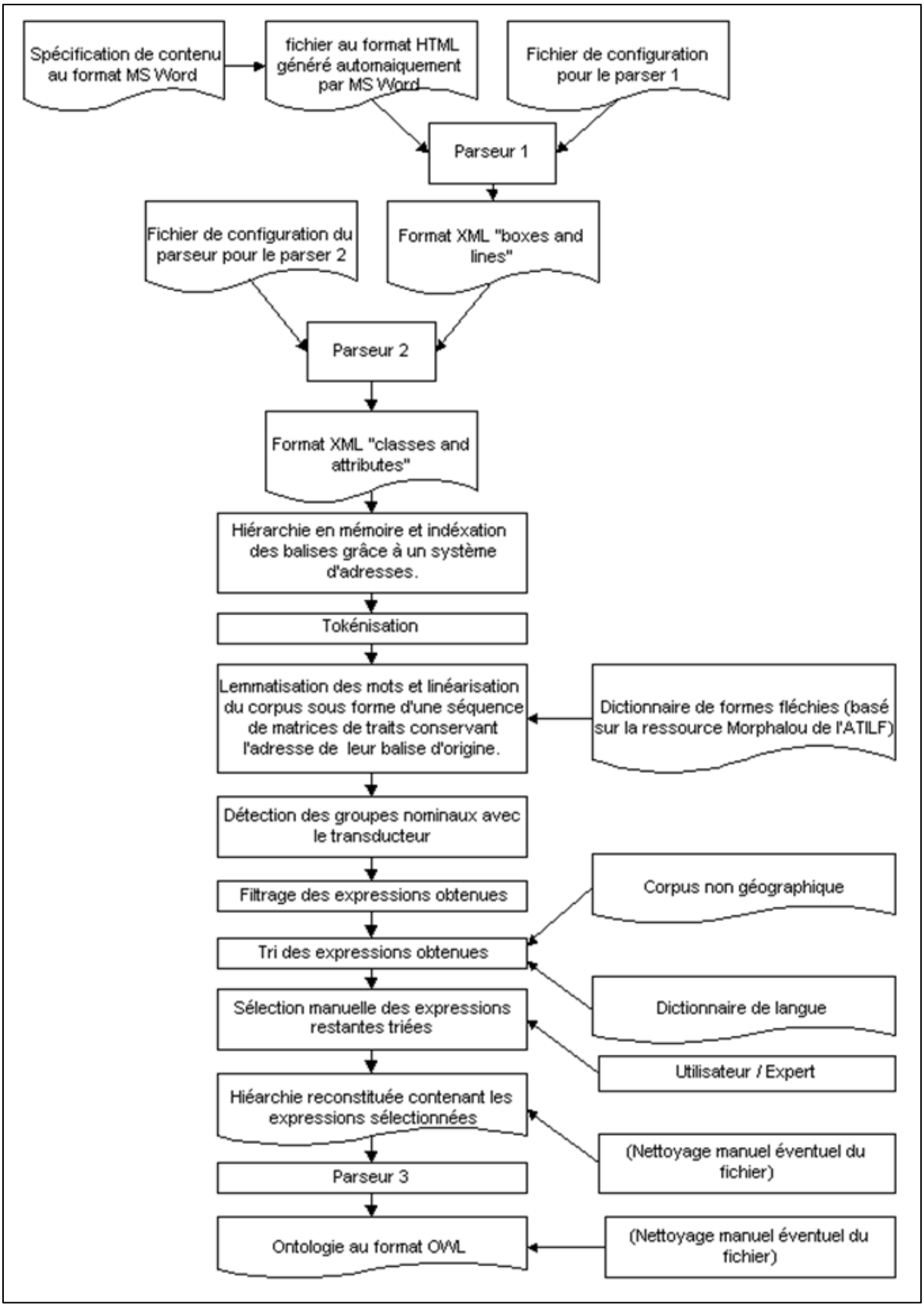


Figure 27: Enchaînement des traitements mis en œuvre pour la constitution d'une ontologie de concepts topographiques à partir des spécifications de base de données topographiques

```
- <class name="C - Transport d'énergie et de fluides">
    <class name="Ligne électrique" />
    <class name="Poste de transformation" />
  - <class name="Canalisation">
      <class name="Nature = « Eau" definition="Définition : Aqueduc
        (ouvrage maçonné et couvert destiné à transporter l'eau potable
        suivant une faible pente) ou conduite forcée (canalisation
        permettant le transfert d'eau en charge (gravitaire et sous
        pression) vers un ouvrage hydraulique). Regroupement : Aqueduc
        | Conduite forcée | Galerie d'amenée d'eau" />
      <class name="Nature = « Autre" definition="Définition : Canalisation
        ou tapis roulant utilisé pour le transport de matière première
        (gaz, hydrocarbure, minerai, etc.) ou canalisation de nature
        inconnue. Regroupement : Conduite de matière première |
        Gazoduc |Oléoduc | Pipe-line | Tapis roulant industriel" />
    </class>
  - <class name="Pylône">
      <class name="Regroupement : Pylône | Portique Sélection :
        Uniquement les pylônes et portiques soutenant des lignes de 63
        KV et plus. Modélisation géométrique : A l'axe et en haut du
        pylône." />
    </class>
  </class>
```

Figure 28: Structuration XML des spécifications de la BDTOPO© Pays 1.2 réalisée par les parseurs 1 et 2, pour permettre l'extraction des termes géographiques

De plus, on émet l'hypothèse selon laquelle les concepts géographiques recherchés seront systématiquement représentés dans le texte par des groupes nominaux. Ainsi, une hiérarchie est créée en mémoire, en indexant chacune des balises du fichier *.xml présenté en figure 28 à l'aide d'un système d'adresses permettant de retrouver la localisation du contenu de chaque balise dans le texte. En effet, le corpus est ensuite transformé en une suite de matrices de traits et sans ce système d'adresses toute information concernant la structure du texte serait perdue. La construction des matrices de traits consiste à associer à chaque élément du texte divers attributs utiles pour l'analyse : son adresse dans le texte, sa forme, son lemme, sa catégorie grammaticale, son genre, son nombre, etc. A l'étape suivante (voir figure 27), un transducteur procède à l'extraction des groupes nominaux issus de cette suite de matrice de traits. Celui-ci procède par règles en recherchant dans le texte des expressions de type « nom » ou « nom + adjectif qualificatif » ou « nom + 'de' + nom », etc. Cependant, la sélection de groupes nominaux effectuée par le transducteur demeure trop large, et doit être filtrée et triée. Ce filtrage est réalisé par comparaison avec un corpus non géographique, constitué de textes relatifs à des domaines très éloignés, comme la chimie ou la mécanique. D'autre part, un dictionnaire de langues permettra de trier cette sélection. Ainsi les groupes nominaux qui semblent être les plus pertinents seront présentés en priorité, pour validation, à un expert.

Enfin, les relations de subsomption entre concepts topographiques sont dérivées à partir de la distribution hiérarchique des groupes nominaux retenus dans le texte des spécifications, accessible grâce aux adresses des éléments constitutifs des groupes nominaux ainsi sélectionnés. Ainsi, au sein de la fiche de spécifications d'une classe donnée, un concept topographique désigné par un groupe nominal issu de la partie « nom de classe » sera considéré comme plus général qu'un concept topographique qualifié par un groupe nominal extrait d'une partie « définition » de cette fiche de spécifications. La hiérarchie de concepts ainsi obtenue est finalement convertie au format standard préconisé par le W3C pour la représentation d'ontologies, OWL.

Résultats obtenus et analyse des limites de l'approche

Ces traitements ont été appliqués avec succès aux spécifications de la BDTOPO© Pays 1.2 et de la BDCARTO© 3.0 afin d'en extraire deux taxonomies au format OWL. En effet, à partir des spécifications de la BDTOPO©, nous avons obtenu une taxonomie de 580 concepts topographiques, répartis sur quelques 11 niveaux hiérarchiques. Un extrait en est présenté en figure 29. Le même processus appliqué aux spécifications de la BDCARTO© a fourni une taxonomie de 505 concepts organisés sur 5 niveaux de hiérarchie (voir figure 30).

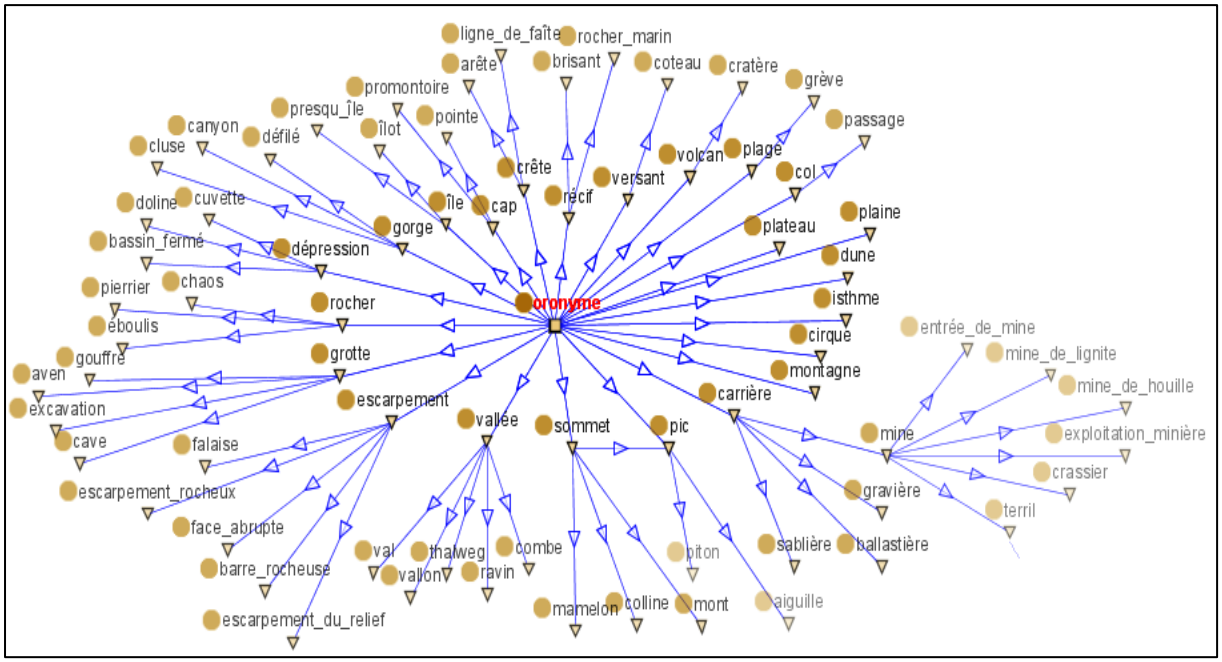


Figure 29: Taxonomie des éléments remarquables du relief obtenue par analyse automatique des spécifications de la BDTOPO© Pays 1.2 (visualisation sous Protégé)

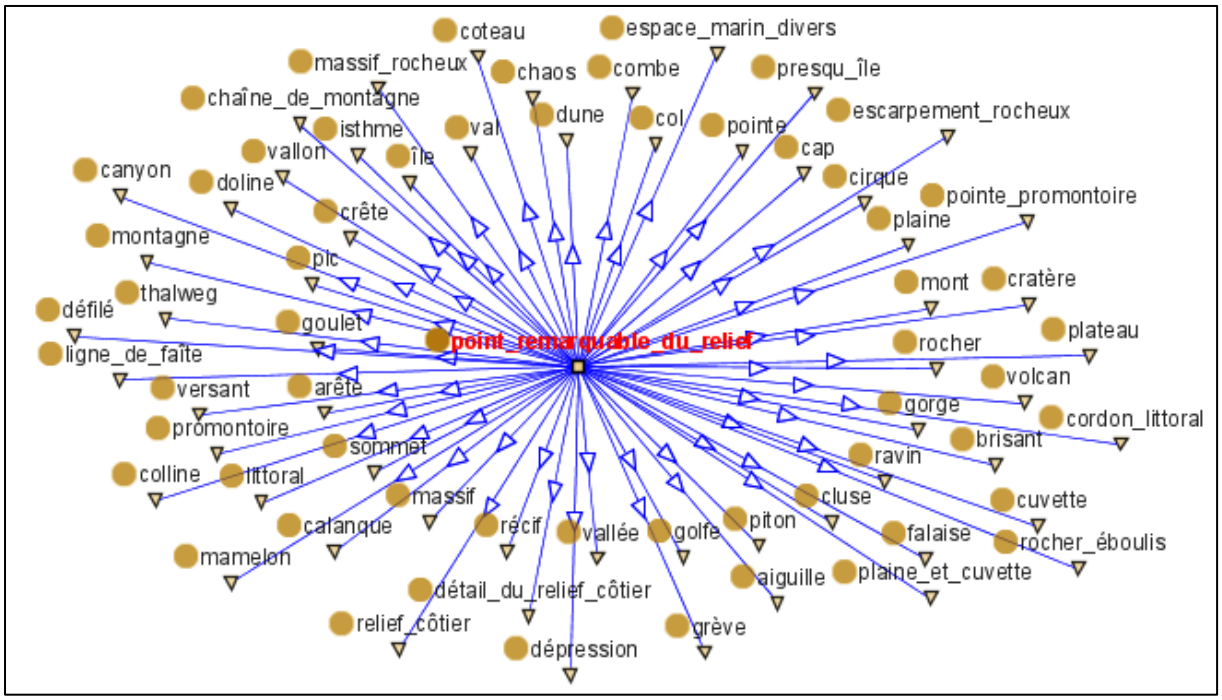


Figure 30: Taxonomie des éléments remarquables du relief obtenue par analyse automatique des spécifications de la BDCARTO© 3.0 (visualisation sous Protégé)

Cependant, en dépit de ces résultats satisfaisants, un post-traitement manuel de l'ordre d'une journée reste nécessaire afin de pallier les manques de l'analyse automatique. En effet, les concepts pertinents peuvent se rencontrer à tous les niveaux de la structure des spécifications - noms de classes, définitions, valeurs d'attributs ou définitions d'attributs, ce qui rend leur détection parfois aléatoire. Ainsi, la partie « définition » des spécifications de la BDTOPO© stipule que « les petits bâtiments isolés (…) de plus de 20 m^2 sont inclus, alors que les petits bâtiments situés en ville ne le sont pas (…) ». Le concept de VILLE va donc être identifié ici, à tort, comme un sous-concept de BÂTIMENT.

De plus, certains mots-clés obtenus traduisent une conceptualisation du monde réel trop proche des préoccupations de modélisation des bases de données géographiques vectorielles. Ainsi, les concepts de CONSTRUCTION SURFACIQUE, CONSTRUCTION LINÉAIRE, et CONSTRUCTION PONCTUELLE devront être fusionnés au profit du concept unique de CONSTRUCTION.

En outre, du fait de la structure même des spécifications, certaines hiérarchies de concepts sont mises à plat, et il est impossible de les reconstituer automatiquement. Ainsi, l'attribut *Nature* de la classe *Point remarquable du relief* pouvant prendre les valeurs *Dune*, *Grotte*, *Sommet* ou *Pic*, l'analyse automatique des spécifications en déduit la hiérarchie de concepts (DUNE, GROTTE, SOMMET, PIC) ISA (POINT REMARQUABLE DU RELIEF) sans que la proximité sémantique entre les deux concepts de PIC et de SOMMET ne soit particulièrement mise en valeur.

Par ailleurs, certains concepts ne peuvent être généralisés sans recourir à un sous-concept représentatif. Le concept de GORGE se spécialise donc à la fois en CANYON, CLUSE, DÉFILÉ et GORGE, faute de terme général plus approprié.

De surcroît, si dans certains cas les sous-concepts extraits sont de réelles spécialisations – c'est le cas pour (COMBE, RAVIN, THALWEG, etc.) ISA (VALLÉE) - il s'agit parfois de synonymes, comme c'est le cas pour la hiérarchie HÔTEL DE VILLE ISA MAIRIE, ou de véritables contresens dans l'exemple suivant (ÎLOT, PRESQU'ÎLE) isA Île

D'autre part, pour des raisons cartographiques, nombre de concepts géographiques sont représentés au sein des bases via leur seul toponyme. Ainsi, les concepts d'EMBOUCHURE, de PÊCHERIE, de LAC, d'AMER et de COURS D'EAU se côtoient indifféremment au sein de la classe *Hydronyme*, sans qu'aucune distinction liée à leurs natures diverses ne soit faite. Cette hiérarchie se retrouve donc nécessairement au sein de la taxonomie issue du traitement automatique des spécifications : les concepts d'EMBOUCHURE, de PÊCHERIE, de LAC, d'AMER et de COURS D'EAU y sont considérés comme des spécialisations du concept d'HYDRONYME, alors même que ce dernier désigne non pas un type d'entités géographiques, mais un label décrivant une entité géographique relative à l'eau (voir extrait en figure 31).

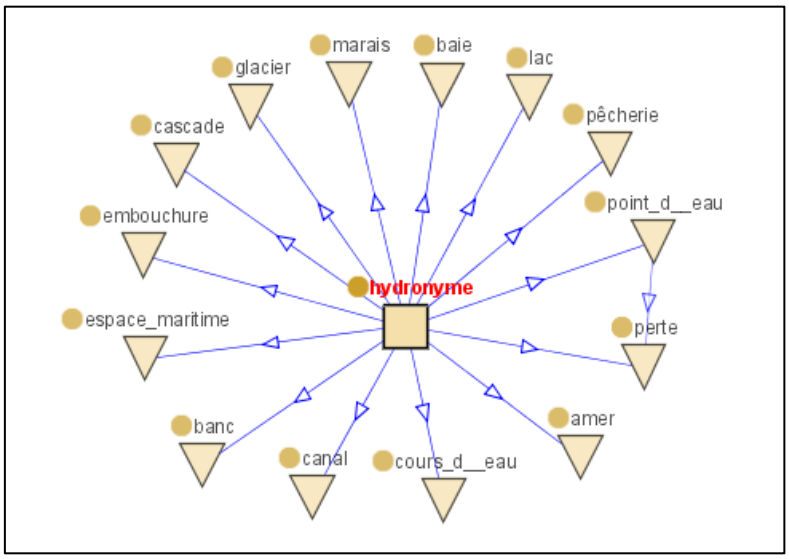


Figure 31: Premier niveau de la hiérarchie obtenue pour les hydronymes

Enfin, les traitements effectués ici ne permettent de générer que des relations de subsomption entre concepts géographiques. Aucune autre forme de relation entre concepts géographiques, telles les relations topologiques, ne figure donc dans notre taxonomie. Ainsi, la présence d'infrastructures permettant le franchissement d'obstacles à la navigation le long d'un cours d'eau (écluses, ponts-canaux, tunnels, etc.) est représentée dans la taxonomie résultante par le concept général de FRANCHISSEMENT, qui lui-même se spécialise en ECLUSE, PONT-CANAL et TUNNEL. Cependant, la relation PASSE PAR, permettant de relier ce concept de FRANCHISSEMENT à celui de COURS D'EAU, devra être ajoutée manuellement.

Dans le cas de la BDTOPO©, la taxonomie obtenue a pu être comparée à un index créé manuellement et présent à la fin des spécifications. Celle-ci comprend 580 concepts au total, dont 380 sont également présents dans l'index. 80 concepts de l'index, en revanche, ne figurent pas dans notre taxonomie. Ceci s'explique en grande partie par le fait que cet index a été créé pour une version des spécifications antérieure à celle utilisée pour l'analyse automatique. Ainsi, une trentaine de termes comme FORTIN, BERGERIE, KIOSQUE ou RÂTELIER PARAVALANCHE, figurant dans cet index, n'ont

pu être détectés dans notre corpus de travail tout simplement parce qu'ils ne s'y trouvent pas. D'autres termes figurant à la fois dans l'index et dans le texte n'ont cependant pas été détectés. Il s'agit pour la plupart de groupes nominaux complexes, plus délicats à traiter comme TREMPLIN DE SAUT À SKI, ou de concepts trop spécifiques pour avoir été retenus lors de l'étape de sélection interactive, comme celui de RUE PIÉTONNE auquel on aura préféré le concept plus générique de RUE. En revanche, 200 concepts ne figurant pas dans l'index ont été extraits du texte des spécifications, ce qui témoigne du bon niveau d'exhaustivité de la taxonomie obtenue, vis-à-vis du corpus de base.

Alignement et fusion semi-automatique des deux taxonomies obtenues

La hiérarchie des concepts des taxonomies obtenues en sortie de ces traitements étant inférée automatiquement à partir de celle des textes des spécifications, un texte bien structuré aboutira nécessairement à une hiérarchie de concepts plus élaborée. Ce processus, initialement élaboré pour les spécifications de la BDTOPO©, produit de bons résultats sur ces textes. En revanche, il a ensuite été transposé directement aux spécifications de la BDCARTO©, sans adaptation préalable à ce nouveau corpus. Ceci explique la faible structure hiérarchique obtenue en sortie des traitements pour ce second corpus : l'organisation des spécifications de la BDCARTO© étant sensiblement différente, la taxonomie qui en découle est presque totalement dénuée de structure hiérarchique. Par ailleurs, les deux taxonomies résultantes ne comportent que 170 concepts rigoureusement identiques. Afin de disposer à la fois des concepts géographiques détectés dans les spécifications des deux bases et de la hiérarchie bien détaillée issue de l'analyse des spécifications de la BDTOPO©, il convient de fusionner ces deux taxonomies en conservant la structure de celle issue des spécifications de la BDTOPO©, et en s'appuyant sur les concepts communs aux deux taxonomies. Les concepts supplémentaires présents dans chacune des deux taxonomies d'origine devront également être intégrés au résultat de cet alignement. Au final, nous disposerons ainsi d'une taxonomie de concepts géographiques plus riche et relativement bien structurée.

Cette opération de fusion des deux taxonomies obtenues a été réalisée à l'aide de Prompt[36], une surcouche de Protégé développée à l'université de Stanford pour permettre la gestion de plusieurs ontologies sous Protégé. En particulier, Prompt propose un outil d'alignement et de fusion d'ontologies, semi-automatique ; des relations de correspondance sont détectées automatiquement à l'aide de techniques terminologiques et sont présentées à l'expert pour validation ou correction. Les résultats de cette opération tendent à prouver qu'un alignement automatique de ces taxonomies est possible, même s'il doit être affiné. La figure 32 présente un extrait de la taxonomie finale, obtenue après post-traitement manuel et fusion des deux taxonomies issues des spécifications des bases de données BDTOPO© Pays 1.2 et BDCARTO© 3.0.

[36] http://protege.stanford.edu/plugins/prompt/prompt.html

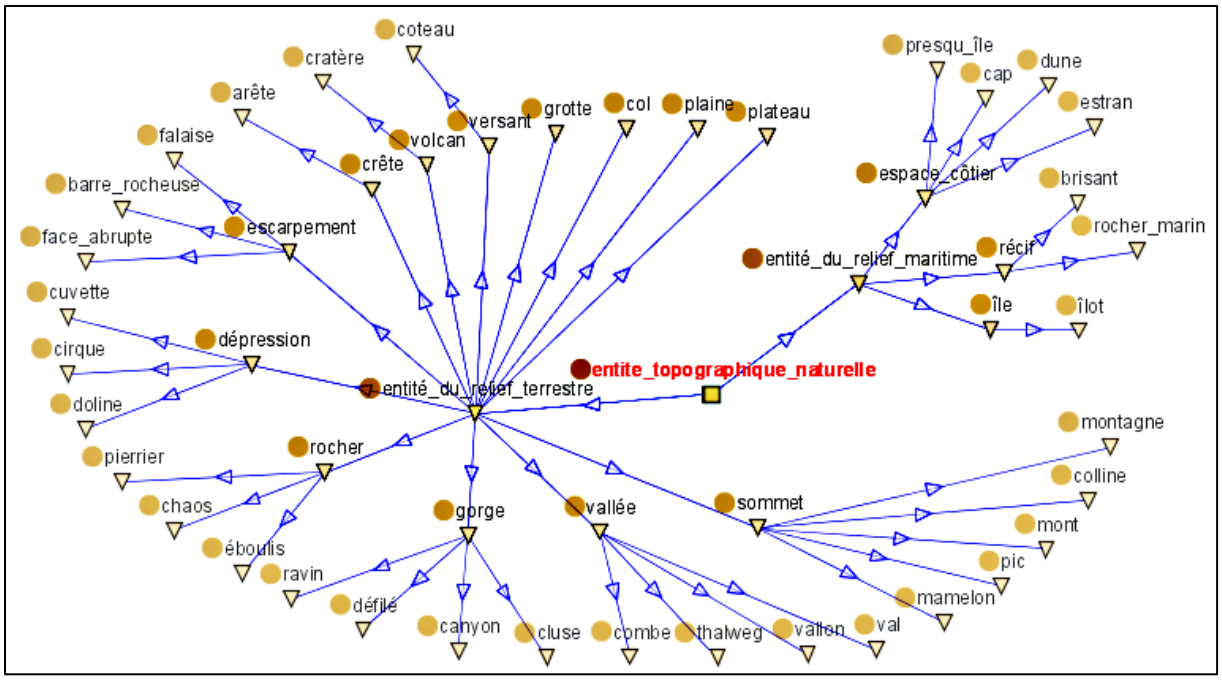


Figure 32: Extrait de la taxonomie obtenue après post-traitement manuel et alignement automatique des taxonomies présentées en figure 29 et figure 30 (visualisation sous Protégé)

4.2.3 Mise en œuvre

L'application de mise en œuvre de notre approche d'appariement automatique de schémas de bases de données topographiques a été implémentée dans le langage Java, au sein de la plateforme de système d'information géographique développée au laboratoire COGIT, GéOxygène. Des tests ont été réalisés sur les schémas des bases de données de l'IGN, BDTOPO© Pays 1.2 et BDCARTO© 3.0 (Abadie, 2009). Quelques résultats sont présentés dans la partie 4.2.3.1. Dans le cadre de notre objectif global d'intégration de données virtuelle, ces résultats sont destinés à venir alimenter une application d'appariement de données. Nous verrons en partie 4.2.3.2 comment ces résultats d'appariement de schémas et la taxonomie du domaine de la topographie créée pour les besoins de notre application peuvent être mis à profit pour apparier des données topographiques vectorielles.

4.2.3.1 Résultats de l'approche proposée pour l'appariement de schémas

Pour des raisons de lisibilité, les résultats d'appariement de schémas seront présentés dans cette section sous la forme suivante : *Nom de la base / Nom de la classe / Nom de l'attribut / Valeur de l'attribut*.

Les résultats obtenus tendent à prouver que le fait d'augmenter le niveau de granularité de la nomenclature des schémas en exploitant les labels de concepts géographiques présents dans certaines valeurs d'attributs énumérés améliore significativement les résultats. Ainsi, des correspondances ont pu être détectées entre *BDTOPO© / Tronçon de chemin / Nature / Sentier* et *BDCARTO© / Tronçon de route / Etat physique de la route / Sentier* ou bien entre *BDTOPO© / Oronyme / Nature / Grotte* et *BDCARTO© / Site et curiosité touristique / Nature / Grotte, gouffre*

aménagés alors même que les noms des classes - et même, dans le premier cas, les noms d'attributs - concernées sont lexicalement et sémantiquement différents. Les figures 33 et 34 présentent ces résultats via l'interface de visualisation de relations de correspondance entre éléments de schémas. Les schémas des bases de données y sont représentés sous la forme d'arbres : les classes occupent le premier niveau, les attributs le deuxième et leurs valeurs possibles le dernier.

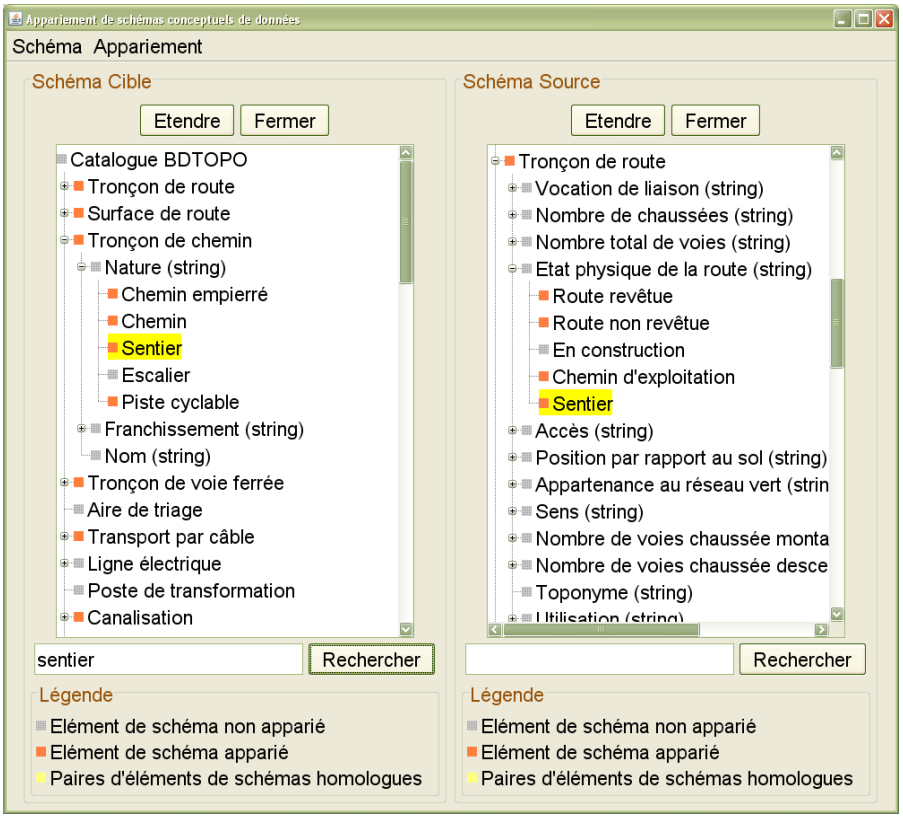


Figure 33: Apports de l'utilisation des « concepts cachés » dans le processus d'appariement: appariement des valeurs d'attributs *Sentier*

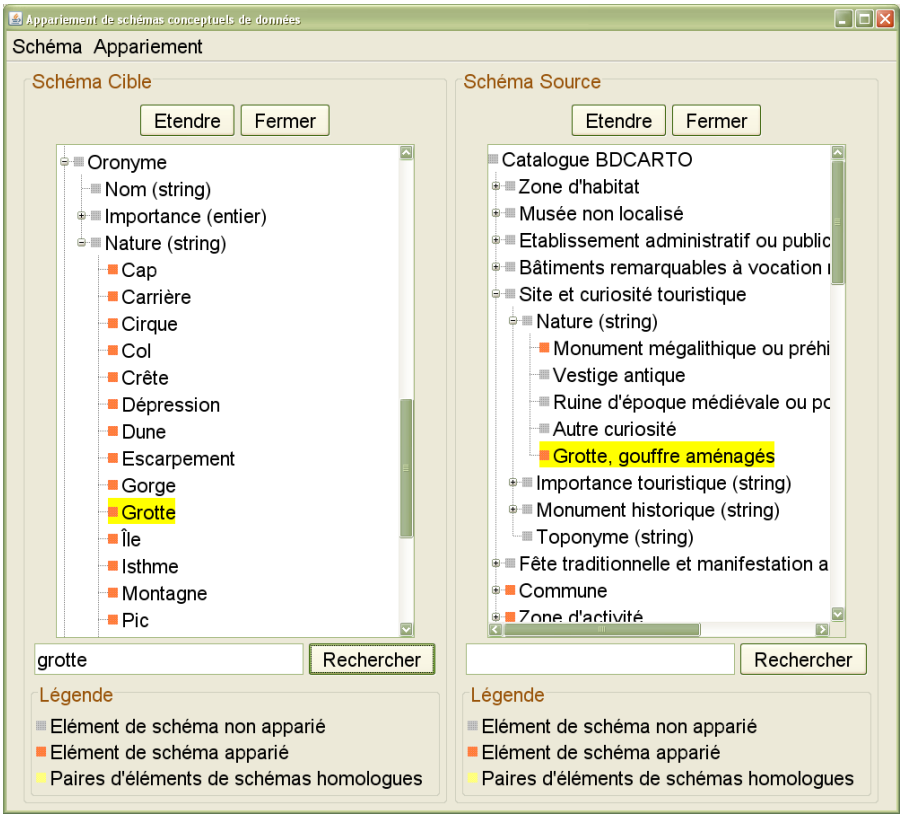


Figure 34: Apports de l'utilisation des « concepts cachés » dans le processus d'appariement: appariement des valeurs d'attributs *Grotte*

De plus, dans le cas où les schémas à apparier présentent des terminologies ou des conceptualisations différentes, le recours à la taxonomie issue des spécifications comme taxonomie de support permet d'obtenir des alignements de concepts qui n'auraient pu être détectés sans cette source de connaissances externe, et donc d'améliorer les résultats d'appariement des schémas. Ainsi, les classes *BDTOPO© / Zone arborée* et *BDCARTO© / Massif boisé* ont pu être appariées en dépit de noms lexicalement différents. En outre, dans les schémas de bases de données utilisés ici, les cluses sont considérées selon des points de vue totalement différents. Le Petit Robert définit, en effet, une cluse comme une « coupure étroite et encaissée creusée perpendiculairement à une chaîne de montagnes ». Ainsi, les spécifications de la BDTOPO© précisent que la valeur *Gorge* de l'attribut *Nature* de la classe *Oronyme* désigne à la fois des gorges, des cluses, des défilés et des canyons. Dans la BDCARTO©, en revanche, les cluses sont représentées au sein de la classe *Point remarquable du relief* sous la valeur *Col, passage, cluse* de l'attribut *Nature*. L'ontologie de support introduisant une proximité sémantique entre les concepts de CLUSE et de GORGE, un appariement entre ces deux valeurs d'attributs pourtant lexicalement différentes a tout de même pu être détecté. Les figures 35 et 36 présentent ces résultats d'appariement.

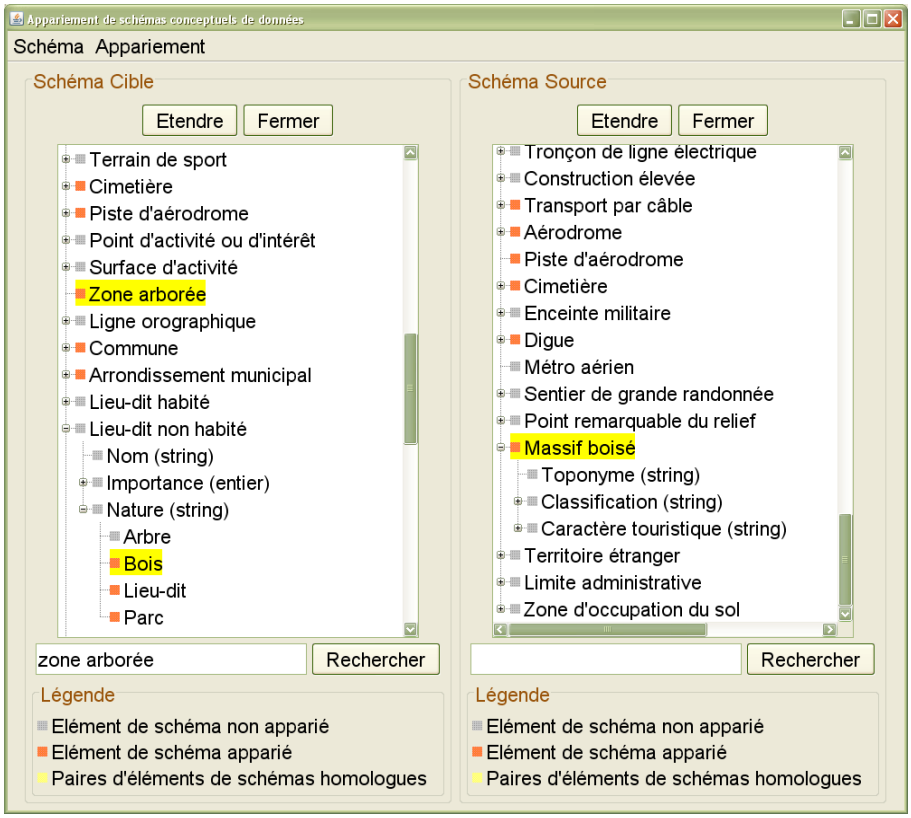


Figure 35: Apports de l'ontologie de support dans le processus d'appariement: appariement des classes *Zone arborée* et *Massif boisé*

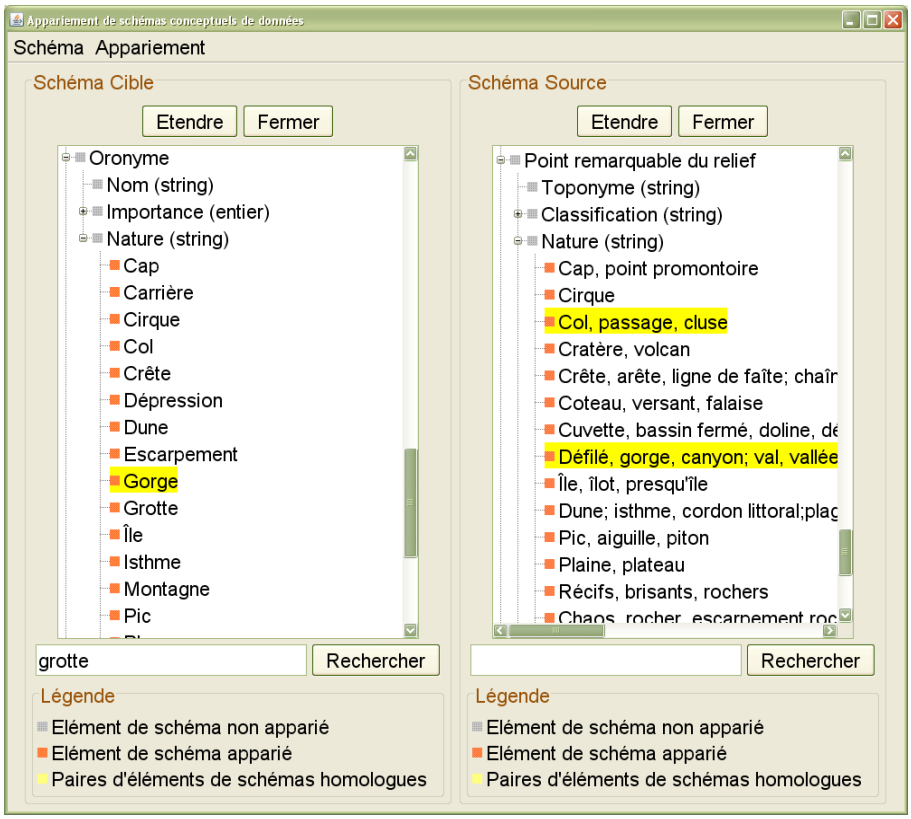


Figure 36: Apports de l'ontologie de support dans le processus d'appariement: appariement valeurs d'attributs *Gorge* et *Col, passage, cluse*

4.2.3.2 Perspectives pour l'appariement de données

Un des objectifs premier de l'appariement de schémas de bases de données géographiques réalisé dans le cadre de ce travail de thèse consiste à permettre le paramétrage d'algorithmes d'appariement de données. A ces fins, les résultats d'appariement de schémas présentés dans la partie précédente peuvent être mis à profit afin de sélectionner, en amont de toute autre opération, les instances de la base source candidates à l'appariement avec les instances d'une classe donnée de la base cible. Cette opération de sélection préalable, qui permet de limiter le nombre d'instances sources candidates à l'appariement en ne conservant que celles représentant des entités topographiques de même catégorie que les instances de la base cible, peut significativement améliorer les résultats d'appariement de données. Un autre avantage est de diminuer les risques d'erreurs d'appariement en éliminant d'emblée les instances de la base source n'appartenant pas à la même catégorie d'entités topographiques que les instances de la base cible traitées, et qui pour des raisons de proximité spatiale ou d'homonymie auraient pu être considérées comme homologues de ces dernières. Enfin, restreindre la quantité d'instances à tester pour les algorithmes d'appariement de données a également comme bénéfice immédiat de réduire les ressources nécessaires et la durée des traitements d'appariement. Cependant, opérée de façon trop stricte, cette sélection peut s'avérer préjudiciable dans la mesure où elle peut rendre la détection de certaines relations de correspondance entre instances homologues impossible, en raison de différences de catégorisation des entités géographiques du monde réel d'une base à l'autre. Un compromis pourrait résider dans l'utilisation de l'ontologie du domaine dont nous disposons pour évaluer la similarité sémantique entre catégories d'entités topographiques afin d'élargir, lorsque cela s'avère judicieux, la sélection des instances sources candidates à l'appariement.

En effet, en l'absence d'identifiant universel pour les entités topographiques, l'appariement de données géographiques est généralement fondé sur la comparaison des géométries des instances, ainsi que sur celle de leurs toponymes ou encore de leurs organisations topologiques. Or, deux instances issues de bases différentes et représentant une même entité du monde réel peuvent être géographiquement proches à la précision des données près et avoir des dénominations différentes selon que l'on utilise leur toponyme en langue locale ou francisé. C'est le cas par exemple de l'instance *le Col des joncs*, de la classe *Point remarquable du relief* de la BDCARTO©, qui a des coordonnées géographiques proches de l'instance *Bizkartzu* de la classe *Oronyme* de la BDTOPO©, mais un toponyme totalement différent, puisque dans ce second cas, il s'agit d'une dénomination basque. A l'inverse, elles peuvent posséder des toponymes identiques mais des coordonnées géographiques relativement éloignées en raison de la précision géométrique des bases auxquelles elles appartiennent, voire de la nature des entités topographiques qu'elles représentent. En l'absence de critère de comparaison supplémentaire permettant de confirmer ou d'infirmer la similarité des instances comparées, ce type de situation, relativement fréquent, rend la prise de décision concernant la mise en correspondance ou non de ces instances complexe. Reprenons l'exemple de l'instance de la BDTOPO© *Bizkartzu*, cité ci-dessus. Celle-ci a pour valeur d'attribut *Nature*, *Col*. Or, si nous considérons les résultats du processus d'appariement de schémas proposé dans ce chapitre, cette valeur d'attribut est appariée à la valeur *Col, passage, cluse* de l'attribut *Nature* de la classe *Point remarquable du relief* de la BDCARTO© (voir figure 36 en partie 4.2.3.1). Or, il s'agit de la valeur d'attribut *Nature* affectée à l'instance *Col des Joncs*. Ce résultat, qui confirme que les deux instances représentent des entités topographiques appartenant à la même catégorie, plaide

donc en faveur d'un appariement de ces deux instances. Ainsi, exploiter les résultats d'appariement de schémas comme source de connaissances sur l'appartenance à une même catégorie d'entités géographiques des instances de la base source et de celles de la base cible, a pour avantage d'apporter un critère de décision supplémentaire aux traitements d'appariement traditionnels.

Cependant, d'une base de données à l'autre, il arrive fréquemment qu'une même entité du monde réel soit répertoriée, au gré des différentes spécifications de saisie, dans des catégories différentes mais néanmoins proches d'un point de vue sémantique. Par exemple, le *Pic de l'Escarpu* est catalogué comme *Pic* au sein de la BDCARTO©, et comme *Sommet* au sein de la BDTOPO©. C'est pourquoi, dans le cadre du processus d'appariement de données, il peut s'avérer préjudiciable d'appliquer de façon trop stricte ce critère d'appartenance à une même catégorie d'entités géographiques fourni par l'appariement des schémas. Une solution consisterait à proposer un appariement des schémas plus souple, avec des correspondances de type n:m, fondées sur la proximité sémantique des éléments des schémas.

Cette solution a été mise en œuvre dans le cadre d'un processus d'appariement de données multicritère (Olteanu, 2008). L'approche proposée permet de considérer conjointement les localisations des instances à apparier, leurs toponymes et également les catégories auxquelles elles appartiennent. La prise en compte de chacun de ces critères est réalisée via des mesures de distance pour les localisations, de similarité de chaînes de caractères pour les toponymes et de similarité sémantique pour les catégories d'entités géographiques. Cette dernière est évaluée en deux étapes. Une première phase d'ancrage consiste à aligner nos éléments de schémas avec les éléments de la taxonomie, à l'aide de techniques d'alignement lexicales. Elle est suivie du calcul de similarité sémantique entre les concepts de la taxonomie auxquels les éléments de schémas ont été ancrés. De nombreuses méthodes d'évaluation de la similarité sémantique existent. Parmi elles, certaines préconisent de s'appuyer sur une taxonomie du domaine, comme celle créée pour notre application d'appariement de schémas. C'est pourquoi nous avons choisi de recourir à une mesure de ce type, la mesure de Wu-Palmer (Wu et Palmer, 1994), choisie pour sa simplicité de mise en œuvre (Abadie et al. 2007). Celle-ci établit la valeur de similarité entre deux concepts d'une même taxonomie grâce à la distance à leur plus petit généralisant commun. En effet, si C est le plus petit généralisant commun de deux concepts C_1 et C_2, prof(C) le nombre d'arcs qui séparent C de la racine de la taxonomie, et $prof_c(C_i)$ le nombre d'arcs qui séparent C_i de la racine de la taxonomie, en passant par C, alors la mesure de similarité de Wu-Palmer entre C_1 et C_2 est donnée par la formule suivante:

$$sim(C_1, C_2) = \frac{2 * prof(C)}{prof_c(C_1) + prof_c(C_2)}$$

Nous l'avons raffinée à deux niveaux pour les besoins de notre application d'appariement. Tout d'abord, pour éviter des mesures de distance faibles, mais non nulles, entre concepts suffisamment éloignés dans la taxonomie pour pouvoir être considérés comme totalement différents sans ambiguïté – c'est le cas par exemple pour les éléments liés à l'espace marin tels que BANC DE SABLE et ceux liés à l'espace terrestre tels que SOMMET - nous avons attribué une similarité nulle aux couples de concepts issus de thématiques différentes ; les couples de concepts dont le plus petit généralisant est la racine de la taxonomie auront donc un score de similarité sémantique nul.

De plus, l'appariement de schémas par similarité sémantique que nous souhaitons réaliser porte sur les valeurs possibles des attributs susceptibles de renfermer des concepts topographiques, tel

l'attribut *Nature.* Or, ces valeurs peuvent regrouper plusieurs concepts topographiques à la fois. C'est le cas, notamment, pour la classe *Oronyme* de la BDTOPO© des valeurs *Plaine ou plateau* ou *Dune ou isthme*, ou pour la classe *Point remarquable du relief* de la BDCARTO© des valeurs *Cap pointe promontoire*, *Dune plage* ou encore *Plaine plateau*. Ces regroupements de concepts topographiques ne correspondent pas toujours à des concepts plus génériques, mais obéissent à des nécessités d'acquisition ou de traitement des données. Cependant, ils compliquent les calculs de similarité sémantique, dans la mesure où il ne s'agit plus de déterminer une valeur de similarité sémantique pour une simple paire de concepts topographiques, mais pour une paire de groupes de concepts topographiques. Or, s'il semble évident que la paire de valeurs d'attributs (*Plaine ou plateau - Plaine, plateau*) devra logiquement prendre une valeur de similarité sémantique égale à 1, le cas de la paire (*Dune ou isthme - Dune, plage*) est déjà plus complexe à traiter. Le choix a été fait, dans nos expériences, de retenir dans ce cas la valeur de similarité maximale prise parmi l'ensemble des valeurs de similarité obtenues pour toutes les combinaisons de concepts possibles. Ainsi, pour les différentes valeurs de similarité (DUNE – DUNE), (DUNE – PLAGE), (ISTHME – DUNE), et (ISTHME – PLAGE), seule la valeur de similarité (DUNE – DUNE) sera conservée, ceci afin de favoriser les tentatives d'appariement entre instances dont les attributs *Nature* possèdent au moins un concept similaire. Le tableau 3 présente les valeurs de similarité sémantique obtenues pour les valeurs d'attributs *Nature* des classes *Point remarquable du relief* de la BDCARTO© et *Oronymes* de la BDTOPO©.

Catégories d'entités topographiques	Cap, pointe	Dune, plage	Ile	Récifs	Cirque	Col, passage	Cuvette, dépression	Pic	Sommet, crête, colline	Plaine, Plateau	Rochers	Vallée	Volcan
Cap	**1**	0,5	0,5	0,5	0	0	0	0	0	0	0	0	0
Dune, Isthme	0,5	**1**	0,5	0,5	0	0	0	0	0	0	0	0	0
Plage	0,4	**1**	0,4	0,4	0	0	0	0	0	0	0	0	0
Ile	0,5	0,5	**1**	0,5	0	0	0	0	0	0	0	0	0
Récifs	0,5	0,5	0,5	**1**	0	0	0	0	0	0	0	0	0
Cirque	0	0	0	0	**1**	0,34	0,34	0,57	0,67	0,34	0,34	0,3	0,34
Col	0	0	0	0	0,28	**1**	0,4	0,34	0,4	0,4	0,4	0,4	0,4
Escarpement	0	0	0	0	0,28	0,6	0,4	0,34	0,4	0,4	0,4	0,4	0,4
Dépression	0	0	0	0	0,28	0,5	**1**	0,4	0,5	0,5	0,5	0,5	0,5
Pic	0	0	0	0	0,57	0,6	0,4	**1**	**0,8**	0,4	0,4	0,4	0,4
Sommet	0	0	0	0	0,67	0,5	0,5	**0,8**	**1**	0,5	0,5	0,5	0,5
Montagne	0	0	0	0	**0,85**	0,6	0,4	0,67	**0,8**	0,4	0,4	0,4	0,4
Crête	0	0	0	0	0,57	0,6	0,4	0,67	**0,8**	0,4	0,4	0,4	0,4
Plaine, plateau	0	0	0	0	0,29	0,5	0,5	0,4	0,5	**1**	0,5	0,5	0,5
Rochers	0	0	0	0	0,34	0,5	0,5	0,4	0,5	0,5	**1**	0,5	0,5
Vallée	0	0	0	0	0,34	0,5	0,5	0,4	0,5	0,5	0,5	**1**	0,5
Gorges	0	0	0	0	0,28	0,6	0,4	0,34	0,4	0,4	0,4	**0,8**	0,4
Volcan	0	0	0	0	0,33	0,4	0,33	0,4	0,5	0,5	0,5	0,5	**1**

Tableau 3: Valeurs de similarité sémantique calculées pour l'appariement de valeurs d'attributs énumérées ; cas des points remarquables du relief

Pour affiner notre mesure de similarité, il serait pertinent de recourir à une mesure prenant également en compte les propriétés des concepts géographiques, et en particulier leur forme. En effet, nous nous attachons ici à établir des correspondances sémantiques entre entités dont la conformation géométrique constitue la principale caractéristique (Smith et Mark, 1998).

L'introduction de cette notion fondamentale en topographie permettrait d'obtenir des résultats plus contrastés selon que les concepts géographiques comparés possèdent ou non des formes semblables. Ainsi, les concepts de PLAINE et de MONTAGNE par exemple, dont la similarité est évaluée ici à 0,4 verraient cette valeur diminuer. A l'inverse, pour les concepts de VOLCAN et de MONTAGNE, qui avec une mesure de similarité de 0,4 sont considérés ici comme sémantiquement aussi proches l'un de l'autre que les deux exemples précédents, cette valeur augmenterait. Les concepts de VOLCAN et de MONTAGNE caractérisés, contrairement au concept de PLAINE, par leur forme proéminente, seraient donc considérés comme plus proches, et donc plus susceptibles de faire l'objet d'éventuels cas d'appariement. Ceci nécessite d'enrichir notre taxonomie pour aller vers une véritable ontologie dotée de propriétés pour les concepts recensés, ce qui est l'objet du projet GéOnto (ANR-07-MDCO-005).

La figure 37 illustre un résultat d'appariement de données en montrant l'apport de l'appariement de schémas par mesure de similarité sémantique. Dans le cas présenté, avant prise en compte de ce critère, le processus d'appariement ne détecte aucune relation de correspondance entre le *Pic de Louesque* et *l'Escarpu*, en raison des différences toponymiques trop importantes. Un appariement strict des valeurs d'attribut *Nature* n'aurait pu permettre d'obtenir une mise en correspondance de ces instances, respectivement dotées des valeurs *Sommet* et *Pic*. En introduisant une mesure de similarité sémantique calculée à partir de la taxonomie dont nous disposons, l'appariement est plus fin et établit à raison une correspondance entre ces instances en raison de leur appartenance à des catégories d'entités topographiques sémantiquement proches.

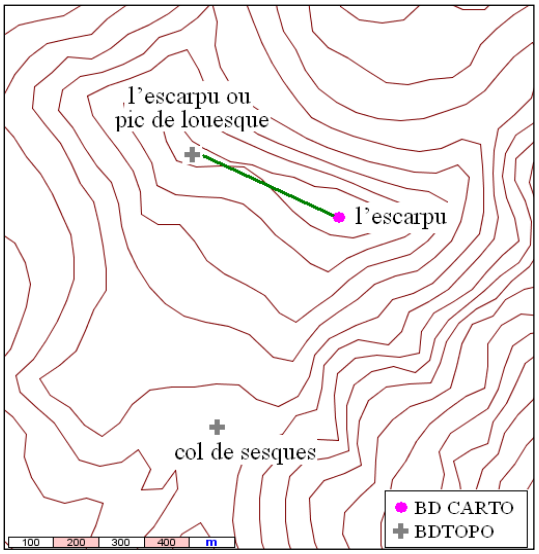


Figure 37: Cas d'appariement montrant l'intérêt du critère sémantique (extrait d'Olteanu, 2008)

Conclusion

Dans cette partie, nous avons proposé une mise en œuvre de notre modèle pour l'intégration virtuelle de bases de données topographiques, en automatisant, autant que faire se peut, l'étape d'annotation des ressources décrivant les données sur laquelle repose l'appariement des schémas. Pour ce faire, nous avons adopté une approche combinant plusieurs techniques d'appariement automatique élémentaires. En effet, afin d'étendre les possibilités de mise en correspondance

d'éléments de schémas et d'éléments de l'ontologie du domaine via des techniques d'appariement terminologiques et des techniques fondées sur les relations taxonomiques, nous avons entrepris d'augmenter le niveau de granularité des schémas à apparier en faisant émerger des concepts topographiques cachés dans des valeurs d'attributs, et constituant une source de connaissances importante sur le contenu des classes dont ils sont extraits. Ceci est réalisé au travers de la dérivation semi-automatique d'ontologies d'applications à partir des schémas de bases de données que l'on cherche à apparier. L'annotation des ontologies d'applications via l'ontologie du domaine est réalisée à l'aide de techniques d'appariement terminologiques et de techniques fondées sur les relations taxonomiques. C'est sur les relations de correspondance établies entre éléments d'ontologies d'applications et éléments d'ontologie du domaine que repose l'alignement des ontologies d'applications. Ceci revient donc à utiliser l'ontologie du domaine comme une ontologie de support pour réaliser l'alignement des ontologies d'applications et compenser ainsi d'éventuelles carences de recouvrement lexical et structurel entre les ontologies d'applications. Enfin, les relations de correspondance entre ontologies d'applications hétérogènes, et celles entre schémas et ontologies d'applications, préalablement conservées lors de l'étape de création de ces dernières, sont exploitées afin de déduire automatiquement les relations de correspondance entre schémas hétérogènes. La mise en œuvre de cette approche a nécessité l'instanciation des principales composantes du modèle.

Celle des ontologies d'applications, d'une part, a été réalisée selon une approche relativement intuitive, qui a consisté à traduire les divers éléments de schémas, représentés selon la norme « ISO 19109 - Rules for application schema », en leurs éléments d'ontologies homologues, au format OWL. Seules les valeurs d'attributs renfermant des concepts topographiques ont bénéficié d'un traitement particulier dans la mesure où elles ont été traduites, non pas par des propriétés mais par des classes, au sein des ontologies générées. Ce choix de conversion des schémas vers un formalisme fondé sur les logiques de description est motivé par les possibilités de raisonnement associées à un tel formalisme, que nous souhaitons intégrer à la suite de nos travaux (cf. partie 4.3). Cette question de la dérivation d'ontologies d'applications à partir de schémas de bases de données géographiques fait d'ailleurs actuellement l'objet de la rédaction d'une norme au sein de l'Organisation Internationale de Standardisation, « ISO 19150 - Ontology — Part 2: Rules for developing ontologies in the Web Ontology Language (OWL)», preuve d'un intérêt grandissant de la communauté de l'information géographique pour ces formalismes de représentation de connaissances et les possibilités qu'ils offrent en matière de description de la sémantique des données et de raisonnement associé à cette description, en comparaison des formalismes de modélisation traditionnels.

Celle de l'ontologie du domaine, d'autre part, a été réalisée par applications successives de techniques de traitement automatique du langage naturel sur les textes des spécifications de deux bases de données de l'IGN, la BDTOPO© et la BDCARTO©, de traitement manuel, et de techniques de fusion d'ontologies. Nous disposons désormais d'une taxonomie du domaine de la topographie, également au format OWL, comportant plus de 700 concepts, qui pourra être mise à profit dans la suite de nos travaux.

Les résultats d'appariement de schémas finalement obtenus tendent à prouver qu'une automatisation du processus d'annotation sémantique des schémas est réalisable, en combinant différentes techniques d'appariement automatique, tout en conservant des résultats satisfaisants au niveau de l'appariement des schémas. L'approche que nous avons proposée repose largement sur

des techniques terminologiques et sur des techniques fondées sur les relations taxonomiques entre éléments d'ontologies. Aussi la qualité des résultats obtenus est-elle directement liée à la richesse lexicale et structurelle des ressources mises en œuvre : augmenter le niveau de granularité des schémas et exploiter une ontologie de support la plus exhaustive possible en termes de vocabulaire, et bien structurée, permet de détecter des relations de correspondances entre ontologies d'applications, et donc entre schémas, qui n'auraient pas pu être établies avec des ressources moins riches. Cependant, nous restons consciente qu'apparier les schémas des bases de données BDTOPO© et BDCARTO© en s'appuyant sur une ontologie du domaine générée à partir des spécifications de ces bases comporte un biais important, dans la mesure où le vocabulaire de cette ontologie est directement lié à ceux des schémas. C'est pourquoi, afin de pouvoir étendre cette approche à d'autres bases de données, un enrichissement de la taxonomie créée, à partir d'autres sources de connaissances, a été engagé au sein du projet GéOnto (ANR-O7-MDCO-005).

Enfin, nous avons introduit, dans un processus d'appariement de données multicritère, des résultats d'appariement de schémas. Ceux-ci permettent en effet d'améliorer les résultats d'appariement de données en introduisant un critère de comparaison supplémentaire entre instances : deux instances de bases de données hétérogènes, représentant des entités géographiques de même catégorie, seront plus susceptibles d'être mises en correspondance. Cependant, nous avons pu observer que l'approche proposée pour la dérivation des relations de correspondances entre ontologies d'applications, présentée en partie 4.2.1.2, pouvait s'avérer trop restrictive quant aux possibilités de détection de relations de correspondance entre éléments d'ontologies d'applications, et donc d'éléments de schémas. Aussi une approche de dérivation fondée, non plus sur l'équivalence ou la parenté directe, mais sur la similarité sémantique des concepts de l'ontologie de support auxquels les concepts des ontologies d'applications sont ancrés, a été proposée afin d'élargir les possibilités de mise en correspondance des éléments d'ontologies d'applications. Cette approche, produisant un appariement de schémas plus souple, s'est avérée efficace pour le cas d'appariement de points remarquables du relief testé.

4.3 Intégration virtuelle de bases de données géographiques fondée sur les spécifications de ces bases

L'approche proposée dans la partie 4.2 pour l'appariement de schémas de bases de données topographiques permet de détecter automatiquement les classes de chacun des schémas en entrée représentant les mêmes catégories d'entités topographiques du monde réel. Elle offre, en outre, un niveau d'appariement plus fin en s'attachant à détecter les sous-ensembles d'instances de ces classes qui représentent une même catégorie d'entités topographiques. Cependant, les relations de correspondance détectées se cantonnent à faire état de l'appartenance des entités topographiques représentées par les instances des classes de schémas à apparier à une même catégorie d'entités topographiques ou, à des catégories d'entités topographiques sémantiquement proches. En effet, cette approche n'intègre pas de connaissances susceptibles de permettre au système de déceler d'éventuels conflits de critères de sélection ou de description géométrique des données (Devogele, 1997) entre les classes mises en relation. Ainsi, si nous reprenons l'exemple des points remarquables du relief, les relations de correspondance établies ne prennent pas en compte les différences entre les critères de sélection qui régissent la saisie des instances de la classe *Oronyme* de la BDTOPO© et

ceux qui sont appliqués pour la classe *Point remarquable du relief* de la BDCARTO©. Or, dans le premier cas, les spécifications stipulent que sont saisis « les oronymes figurant sur la carte au 1 : 25 000 en service ». Dans le second, en revanche, les spécifications de la BDCARTO© précisent que sont retenus « […] les sommets les plus élevés ou les plus caractéristiques, les cols les plus caractéristiques en assurant une cohérence avec le thème routier, les massifs, vallées cluses, passages, plateaux, plaines et cuvettes les plus caractéristiques, les détails du relief côtier […] ». L'analyse comparée de ces deux extraits de spécifications portant sur les critères de sélection des classes *Oronyme* et *Point remarquable du relief* révèle des critères de sélection différents d'une classe à l'autre et nous renseigne donc sur une possible différence d'exhaustivité entre ces deux classes. De même, si l'on considère les spécifications de saisie de la géométrie des instances de ces classes, on apprend que dans les deux cas, chaque entité topographique du monde réel sélectionnée est représentée dans l'une ou l'autre des bases à l'aide d'un point saisi en son centre, information également utile à la fois pour un utilisateur potentiel souhaitant comparer les deux bases de données et pour guider le choix d'un algorithme d'appariement des données.

C'est pourquoi nous nous attachons ici à proposer une mise en œuvre de notre modèle permettant, autant que faire se peut, la détection automatique de relations de correspondance fines entre éléments de schémas de bases de données géographiques. Nous souhaitons pouvoir apparier des schémas de bases de données géographiques en rendant compte, dans les résultats d'appariement, des éventuels conflits de sélection ou de description géométrique de données existant entre ces bases. Pour atteindre cet objectif de mise en correspondance fine des éléments de schémas conceptuels des bases de données à intégrer nous nous sommes inspirée des techniques d'alignement d'ontologies fondées sur la sémantique présentées au chapitre 3.1.2, dont Klien (2008) et Schade (2010) proposent deux exemples de mise en œuvre dans le domaine de l'intégration d'information géographique, ainsi que des travaux de Gesbert (2005) sur la formalisation des spécifications de bases de données géographiques. Nous proposons, en effet, de formaliser les connaissances sur les règles de sélection et de représentation géométrique des entités géographiques issues des spécifications à l'aide d'un langage de représentation de connaissances standard, OWL, afin de pouvoir mettre à profit les systèmes de raisonnement associés pour détecter automatiquement des relations de correspondance entre éléments de schémas. En effet, ce langage de représentation de connaissances est fondé sur une logique de description, de sorte que la sémantique des classes représentées au sein d'une ontologie peut être définie à l'aide d'axiomes logiques. Ces axiomes constituent un ensemble de conditions nécessaires et de conditions nécessaires et suffisantes que doivent vérifier des instances pour appartenir à une classe donnée. Ces axiomes peuvent être interprétés par divers systèmes de raisonnement, comme Pellet[37] ou FACT++[38], fondés sur cette même logique de description (Horrocks, 2008). Dans le cadre de notre objectif de formalisation de connaissances pour la mise en œuvre d'un processus d'appariement de schémas, les raisonnements effectués sur la base de ces axiomes présentent plusieurs avantages. Ils permettent d'une part, de vérifier la cohérence des définitions fournies pour décrire les classes d'une ontologie : une classe décrite par des axiomes contradictoires ne pourra pas posséder d'instances, et sera donc considérée comme incohérente. Ainsi, nous pourrons nous assurer de la cohérence des règles de saisie formelles sur lesquelles repose notre processus d'appariement. D'autre part, ils permettent de détecter des relations d'équivalence ou de subsomption entre classes : deux classes

[37] http://clarkparsia.com/pellet/

[38] http://owl.man.ac.uk/factplusplus/

sont considérées comme équivalentes si les axiomes les définissant impliquent qu'elles possèdent nécessairement les mêmes instances, et une classe est considérée comme une spécialisation d'une autre classe si les axiomes définissant la première impliquent que ses instances constituent un sous-ensemble des instances de la seconde. C'est précisément ce deuxième aspect des possibilités de raisonnement associées aux ontologies OWL que nous souhaitons mettre à profit dans le cadre de notre approche d'appariement de schémas. Pour des raisons d'expressivité, nous adoptons le langage OWL 2, qui offre la possibilité de créer des restrictions sur les types de données XML, ou encore de définir des restrictions de cardinalité qualifiées sur les propriétés (W3C, 2009). Nous présentons donc, dans la partie 4.3.1, notre approche globale pour l'appariement fin de bases de données topographiques. La partie 4.3.2 décrit les approches adoptées pour l'instanciation des principales composantes de notre modèle : le cadre de référence sémantique et les ontologies d'applications. Enfin, la partie 4.3.3 présente une première mise en œuvre de notre approche pour l'appariement de schémas de bases de données de l'IGN, et discute des perspectives qu'offrent les résultats obtenus pour l'appariement de données topographiques. Cette partie se conclut sur une seconde mise en œuvre de cette approche dédiée à la découverte de bases de données topographiques.

4.3.1 Approche globale

L'architecture et le processus global mis en œuvre dans cette approche d'appariement de schémas de bases de données géographiques sont présentés en figure 38. L'architecture correspond à celle présentée en figure 23, et inclut un cadre de référence sémantique. Une première étape de notre approche d'appariement automatique de schémas de bases de données géographiques va consister à créer les ontologies d'application destinées à représenter les structures des différentes bases de données à intégrer, à partir des schémas de ces bases représentés selon la norme « ISO 19109 – Rules for application schema » (figure 38, numéro 1). Cette étape de création de ressources descriptives intermédiaires, intervenant entre les schémas de bases de données et le cadre de référence sémantique, va nous fournir un support pour la représentation des connaissances issues des spécifications nécessaires à la réalisation d'un appariement de schémas fin. En effet, détecter des relations de correspondance fines entre éléments de schémas, incluant des restrictions relatives à des différences de critères de sélection ou de représentation des entités topographiques du monde réel au sein des bases de données à intégrer, suppose de comparer, outre ces éléments de schémas, les éléments de spécifications qui les décrivent. Le processus d'appariement de schémas que nous souhaitons mettre en œuvre ici revient donc à apparier, en premier lieu, des éléments de spécifications. Les relations de correspondance entre éléments de schémas sont ensuite dérivées des relations de correspondance préalablement établies entre leurs éléments de spécifications respectifs. Pour ce faire deux niveaux de description des données sont générés: l'un pour les éléments de schémas, l'autre pour les éléments de spécifications. Lors de cette étape, les éléments de schémas et les éléments de spécifications à apparier sont réifiés afin de permettre la formalisation des règles de saisie fournies par les spécifications sous la forme d'axiomes portés par les classes destinées à représenter les divers éléments de schémas et de spécifications. Ces axiomes utilisent des concepts du domaine et des concepts de haut niveau fournis par le cadre de référence sémantique. Nous nous appuyons ici sur les travaux de Klien (2008) et Schade (2010) en proposant un cadre conceptuel décrivant les primitives de base nécessaires à la représentation des règles de

saisie des entités topographiques au sein d'une base de données. Cette étape d'axiomatisation des ontologies d'applications est réalisée manuellement par analyse des textes des spécifications (figure 38, numéro 2). Ces axiomes sont ensuite exploités à l'aide d'un système de raisonnement afin de déduire des relations d'équivalence et de subsomption entre concepts des ontologies d'applications décrivant des éléments de spécifications et des éléments de schémas (figure 38, numéro 3).

L'ensemble de ce processus est détaillé dans la suite de cette partie et une mise en œuvre est proposée. En outre, une mise en œuvre de l'architecture globale de notre approche pour la découverte de bases de données topographiques hétérogènes est également proposée.

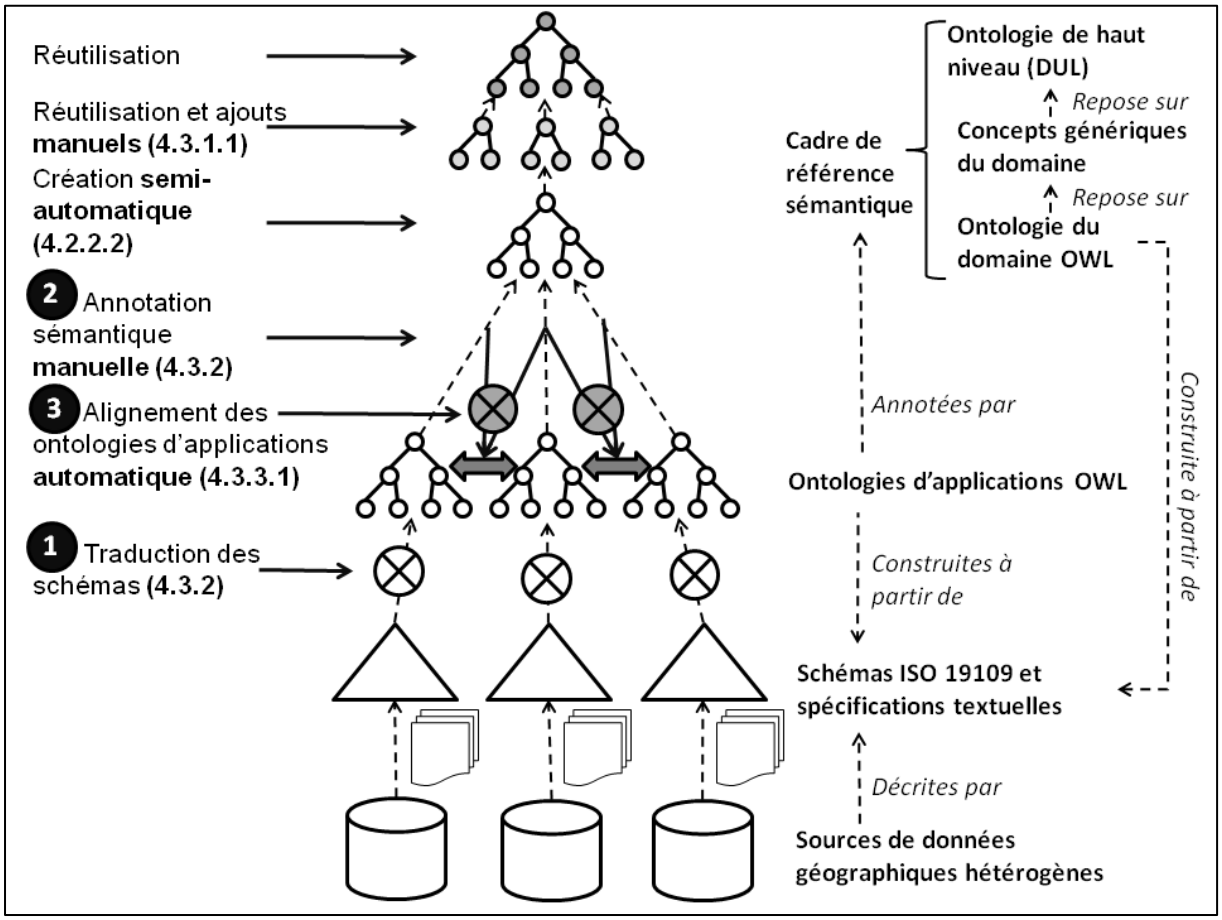


Figure 38: Architecture globale pour l'intégration virtuelle de bases de données géographiques fondées sur les spécifications de ces bases ; processus global d'appariement des schémas

4.3.1.1 Création du cadre de référence sémantique

La description de la sémantique des classes de schémas de bases de données topographiques dépasse la seule explicitation des catégories d'entités topographiques représentées par ces classes. En effet, elle comprend, en outre, la représentation de connaissances sur le processus de saisie des données, issues des spécifications de ces bases de données. Ces connaissances, utiles à la réalisation de notre objectif d'intégration virtuelle de bases de données topographiques, ont été identifiées et formalisées à l'aide d'un langage dédié dans les travaux de Gesbert (2005). Nous nous sommes donc inspirée de ces travaux pour la définition d'un cadre de référence sémantique décrivant l'ensemble

des concepts nécessaires à la représentation des règles de saisie des bases de données topographiques vectorielles. En accord avec les propositions de (Kuhn, 2003), Klien (2008) et Schade (2010) définissent chacun un cadre de référence sémantique destiné à décrire les concepts de base pour l'annotation sémantique de schémas de bases de données géographiques. Nous suivons l'approche proposée dans leurs travaux pour la création d'un cadre de référence sémantique adapté à notre objectif de formalisation de spécifications.

Kuhn (2003) définit cette notion de cadre de référence sémantique comme une structure conceptuelle indépendante de toute application et définie de façon formelle, pouvant être assimilée à une ontologie de haut niveau (Guarino, 1998). Pour Schade (2010), un cadre de référence sémantique peut également comporter des concepts du domaine dans la mesure où ceux-ci reposent sur une ontologie de haut niveau. Nous adoptons ce second point de vue pour la création du cadre de référence sémantique sur lequel repose notre proposition : celui-ci comporte l'ensemble des concepts topographiques issus de l'ontologie du domaine créée pour l'approche proposée dans la partie 4.2, ainsi qu'un ensemble de concepts, propriétés et relations utiles à l'expression des règles de saisie des données topographiques. Conformément aux recommandations de Kuhn (2003) et aux travaux de Klien (2008) et Schade (2010), ces concepts, propriétés et relations sont ancrés à une ontologie de haut niveau via des relations de généralisation-spécialisation. Cependant, notre proposition se distingue de celles de Klien (2008) et Schade (2010) par le choix de l'ontologie de haut niveau : dans un souci de simplicité, nous avons choisi d'adopter Dolce+DnS Ultra Lite (DUL[39]) au lieu de Dolce. Tout comme Dolce, celle-ci vise à fournir un ensemble de concepts de haut niveau sur lesquels diverses ontologies de domaine ou ontologies d'applications peuvent ancrer les concepts qu'elles définissent, mais elle décrit ces concepts à l'aide d'un vocabulaire et d'une modélisation simplifiés. Enfin, elle présente l'avantage d'être publiée sur le Web, dans le langage OWL.

Description des éléments de schémas de bases de données topographiques

Les systèmes de raisonnement disponibles pour les ontologies implémentées dans le langage OWL permettent de détecter des relations d'équivalence ou de subsomption entre classes : deux classes sont considérées comme équivalentes si les axiomes les définissant impliquent qu'elles possèdent nécessairement les mêmes instances, et une classe est considérée comme une spécialisation d'une autre classe si les axiomes définissant la première impliquent que ses instances constituent un sous-ensemble des instances de la seconde. Nous souhaitons mettre à profit ces possibilités de détection de relations de correspondance fondées sur la sémantique afin d'établir des relations de correspondance entre éléments de schémas de bases de données topographiques hétérogènes. Une condition indispensable à la réalisation de cet objectif va donc être de réifier les éléments de schémas que nous souhaitons pouvoir apparier, et de les doter de définitions axiomatisées afin de permettre aux systèmes de raisonnement disponibles d'inférer des relations de correspondance entre les classes ainsi définies. Nous proposons donc de représenter les éléments de schémas de bases de données topographiques à intégrer comme des spécialisations des classes du modèle présenté en figure 39. Ce modèle fait partie intégrante du cadre de référence sémantique que nous proposons. Les classes le composant sont insérées en tant que spécialisations du concept INFORMATIONOBJECT - objet informationnel - de l'ontologie DUL. Celui-ci vise en effet à décrire des sources d'informations, comme des textes ou des images, considérées indépendamment de leur

[39] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

réalisation concrète. Les classes FEATURE, qui désigne une instance de base de données topographique dotée d'un type dit *FeatureType*, et GEOMETRY, qui définit les types géométriques pouvant être utilisés pour la représentation des entités géographiques au sein de la base, sont directement issues de l'ontologie GeoOWL[40]. Celle-ci est une implémentation en OWL du modèle de format de données géographiques dérivé de GML, GeoRSS[41], destiné à la syndication de contenus du Web. Cette ontologie résulte d'une amélioration d'un premier vocabulaire développé par le W3C pour la géolocalisation et la description des propriétés géographiques des ressources du Web. Elle a été réalisée par le groupe d'incubation « Vocabulaire Géographique » - Geospatial Vocabulary - du W3C. Le choix d'importer cette ontologie pour modéliser les éléments de schémas à apparier est motivé par une volonté de réutilisation de ressources existantes, et de conformité aux standards de représentation de l'information géographique. En outre, elle présente l'avantage, par rapport à une transposition directe des modèles GML ou ISO en OWL, de fournir un modèle relativement simple. Dans le cadre de notre application, elle est complétée par l'ajout de deux classes. La première, ATTRIBUTE - Attribut, représente un attribut de classe de base de données. La seconde, ATTRIBUTEVALUE - Valeur d'attribut, représente une valeur que peut prendre un attribut sous certaines conditions précisées dans les spécifications. Ces classes sont ajoutées manuellement afin de permettre la description des règles de saisie de ces types d'éléments de schémas à l'aide d'axiomes.

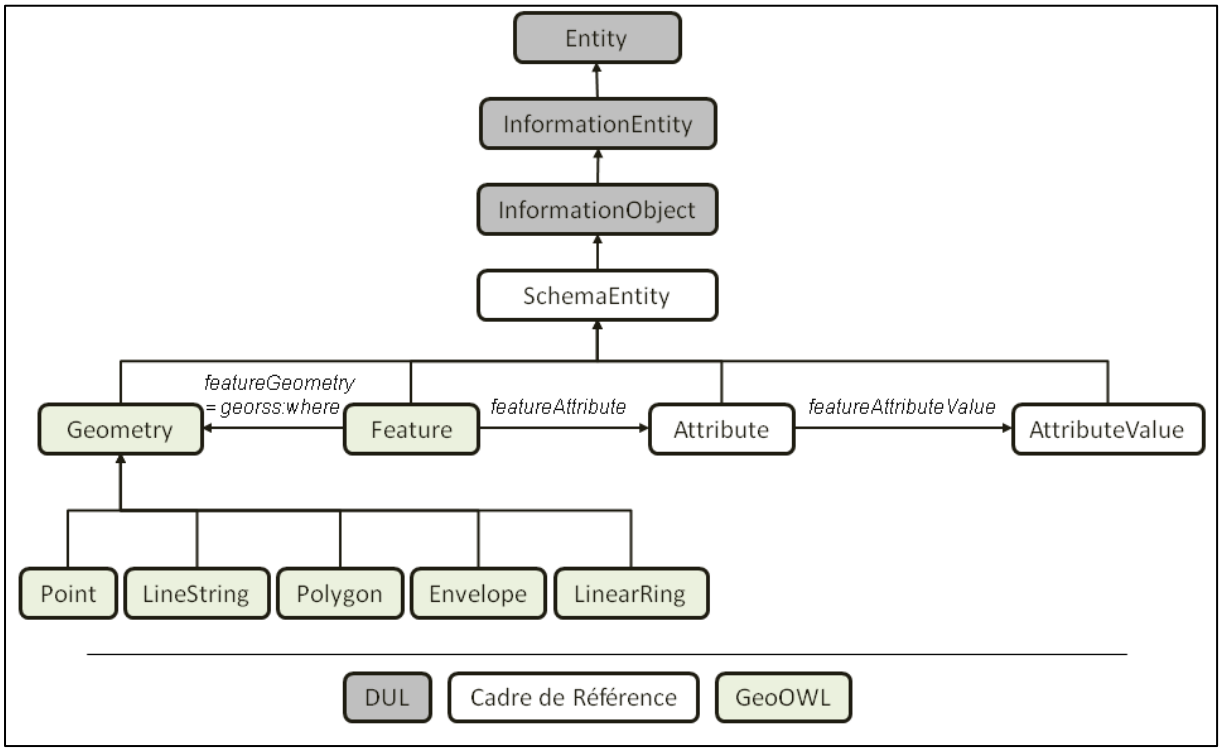


Figure 39: Partie du cadre de référence sémantique destinée à la représentation des éléments de schémas de bases de données topographiques

Description des éléments de spécifications de bases de données topographiques

[40] http://www.w3.org/2005/Incubator/geo/XGR-geo/

[41] http://www.georss.org/Main_Page

La représentation des éléments de spécifications que nous souhaitons pouvoir formaliser et comparer automatiquement suit les mêmes principes que celle des éléments de schémas. Nous distinguons en priorité quatre catégories d'éléments de spécifications, présentées en figure 40. La première, que nous nommons « processus de saisie d'un *Feature* » - FEATURECAPTUREPROCESS dans notre cadre de référence sémantique – constitue l'élément central du modèle. Elle désigne l'ensemble du processus de saisie d'une classe de base de données topographique, et sert de point d'ancrage à l'ensemble des règles de saisie des instances d'une classe, disséminées au sein des divers éléments de spécifications de cette classe. En tant que modèle du processus de saisie d'une classe de base de données, FEATURECAPTUREPROCESS est donc destiné à porter les règles de sélection des entités géographiques devant figurer dans cette classe. Les trois autres catégories majeures d'éléments de spécifications sont destinées à la représentation des règles de modélisation géométrique, de définition des attributs et de saisie conditionnelle des valeurs d'attributs énumérées des instances d'une classe de base de données topographique. Chacune de ces catégories est représentée par une classe. Au sein de chaque classe, les règles de saisie sont représentées sous la forme d'axiomes qui définissent les propriétés fondamentales des instances de la classe. C'est pourquoi ces trois classes sont nommées par des termes désignant le résultat du processus de saisie qui les définit. A titre d'exemple, « Géométrie modélisée » - MODELEDGEOMETRY - est définie par les règles de saisie de la géométrie des instances de la base. Les classes composant ce modèle sont également insérées, dans notre cadre de référence sémantique, en tant que spécialisations du concept INFORMATIONOBJECT - objet informationnel - de l'ontologie DUL.

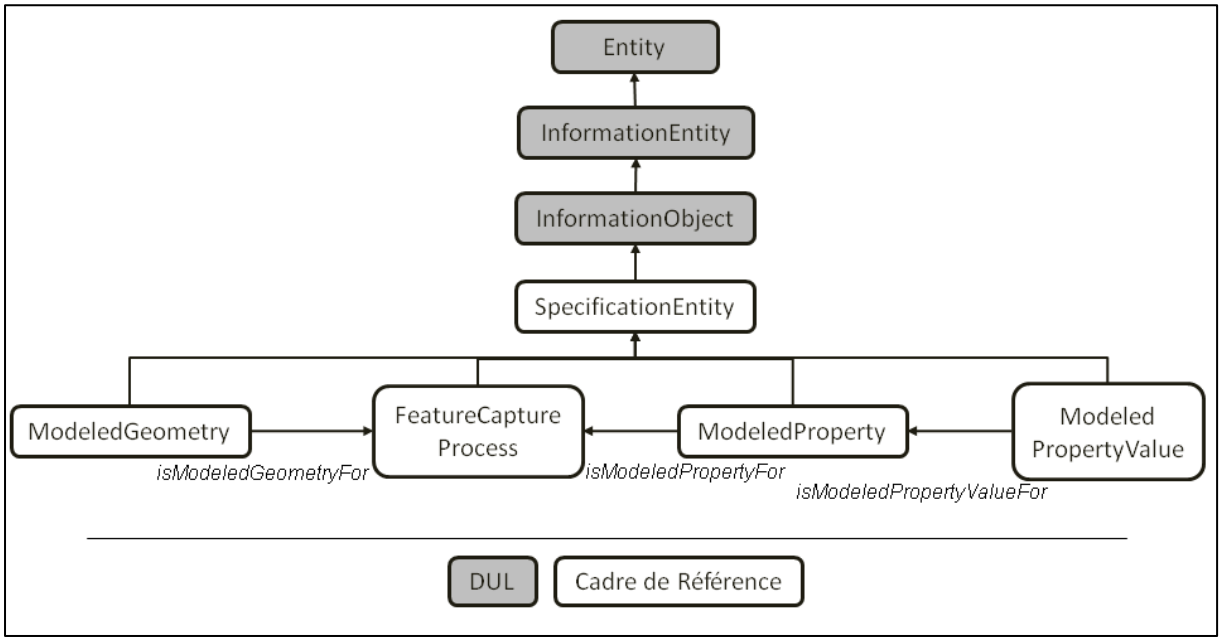


Figure 40: Partie du cadre de référence sémantique destinée à la représentation des éléments de spécifications de bases de données topographiques – Classes principales

Description des catégories d'entités topographiques

La représentation, sous la forme d'axiomes, des règles de saisie des entités topographiques du monde réel au sein d'une base de données topographique nécessite de pouvoir faire référence aux catégories d'entités topographiques sélectionnées pour figurer dans l'une ou l'autre des classes de la base, ainsi qu'à leurs propriétés et relations. C'est pourquoi nous intégrons à notre cadre de

référence sémantique la taxonomie du domaine de la topographie décrite dans la partie 4.2. Celle-ci est importée de telle sorte que ses deux concepts les plus généraux, ENTITÉTOPOGRAPHIQUENATURELLE et ENTITÉTOPOGRAPHIQUEARTIFICIELLE viennent spécialiser le concept d'objet géographique – GEOGRAPHICOBJECT (voir figure 41). Ce concept d'objet géographique est lui-même ajouté manuellement à DUL, en tant que spécialisation du concept de lieu physique - PHYSICALPLACE. Suivant les cadres de référence proposés par Klien (2008) et Schade (2010), nous associons au concept d'objet géographique celui de localisation, modélisé ici comme une qualité associée à ce premier concept via une relation HASQUALITY. Une localisation est définie comme une région de l'espace – SPACEREGION – dont la position est définie dans un système de référence. Les coordonnées indiquant cette position sont rattachées à la région de l'espace par la relation HASREGIONDATAVALUE. De plus, un objet géographique peut jouer un rôle particulier, représenté par le concept de rôle d'objet géographique – GEOGRAPHICOBJECTROLE.

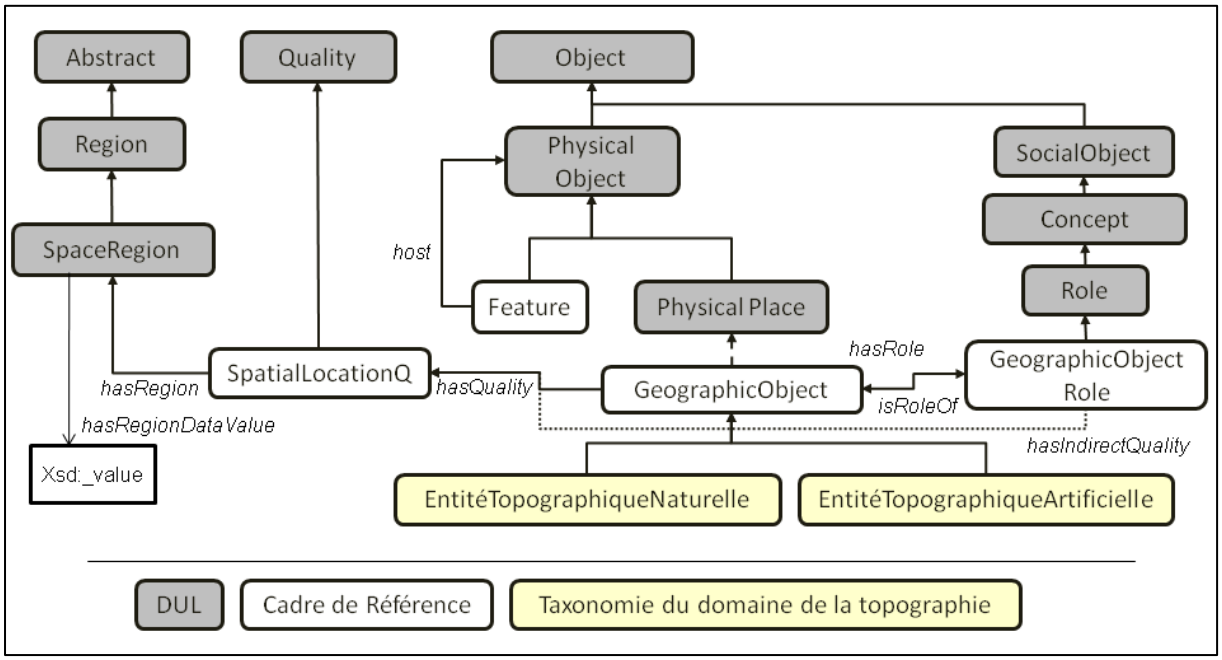


Figure 41: Cadre de référence sémantique pour la représentation des catégories d'entités topographiques, de leurs rôles et de leurs propriétés

Les propriétés des objets géographiques sont représentées ici en tant que qualités - QUALITY. Elles peuvent être de deux types : les qualités physiques, relatives aux objets géographiques en tant qu'objets physiques -PHYSICALQUALITY (voir figure 42)- et les qualités sociales, relatives aux objets géographiques en tant qu'objets sociaux - SOCIALQUALITY (voir figure 43). Ces qualités peuvent prendre leurs valeurs dans un espace bien défini - PHYSICALATTRIBUTE ou SOCIALOBJECTATTRIBUTE, valeurs affectées via la relation HASREGIONDATAVALUE. Elles peuvent également jouer un rôle particulier dans la description d'un objet géographique. Par exemple, WIDTH - largeur - peut être l'un des rôles possibles pour la qualité SIZE1D - taille - dans la représentation des dimensions d'un objet.

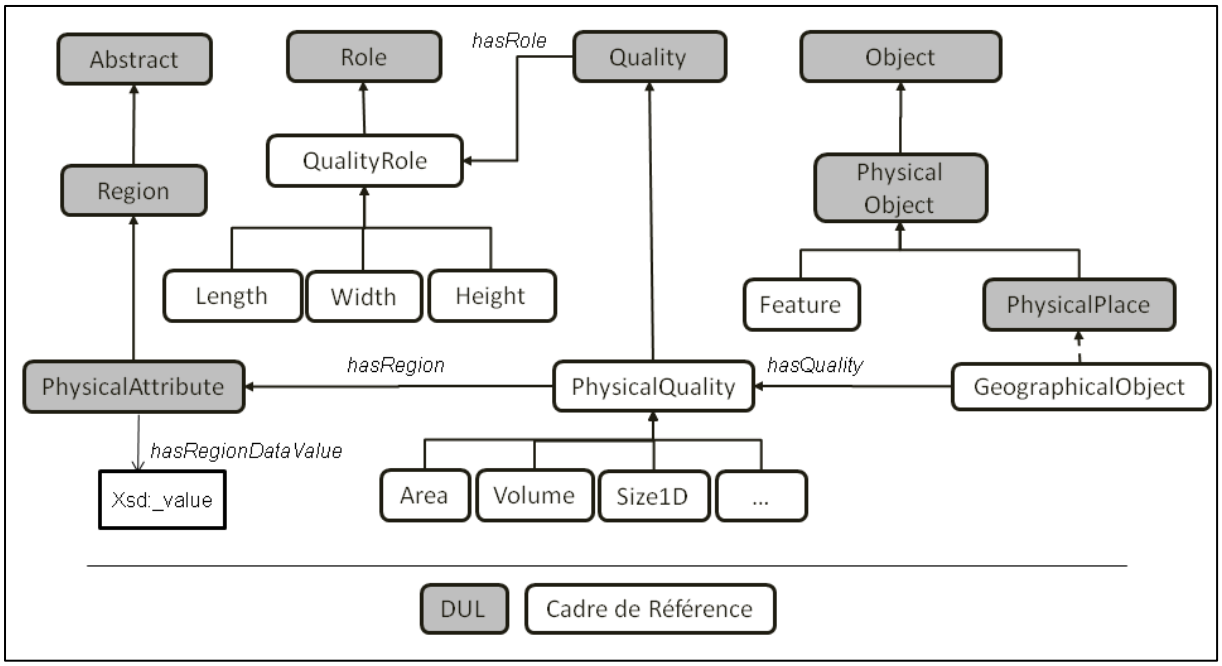


Figure 42: Cadre de référence sémantique pour la représentation des propriétés physiques des catégories d'entités topographiques

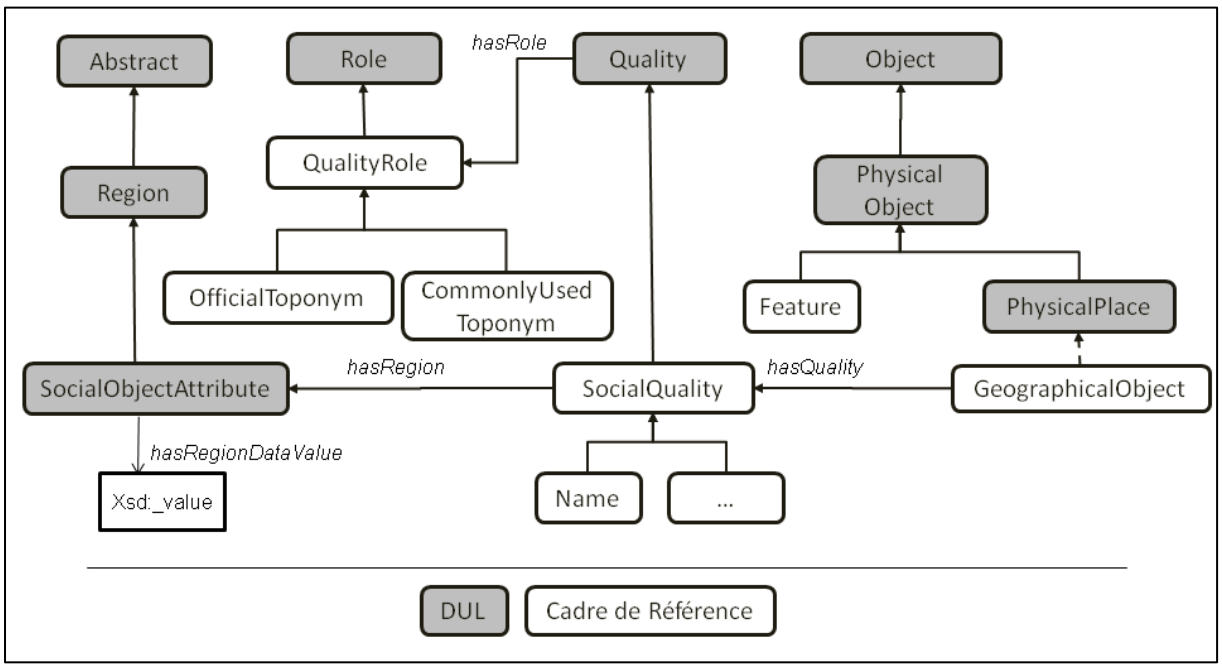


Figure 43: Cadre de référence sémantique pour la représentation des propriétés sociales des catégories d'entités topographiques

Les relations entre objets géographiques jouent également un rôle important dans les spécifications des bases de données topographiques. C'est pourquoi nous devons inclure ces relations à notre cadre de référence sémantique afin de pouvoir exprimer les contraintes issues des spécifications qui y font référence. La littérature relative aux relations spatiales et à leur représentation au sein d'ontologies est très riche. En première approche nous n'intégrons à notre cadre de référence que les relations utiles à la représentation des spécifications. Ainsi, nous ajoutons les relations de

DISTANCE et de ELEVATIONDIFFERENCE. La représentation de relations en OWL se fait généralement sous la forme d'ObjectProperties. Or, DISTANCE et ELEVATIONDIFFERENCE - distance et dénivelé - sont des relations ternaires qui ne peuvent pas être représentées de cette façon. C'est pourquoi nous proposons de les réifier (voir figure 44). La représentation des relations topologiques, en revanche, est plus simple dans la mesure où il s'agit de relations binaires qui pourront être implémentées comme des ObjectProperties. Dans un souci de réutilisation de l'existant, nous importons une ontologie de relations topologiques implémentée par l'Ordnance Survey : SpatialRelations[42]. Pour des raisons de lisibilité nous ne représentons pas sur la figure 44 les nombreuses relations spatiales binaires entre objets géographiques qu'elle décrit : ADJACENTTO, BUILTOVER, LOCATED AT, LOCATEDNEAR, SPATIALLYCONTAINS, SPATIALLYEQUAL, SPATIALLYTOUCHING, etc.

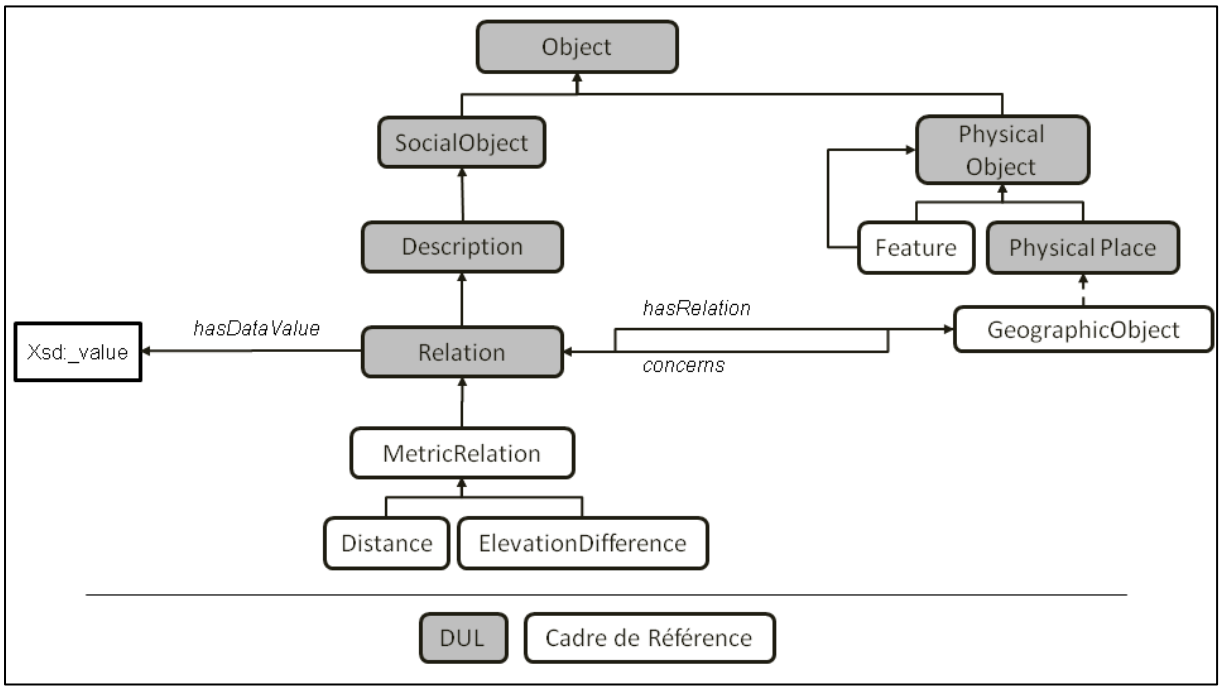


Figure 44: Cadre de référence sémantique pour la représentation des relations spatiales entre entités topographiques

Enfin, nous ajoutons à ce cadre de référence sémantique le concept de FEATURE, directement issu de Dolce. Dans ce cas précis, le terme *Feature* ne fait pas référence à une instance de base de données topographique, mais désigne une entité physique « parasite » dont l'existence dépend de celle de son hôte. Dans le cas des objets géographiques, nous introduisons le concept de SPATIALFEATURE qui englobe l'ensemble des éléments caractéristiques de la forme des objets topographiques ou dépendant de ces objets, comme leur contour ou leur centre. Ces concepts étant fondamentaux pour la description de la saisie de la géométrie des instances de classes de bases de données topographiques, nous spécialisons FEATURE afin de disposer de l'ensemble des concepts de ce type, utiles à notre objectif de formalisation de spécifications. Pour ce faire, nous suivons le modèle proposé par Schade (2010), en accord avec les travaux de Smith et Mark (1998) sur la catégorisation des concepts topographiques via leurs propriété méréologiques et topologiques. En effet, cette approche de caractérisation des catégories d'entités topographiques à l'aide des éléments

[42] http://www.ordnancesurvey.co.uk/oswebsite/ontology/

représentatifs de leur forme, relève d'une interprétation très cartographique des entités topographiques du monde réel. Elle s'adapte donc bien à notre objectif de représentation des règles de saisie de la géométrie des instances de base de données topographiques. En outre, elle introduit des connaissances pouvant être mises à profit dans le choix d'un algorithme d'appariement géométrique de données adapté (cf. partie 4.3.3.3). En effet, SPATIALFEATURE est associé à une localisation. Dans la mesure où les instances de base de données topographiques sont saisies au niveau d'un (ou plusieurs) élément(s) caractéristique(s) de la forme des objets géographiques, c'est la localisation de ces éléments caractéristiques qui leur sera attribuée et sera donc utilisée comme critère d'appariement géométrique. Le premier niveau de ce modèle est présenté en figure 45. SPATIALFEATURE se spécialise en quatre concepts que nous détaillons plus bas.

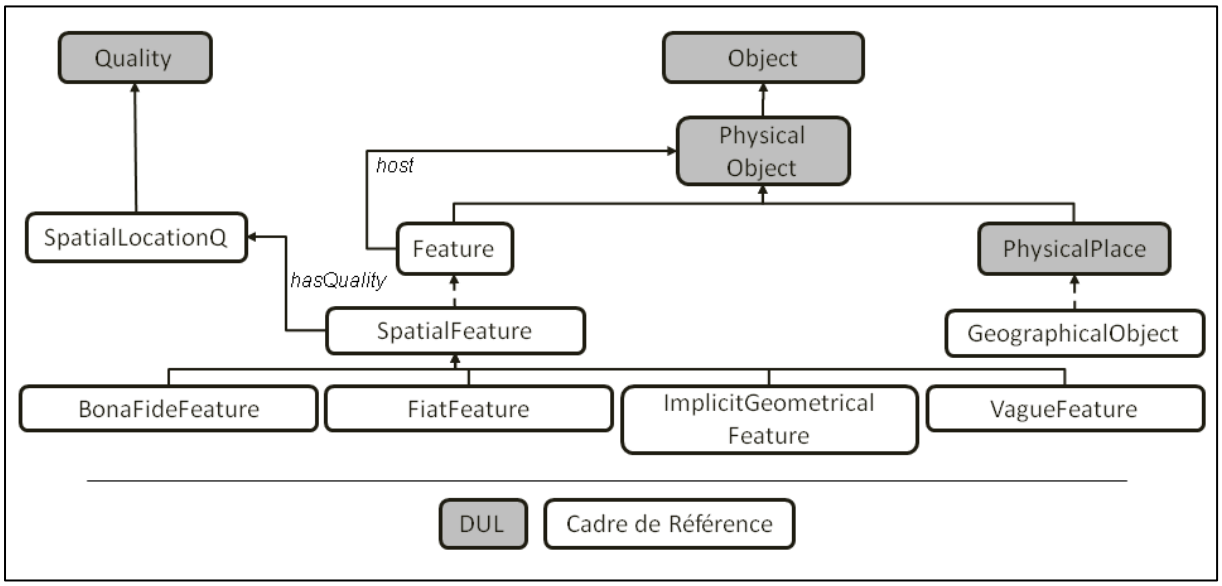


Figure 45: Cadre de référence sémantique pour la représentation des éléments caractérisant la forme des entités géographiques

La figure 46 présente le concept de BONAFIDEFEATURE et ses différentes spécialisations. Ce concept désigne l'ensemble des éléments caractéristiques de la forme d'un objet géographique, matérialisés par une discontinuité physique, et donc aisément identifiables ; c'est le cas par exemple de l'angle - CORNER - d'un bâtiment ou du contour de son toit qui correspond à une limite de type *Bona Fide* - BONAFIDEBOUNDARY.

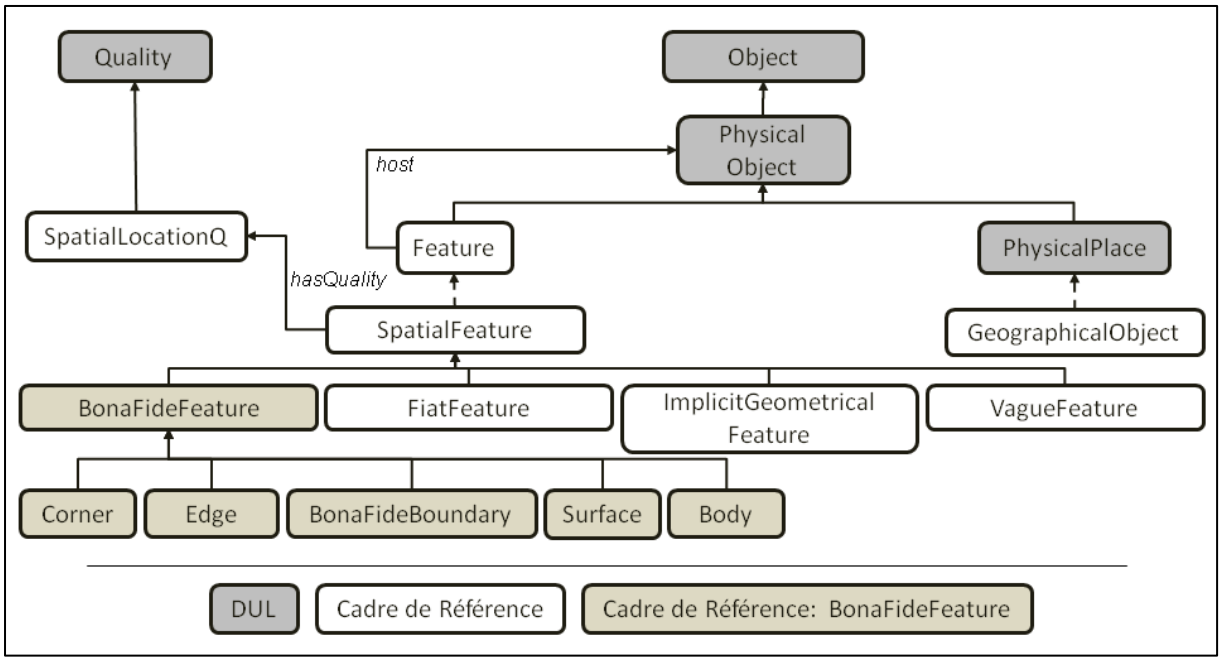


Figure 46: Cadre de référence sémantique pour la représentation des éléments caractérisant la forme des entités géographiques de type *Bona Fide*

La figure 47 présente le concept de FIATFEATURE et ses différentes spécialisations. Ce concept désigne l'ensemble des éléments caractéristiques de la forme d'un objet géographique, découlant d'un choix subjectif, voire arbitraire. Ces éléments ne sont donc pas nécessairement simples à identifier et leur interprétation peut varier d'un opérateur de saisie à l'autre; c'est le cas, par exemple, des limites administratives qui sont du type *Fiat* - FIATBOUNDARY.

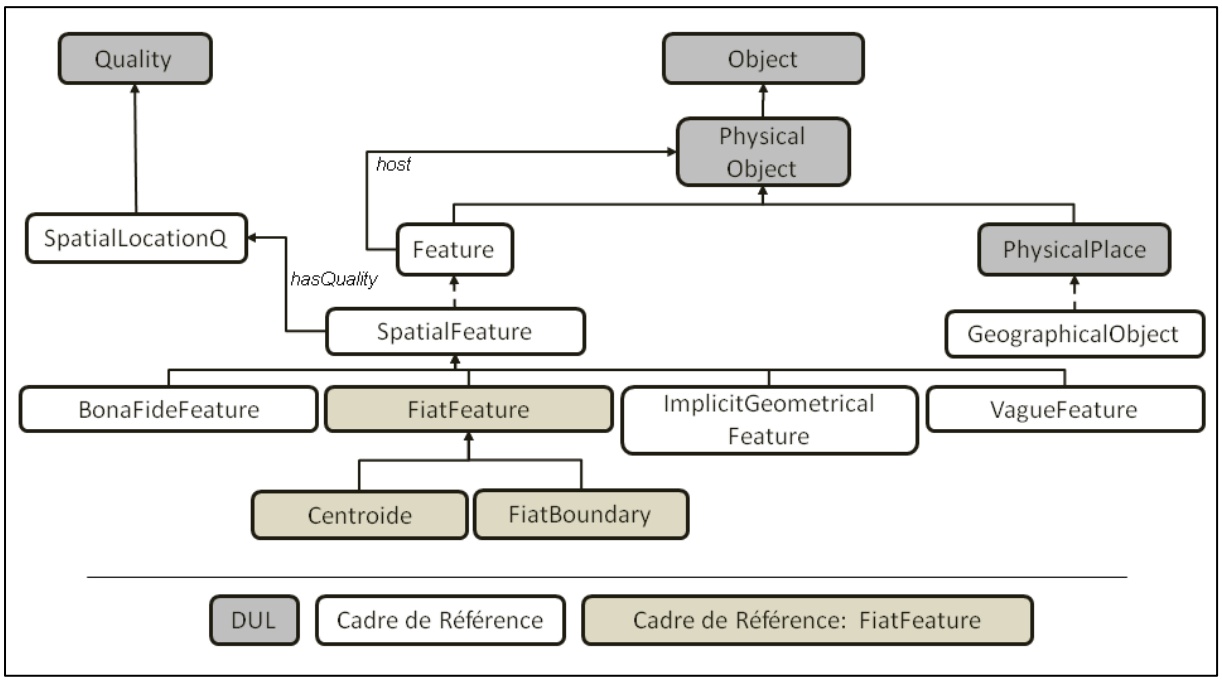


Figure 47: Cadre de référence sémantique pour la représentation des éléments caractérisant la forme des entités géographiques de type *Fiat*

La figure 48 présente le concept d' IMPLICITGEOMETRICALFEATURE et ses différentes spécialisations. Ce concept désigne l'ensemble des éléments caractéristiques de la forme d'un objet géographique pouvant être déduits pas une opération géométrique ; c'est le cas, par exemple, du centre - CENTRE - d'un bâtiment ou de l'axe - CENTRELINE -d'une route.

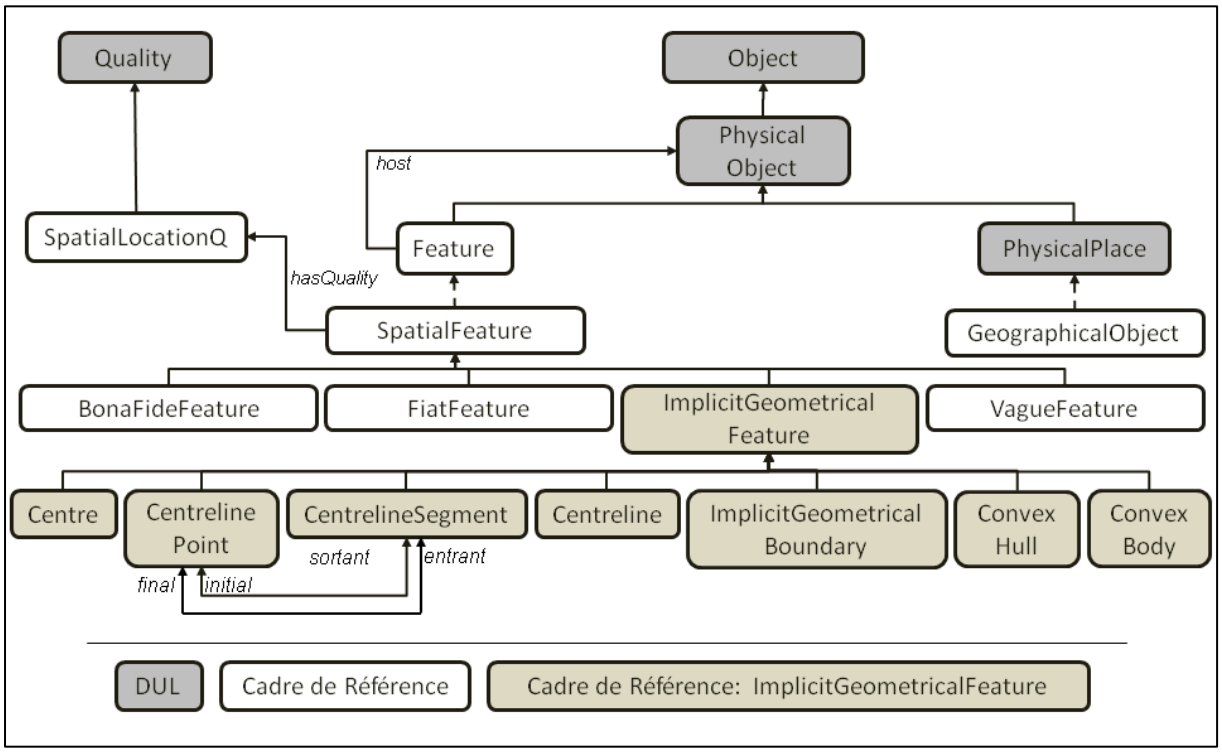


Figure 48: Cadre de référence sémantique pour la représentation des éléments géométriques implicites caractérisant la forme des entités géographiques

Enfin, la figure 49 présente le concept de VAGUEFEATURE qui désigne l'ensemble des éléments caractéristiques de la forme d'un objet géographique dont la détermination précise est impossible par essence; c'est le cas, par exemple, des limites d'une vallée.

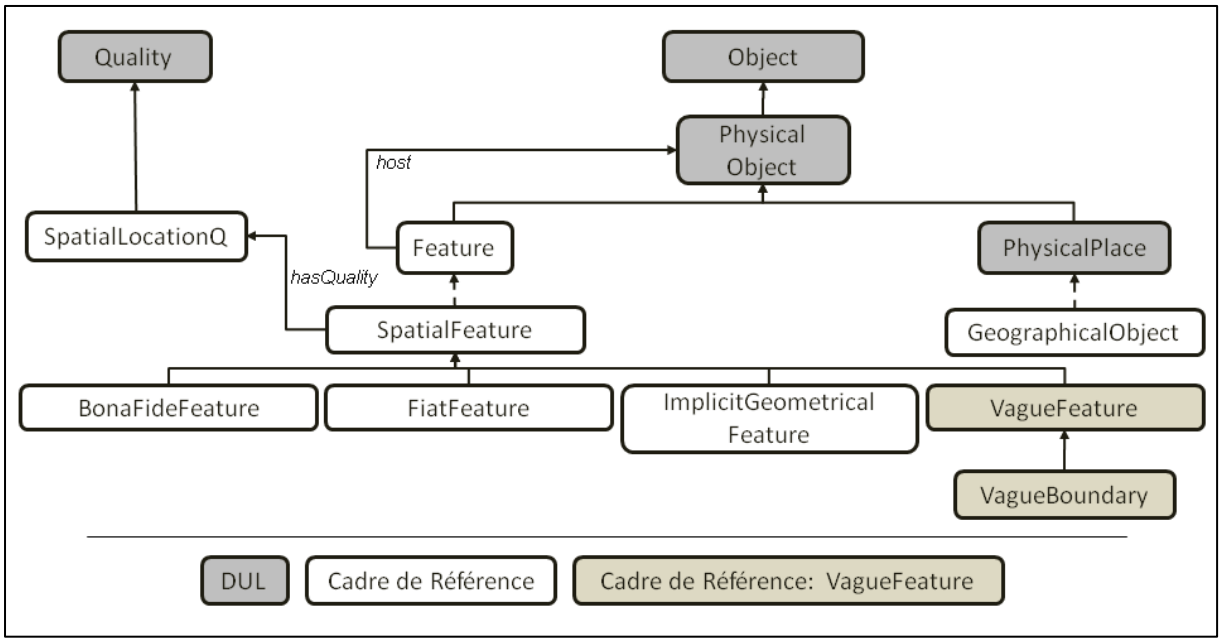


Figure 49: Cadre de référence sémantique pour la représentation des éléments caractérisant la forme des entités géographiques ne pouvant être déterminés avec précision

4.3.1.2 Annotation sémantique des éléments de schémas et de spécifications pour l'appariement des schémas

Ces trois niveaux de description, respectivement dédiés aux schémas de bases de données topographiques, à leurs spécifications, et aux entités topographiques du monde réel sont reliés par des axiomes utilisant les relations présentées en figure 50. Ainsi, les classes FEATURE et FEATURECAPTUREPROCESS, sont liées l'une à l'autre par la relation ISDEFINEDBY. La classe FEATURECAPTUREPROCESS est elle-même reliée à GEOGRAPHICALOBJECT par la relation ISPERFORMEDON. Ce modèle vise à décrire une classe de base de données topographique : cette classe est définie par un processus de saisie qui s'applique à certaines catégories d'entités topographique. Les classes GEOMETRY et MODELEDGEOMETRY, ATTRIBUTE et MODELEDPROPERTY, et ATTRIBUTEVALUE et MODELEDPROPERTYVALUE suivent le même modèle. Ainsi, les instances d'une classe de base de données topographique possèdent chacune une géométrie représentant une géométrie modélisée définie par rapport aux éléments caractéristiques de la forme des entités topographiques sélectionnées pour être saisies dans cette classe. Elles possèdent également des attributs représentant des propriétés modélisées des entités topographiques sélectionnées pour être saisies dans cette classe. Ces propriétés modélisées sont elles-mêmes définies par rapport aux propriétés réelles des entités topographiques qu'elles visent à représenter. Elles peuvent prendre un certain nombre de valeurs modélisées, définies par rapport aux valeurs réelles des propriétés des entités topographiques sélectionnées pour figurer dans la classe, dont elles cherchent à fournir une estimation dont la précision pourra varier en fonction du niveau de détail de la base de données et des sources de données disponibles.

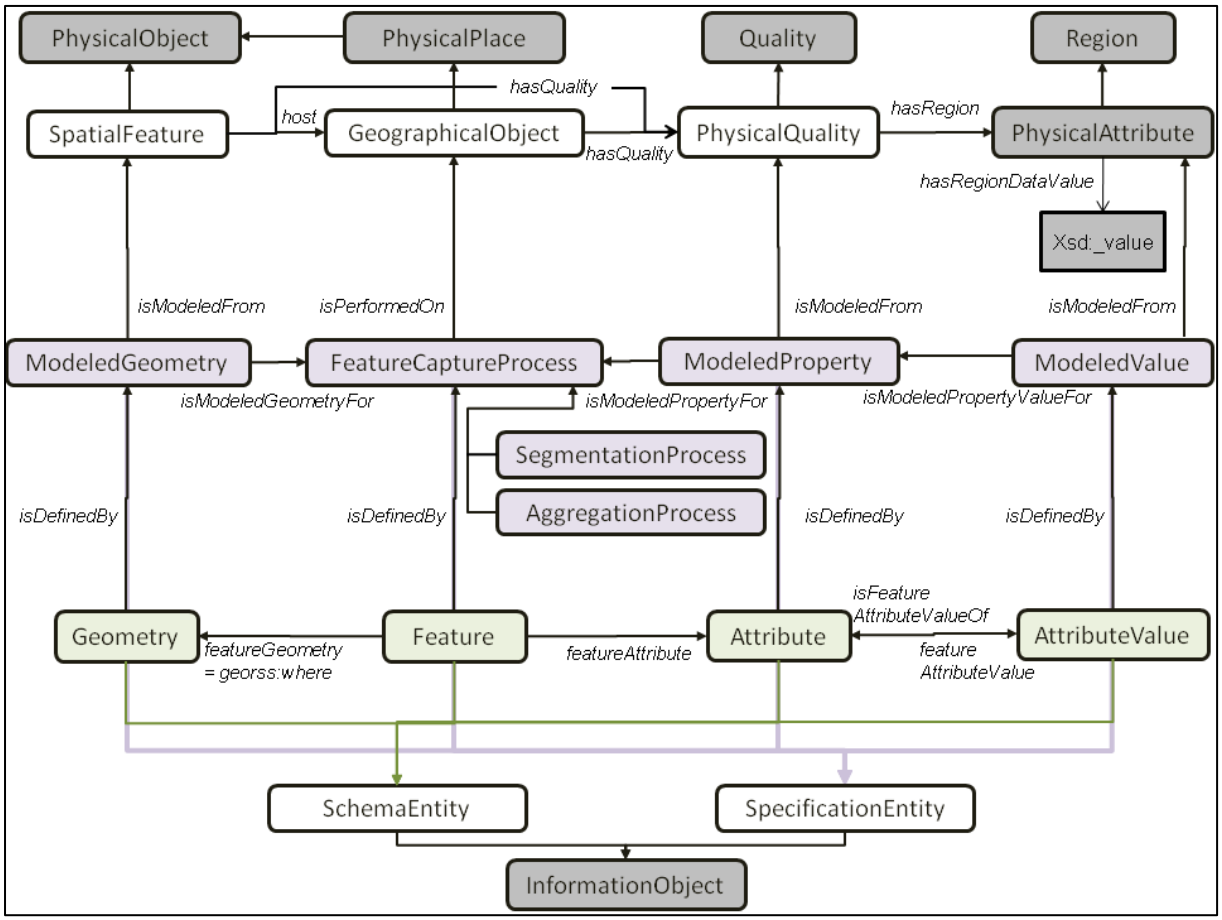


Figure 50: Modèle global pour la représentation des spécifications

4.3.1.3 Éléments de spécifications particuliers : segmentation et agrégation d'entités topographiques

Le modèle global présenté en figure 50 inclut deux classes que nous n'avons pas encore définies : SEGMENTATIONPROCESS - processus de segmentation - et AGGREGATIONPROCESS - processus d'agrégation. Il arrive en effet fréquemment que la géométrie saisie pour représenter les entités topographiques du monde réel au sein d'une base de données ne corresponde pas directement à une entité topographique directement identifiable sur le terrain, mais à une portion d'entité topographique, ou bien à un agrégat de plusieurs entités.

La segmentation d'entités topographiques est couramment pratiquée pour la représentation des réseaux de toutes sortes : routes, chemin de fer, réseaux hydrographiques ou encore électriques. L'un des objectifs premiers de cette pratique de modélisation géométrique est de permettre une représentation des réseaux, au sein des bases de données topographiques, prenant en compte leur organisation topologique. Dans le cas des réseaux routiers, ceci permet, entre autres, d'effectuer des calculs d'itinéraires. Ce procédé présente en outre l'avantage de permettre la description des réseaux à l'aide d'attributs dont les valeurs varient le long du réseau : la segmentation du réseau en tronçons, en fonction de la variation d'une propriété donnée, permet de faire porter à chaque tronçon un attribut doté d'une valeur constante reflétant la valeur locale de cette propriété. Ce découpage des réseaux est opéré en des points bien particuliers, qui constitueront les points initiaux

et finaux des tronçons saisis dans la base. Gesbert (2005) identifie quatre principaux types de points de découpage : les points de découpage correspondant à des intersections du réseau, comme les carrefours, les points de découpage liés à la présence d'obstacles sur le réseau, comme les îles situées au milieu des cours d'eau, les culs-de-sac, comme les fins d'impasses, et enfin les points au niveau desquels la valeur d'une propriété du réseau change. Nous proposons donc d'intégrer ces notions de tronçon modélisé - MODELEDSEGMENT - et de point de découpage - SEGMENTATIONNODE - à notre cadre de référence sémantique afin de permettre la représentation des règles de saisie relative au découpage des entités de type réseau.

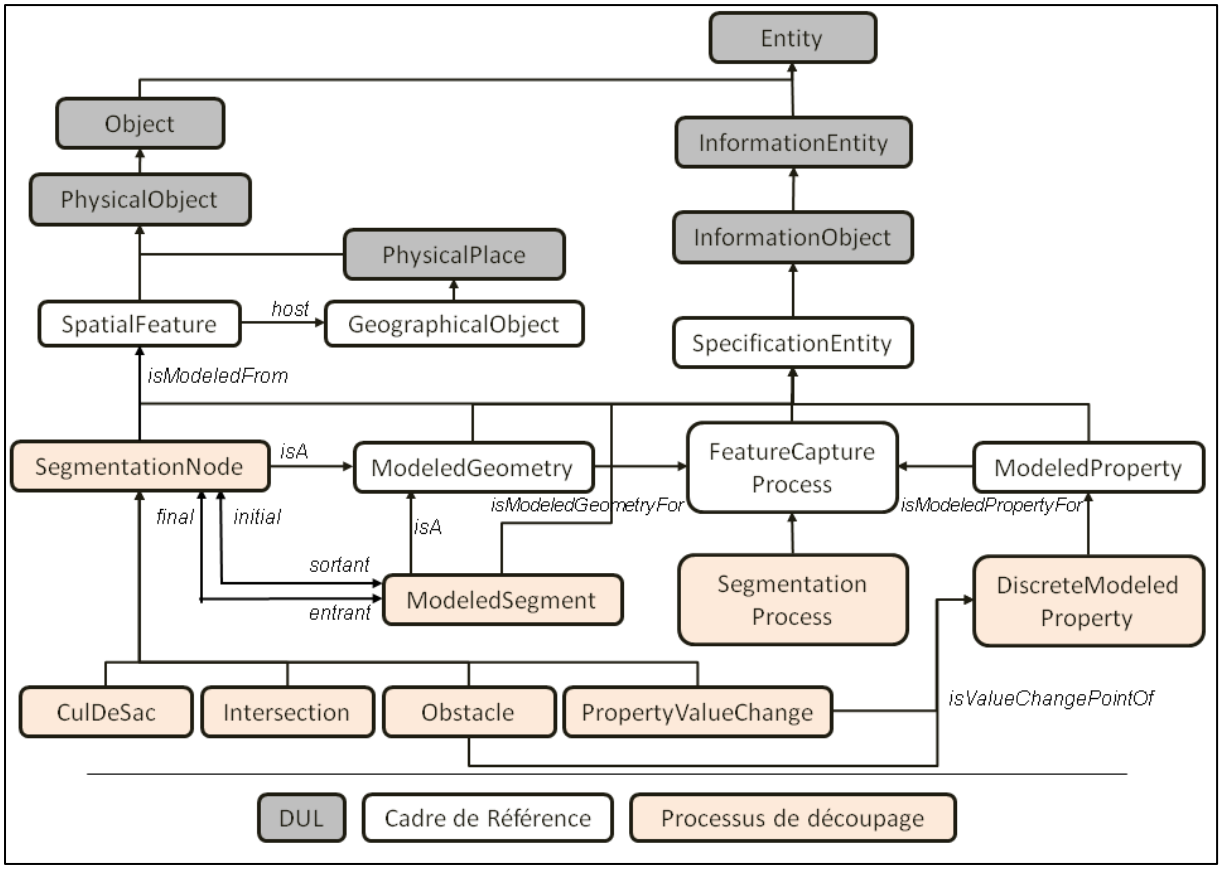


Figure 51: Cadre de référence sémantique pour la représentation du processus de découpage des entités topographiques de type réseau

Ainsi, un processus de saisie impliquant de découper certaines entités topographiques de type réseau définit une géométrie modélisée composée de segments modélisés - MODELEDSEGMENT - délimités par des points de découpage - SEGMENTATIONNODE.

Les points de découpage sont définis à l'aide d'axiomes faisant référence aux éléments caractéristiques de la forme des entités géographiques au niveau desquels un découpage doit être réalisé. Les éléments caractéristiques que nous citons dans les définitions de classes présentées ci-dessous[43] correspondent à ceux couramment pris comme référence pour la saisie des entités de type réseau, dans les spécifications de l'IGN. Ces définitions reposent sur la topologie des réseaux. Ainsi,

[43] Pour des raisons de lisibilité, les axiomes présentés dans ce mémoire seront rédigés dans la syntaxe OWL Manchester

un cul-de-sac correspond à un point situé sur l'axe de l'entité géographique saisie, et ne possédant qu'un segment entrant ou sortant. Toutefois, d'autres définitions pourront s'ajouter à celles-ci, en fonction des données à décrire.

```
Class: SegmentationNode

      EquivalentTo: (entrant some ModeledSegment) and (sortant some
      ModeledSegment)


Class: CulDeSac

      EquivalentTo:

            isModeledFrom some (CentrelinePoint and (entrant exactly 1
            (CentrelineSegment and (locatedAt some (Centreline and (host
            some GeographicalObject))))) and (sortant exactly 0
            CentrelineSegment)),

            isModeledFrom some (CentrelinePoint and (entrant exactly 0
            CentrelineSegment) and (sortant exactly 1 (CentrelineSegment
            and (locatedAt some (Centreline and (host some
            GeographicalObject))))))

      SubClassOf:

            SegmentationNode


Class: Intersection

      EquivalentTo:

            isModeledFrom some (CentrelinePoint and (((entrant min 1
            CentrelineSegment) and (sortant min 2 CentrelineSegment)) or
            ((entrant min 2 CentrelineSegment) and (sortant min 1
            CentrelineSegment))) and (locatedAt some ((Centre or
            CentrelinePoint) and (host some GeographicalObject))))

      SubClassOf:

            SegmentationNode


Class: Obstacle

      EquivalentTo:

            isModeledFrom some (CentrelinePoint and ((locatedAt some
            (Centre and (host some GeographicalObject))) and (entrant
            exactly 1 CentrelineSegment) and (sortant exactly 1
            CentrelineSegment)))

      SubClassOf:

            SegmentationNode


Class: PropertyValueChange

      EquivalentTo:
```

```
        isModeledFrom some (CentrelinePoint and ((isValueChangePointOf
        some DiscreteModeledProperty) and (entrant exactly 1
        CentrelineSegment) and (sortant exactly 1 CentrelineSegment)))

    SubClassOf: SegmentationNode
```

Si la segmentation d'entités topographiques est pratiquée à l'échelle de l'ensemble des instances d'une classe, il n'en va pas systématiquement de même pour l'agrégation qui peut ne concerner qu'une partie des instances d'une classe, choisies en fonction de critères définis dans les spécifications de cette classe. Ainsi, les spécifications de la classe Surface d'eau de la BDTOPO© 1.2 précisent que « des bassins très proches les uns des autres (séparation < 10 m […]) […] peuvent, dans certains cas, être modélisés par un seul objet englobant la zone de bassins ». Ainsi, une classe dont les spécifications font état de conditions d'agrégation d'entités topographiques sera décrite par deux classes spécialisant FEATURECAPTUREPROCESS ; l'une sera consacrée aux entités saisies indépendamment les unes des autres, l'autres aux agrégats. Gesbert (2005) identifie cinq types de critères utilisés dans les spécifications pour motiver l'agrégation de plusieurs entités géographiques au sein d'une base de données. Il s'agit de conditions sur des valeurs de propriétés que doivent vérifier chacune des entités topographiques ou bien un groupe d'entités topographiques, de critères de similitude ou de différence ou enfin de relations de proximité spatiale. Ceux-ci peuvent se traduire par des axiomes portés par les classes spécialisant AGGREGATIONPROCESS - processus d'agrégation- et AGGREGATEDGEOMETRY -géométrie agrégée, présentées en figure 52.

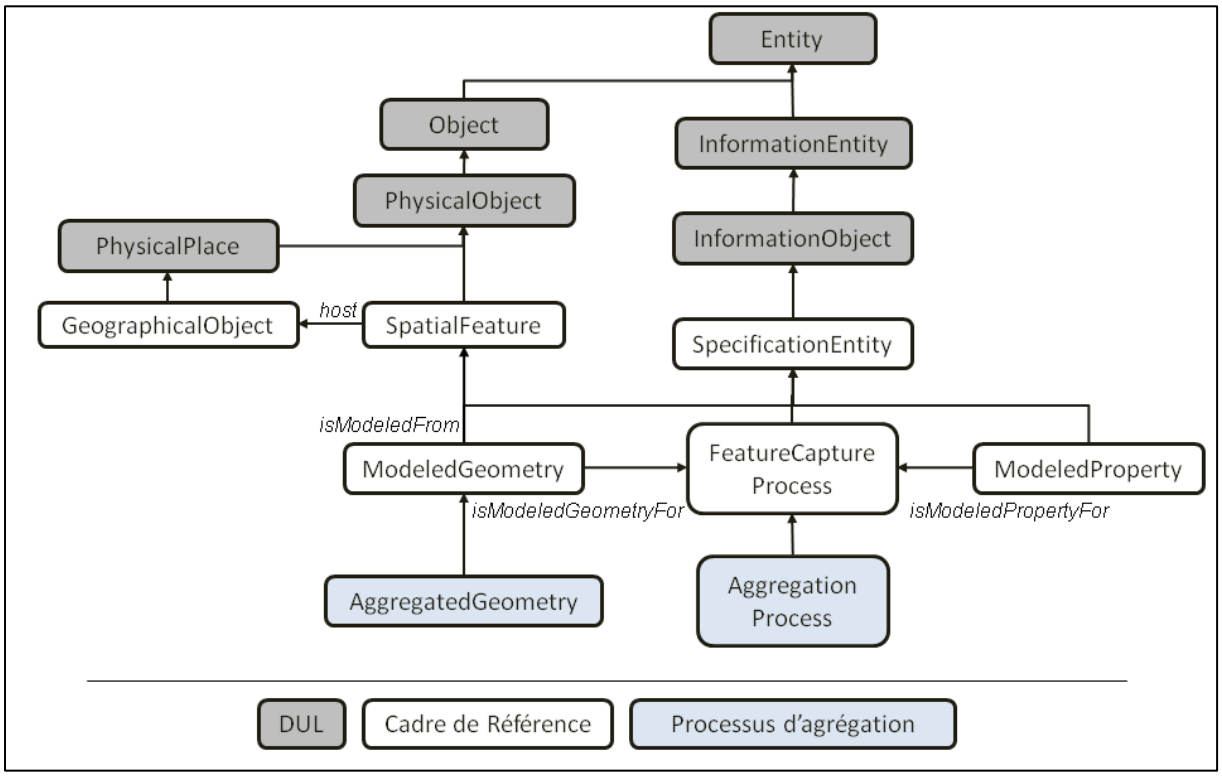


Figure 52: Cadre de référence sémantique pour la représentation du processus d'agrégation des entités topographiques

4.3.2 Instanciation du modèle : Création des ontologies d'applications et formalisation des connaissances issues des spécifications

Pour chaque base de données à intégrer, nous proposons de créer une ontologie d'application décrivant les éléments constitutifs de son schéma ainsi que les éléments de spécifications qui décrivent ces derniers.

4.3.2.1 Cas des classes simples

L'étape de création des classes constituant une ontologie d'application suit le modèle présenté en figure 50. Afin de disposer des concepts utiles à la représentation des spécifications de bases de données topographiques, chaque ontologie d'application à créer importe le cadre de référence sémantique. Chaque classe du schéma de la base est traduite sous la forme d'une classe, dotée d'un label correspondant au nom de la classe de base de données, au sein de l'ontologie d'application. Elle est ancrée au cadre de référence en tant que spécialisation de la classe FEATURE de GeoOWL. Ses attributs et valeurs énumérées sont également traduits sous la forme de classes, dotées de labels correspondant à leurs noms respectifs au sein de la base de données. Les classes représentant des attributs sont ancrées comme spécialisations de la classe ATTRIBUTE, et celles représentant des valeurs d'attributs énumérées comme spécialisations de ATTRIBUTEVALUE. Ces trois types de classes sont associées les unes aux autres à l'aide des relations FEATUREATTRIBUTE et ATTRIBUTEVALUE. La géométrie associée à chaque classe de base de données est ajoutée manuellement sous la forme d'une classe spécialisant l'une des classes géométriques de GeoOWL, et associée à FEATURE via la relation WHERE.

A chacune des classes ainsi créées pour représenter les éléments de schémas à apparier va correspondre une classe destinée à représenter les spécifications de respectives de chacun de ces éléments de schémas. Considérons la classe *Massif_Boisé* de la BDCARTO© 3.1. Elle comporte deux attributs, dont l'un est un identifiant interne dont la description n'est pas nécessaire ici. Elle est également dotée d'une géométrie de type ponctuel. La représentation de cette classe et de ses spécifications au sein d'une ontologie d'application donnera lieu à la création des classes MASSIFBOISE qui spécialise FEATURE, TOPONYME qui spécialise ATTRIBUTE, POINTMASSIFBOISE qui spécialise POINT, MASSIFBOISECP qui spécialise FEATURECAPTUREPROCESS, TOPONYMEMODELISE qui spécialise MODELEDPROPERTY et POINTMASSIFBOISEMODELISE qui spécialise MODELEDGEOMETRY. MASSIFBOISE est définie de la façon suivante :

```
Class: MassifBoise
      EquivalentTo : isDefinedBy some MassifBoiseCP
      SubClassOf :
      featureAttribute some Toponyme,
      where some PointMassifBoise
```

Avec:

```
Class: Toponyme
      EquivalentTo: isDefinedBy some ToponymeModelise
```

```
Class: PointMassifBoise
      EquivalentTo: isDefinedBy some PointMassifBoiseModelise
```

L'étape de formalisation des règles de saisie issues des spécifications est réalisée de façon manuelle, par analyse des spécifications. Ainsi, les critères de sélection des entités géographiques devant figurer au sein d'une classe de la base de données sont représentés sous la forme d'axiomes au sein de la classe spécialisant FEATURECAPTUREPROCESS. Il s'agit donc ici de traduire les contraintes de sélection sur nature, sur propriété ou sur relation identifiées par Gesbert (2005). Dans le cas de l'exemple de la classe *Massif_Boisé*, celles-ci stipulent que sont sélectionnés «[…] les bois et forêts d'une superficie supérieur à 500ha […] ». Nous proposons de le traduire de la façon suivante :

```
Class: MassifBoiseCP
      EquivalentTo :
            isPerformedOn some ((Bois or Foret)and hasRegion some
            (AreaReferenceRegion and hasRegionDataValue some
            double[>5000000]))
```

Le même procédé est appliqué pour la représentation des définitions des attributs et des règles de saisie de leurs valeurs. Il permet ici de traduire les règles d'instanciation des valeurs d'attributs identifiées par Gesbert (2005). La gestion des règles d'instanciations conditionnelles pour les valeurs énumérées sera présentée plus bas. Ainsi, l'attribut *Toponyme* qui représente le « […] toponyme éventuellement associé au massif boisé […] », sera défini de la façon suivante :

```
Class: ToponymeModelise
      EquivalentTo :
            isModeledPropertyFor some MassifBoiseCP and isModeledFrom some
            (Nom and isQualityOf some Bois)
```

Les contraintes de modélisation géométrique, que Gesbert (2005) représente également comme des règles d'instanciation propres à l'attribut *Géométrie* des instance d'une classe, sont portées par des axiomes définissant les instances de la classe spécialisant la classe ModeledGeometry. Ainsi, pour la classe *Massif_Boisé*, la contrainte de modélisation géométrique « Un massif boisé est localisé par un sommet géométrique situé en son centre géographique » se traduit :

```
Class: PointMassifBoiseModelise
      EquivalentTo:
            isModeledGeometryFor some MassifBoiseCP and isModeledFrom some
            Centre
```

La référence à la classe définissant les critères de sélection des instances de *Massif_Boisé* permet d'éviter de répéter ces critères de sélection lors de la description des règles de modélisation géométrique.

4.3.2.2 Cas des attributs cachant des concepts topographiques

Nous avons vu dans la partie 4.2 qu'il arrive fréquemment que des contraintes de modélisation conduisent à regrouper au sein d'une même classe de base de données topographique des entités topographiques appartenant à des catégories différentes, mais néanmoins sémantiquement proches.

Ces diverses entités topographiques sont représentées dans une même classe nommée à l'aide d'un terme faisant référence à une catégorie d'entités topographiques plus générique. La catégorie exacte à laquelle appartient chaque instance de la classe est précisée à l'aide d'un attribut généralement appelé *Nature*, dans les bases de données de l'IGN, dont les valeurs possibles sont précisées dans les spécifications de la base. Si diverses catégories d'entités topographiques sont regroupées au sein d'une seule et même classe, et sont donc décrites à l'aide des mêmes attributs et du même type de géométrie, certaines règles de saisie qui leurs sont appliquées peuvent néanmoins varier. En effet, il arrive fréquemment que les spécifications précisent des règles de sélection différentes pour les entités topographiques devant figurer au sein d'une classe avec différentes valeurs d'attribut *Nature*. De la même façon, celles-ci peuvent faire l'objet de contraintes de modélisation géométrique différentes. C'est le cas, par exemple, de la classe *Zone_végétation* de la BDTOPO© 2.1 qui renferme à la fois des zones arborées, des forêts fermées de feuillus, des forêts fermées mixtes, des forêts fermées de conifères, des forêts ouvertes, des peupleraies, des haies, des landes ligneuses, des vergers, des vignes, des bois, des bananeraies, des mangroves, ou des plantations de canne à sucre. Ainsi, les bananeraies doivent couvrir une « […] superficie supérieure à 30 000 m^2 […] » pour être saisies, tandis qu'une « […] superficie supérieure à 5 000 m^2 […] » suffit pour que des vignes figurent dans la classe *Zone_végétation*.

La représentation de ces différents critères de sélection et de modélisation géométrique nécessite une modélisation particulière. Pour ce faire, nous proposons d'adopter de nouveau l'approche, proposée dans la partie 4.2, qui consiste à augmenter le niveau de granularité du schéma en traitant ces valeurs d'attributs faisant directement référence à des catégories d'entités topographiques comme des classes. Ces classes sont définies comme des spécialisations de la classe possédant les valeurs d'attributs dont elles sont issues. Dans le cas de la classe *Zone_végétation*, cela conduit à définir quatorze classes filles définies de la façon suivante :

```
Class: ZoneVegetation
      SubClassOf:
      gml: Feature,
      featureAttribute some Nature


Class: Bananeraie
      EquivalentTo:
            ZoneVegetation and hasAttribute some (Nature and hasDataValue
            value "bananeraie")
      SubClassOf: where some PolygoneBananeraie


Class: Vigne
      EquivalentTo:
            ZoneVegetation and hasAttribute some (Nature and hasDataValue
            value "vigne")
      SubClassOf: where some PolygoneVigne
Etc.
```

Pour chacune de ces quatorze classes sont définies des classes décrivant leurs éléments de spécifications propres. Ainsi, les critères de sélection de la classe des bananeraies sont représentés de la façon suivante :

```
Class: Bananeraie
      EquivalentTo: isDefinedBy some BananeraieCP
Class: BananeraieCP
      EquivalentTo:
            isPerformedOn some (Bananeraie and (hasRegion only
            (AreaReferenceRegion and (hasRegionDataValue some xsd:double[>
            30000.0])))))
```

Les règles de modélisation géométrique sont représentées de la même façon que pour les classes simples.

```
Class: PolygoneBananeraieModelise
      EquivalentTo:
            isModeledGeometryFor some BananeraieCP and isModeledFrom some
            FrontièreBonaFide
```

L'exemple des bananeraies est relativement simple. D'autres classes, en revanche, présentent des critères de sélection ou de modélisation géométrique plus complexes. C'est le cas des forêts fermées de feuillus dont la représentation géométrique prend en compte la présence de clairières lorsque celles-ci remplissent des conditions précises. Nous représentons cette condition en précisant que la limite de la forêt devant être saisie peut concerner une clairière ayant diverses propriétés. Ainsi, dans l'exemple ci-dessous, les clairières saisies doivent avoir une superficie supérieure à 5000 m^2.

```
Class: PolygoneForetFermeeFeuillusModelise
      EquivalentTo:
            (isModeledGeometryFor some ForetFermeeFeuillusCP) and
            (isModeledFrom some (BonaFideBoundary and (host some (Clairière
            and (hasRegion only (AreaReferenceRegion and
            (hasRegionDataValue some xsd:double[> 5000.0])))))))
```

L'exemple suivant représenta la règle de modélisation géométrique des forêts fermées de feuillus prenant en compte les clairières bâties de plus de 500 m^2.

```
Class: PolygoneForetFermeeFeuillusModelise
      EquivalentTo:
            (isModeledGeometryFor some ForetFermeeFeuillusCP) and
            (isModeledFrom some(BonaFideBoundary and (host some (Clairière
            and (hasRegion only (AreaReferenceRegion and
            (hasRegionDataValue some xsd:double[> 500.0]))) and
            spatiallyContains some Bâti))))
```

Les attributs de la classe *Zone_Végétation* ayant des règles de saisie identiques pour l'ensemble des instances de la classe sont directement associés à son processus de saisie. C'est le cas pour l'attribut *Prec_Plani* qui indique la précision planimétrique des géométries des instances de la classe. Il s'agit d'un attribut doté d'un ensemble fini de valeurs possibles, dont les règles d'instanciation

conditionnelles sont définies dans les spécifications. La représentation de ces règles est réalisée de la façon suivante :

```
Class: ZoneVegetation
      EquivalentTo: isDefinedBy some ZoneVegetationCP
      SubClassOf:
      gml: Feature,
      featureAttribute some Prec_Plani

Class: Prec_Plani
      EquivalentTo: isDefinedBy some Prec_Plani_Modelise
      SubClassOf:
      hasAttributeValue some ZeroVCinq_ZV,
      hasAttributeValue some DeuxVCinq_ZV
      […]

Class: ZeroVCinq_ZV
      EquivalentTo: isDefinedBy some ZeroVCinq_ZV_Modelise
Class: DeuxVCinq_ZV
      EquivalentTo: isDefinedBy some DeuxVCinq_ZV_Modelise
      […]

Class: Prec_Plani_ Modelise
      EquivalentTo:
            isModeledPropertyFor some ZoneVegetationCP and isModeledFrom
            some (PrecisionQuality and (isQualityOf some (DataSource)))
Class: ZeroVCinq_ZV _ Modelise
      EquivalentTo:
            (isModeledPropertyValueFor some Prec_Plani_ Modelise) and
            (isModeledFrom some ((PrecisionReferenceRegion and (isRegionFor
            some (PrecisionQuality and (isQualityOf some LeveGPS))))
            and (hasRegionDataValue some xsd:double[< 0.5])))
Class: UnVCinq_ZV _ Modelise
      EquivalentTo:
            (isModeledPropertyValueFor some Prec_Plani_ Modelise) and
            (isModeledFrom some ((PrecisionReferenceRegion and (isRegionFor
            some (PrecisionQuality and (isQualityOf some
            Photogrammetrie))))     and (hasRegionDataValue some
            xsd:double[> 0.5, < 1.5])))
[…]
```

4.3.2.3 Cas des entités topographiques découpées

La représentation des règles de saisie des entités géographiques de type dotées d'une organisation de type réseau nécessite de définir les points du réseau au niveau desquels ces entités seront découpées afin de créer les objets qui seront instanciés dans la base de données. Nous proposons de décrire ces points à l'aide d'axiomes faisant référence aux éléments caractéristiques de la forme des entités géographiques, au niveau desquels un découpage doit être réalisé. Considérons l'exemple de la classe *Tronçon de cours d'eau* de la BDTOPO© Pays 1.2. Cette classe possède un attribut nommé *Franchissement* dont les valeurs font référence à des concepts topographiques tels les BARRAGES ou les CASCADES. Cet attribut est destiné à indiquer la nature exacte de la portion de cours d'eau considérée. Dans les exemples que nous présentons plus bas, nous ne traiterons que le cas général des tronçons de cours d'eau, de type RIVIÈRE ou FLEUVE. Ainsi, si l'on ne considère que les cours d'eau indifférenciés, les règles de sélection des instances de la classe précisent que sont saisis « les cours d'eau nommés, permanent ou intermittents »

```
Class : TronconCoursDEauCP
      EquivalentTo :
            isPerformedOn some (CoursDEau and (hasQuality some NomQuality)
            and (hasQuality some (RegimeDesEauxQuality and (hasRegion only
            (RegimeDesEauxReferenceRegion and (hasRegionDataValue value
            "permanent")))))),
            isPerformedOn some (CoursDEau and (hasQuality some NomQuality)
            and (hasQuality some (RegimeDesEauxQuality and (hasRegion only
            (RegimeDesEauxReferenceRegion and (hasRegionDataValue value
            "intermittent")))))),
            isPerformedOn some (CoursDEauArtificiel and (not (adjacentTo
            some Route)) and (hasQuality some NameQuality) and (hasQuality
            some (RegimeDesEauxQuality and (hasRegion only
            (RegimeDesEauxReferenceRegion and (hasRegionDataValue value
            "intermittent"))))))
```

De plus, l'analyse des spécifications de cette classe permet de déduire que trois types de points de découpage sont appliqués lors de la saisie des instances de la classe. Il s'agit tout d'abord d'intersections avec d'autres cours d'eau, c'est-à-dire de confluences. On définit donc un point de découpage POINTINTERSECTIONCOURSDEAU décrivant ce type de points de découpage :

```
Class: PointIntersectionCoursDEau
      EquivalentTo:
            isModeledFrom some (CentrelinePoint and (((entrant min 1
            CentrelineSegment) and (sortant min 2 CentrelineSegment)) or
            ((entrant min 2 CentrelineSegment) and (sortant min 1
            CentrelineSegment))) and (locatedAt some (Centre and (host some
            Confluence))))
```

Le deuxième type de point de découpage concerne les culs-de-sac, c'est-à-dire les extrémités du réseau hydrographique, qu'il s'agisse de sources ou bien d'embouchures ou de pertes. Leur représentation est la suivante :

```
Class : PointCulDeSacCoursDEau
```

```
EquivalentTo :

    isModeledFrom some (CentrelinePoint and (entrant exactly 1
    (CentrelineSegment and (locatedAt some (Centreline and (host
    some CoursDEau))))) and (sortant exactly 0 CentrelineSegment)),

    isModeledFrom some (CentrelinePoint and (entrant exactly 0
    CentrelineSegment)and (sortant exactly 1 (CentrelineSegment
    and (locatedAt some (Centreline and (host some CoursDEau))))))
```

Enfin, le troisième type de points de découpage décrit par les spécifications concerne les changements de valeurs de propriétés des cours d'eau. Les propriétés concernées sont relativement nombreuses, ce qui implique un découpage relativement fin du réseau :

```
Class : PointChangementPpteCoursDEau

EquivalentTo :

    isModeledFrom some (CentrelinePoint and ((isValueChangePointOf
    some ModeledArtif) and (entrant exactly 1 CentrelineSegment)
    and (sortant exactly 1 CentrelineSegment))),

    isModeledFrom some (CentrelinePoint and ((isValueChangePointOf
    some ModeledFictif) and (entrant exactly 1 CentrelineSegment)
    and (sortant exactly 1 CentrelineSegment))),

    isModeledFrom some (CentrelinePoint and ((isValueChangePointOf
    some ModeledPos_Sol) and (entrant exactly 1 CentrelineSegment)
    and (sortant exactly 1 CentrelineSegment))),

    isModeledFrom some (CentrelinePoint and ((isValueChangePointOf
    some ModeledRegime) and (entrant exactly 1 CentrelineSegment)
    and (sortant exactly 1 CentrelineSegment)))
```

La représentation des divers types de points de découpage possibles nous permet de décrire les types de segments de cours d'eau définis par le processus de découpage. Nous identifions deux types de segments de cours d'eau. D'une part, les tronçons de cours d'eau situés aux extrémités du réseau :

```
Class : TronconCoursDEauCulsDeSacModelise

EquivalentTo :

    ((final exactly 1 CulDeSacTCE) and (initial exactly 1
    IntersectionTCE)),

    ((final exactly 1 IntersectionTCE) and (initial exactly 1
    CulDeSacTCE)),

    (((final exactly 1 PropertyValueChangeTCE) and (initial exactly
    1 CulDeSacTCE)),

    ((final exactly 1 PropertyValueChangeTCE) and (initial exactly
    1 CulDeSacTCE))
```

D'autre part, les autres tronçons constituent le corps des cours d'eau:

```
Class : TronconCoursDEauModelise

EquivalentTo :
```

```
(isModeledGeometryFor some TronconCoursDEauCP) and ((final
exactly 1 (IntersectionTCE or PropertyValueChangeTCE))and
(initial exactly 1 (IntersectionTCE or
PropertyValueChangeTCE)))
```

Ainsi, la géométrie des instances de la classe *Tronçon de cours d'eau* sera saisie en accord avec les segments de cours d'eau définis par les spécifications :

```
Class : LigneTronconCoursDEau
      EquivalentTo :
            isDefinedBy some (TronconCoursDEauCulsDeSacModelise and
            (hasRegion only (LengthReferenceRegion and (hasRegionDataValue
            some double[> 100.0]))))),
            isDefinedBy some TronconCoursDEauModelise
```

Notons que la description des règles de saisie de la géométrie des tronçons de cours d'eau comporte une restriction sur la taille des segments de cours d'eau modélisés devant être instanciés dans la classe. Afin de produire des données dotées d'une granularité conforme au niveau de détail de la base de données, certains segments modélisés, trop petits par rapport au niveau de détail de la base, sont en effet agrégés avec leurs voisins lors de la saisie de la géométrie des instances.

Les attributs et leurs éventuelles valeurs énumérées sont décrits selon l'approche présentée dans les parties précédentes (4.3.2.1 et 4.3.2.2).

4.3.2.4 Cas des entités topographiques agrégées

La représentation des règles de saisie des entités géographiques devant être agrégées dans la base nécessite de définir deux processus de saisie. Le premier est destiné à la description des règles de sélection des entités qui ne subissent pas d'agrégation, le second à celles des entités devant être agrégées. Deux géométries modélisées seront donc utiles pour décrire les règles de modélisation géométrique des instances de la classe. La première est consacrée aux géométries des entités non agrégées, et la seconde à celles des entités agrégées dont la saisie peut parfois faire l'objet de règles de modélisation différentes de celles appliquées à la saisie des entités non agrégées. Considérons l'exemple de la classe *Surface d'eau* de la BDTOPO© Pays 1.2. Cette classe possède un attribut nommé *Nature* dont les valeurs font référence à des concepts topographiques tels les BASSINS ou les SURFACES D'EAU. Cet attribut est destiné à indiquer la nature exacte de l'étendue d'eau considérée. Ainsi, chacune de ces valeurs d'attribut *Nature* sera traité comme une classe fille de FEATURE lors de l'instanciation de l'ontologie d'application décrivant les *Surfaces d'eau*. Dans le cas des *Bassins*, les règles de sélection des instances de la classe précisent que sont saisis « Bassins d'élevage piscicole, le bassins d'épuration, les bassins de décantation, les bassins de filtrage […], les retenues collinaires, les salines et les viviers de plus de 10m de long et 5m de large ». Ces instances sont représentées par des polygones saisis au niveau du « rebord extérieur du bassin ». Une contrainte de modélisation géométrique précise que « des bassins très proches les uns des autres (séparation < 10m […])[…] peuvent être modélisés par un seul objet englobant la zone de bassins ». La représentation de ces règles de saisie dans notre modèle donne les axiomes suivants :

```
Class: Bassin
```

```
    EquivalentTo:
        isDefinedBy some BassinCP,
        isDefinedBy some BassinAgregeCP
    SubClassOf:
        where some PolygoneBassin

Class: PolygoneBassin
    EquivalentTo:
        isDefinedBy some BassinAgregeModelise,
        isDefinedBy some BassinModelise

Class: BassinCP
    EquivalentTo:
      isPerformedOn some (Réservoir and (hasRegion only
      (LengthReferenceRegion and (hasRegionDataValue some xsd:double[>
      10.0]))) and (hasRegion only (WidthReferenceRegion
      and (hasRegionDataValue some xsd:double[> 5.0])))),
      isPerformedOn some (BassinDÉpuration and (hasRegion only
      (LengthReferenceRegion and (hasRegionDataValue some xsd:double[>
      10.0]))) and (hasRegion only (WidthReferenceRegion
      and (hasRegionDataValue some xsd:double[> 5.0])))
    SubClassOf: FeatureCaptureProcess

Class: BassinAgregeCP
    EquivalentTo:
      isPerformedOn some (Réservoir and ((hasSpatialRelation some (Distance
      and (concerns some Bassin))) and (hasDataValue some xsd:double[> 0.0
      ,< 10.0])) and (hasRegion only  (LengthReferenceRegion and
      (hasRegionDataValue some xsd:double[> 10.0]))) and (hasRegion only
      (WidthReferenceRegion and (hasRegionDataValue some xsd:double[>
      5.0])))),
      isPerformedOn some (BassinDÉpuration and ((hasSpatialRelation some
      (Distance and (concerns some Bassin))) and (hasDataValue some
      xsd:double[> 0.0 , < 10.0])) and (hasRegion only
      (LengthReferenceRegion and (hasRegionDataValue some xsd:double[>
      10.0]))) and (hasRegion only (WidthReferenceRegion and
      (hasRegionDataValue some xsd:double[> 5.0]))))
    SubClassOf: AggregationProcess

Class: BassinModelise
    EquivalentTo:
```

```
      (isModeledFrom some BonaFideBoundary) and (isModeledGeometryFor some
      BassinCP)
    SubClassOf: ModeledGeometry


Class: BassinAgregeModelise
    EquivalentTo:
      (isModeledFrom some BonaFideBoundary) and (isModeledGeometryFor some
      BassinCP)
    SubClassOf: referenceFrame:AggregatedGeometry
```

4.3.3 Exploitation du modèle

Disposer de spécifications formelles décrivant précisément la sémantique du contenu de bases de données topographiques hétérogènes est utile pour diverses applications relevant du domaine de l'intégration virtuelle de bases de données topographiques. Nous présentons, dans la suite de cette partie, deux applications s'appuyant sur des spécifications formalisées au sein d'ontologies d'applications. La première met en œuvre l'approche proposée dans cette partie pour l'appariement fin de schémas de bases de données topographiques, la seconde concerne la découverte de bases de données topographiques hétérogènes.

4.3.3.1 Exploitation du modèle pour l'appariement de schémas

L'exploitation des règles de saisie des classes *Massif Boisé* de la BDCARTO© 3.1, *Zone_végétation* de la BDTOPO© 2.1, instanciées conformément à l'approche présentée dans les parties 4.3.2.1 et 4.3.2.2, pour l'appariement fin de schémas de bases de données topographiques repose sur les possibilités de raisonnement associées à la logique de description sur laquelle repose le langage de représentation de connaissances OWL 2 que nous avons choisi d'utiliser ici.

Ainsi, pour chaque classe de base de données, une ontologie d'application décrivant l'ensemble des règles de saisie des instances de la classe est instancié manuellement, à l'aide du logiciel d'édition d'ontologies Protégé. Les ontologies d'applications ainsi créées sont ensuite chargées dans une application Java utilisant l'API OWL [44] pour la manipulation des ontologies. Elles sont fusionnées deux à deux, puis, nous traitons chaque ontologie issue de la fusion d'une paire d'ontologies d'applications à l'aide de l'un des systèmes de raisonnement disponibles pour les ontologies OWL 2 : Hermit[45]. Les relations d'équivalence ou de subsomption détectées par le système de raisonnement sont ensuite sauvegardées dans une troisième ontologie. La figure 53 présente l'affichage sous Protégé des résultats obtenus par le système de raisonnement (les axiomes inférés sont affichés en surbrillance, en jaune).

[44] http://owlapi.sourceforge.net/
[45] http://hermit-reasoner.com/

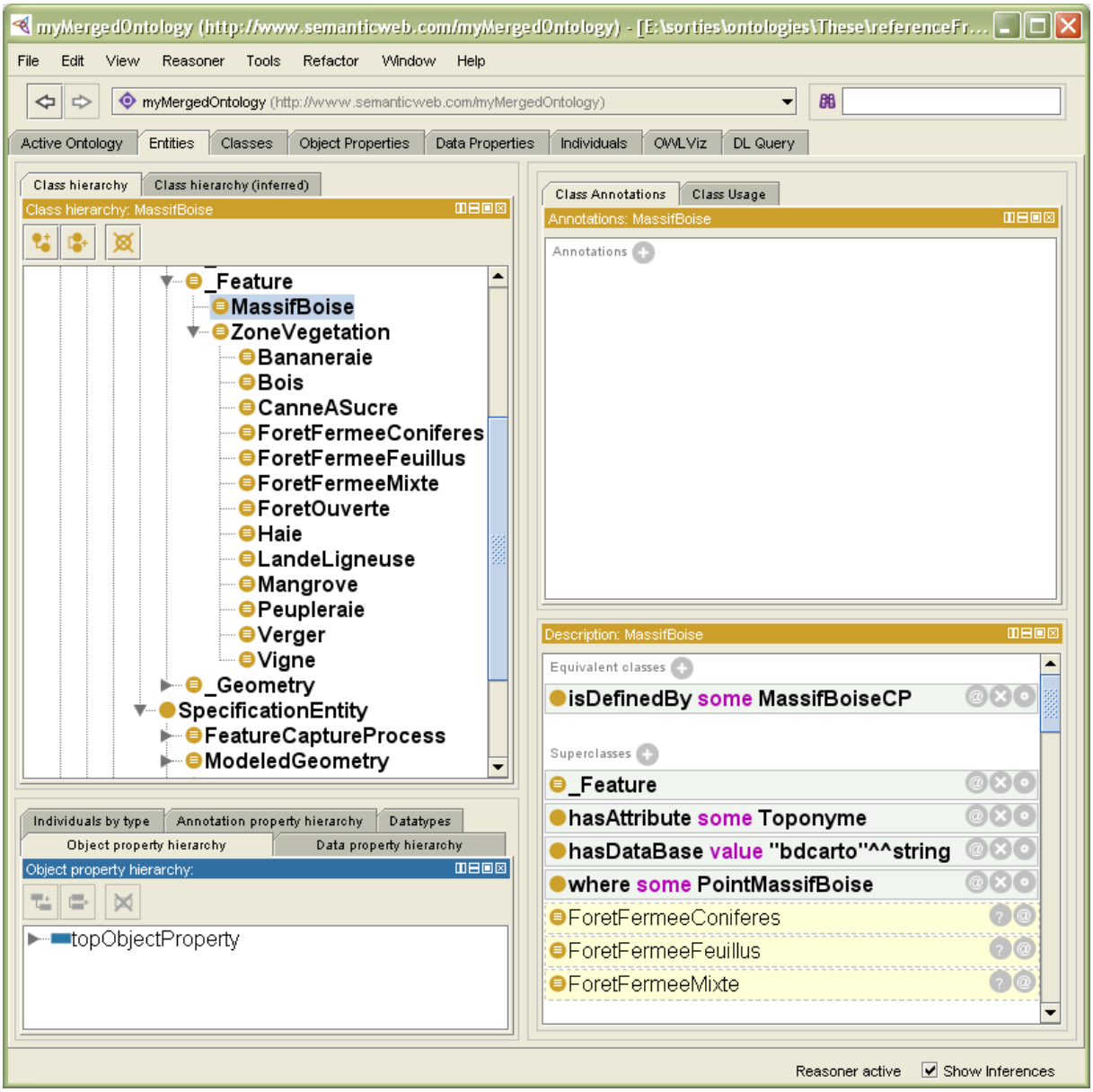


Figure 53: Utilisation d'un système de raisonnement pour ontologies pour l'appariement des classes *Massif Boisé* et *Zone_Végétation* (Visualisation sous Protégé)

La classe *MassifBoise* de la BDCARTO© 3.1 a été identifiée comme une spécialisation des classes *ForetFermeeConiferes*, *ForetFermeeFeuillus* et *ForetFermeeMixte*, qui désignent respectivement les sous-ensembles d'instances de la classe *Zone_Végétation* de la BDTOPO© 2.1 ayant comme valeurs d'attribut *Nature*, *Forêt fermée de conifères, Forêt fermée de feuillus* et *Forêt fermée mixte*. Ce résultat est directement calculé à partir des axiomes représentant les règles de sélection des entités topographiques devant figure dans ces classes. Les instances de *Massif Boisé* représentant des entités topographiques appartenant aux mêmes catégories que les entités représentées par *ForetFermeeConiferes*, *ForetFermeeFeuillus* et *ForetFermeeMixte*, mais dotées d'une superficie supérieure, elles constituent *a priori* un sous-ensemble des instances de ces trois classes.

Notons que nous n'avons pas représenté ici l'une des valeurs énumérées de *Zone_Végétation* , « *Zone Arborée* ». Celle-ci correspond en fait à une classe présente dans une version antérieure de la BDTOPO©, qui doit être progressivement remplacée par *Zone_Végétation*, par répartition de ses

instances dans les autres valeurs d'attribut *Nature*. L'appariement de cette classe avec *Massif Boisé* montre que les instances de *Massif Boisé* constituent un sous ensemble de celles de *Zone Arborée*. (voir figure 54).

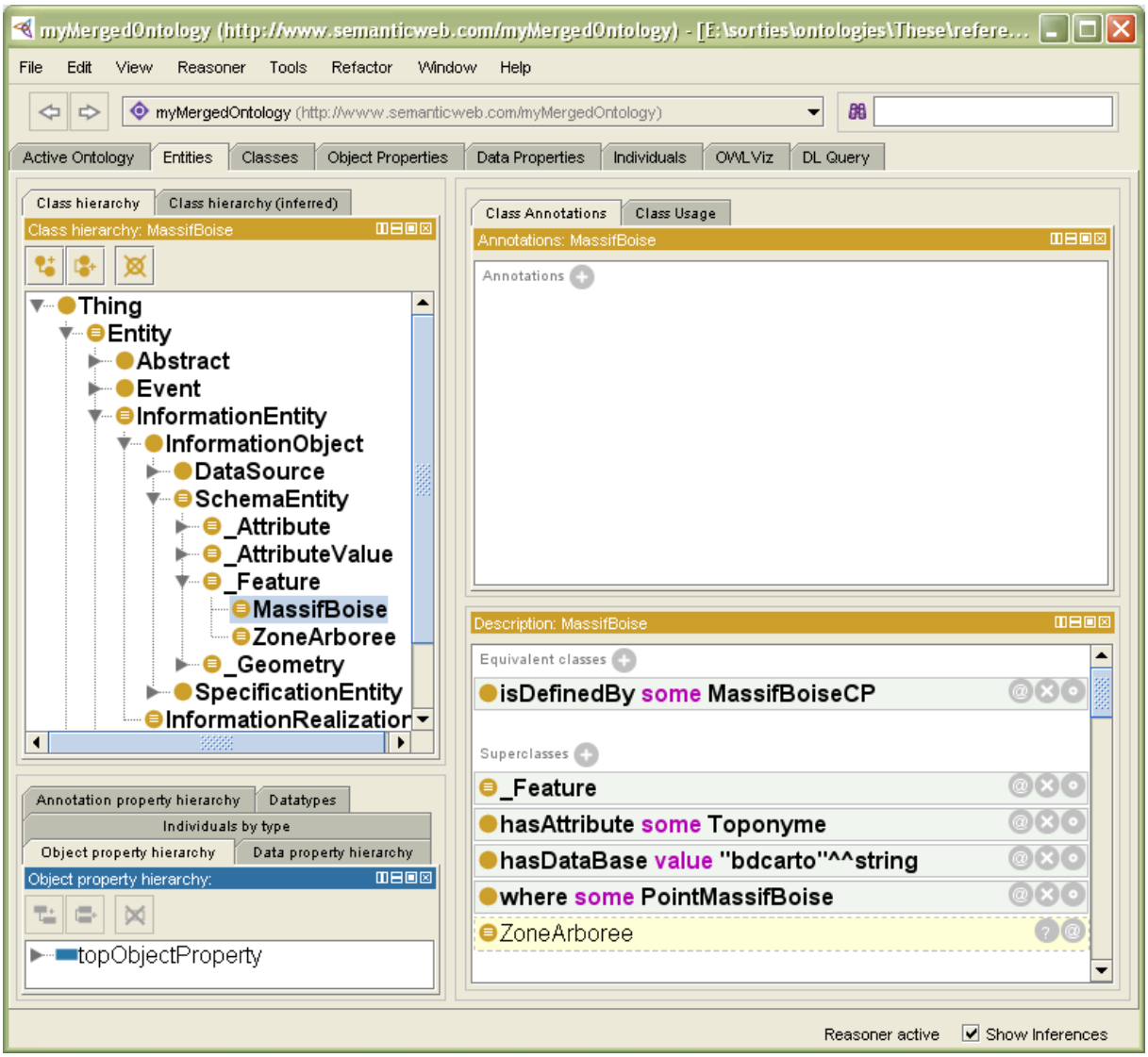


Figure 54: Utilisation d'un système de raisonnement pour ontologies pour l'appariement des classes *Massif Boisé* et *ZoneArborée* (Visualisation sous Protégé)

L'obtention de relations de correspondance plus fines nécessite d'analyser les axiomes des classes mises en relation par le système de raisonnement. Pour ce faire, nous avons étendu cette première application pour permettre l'extraction de relations de correspondances fines entre éléments de schémas de bases de données topographiques. Celle-ci prend, en entrée, l'ontologie inférée obtenue à l'étape précédente, compare les axiomes décrivant les éléments de spécifications de chacune des classes identifiées comme ayant une relation de correspondance, et fournit en sortie une liste de relations de correspondances. Nous obtenons donc, pour l'exemple des classes *Massif Boisé* et *Zone_Végétation* les résultats suivants :

- Les instances de la classe *Zone_Végétation* ayant comme valeur d'attribut *Nature* « *Forêt fermée de feuillus »* et représentant des entités topographiques d'une superficie supérieure à

500 hectares représentent les mêmes entités topographiques que les instances de la classe *Massif Boisé*. Les premières les représentent sous la forme de polygones saisis au niveau de leurs frontières de type *bona fide*, les secondes sous la forme de points saisis au niveau de leurs centres respectifs.

- Les instances de la classe *Zone_Végétation* ayant comme valeur d'attribut *Nature* « *Forêt fermée mixte »* et représentant des entités topographiques d'une superficie supérieure à 500 hectares représentent les mêmes entités topographiques que les instances de la classe *Massif Boisé*. Les premières les représentent sous la forme de polygones saisis au niveau de leurs frontières de type *bona fide*, les secondes sous la forme de points saisis au niveau de leurs centres respectifs.

- Les instances de la classe *Zone_Végétation* ayant comme valeur d'attribut *Nature* « *Forêt fermée de conifères »* et représentant des entités topographiques d'une superficie supérieure à 500 hectares représentent les mêmes entités topographiques que les instances de la classe *Massif Boisé*. Les premières les représentent sous la forme de polygones saisis au niveau de leurs frontières de type *bona fide*, les secondes sous la forme de points saisis au niveau de leurs centres respectifs.

Ces relations de correspondances sont fournies par notre application sous la forme d'un fichier XML dont un extrait, correspondant aux résultats décrits dans ce paragraphe, est présenté ci-dessous.

```
[…]
<AppariementFeature>
      <Cible>
            <FeatureCible> ZoneVegetation </FeatureCible>
            <RestrictionSurAttribut>
                  <Attribut> Nature </Attribut>
                  <Comparateur> egal </Comparateur>
                  <Valeur> Forêt fermée de feuillus </Valeur>
            </RestrictionSurAttribut>
            <RestrictionSurPropriete>
                  <Propriete> AreaReferenceRegion </Propriete>
                  <Comparateur> superieur </Comparateur>
                  <Valeur> 5000000.0 </Valeur>
            </RestrictionSurPropriete>
      </Cible>
      <Source>
            <FeatureSource> MassifBoise </FeatureSource>
      </Source>
      <Relation> isEq </Relation>
      <DescriptionGeometries>
            <GeometrieCible> Polygon </GeometrieCible>
            <FormeCaracteristiqueCible>
                  BonaFideBoundary
            </FormeCaracteristiqueCible>
            <GeometrieSource> Point </GeometrieSource>
            <FormeCaracteristiqueSource>
                  Centre
            </FormeCaracteristiqueSource>
```

```
      </DescriptionGeometries>
</AppariementFeature>
<AppariementFeature>
      <Cible>
            <FeatureCible> ZoneVegetation </FeatureCible>
            <RestrictionSurAttribut>
                  <Attribut> Nature </Attribut>
                  <Comparateur> egal </Comparateur>
                  <Valeur> Forêt fermée mixte </Valeur>
            </RestrictionSurAttribut>
            <RestrictionSurPropriete>
                  <Propriete> AreaReferenceRegion </Propriete>
                  <Comparateur> superieur </Comparateur>
                  <Valeur> 5000000.0 </Valeur>
            </RestrictionSurPropriete>
      </Cible>
      <Source>
            <FeatureSource> MassifBoise </FeatureSource>
      </Source>
      <Relation> isEq </Relation>
      <DescriptionGeometries>
            <GeometrieCible> Polygon </GeometrieCible>
            <FormeCaracteristiqueCible>
                  BonaFideBoundary
            </FormeCaracteristiqueCible>
            <GeometrieSource> Point </GeometrieSource>
            <FormeCaracteristiqueSource>
                  Centre
            </FormeCaracteristiqueSource>
      </DescriptionGeometries>
</AppariementFeature>
<AppariementFeature>
      <Cible>
            <FeatureCible> ZoneVegetation </FeatureCible>
            <RestrictionSurAttribut>
                  <Attribut> Nature </Attribut>
                  <Comparateur> egal </Comparateur>
                  <Valeur> Forêt fermée de conifères </Valeur>
            </RestrictionSurAttribut>
            <RestrictionSurPropriete>
                  <Propriete> AreaReferenceRegion </Propriete>
                  <Comparateur> superieur </Comparateur>
                  <Valeur> 5000000.0 </Valeur>
            </RestrictionSurPropriete>
      </Cible>
      <Source>
            <FeatureSource> MassifBoise </FeatureSource>
      </Source>
      <Relation> isEq </Relation>
      <DescriptionGeometries>
            <GeometrieCible> Polygon </GeometrieCible>
```

```
            <FormeCaracteristiqueCible>
                  BonaFideBoundary
            </FormeCaracteristiqueCible>
            <GeometrieSource> Point </GeometrieSource>
            <FormeCaracteristiqueSource>
                  Centre
            </FormeCaracteristiqueSource>
      </DescriptionGeometries>
</AppariementFeature>
[…]
```

Si l'on traite la valeur d'attribut *Zone arborée* en tant que classe de la BDTOPO© Pays 1.2, on obtient alors:

- Les instances de la classe *Zone arborée* représentant des entités topographiques d'une superficie supérieure à 500 hectares représentent les mêmes entités topographiques que les instances de la classe *Massif Boisé*. Les premières les représentent sous la forme de polygones saisis au niveau de leurs frontières de type *bona fide*, les secondes sous la forme de points saisis au niveau de leurs centres respectifs.

```
[…]
<AppariementFeature>
      <Cible>
            <FeatureCible> ZoneArboree </FeatureCible>
            <RestrictionSurPropriete>
                  <Propriete> AreaReferenceRegion </Propriete>
                  <Comparateur> superieur </Comparateur>
                  <Valeur> 5000000.0 </Valeur>
            </RestrictionSurPropriete>
      </Cible>
      <Source>
            <FeatureSource> MassifBoise </FeatureSource>
      </Source>
      <Relation> isEq </Relation>
      <DescriptionGeometries>
            <GeometrieCible> Polygon </GeometrieCible>
            <FormeCaracteristiqueCible>
                  BonaFideBoundary
            </FormeCaracteristiqueCible>
            <GeometrieSource> Point </GeometrieSource>
            <FormeCaracteristiqueSource>
                  Centre
            </FormeCaracteristiqueSource>
      </DescriptionGeometries>
</AppariementFeature>
[…]
```

4.3.3.2 Découverte du contenu de bases de données topographiques hétérogènes

En raison de la variété des bases de 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; 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. L'objectif de l'application décrite dans cette partie est de fournir à un utilisateur une application lui permettant de découvrir de façon simple les données les plus appropriées à 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. Cette application vise donc à aider les utilisateurs à retrouver simplement les catégories d'entités géographiques qu'ils recherchent, en leur proposant d'utiliser des termes courants issus de l'ontologie du domaine présentée dans la partie 4.2, au lieu des termes techniques utilisés dans les schémas conceptuels de bases de données. De plus, elle doit leur permettre de retrouver automatiquement les données qui les intéressent dans les diverses bases disponibles et leur fournir des informations supplémentaires sur les données : quels types d'entités géographiques 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 (i.e. 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 ?) ? Enfin, elle doit leur permettre de visualiser les données correspondant à leurs besoins 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 (cf. partie 3.1.1). 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*, mais également que les zones arborées représentées dans cette classe ont nécessairement une superficie supérieure à 5 hectares. L'approche d'annotation sémantique fondée sur les spécifications mise en œuvre dans le cadre de cette application a constitué une première étape vers la définition de celle présentée ci-dessus, et est présentée plus en détail dans (Mechouche et al., 2012).

Architecture globale

L'architecture globale de notre système est présentée en figure 55. 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 côté 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 d'applications - dites « ontologies locales » dans la figure 55 - visant à formaliser 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 d'applications, 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 instances de bases de données géographique, etc.) est renvoyé à l'utilisateur via l'interface de réponse.

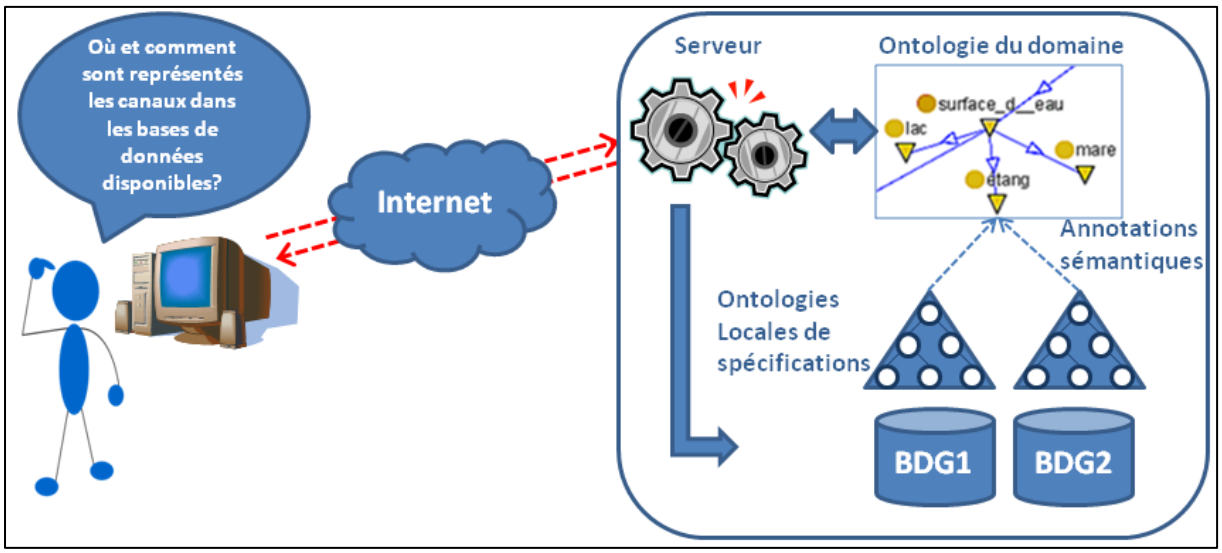


Figure 55: Architecture globale du système pour la découverte de bases de données topographiques hétérogènes

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.

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.

Implémentation du système

Le système proposé a été implémenté sous forme d'une application Web. Les différents langages et systèmes utilisés pour son implémentation sont détaillés en figure 56.

Deux bases de données géographiques sont référencées par le système (la BDTOPO© 1.2 et la BDCARTO© 3.0), avec leurs ontologies d'applications 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 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 Geoserver[46] 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 PostgreSQL couplé à son extension spatiale PostGIS. Le système utilise des services WMS pour l'affichage de données et l'API du GéoPortail[47] afin de disposer de fonds de cartes pour le module d'affichage cartographique.

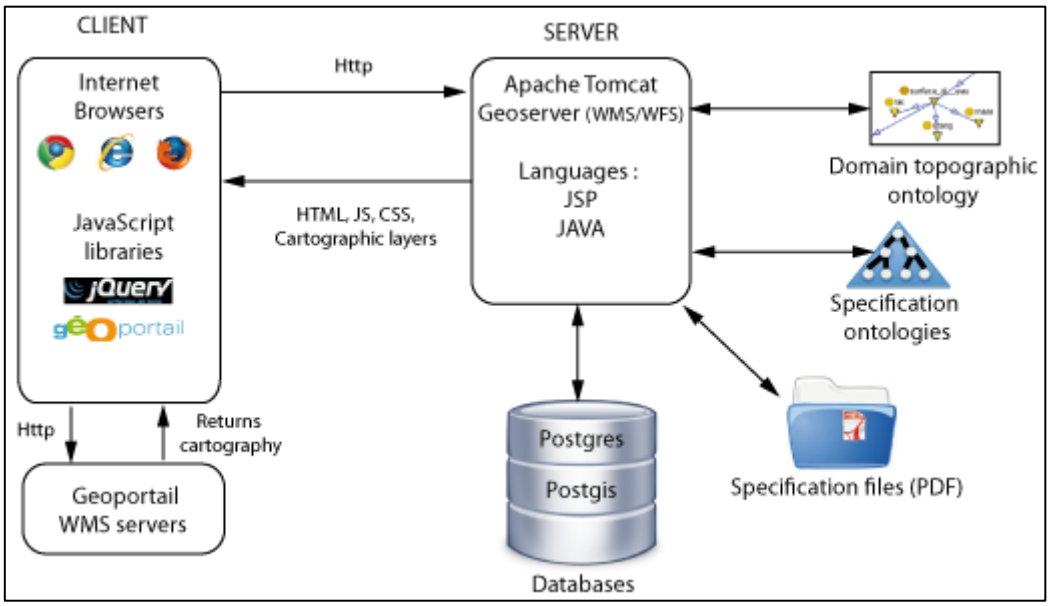


Figure 56: Implémentation de l'application de découverte de bases de données topographiques hétérogènes

L'interface Web, composée de trois parties, est présentée en figure 57. 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, des canaux. 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.

[46] http://geoserver.org/display/GEOS/Welcome
[47] http://api.ign.fr/accueil

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 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 Levenshtein[48] normalisée (Yujian et Bo, 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 spécialisant le concept désigné par le terme initialement demandé.

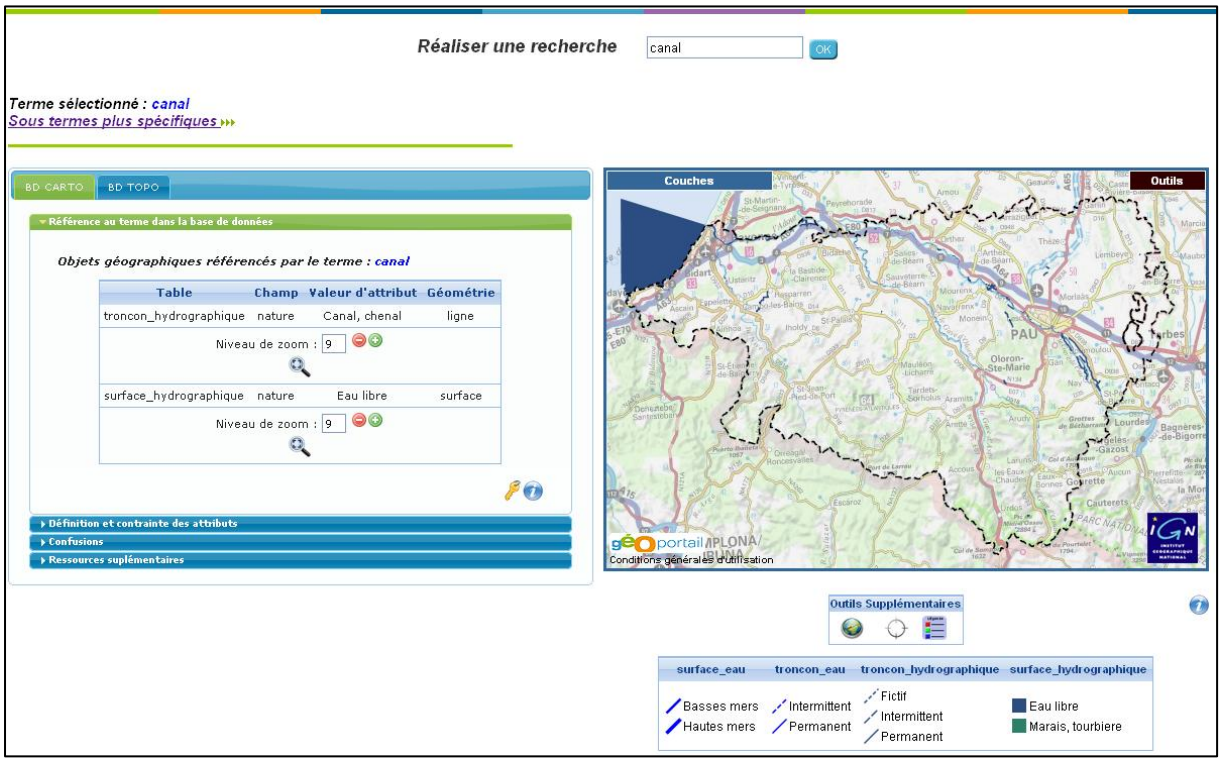


Figure 57: Interface Web de l'application de découverte de bases de données topographiques hétérogènes

Le module d'extraction d'informations extrait des ontologies d'applications 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 d'applications 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 58).

[48] http://fr.wikipedia.org/wiki/Distance_de_Levenshtein

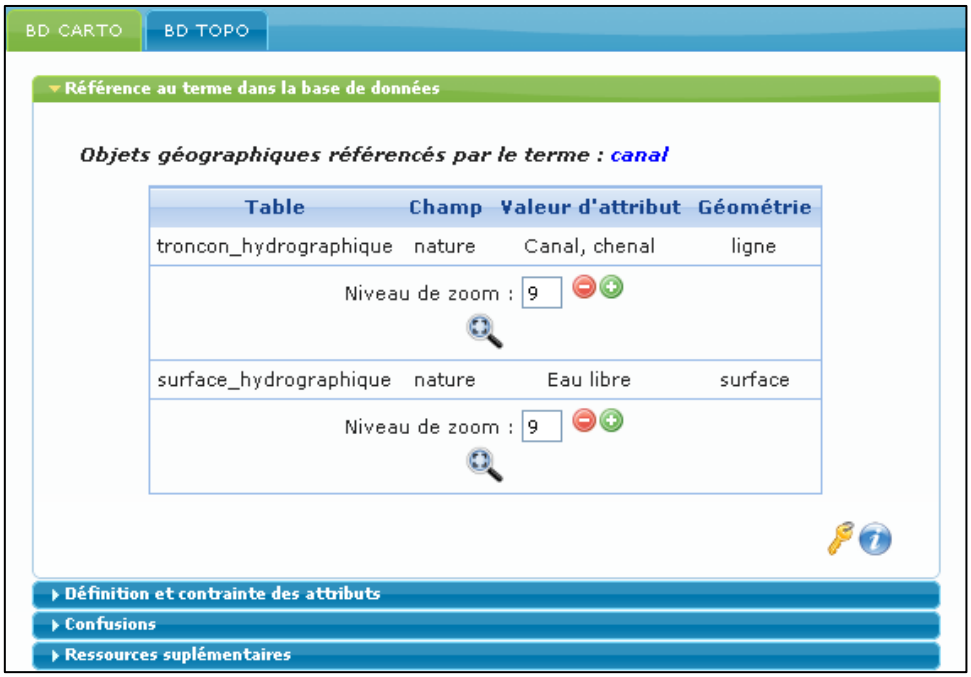


Figure 58: Informations sur les données renvoyées par le système à l'utilisateur

Lorsque plusieurs bases disposent de données sur la catégorie d'entités topographiques recherchée par l'utilisateur, un onglet par base est proposé. Un onglet se présente sous la forme d'un accordéon composé de quatre sections. La première section (voir figure 59) indique :

- La(les) classe(s) de la base de données où sont représentées les entités géographiques désignées par le terme choisi par l'utilisateur.
- 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*.
- Le nom de l'attribut pouvant prendre ces valeurs.
- 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 : 9	
troncon_eau	franchisst	Tunnel	ligne
troncon_eau	artif	1	ligne
troncon_eau	franchisst	Sans objet	ligne
		Niveau de zoom : 9	

Figure 59: Identification des données recherchées par l'utilisateur au sein d'une base de données

La deuxième section (voir figure 60) 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 tracéa 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

Figure 60: Définitions et règles de saisie des données de la base

La troisième section (voir figure 61) 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 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	*canal bief cours d'eau*
surface_eau	nature	Surface d'eau	*mare étang rivière lac canal surface d'eau fleuve*
troncon_eau	franchisst	Sans objet	*torrent rivière canal fossé bief cours d'eau ruisseau fleuve*
troncon_eau	franchisst	Tunnel	*canal*
surface_eau	regime	Permanent	*mare étang rivière lac canal surface d'eau fleuve*

Figure 61: Liste des regroupements d'entités topographiques effectués 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 s 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.

Cette application permet donc 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 62 et 63).

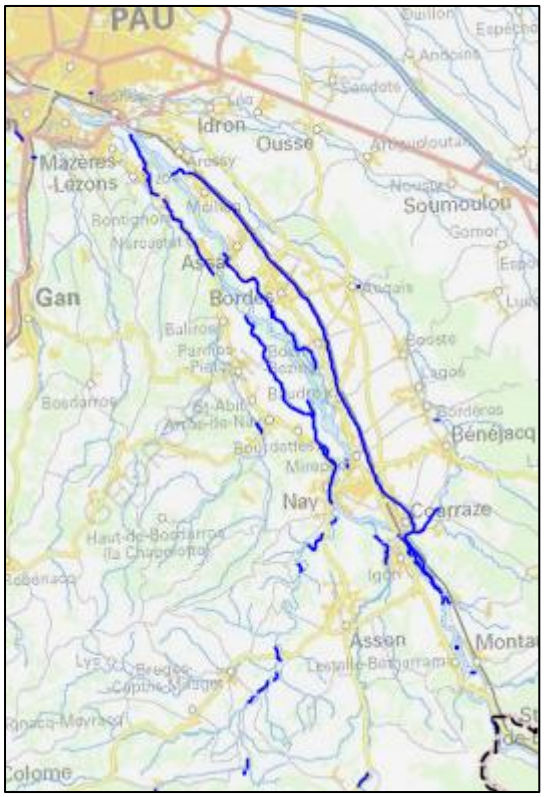


Figure 62: Représentation des canaux dans la BDTOPO© 1.2

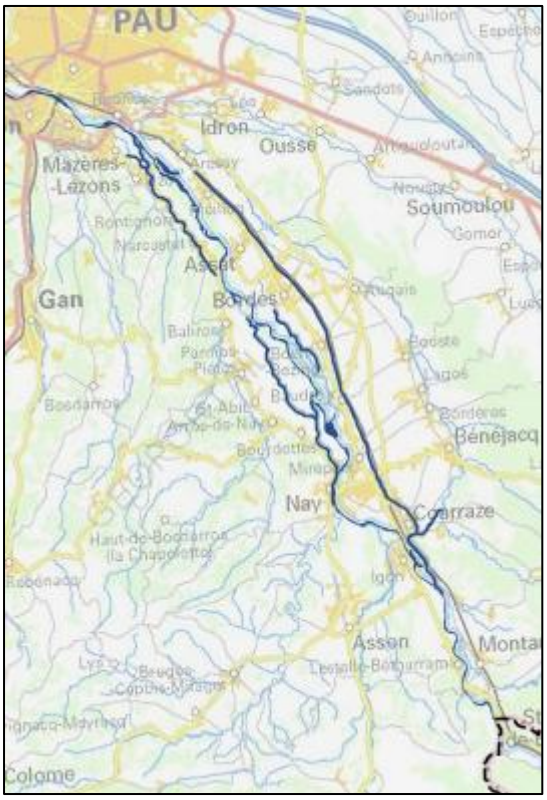


Figure 63: Représentation des canaux dans la BDCARTO© 3.0

4.3.3.3 Perspectives pour l'appariement des données

L'objectif premier de l'approche que nous proposons pour l'appariement fin de schémas de bases de données topographiques est d'obtenir des relations de correspondance entre éléments de schémas rendant compte des divers types d'hétérogénéités pouvant intervenir entre ces bases. Ces relations de correspondance sont destinées à guider l'appariement des données elles-mêmes en fournissant les connaissances nécessaires au choix et au paramétrage d'algorithmes d'appariement de données pertinents.

Considérons le cas où une relation de correspondance est détectée entre une classe d'une base de données et une classe d'une autre base dotée d'un attribut *Nature* permettant de distinguer la catégorie exacte à laquelle appartiennent les entités saisies dans cette seconde classe, ou bien entre deux classes de ce second type. Cette relation de correspondance comportera alors une (ou plusieurs) restriction(s) sur la(les) valeur(s) de cet (ces) attribut(s) précisant quelle(s) valeurs(s) les instances de l'une ou l'autre des classes appariées devront prendre pour être sélectionnées comme candidates à l'appariement. C'est le cas, par exemple, pour l'appariement des classes *Zone_végétation* de la BDTOPO© 2.1 et *Massif_Boisé* de la BDCARTO© 3.1 : la classe *Massif_Boisé* a une relation de correspondance avec le sous-ensemble de la classe *Zone_végétation* composé des instances dont l'attribut *Nature* a pour valeurs *Forêt fermée de conifères*, *Forêt fermée de feuillus*, et *Forêt fermée mixte*. Disposer de relations de correspondance précisant ce type de restrictions sur valeurs d'attributs nous permettra donc de sélectionner au préalable les instances de bases de données candidates à l'appariement. Cette opération de sélection préalable permet de limiter le nombre d'instances cibles et sources candidates à l'appariement, et donc de diminuer le nombre de tests à effectuer pour les algorithmes d'appariement utilisés. Un autre avantage est de diminuer les risques d'erreurs d'appariement en éliminant d'emblée les instances ne se conformant pas aux critères établis par la (les) restriction(s) sur valeur(s) d'attribut(s) établie(s) lors de l'appariement des schémas, et qui auraient pu causer la détection de relations de correspondance entre instances erronées (cf. partie 4.2.3.2).

Ce cas de relation de correspondance dotée d'une restriction sur valeurs d'attribut peut être assimilé, lorsque les deux classes mises en relation présentent ce type de restrictions, aux cas des relations de correspondance entre attributs et éventuellement entre leurs valeurs énumérées. Nous le traitons comme une restriction sur valeurs d'attribut afin de pouvoir conserver les cas où la restriction ne concerne que l'une des deux classes. En effet, il s'agit d'une information importante si l'on souhaite pouvoir sélectionner au préalable les instances de cette classe candidates à l'appariement. De façon générale, les relations de correspondance entre attributs et éventuellement entre leurs valeurs énumérées pourront être mises à profit comme critères d'appariement supplémentaires dans le cadre d'un processus d'appariement de données multicritère (cf. partie 4.2.3.2).

La comparaison automatique des critères de sélection des entités topographiques du monde réel devant figurer au sein de chaque classe de base de données peut conduire à la détection de relations de correspondance entre classes comprenant une (ou plusieurs) restriction(s) sur la(les) valeur(s) d'une (ou plusieurs) des propriété(s) des entités topographiques représentées par les instances de ces classes. Ainsi, si les données à apparier présentent des informations attributaires suffisantes ou

des géométries suffisamment détaillées, de telles restrictions pourront être traduites en restrictions sur valeurs d'attributs ou en restrictions sur la géométrie des instances. Considérons de nouveau le cas de l'appariement des classes *Zone_végétation* et *Massif_Boisé.* La classe *Massif_Boisé* a une relation de correspondance avec les instances de la classe *Zone_Végétation* ayant pour valeur d'attribut *Nature, Forêt fermée de conifères*. Cette relation présente également une restriction sur valeurs de propriété précisant que seules les instances de *Zone_Végétation* représentant des forêts fermées de conifères dotées d'une superficie supérieure à 5000000 m^2 pourront avoir une instance homologue dans la classe *Massif_Boisé.* Dans la mesure où les instances de *Zone_Végétation* représentent les entités topographiques sous la forme de polygones, cette restriction sur propriété pourra se traduire par une restriction sur l'aire des polygones associés aux instances de *Zone_Végétation*. Ainsi, une sélection préalable des instances de *Zone_Végétation* pourra être effectuée afin de n'inclure, dans le processus d'appariement de données, que celles dont la géométrie présente une aire de plus 5000000 m^2.

Les relations de correspondance entre géométries indiquent, pour chaque classe, le type de géométrie associé à la classe, ainsi que les éléments caractéristiques de la forme des entités topographiques représentées, au niveau desquels la géométrie des instances de la classe est saisie. Les algorithmes d'appariement de données géographiques reposant pour une large part sur la géométrie des données, disposer de telles indications est donc particulièrement utile. D'une part, la comparaison des types de géométrie associés aux classes mises en relation permet de choisir un algorithme d'appariement géométrique adapté. En effet, les traitements d'appariement diffèreront selon que l'on souhaite mettre en correspondance des instances de classes dotées de géométries ponctuelles, surfaciques ou linéaires, voire de géométries différentes d'une classe à l'autre. D'autre part, les informations sur les éléments caractéristiques de la forme des entités géographiques saisis peuvent également être mises à profit pour le paramétrage ou le choix des algorithmes d'appariement. En effet, ceux-ci incluent généralement des seuils de distances maximales entre géométries comparées. Ces seuils sont le plus souvent définis en fonction des précisions géométriques respectives des bases à intégrer. Une telle approche suppose que le seul décalage possible entre les données des deux bases résulte de leur différence de précision géométrique. Or, celui-ci peut également provenir des règles de saisie de la géométrie. Ainsi, pour deux classes de bases de données représentant des bâtiments sous la forme de points, les géométries des instances de la première classe pourront avoir été saisies devant l'entrée de chaque bâtiment, tandis que celles des instances de la deuxième base représenteront le centre de ces bâtiments, causant ainsi un décalage supplémentaire entre les localisations des bâtiments fournies par les deux bases. En outre, dans le cas où les géométries saisies sont différentes, comme c'est le cas pour les instances de la classe *Massif_Boisé,* représentées par des points situés au centre des zones arborées représentées, et les instances de *Zone_Végétation,* représentées par des polygones saisis au niveau du contour de ces zones, l'indication des éléments caractéristiques de la forme des entités topographiques saisis permet de choisir un algorithme d'appariement géométrique adapté. Ainsi, pour l'exemple auquel nous nous intéressons, il conviendra de choisir un algorithme d'appariement testant l'intersection des géométries des instances à apparier, et vérifiant éventuellement la localisation des points représentant des instances de *Massif_Boisé* par rapport au centre géométrique des polygones représentant des instances de *Zone_Végétation*. L'automatisation de ce type de choix pourrait d'ailleurs être envisagée en s'appuyant sur le cadre de référence sémantique proposé dans cette partie. Celui-ci pourrait, en effet, être affiné et enrichi de connaissances sur les relations

topologiques existant entre les classes spécialisant SpatialFeature. A titre d'exemple, le concept de Centre pourrait être défini de la façon suivante :

```
Class: Centre
      SubClassOf : within some BonaFideBoundary
```

Disposer de telles connaissances sur les relations spatiales existant entre éléments caractéristiques de la forme des entités topographiques permettrait alors de déduire des relations géométriques entre les instances de classes de bases de données topographiques. Cette mise en œuvre de connaissances externes dans le choix des algorithmes d'appariement de données pourrait également porter sur la détermination des seuils d'appariement géométrique. En effet, certaines classes de bases de données topographiques visent à représenter des catégories d'entités topographiques dont la forme est vague par essence. Les géométries saisies pour représenter de telles entités ne pourront pas avoir le même niveau de précision que celles représentant des entités topographiques naturellement bien délimitées. L'indication de l'élément caractéristique de la forme des entités topographiques saisi constitue donc également une bonne indication du niveau de précision raisonnablement atteignable pour les données saisies. Reprenons l'exemple des classes *Oronyme* de la BDTOPO©et *Point remarquable du relief* de la BDCARTO©. Ces deux classes représentent diverses catégories d'entités topographiques sous la forme de points. Il peut s'agir aussi bien de sommets que de vallées. Ces dernières sont localisées par un point situé en leur centre. Or, dans la mesure où il est extrêmement difficile, voire impossible, de s'accorder sur la localisation exacte des limites d'une vallée, il est très probable que les points saisis pour représenter le centre d'une vallée présentent des coordonnées relativement éloignées d'une base de données à l'autre. Il conviendra donc de tenir compte de ce paramètre dans la détermination des seuils d'appariement géométrique.

Enfin lorsque les bases de données à apparier possèdent des niveaux de détails différents, il arrive que les relations de correspondance à détecter soient de type n :m. C'est le cas notamment lorsque des réseaux sont découpés selon des critères différents, ou lorsque des entités sont agrégées pour ne créer qu'une instance au sein d'une base de données. Le modèle que nous proposons pour la représentation des règles de saisies des données topographiques inclut ces cas et permet leur détection afin de permettre le choix d'un algorithme adapté.

Conclusion

Dans cette partie, nous avons proposé une mise en œuvre de notre modèle pour l'intégration virtuelle de bases de données topographiques permettant la détection automatique de relations de correspondance fines entre éléments de schémas des bases de données à intégrer. Pour ce faire, nous nous sommes inspirée des techniques d'alignement d'ontologies fondées sur la sémantique présentées au chapitre 3.1.2, dont Klien (2008) et Schade (2010) proposent deux exemples de mise en œuvre dans le domaine de l'intégration d'information géographique, ainsi que des travaux de Gesbert (2005) sur la formalisation des spécifications de bases de données géographiques. Nous avons donc proposé de formaliser les connaissances sur les règles de sélection et de représentation géométrique des entités géographiques issues des spécifications à l'aide d'un langage de représentation de connaissances standard, OWL 2, afin de pouvoir mettre à profit les systèmes de raisonnement associés pour détecter automatiquement des relations de correspondance entre

éléments de schémas. Ces connaissances sont formalisées au sein d'ontologies d'applications dans lesquelles les éléments de schémas et les éléments de spécifications à apparier sont réifiés afin de permettre la représentation des règles de saisie fournies par les spécifications sous la forme d'axiomes portés par les classes destinées à représenter les divers éléments de schémas et de spécifications. L'appariement des éléments de spécifications et des éléments de schémas est effectué par une application mettant en œuvre un système de raisonnement pour ontologies, Hermit. Cette approche visant à représenter et exploiter des spécifications de bases de données topographiques au sein d'ontologies d'applications a également été mise en œuvre dans le cadre d'une application de découverte de base de données topographiques, permettant à un utilisateur de comprendre aisément les différences de contenus entre bases de données hétérogènes. La mise en œuvre de cette approche a nécessité l'instanciation des principales composantes du modèle.

La représentation des règles de saisie des données topographiques sous la forme d'axiomes nécessite tout d'abord de disposer d'un certain nombre de concepts propres à ce domaine auxquels les axiomes devront faire référence. Nous proposons donc un cadre de référence sémantique qui comporte l'ensemble des concepts topographiques issus de l'ontologie du domaine créée pour l'approche proposée dans la partie 4.2, ainsi qu'un ensemble de concepts, propriétés et relations utiles à l'expression des règles de saisie des données topographiques (voir figure 64). Conformément aux recommandations de Kuhn (2003) et aux travaux de Klien (2008) et Schade (2010), ces concepts, propriétés et relations sont ancrés à une ontologie de haut niveau, DUL, via des relations de généralisation-spécialisation. L'instanciation de ce cadre de référence sémantique est réalisée manuellement, en réutilisant, autant que faire se peut, certaines ontologies existantes couvrant une partie des thématiques utiles. Celles-ci sont directement importées par le cadre de référence sémantique et intégrées manuellement à l'ontologie de haut niveau via des relations de généralisation-spécialisation définies entre leurs concepts les plus génériques et les concepts les plus spécifiques de DUL. Les concepts et propriétés manquants sont définis et ajoutés manuellement, en s'appuyant sur les travaux de Klien(2008), Schade(2010) et Gesbert(2005). Le cadre de référence sémantique proposé suffit à la représentation de l'essentiel des règles de saisies rencontrées dans les spécifications. Cependant, il pourrait encore être enrichi d'autres connaissances, en particulier concernant les relations spatiales entre éléments caractéristiques de la forme des entités topographiques – les classes filles de SPATIALFEATURE – et leur traduction en opérations d'analyse spatiale entre géométries des instances de base de données, pour l'appariement des données.

Les ontologies d'applications sont créées manuellement. Elles importent le cadre de référence sémantique au sein duquel elles s'intègrent (voir figure 64); les classes visant à représenter les éléments de schémas et les éléments de spécifications y sont définies comme des spécialisations du concept INFORMATIONOBJECT. Elles sont définies par des axiomes faisant référence aux concepts, propriétés et relations du cadre de référence sémantique et représentant les règles de sélection et de représentation géométrique des entités topographiques au sein de la base. Ceux-ci sont écrits conformément à l'approche proposée en partie 4.3.2. Cette étape d'instanciation des ontologies d'applications est relativement longue et complexe. Elle requiert une bonne connaissance du cadre de référence sémantique, du langage OWL 2 et une analyse approfondie des spécifications afin d'en extraire les règles de saisie les plus représentatives et de les synthétiser. Les exemples présentés dans cette partie ont été entièrement réalisés manuellement. Cependant, comme dans l'approche présenté en partie 4.2, il est tout à fait envisageable d'automatiser en partie cette étape. En effet, les classes représentant les éléments de schémas et de spécifications peuvent être générées

automatiquement à partir des schémas ISO 19109. Il resterait alors à les compléter avec les axiomes décrivant les règles de saisie des données.

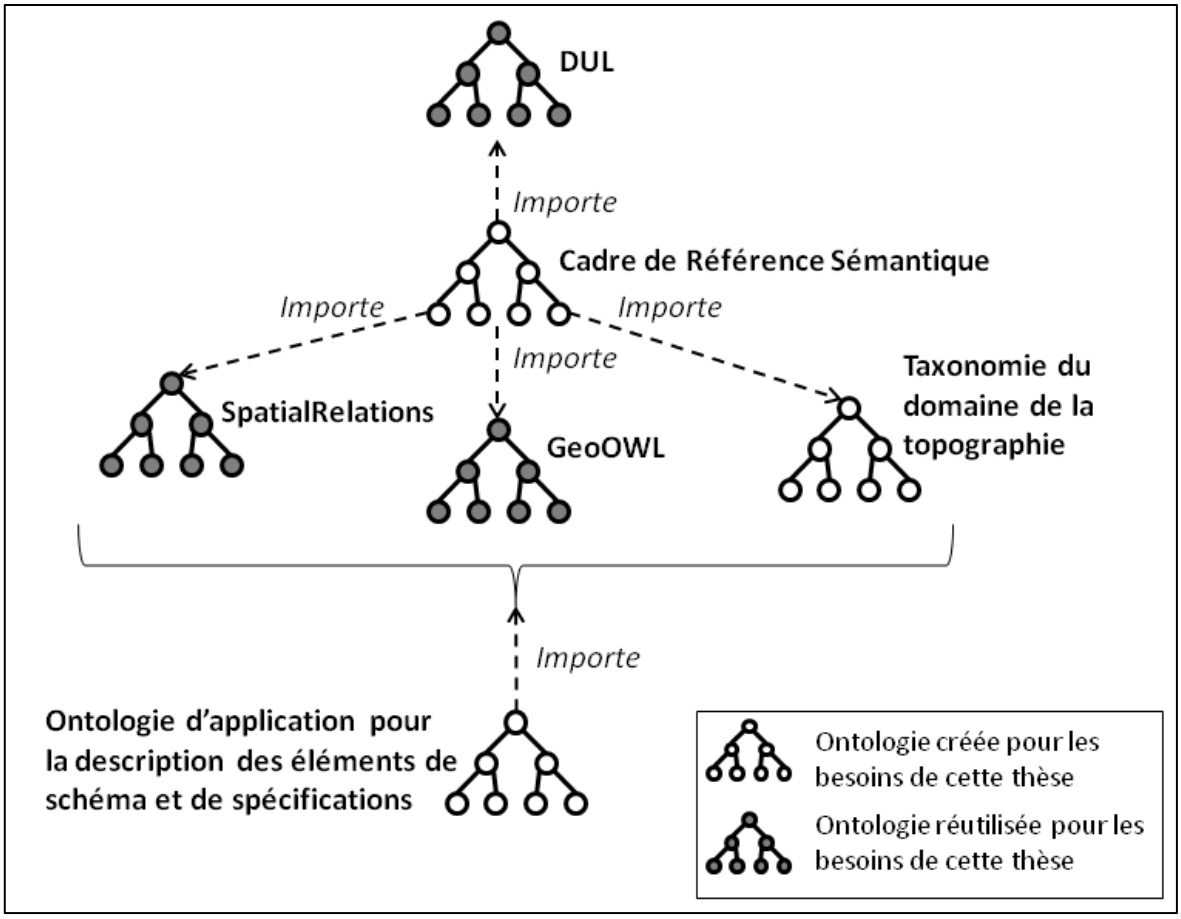


Figure 64: Articulation des différentes ontologies mises en œuvre dans l'approche fondée sur les spécifications (partie 4.3)

Une application Java a été développée afin de traiter les ontologies d'applications créées et d'en déduire des relations d'appariement entre les éléments de schémas qu'elles décrivent. La détection des correspondances est réalisée par un système de raisonnement pour ontologies. L'extraction de relations de correspondances fines est effectuée par comparaison des axiomes des classes mises en relation par le système de raisonnement. Un schéma conceptuel a été défini afin de permettre l'instanciation des relations de correspondance détectées. Celui-ci est défini dans un objectif d'appariement de données. Il inclut, pour les relations de correspondance entre classes, la représentation de restrictions sur des valeurs d'attributs ou sur des valeurs de propriétés, afin de permettre la sélection préalable des instances candidates à l'appariement. Les relations entre géométries conservent une référence aux éléments caractéristiques de la forme des entités topographiques saisis, afin d'en conserver la trace pour guider l'étape d'appariement des données.

Les résultats d'appariement de schémas finalement obtenus tendent à prouver que des techniques d'appariement fondées sur la sémantique peuvent être mises en œuvre avec succès pour l'appariement de schémas de bases de données topographiques. Cependant, la détection de relations de correspondance fines a nécessité le développement d'une application dédiée à la comparaison des axiomes décrivant les éléments de schémas et de spécifications formalisés. De plus,

la mise en œuvre de cette approche pour les exemples que nous avons présentés a nécessité la création d'une ontologie d'application par classe de base de données afin de permettre le traitement de ces classes deux par deux. En effet, le système de raisonnement devant traiter des axiomes relativement complexes, la durée des traitements s'est parfois avérée très longue. Ainsi, la mise en œuvre de cette approche d'appariement de schémas à l'échelle de bases de données topographiques entières ne semble envisageable que dans le cas d'applications statiques et nécessiterait probablement une optimisation du modèle proposé.

Enfin, la mise en œuvre de l'architecture globale proposée dans cette approche pour la découverte de bases de données topographiques hétérogènes met à disposition de l'utilisateur, par rapport aux applications ce type présentées en partie 3.1.1, des connaissances sur le contenu des bases de données disponibles auxquelles il n'avait que difficilement accès jusque là. Celles-ci lui permettent de comparer les règles de saisie des classes des différentes bases de données référencées par le système, et d'en connaître rapidement et simplement les différences et les points communs afin de pouvoir déterminer quelle base de données présente les données les plus adaptées à son application. A ce titre, l'affichage conjoint des règles de saisie des données et d'extraits de ces mêmes données semble constituer une approche efficace pour faciliter le choix de l'utilisateur.

5 Conclusion et perspectives

Ce travail s'inscrit dans la continuité du travail de thèse de Nils Gesbert sur la « Formalisation des spécifications de bases de données géographiques en vue de leur intégration » (2005). La principale proposition de Gesbert (2005) réside dans la formalisation des spécifications de bases de données topographiques en tant que liens entre schémas conceptuels de ces bases de données et une ontologie du domaine. L'architecture globale du modèle proposé par Gesbert (2005) s'apparente donc aux diverses architectures d'intégration virtuelle présentées au chapitre 3.1 de ce mémoire, mais s'en démarque par l'utilisation d'annotations sémantiques complexes, appelées *procédures de représentation*, pour relier schémas conceptuels et ontologie du domaine. Cette approche, qui intègre la complexité de la relation entre le terrain réel et sa représentation au sein d'une base de données topographique, vise à en rendre les détails accessibles à une application d'intégration virtuelle de bases de données topographiques. Le bénéfice attendu est de permettre la détection automatique des divers types d'hétérogénéités pouvant intervenir entre bases de données topographiques, lors de l'étape d'appariement de leurs schémas, à des fins d'appariement de données.

Nous nous sommes attachée, dans cette thèse, à la réalisation de cet objectif d'appariement de schémas conceptuels de bases de données topographiques, en vue de guider l'appariement des données.

Contributions

Une première étape de nos travaux a consisté à appréhender l'appariement automatique de schémas de bases de données topographiques à l'aide d'annotations sémantiques simples. Nous avons donc établi des relations de correspondance entre schémas conceptuels de bases de données topographiques et concepts d'une ontologie du domaine, afin de désigner, pour chaque classe des schémas à apparier, les catégories d'entités topographiques représentées. Les relations de correspondance entre éléments de schémas à apparier ont ensuite pu être dérivées par analyse des relations d'annotation sémantique. L'établissement des annotations sémantiques entre schémas et ontologie a été réalisé de façon automatique, à l'aide de techniques d'appariement terminologiques et structurelles. Dans cette première approche, l'ontologie du domaine est utilisée comme une ontologie de support destinée à pallier d'éventuels manques de recouvrement lexical et structurel entre les schémas à apparier. Si cette première approche se cantonne à des annotations de schémas simples, elle nous a conduite à traiter la question de l'instanciation des principales composantes du modèle proposé (voir figure 65).

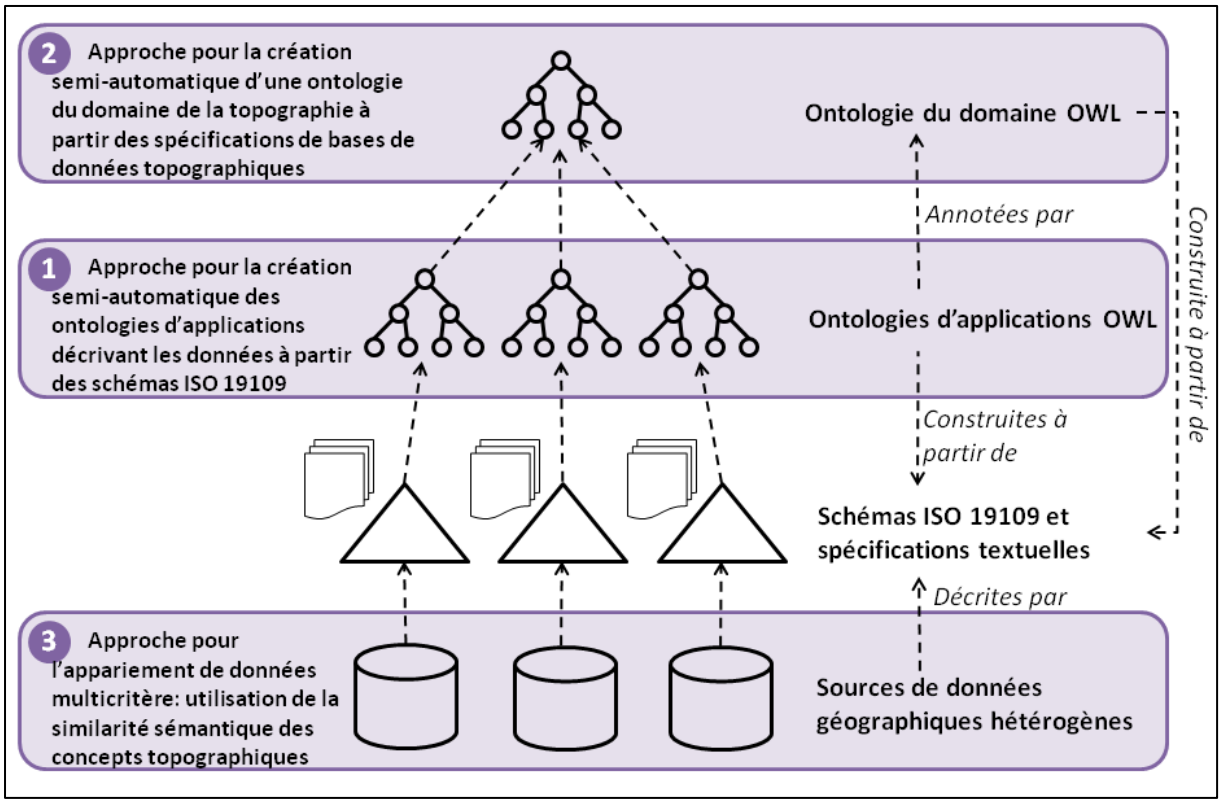


Figure 65: Principales contributions pour l'approche d'appariement fondée sur les valeurs d'attributs et une ontologie de support

Ainsi, nous avons proposé une approche pour la création semi-automatique d'ontologies d'applications au format OWL, à partir de schémas de bases de données topographiques représentés selon la norme ISO 19109 (figure 65, numéro 1). Ces ontologies visent à décrire la structure des bases de données à intégrer, en apportant des connaissances sur le contenu de ces bases qui n'apparaissent pas directement dans leurs schémas. En effet, l'approche de dérivation semi-automatique adoptée intègre des connaissances issues des spécifications sur l'existence de regroupements de catégories d'entités topographiques éventuellement très différentes au sein d'une même classe de base de données topographique et permet leur explicitation au sein de l'ontologie d'application générée.

Nous avons également proposé une approche semi-automatique pour la création de l'ontologie du domaine nécessaire à la mise en œuvre de notre modèle (figure 65, numéro 2). Celle-ci s'inspire des travaux de Gesbert (2005), qui préconise d'extraire manuellement le vocabulaire et les connaissances nécessaires à la création d'une telle ontologie des textes des spécifications de bases de données topographiques de référence. Cependant, dans la mesure où les textes à traiter sont relativement volumineux et les vocabulaires utilisés, variés et complémentaires, nous nous sommes attachée à proposer une approche semi-automatique fondée sur la mise en œuvre de techniques de traitement automatique du langage naturel, de traitement manuel, et de techniques de fusion d'ontologies.

Enfin, nous avons proposé une première approche pour l'exploitation des résultats d'appariement de schémas dans la mise en œuvre d'un appariement de données. Ainsi, nous avons introduit, dans un processus d'appariement de données multicritère, un critère vérifiant l'appartenance des instances candidates à l'appariement à une même catégorie d'entités topographiques. Les tests d'appariement

de données effectués nous ont conduite à proposer une seconde approche d'appariement de schémas, fondée sur la similarité sémantique des concepts de l'ontologie de support auxquels les concepts des ontologies d'applications sont ancrés, afin d'élargir les possibilités de mise en correspondance des éléments de schémas (figure 65, numéro 3). Ceci a permis d'assouplir l'application de ce critère d'appartenance à une même catégorie d'entités topographiques en lui substituant un critère d'appartenance à des catégories d'entités topographiques sémantiquement proches. Les résultats d'appariement de données s'en sont trouvés améliorés.

Cette première approche pour l'appariement de schémas de bases de données topographiques nous a donc permis de détecter automatiquement les classes de chacun des schémas en entrée, ainsi que les sous-ensembles d'instances de ces classes, représentant les mêmes catégories d'entités topographiques du monde réel. Cependant, les relations de correspondance détectées se cantonnent à faire état de l'appartenance des entités topographiques représentées par les instances des classes de schémas à apparier à une même catégorie d'entités topographiques ou, à des catégories d'entités topographiques sémantiquement proches. En effet, cette approche n'intègre pas de connaissances susceptibles de permettre au système de déceler d'éventuels conflits de critères de sélection ou de description géométrique des données entre les classes mises en relation.

Une seconde étape dans nos travaux à consisté à reprendre l'approche d'annotation complexe de schémas conceptuels de bases de données topographiques proposée par Gesbert (2005), en l'adaptant aux standards et normes actuellement recommandés en matière de gestion de l'information géographique et de représentation de connaissances. Au-delà des aspects de représentation des schémas et de l'ontologie selon des schémas ou des formalismes standards, nous nous sommes attachée à proposer une approche mettant à profit ces formalismes, à la fois pour la représentation des connaissances issues des spécifications et pour l'exploitation de ces connaissances par les systèmes de raisonnement associés, pour détecter automatiquement des relations de correspondance entre éléments de schémas (voir figure 66).

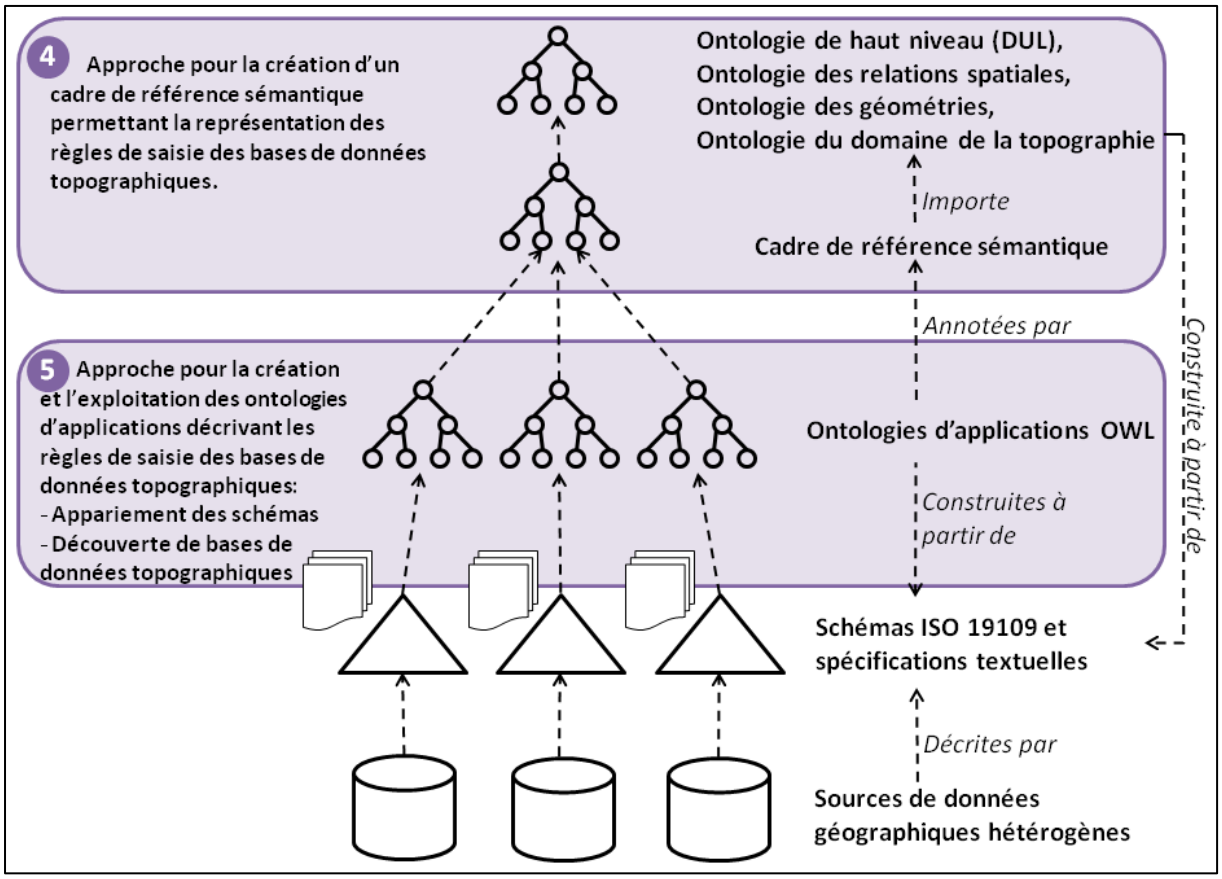


Figure 66: Principales contributions pour l'approche d'appariement fondée sur les spécifications des bases de données topographiques

Pour ce faire, nous avons adopté l'approche proposée par Kuhn (2003) pour l'explicitation de la sémantique des données géographiques. Nous avons donc défini et implémenté, dans le langage OWL 2, un cadre de référence sémantique (figure 66, numéro 1). Ce dernier regroupe l'ensemble des concepts topographiques issus de l'ontologie du domaine créée pour l'approche proposée dans la partie 4.2, ainsi que les concepts, propriétés et relations utiles à l'expression des règles de saisie des données topographiques, définis manuellement à partir des travaux de Gesbert (2005). Ces concepts, propriétés et relations sont ancrés à une ontologie de haut niveau via des relations de généralisation-spécialisation. Ce cadre de référence sémantique comprend un modèle pour la représentation des éléments de schémas et des éléments de spécifications des bases de données topographiques. Dans ce modèle, les éléments de schémas et de spécifications que nous souhaitons apparier sont réifiés, afin de pouvoir les doter de définitions axiomatisées pour permettre aux systèmes de raisonnement disponibles d'inférer des relations de correspondance entre les classes ainsi définies.

Nous avons proposé une approche pour l'instanciation de ce modèle, permettant ainsi la formalisation de l'essentiel des règles de saisie des bases de données topographiques (figure 66, numéro 2). Nous avons illustré cette approche par quelques exemples d'instanciation de règles de saisies pour diverses classes de bases de données hétérogènes, choisis pour leur caractère représentatif. Puis nous avons exploité les connaissances ainsi formalisées au sein d'une application permettant, d'une part la détection automatique des relations de correspondance entre éléments de schémas des bases de données topographiques, et d'autre part la comparaison automatique des axiomes les décrivant afin d'en extraire des relations d'appariement fines.

Enfin, nous avons proposé une application Web pour la découverte du contenu de bases de données topographiques hétérogènes s'appuyant sur le modèle que nous avons proposé pour l'appariement des schémas de ces bases. Cette application permet à un utilisateur non expert d'interroger le système dans les termes de l'ontologie du domaine afin de savoir si l'une ou l'autre des bases de données référencées par le système dispose de données sur le thème qui l'intéresse. Lorsque plusieurs bases de données représentent la catégorie d'entités topographiques recherchée par l'utilisateur, l'application lui permet de visualiser simplement les règles de saisie qui ont régi la création de chacune des classes représentant cette catégorie. Ainsi, l'utilisateur peut comparer ces différentes règles de saisie et en appréhender aisément les similitudes et les différences. Une interface cartographique lui permet également de visualiser un échantillon de données pour chacune de ces classes, afin de faciliter l'interprétation des règles de saisie affichées par l'interface générale de l'application.

Discussion et perspectives

De façon plus générale, notre travail pose la question de l'intégration de données géographiques. Cette notion recouvre en effet de nombreuses applications, dont la réalisation requiert la mise en œuvre de connaissances et de techniques éventuellement différentes. En effet, des applications de géocodage, de mutualisation de mises à jour, de contrôle qualité, de découverte et d'accès aux données, de transformation de schémas, de gestion de versions successives de bases de données, etc., ne nécessiteront pas obligatoirement la mise en œuvre des mêmes connaissances et des mêmes techniques. Cependant, toutes requièrent de disposer de connaissances décrivant la sémantique des schémas des bases de données topographiques, connaissances dont les spécifications de saisie de ces bases restent la principale source.

Nous nous sommes principalement intéressée ici à l'appariement de schémas de bases de données topographiques hétérogènes, vu comme une première étape d'un processus global d'intégration ayant comme finalité l'appariement des données elles-mêmes. Aussi, les choix effectués en matière de formalisation et d'acquisition des connaissances nécessaires pour la mise en œuvre de cette étape d'appariement de schémas de bases de données topographiques ont-ils été orientés en vue de cet objectif final : nous souhaitions pouvoir détecter des relations de correspondance entre éléments de schémas nous apportant des connaissances sur l'origine d'éventuels cas d'hétérogénéités pouvant survenir entre les données, à des fins d'appariement de données.

Nous avons successivement proposé deux approches pour l'appariement de schémas de bases de données topographiques. La première, fondée sur des techniques d'appariement terminologiques et structurelles, présente l'avantage de ne nécessiter aucune annotation manuelle des schémas à apparier, ce qui constitue un avantage certain. Les premiers résultats obtenus prouvent qu'une telle approche peut fournir des résultats exploitables et pourrait être étendue avec succès. En particulier, pour cette première approche, nous n'avons cherché à mettre en correspondance que des classes ou des valeurs d'attributs faisant référence à des concepts topographiques. Disposant, au terme du projet GéOnto (ANR-O7-MDCO-005), d'une ontologie du domaine de la topographie dotée de propriétés et de relations, une extension de notre approche à l'appariement des attributs, de leurs valeurs énumérées et des associations entre classes est envisageable. En outre, la mise en œuvre de techniques d'appariement prenant en compte la mise en correspondance d'attributs et de relations pour décider de la mise en correspondance de classes pourrait venir conforter les résultats initiaux.

Enfin, la géométrie des données topographiques, qui constitue une composante fondamentale de ces données, n'a pas été traitée dans cette approche. Sur ce point particulier, il semblerait profitable de s'inspirer des travaux sur la représentation et la description de la géométrie de données topographiques réalisés dans la seconde approche que nous avons proposée.

La seconde approche proposée, repose sur des techniques d'appariement fondées sur la sémantique. Elle présente l'avantage, par rapport à l'approche précédente, de permettre la détection de divers types d'hétérogénéités entre bases de données topographiques, grâce aux connaissances sur les règles de saisie des données sur lesquelles elle repose. En revanche, elle nécessite un effort préalable d'annotation manuelle des schémas de bases de données à apparier non négligeable. Cette question du coût de l'instanciation de ce type de modèle, nécessitant la création d'annotations sémantiques complexes, s'était déjà posée lors de nos premiers travaux sur le modèle proposé par Gesbert (2005). Une solution visant à extraire automatiquement les connaissances nécessaires à l'instanciation de ces annotations sémantiques à partir des textes des spécifications, à l'aide d'outils de traitement automatique du langage naturel, a donc été testée lors d'un stage (Picard, 2006). L'approche adoptée mettait à profit la structure du texte des spécifications de la BDTOPO© Pays 1.2 pour identifier des portions de texte susceptibles de renfermer des règles de sélection ou de modélisation géométrique des entités topographiques. Les règles de saisie elles-mêmes étaient ensuite extraites à l'aide de patrons lexico-syntaxiques. Si les résultats obtenus se sont avérés encourageants, l'absence d'adoption d'un modèle standard pour la rédaction de spécifications de bases de données topographiques demeure un frein important à l'extension de cette solution à d'autres bases de données. Une autre solution envisageable consisterait à tenter de retrouver automatiquement les règles de saisie ayant régi la production d'une base de données par comparaison et analyse des données de cette base par rapport à une base de données de référence. En outre, la mise en œuvre pratique, à grande échelle, de cette approche fondée sur la sémantique, semble nécessiter un allègement du modèle proposé dans la mesure où les systèmes de raisonnement pour ontologies mis en œuvre ici peinent à traiter de trop nombreux éléments de schémas et les règles de saisie qui leurs sont associées. Cet allègement pourrait passer par la suppression du modèle intermédiaire, entre schémas et concepts du mode réel, servant à la représentation des éléments de spécifications, et au report des connaissances qu'il renferme au niveau des éléments de schémas. Ceci impliquerait alors de trouver une solution de substitution pour la représentation des règles de découpage et d'agrégation des entités topographiques qui sont à l'origine de la définition de ce niveau intermédiaire de description des données.

Si cette seconde approche s'inscrit plus directement dans la lignée des travaux de Gesbert (2005), et fournit des résultats d'appariement de schémas comportant des connaissances importantes pour l'appariement de données, nous ne souhaitons pour autant privilégier aucune des deux approches proposées, que nous considérons comme complémentaires dans notre réflexion sur l'intégration de bases de données topographiques. En effet, la mise en œuvre de l'approche d'appariement de schémas lexicale et structurelle demeure, pour peu que l'on dispose d'une représentation des schémas à apparier conforme à la norme ISO 19109, extrêmement rapide. A ce titre, elle peut d'ailleurs être vue comme un préalable à la mise en œuvre de l'approche fondée sur la sémantique en fournissant un premier appariement permettant d'évaluer la complémentarité de deux bases de données hétérogènes et donc de décider de la pertinence de leur mise en correspondance selon l'application visée. Cette question de la comparaison globale de schémas ou d'ontologies a d'ailleurs fait l'objet de travaux dans le cadre du projet GéOnto (Mechouche et al., 2010). En outre, nous

voyons dans la question de la représentation et du traitement de la géométrie pour la première approche, une possibilité de convergence des deux approches.

Les tests présentés en partie 4.2, visant à exploiter les résultats d'appariement obtenus via la première approche au sein d'une application d'appariement de données, ont mis en évidence la nécessité de la prise en compte d'une source d'hétérogénéité entre bases de données topographiques dont les spécifications ne font pas état : l'absence de catégorisation universelle des entités topographiques. En effet, d'une base de données à l'autre, il arrive qu'une même entité du monde réel soit répertoriée, au gré des différentes interprétations qui en sont faites par les opérateurs de saisie, dans des classes ou sous des valeurs d'attribut *Nature* différentes mais néanmoins proches d'un point de vue sémantique. C'est notamment le cas pour les instance des classes *Oronyme* de la BDTOPO© et *Point remarquable du relief* de la BDCARTO© qui peuvent prendre comme valeurs d'attribut *Nature*, *Pic* ou *Sommet* ; il arrive fréquemment qu'une instance saisie en tant que *Pic* dans l'une des bases soit considérée comme un *Sommet* dans l'autre base et inversement. Or, l'appariement des schémas des deux bases fournit les relations de correspondance suivantes : *BDTOPO© / Oronyme / Nature / Pic* équivalent à *BDCARTO© / Point remarquable du relief / Nature /Pic*, *BDTOPO© / Oronyme / Nature / Sommet* équivalent à *BDCARTO© / Point remarquable du relief / Nature /Sommet, crête, colline.* L'interprétation directe et stricte d'un tel résultat pour la sélection préalable des instances candidates à l'appariement, dans le cadre d'un processus d'appariement de données, conduira nécessairement à la perte d'un grand nombre d'instances candidates pourtant pertinentes, et donc à une dégradation possible des résultats d'appariement de données. Pour pallier ce problème, nous avons proposé une solution d'appariement des schémas plus souple, avec des correspondances de type n:m, fondées sur la proximité sémantique des éléments des schémas. De façon générale, il semble que la réalisation d'un appariement de schémas en vue de guider un appariement de données doive, dans le cas des bases de données topographiques, privilégier des approches conduisant à un « sur-appariement » et intégrant aux relations de correspondance détectées un score de probabilité d'existence d'instances communes entre les classes ou sous-ensembles de classes mis en relation.

En outre, nous avons présenté, dans la partie 4.3, quelques perspectives de mise en œuvre des résultats d'appariement de schémas obtenus. Celles-ci comprennent la sélection préalable des instances de bases de données candidates à l'appariement sur la base de critères géométriques. Or, les seuils concernant la taille minimale que doivent présenter les entités topographiques du monde réel pour être saisies au sein d'une classe de base de données, précisés dans les spécifications, doivent être interprétés avant tout comme une indication sur le niveau de détail de la base de données. En effet, les délais de production des données ainsi que les outils de mesure disponibles pour l'évaluation de la taille des entités topographiques à partir de sources de données choisies, ne permettent pas l'application stricte de ces critères de sélection. En outre, il arrive fréquemment que les spécifications autorisent la saisie de certaines entités topographiques, non conformes aux critères de sélection qu'elles édictent, en raison de leur caractère exceptionnel ou de leur pertinence vis-à-vis de la description du terrain réel. Ainsi, l'application stricte des critères de taille pour la sélection préalable des instances de bases de données candidates à l'appariement pourrait affecter les résultats d'appariement de données en réduisant de façon trop rigoureuse le nombre d'instances candidates. La mise en œuvre des résultats d'appariement de schémas obtenus via notre seconde approche pour l'appariement de données sera l'occasion d'évaluer les effets de ces critères de sélection sur les résultats finaux.

Le schéma conceptuel que nous avons défini pour la représentation des relations de correspondance entre éléments de schémas de bases de données topographiques prévoit de représenter systématiquement les géométries des classes appariées, y compris lorsque ces géométries ne sont pas appariées elles-mêmes. Ceci s'explique par le fait que la géométrie des instances de bases de données topographiques hétérogènes demeure le critère d'appariement privilégié pour ces instances : que les géométries des instances de deux classes appariées soient du même type ou pas, qu'elles soient saisies selon les mêmes règles de modélisation ou pas, elles portent toujours une information fondamentale sur les données qui est la localisation de l'entité topographique représentée. A ce titre, la géométrie sera toujours prise en compte pour l'appariement des données. Les indications supplémentaires que notre modèle permet d'obtenir concernant les règles de saisie de la géométrie des instances des classes mises en relation permettent de guider le choix des opérations effectuées par les algorithmes d'appariement géométrique et d'affiner les seuils de distance souvent utilisés par ces algorithmes comme critères d'appariement. A ce jour, notre modèle permet de représenter des connaissances sur les relations spatiales entre éléments caractéristiques de la forme des entités topographiques au niveau desquels sont saisies les géométries des données. En revanche, il ne prend pas en charge l'interprétation automatique de ces connaissances en termes de relations géométriques entre instances de bases de données hétérogènes, qui nous semble être l'un des aspects à explorer prioritairement à l'avenir, pour l'appariement de données.

BIBLIOGRAPHIE

d'Aquin, M. & Lewen, H. *The NeOn Methodology Handbook - NeOn Methodology in a Nutshell.* Chapitre 6 - Searching Ontologies. **2009**. Disponible sur: http://www.neon-project.org/web-content/media/book-chapters/Chapter-06.pdf

Abadie, N. *Schema matching based on attribute values and background ontology.* 12th International Conference on Geographic Information Science (AGILE'09). **2009.** Disponible sur : http://www.ikg.uni-hannover.de/agile/fileadmin/agile/paper/138.pdf

Abadie, N. & Mustière, S. *Constitution et exploitation d'une taxonomie géographique à partir des spécifications de bases de données.* Revue Internationale de Géomatique, **2010**, Vol. 20(2), pp. 145-174

Abadie, N., Olteanu, A.-M. & Mustière, S. *Comparaison de la nature d'objets géographiques.* Conférence Ingénierie des Connaissances, journée "Ontologies et Gestion de l'Hétérogénéité Sémantique". **2007**

Agarwal, P. *Ontological Considerations in GIScience.* International Journal of Geographical Information Science, Taylor & Francis, **2005**, Vol. 19(5), pp. 501-536

Aleksovski, Z., ten Kate, W. & van Harmelen, F. *Exploiting the structure of background knowledge used in ontology matching.* Ontology Matching Workshop, 5th International Semantic Web Conference. **2006**, pp. 13-24

Aussenac-Gilles, N., Despres, S. & Szulman, S. *The TERMINAE Method and Platform for Ontology Engineering from Texts.* Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, **2008**, pp. 199-223

Aussenac-Gilles, N., Fernandez-Breis, J.T. & Stevens, R. *Ontology Quality.* EKAW 2010 - Workshop W4. **2010**

Baglioni, M., Masserotti, M.V., Renso, C. & Spinsanti, L. *Building geospatial ontologies from geographical databases.* Proceedings of the 2nd international conference on GeoSpatial semantics. Springer-Verlag, **2007**, pp. 195-209

Balley, S. *Aide à la restructuration de données géographiques sur le Web - Vers la diffusion à la carte d'information géographique.* Thèse de doctorat. Université de Marne-La-Vallée, **2007**

Batini, C., Lenzerini, M. & Navathe, S.B. *A comparative analysis of methodologies for database schema integration.* ACM Comput. Surv., ACM, **1986**, Vol. 18, pp. 323-364

Bishr, Y. *Overcoming the Semantic and Other Barriers to GIS Interoperability.* International Journal of Geographical Information Science, Taylor & Francis, **1998**, Vol. 12, pp. 299-314

Bucher, B. *Vers la diffusion en ligne d'information géographique sur mesure.* Habilitation à Diriger des Recherches. Laboratoire COGIT, Institut Géographique National, Université Paris-Est Marne-la-Vallée, **2009**

Bügel, U., Hilbring, D. & Denzer, R. *Application of Semantic Services in ORCHESTRA.* ISESS 2007, International Symposium on Environmental Software Systems. **2007**

Casati, R., Smith, B. & Varzi, A.C. *Ontological Tools for Geographic Representation.* Formal Ontology in Information Systems. **1998**, pp. 77-85

Charlet, J., Bachimont, B. & Troncy, R. *Ontologies pour le web sémantique.* Chapitre 4, **2005**, In : Charlet J., Laublet P. et Reynaud C., coordinateurs, Le Web sémantique. Cépaduès, Toulouse, Hors série de la revue Information - Interaction – Intelligence, Vol. 4(1), ISBN 2.85428.666.9, reprenant et réorganisant le rapport de l'AS « Web sémantique ».

Charte du portail de l'information géographique publique. 2006 [en ligne]. Disponible sur : http://www.geoportail.fr/pub-adm-fw3/display/000/506/187/5061875.pdf. (Consulté le 8 juillet 2011). **2006**

Craglia, M., Goodchild, M.F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maquire, D., Liang, S. & Parsons, E. *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, Revue en ligne publiée par le Joint Research Centre (European Commission), **2008**, Vol. 3, pp. 146 - 167

Cruz, I.F., Sunna, W., Makar, N. & Bathala, S. *A visual tool for ontology alignment to enable geospatial interoperability.* Journal of Visual Languages and Computing, Academic Press, Inc., **2007**, Vol. 18(3), pp. 230-254

Desconnets, J.-C., Libourel, T., Clerc, S. & Granouillac, B. *Cataloguing for Distribution of Environmental Resources.* AGILE'07: 10th International Conference on Geographic Information Science. **2007**

Devogele, T. *Processus d'intégration et d'appariement des bases de données géographiques. Application à une base de données routière multi-échelles.* Thèse de doctorat. Université de Versailles, Laboratoire COGIT, Institut Géographique National, **1997**

Dewynter, B. & Ladurelle-Tikry, E. *Catalogue des infrastructures de données géographiques françaises.* AFIGéO / EUROGI / eSDI-NET +, **2010**

Doucet, A. & Gançarski, S. *Bases de données et Internet, Modèles, Langages et Système.* Chapitre : Entrepôts de données et bases de données multidimensionnelles. Hermès- Lavoisier - Sous la direction de A. Doucet et G. Jomier, **2001**

Duque-Ramos, A., López, U., Fernández-Breis, J.T. & Stevens, R. *Towards an SQUaRE-based Quality Evaluation Framework for Ontologies.* Nathalie Aussenac-Gilles, J. T. F.-B. & Stevens, R. *(ed.).* EKAW 2010 - Workshop W4 - Ontology Quality. **2010**

eSDI Net+ Consortium. *eSDI-NET+ Final Report 2010.* Technische Universität Darmstadt Department of Computer Science Interactive Graphics Systems Group, **2010**

Euzenat, J., Scharffe, F. & Zimmermann, A. *D2.2.10: Expressive alignment language and implementation.* Knowledge Web NoE (FP6-507482), **2007**(1.0 Final)

Euzenat, J. & Shvaiko, P. *Ontology Matching.* Springer, **2007**, pp. 333

Fichtinger, A., Klien, E. & Giger, C. *Challenges in Geospatial Data Harmonisation: Examples and Approaches from the HUMBOLDT project.* 12th International Conference on Geographic Information Science (AGILE'09) - Pre-conference workshop "Challenges in spatial data harmonisation". **2009**

Fonseca, F., Davis, C. & Câmara, G. *Bridging Ontologies and Conceptual Schemas in Geographic Information Integration.* GeoInformatica, **2003**, Vol. 7(4), pp. 355 - 378

Fonseca, F., Egenhofer, M., Davis, C. & Câmara, G. *Semantic Granularity in Ontology-Driven Geographic Information Systems.* Annals of Mathematics and Artificial Intelligence, **2002**, Vol. 36(1-2), pp. 121 - 151

Frank, A. *Ontology for GIS – Draft.* Version 4. **2005**, pp. 450. Disponible sur: http://www.geoinfo.tuwien.ac.at/publications/index.php?by_author:Frank%2C_Andrew_U.:Drafts

Frank, A. *Tiers of Ontology and Consistency Constraints in Geographic Information Systems.* International Journal of Geographic Information Science, Taylor & Francis, **2001**, Vol. 15(7), pp. 667 - 678

Frank, A. *Spatial Ontology: A Geographical Information Point of View*. Spatial and Temporal Reasoning. Springer Netherlands, **1997**, pp. 135-153

Gangemi, A., Presutti, V. *Ontology Design Patterns*, in Staab S. et al. (eds.): Handbook of Ontologies (2nd edition), Springer, **2009**

Gesbert, N. *Etude de la formalisation des spécifications de bases de données géographiques en vue de leur intégration.* Thèse de doctorat. Université de Marne-la-Vallée, **2005**

Goodchild, M.F. *Citizens as Voluntary Sensors : Spatial Data Infrastructures in the World of Web 2.0.* International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission), **2007**, Vol. 2, pp. 24 – 32. Disponible sur: http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/viewFile/28/22

Goodwin, J. *A Methodology for Converting Conceptual Ontologies to OWL.* Version 1.0. Rapport technique, Ordnance Survey Research. **2007**. pp. 17

Grenon, P. & Smith, B. *SNAP and SPAN: Towards Dynamic Spatial Ontology.* Spatial Cognition and Computation, **2004**, Vol. 4(1), pp. 69-103

Guarino, N. *Formal Ontology and Information Systems.* Proceedings of Formal Ontology in Information System. IOS Press, **1998**, pp. 3-15

Guarino, N. & Welty, C. *Evaluating ontological decisions with OntoClean.* Communication ACM, **2002**, Vol. 45(2), pp. 61-65

Gómez-Pérez, A., Fernández, M. & de Vicente, A.J. *Towards a Method to Conceptualize Domain Ontologies.* 12th European Conference on Artificial Intelligence (ECAI'96). **1996**

Gómez-Pérez, A., Fernández-López, M. & Corcho, O. *Ontological Engineering.* Chapitre 3 - Methodologies and Methods for Building Ontologies. Springer, **2003**, pp. 107 – 197

Gruber, T. R., *Toward principles for the design of ontologies used for knowledge sharing*. International Journal of Human-Computer Studies, Academic Press, Inc., **1995**, 43(5-6), pp. 907-928

Hacid, M.-S. & Reynaud, C. *L'intégration de sources de données.* Revue Information - Interaction - Intelligence (R I3), **2004**, Vol. 4(2)

Hamdi, F., Zargayouna, H. & Reynaud, C. *TaxoMap in the OAEI 2008 alignment contest.* Proceedings of the Ontology Matching Workshop, 7th International Semantic Web. Conference. **2008**, pp. 206-213

Hart, G. & Goodwin, J. *Modelling Guidelines for Constructing Domain Ontologies.* **2007**

Horrocks, I. *Ontologies and the semantic web.* Communications of the ACM, **2008**, Vol. 51(12), pp. 58-67

Hotho, A. & Jäschke, R. *Ontology Learning from Folksonomies.* EKAW 2010 - Tutorial 1. **2010**, pp. 44

IGN, *BD Topo Pays, Version 1.2, Descriptif de Contenu,* Edition 1, Institut Géographique National, Paris, France, **2002**, 118 p.

IGN, *BD Carto, Version 3.0, Descriptif de Contenu*, Edition 8, Institut Géographique National, Paris, France, **2006**, 175 p.

IGN, *BD Topo, Version 2.1, Descriptif de Contenu*, Institut Géographique National, Paris, France, **2011 a**, 172 p.

IGN, *BD Carto, Version 3.1, Descriptif de Contenu*, Institut Géographique National, Paris, France, **2011 b**, 54 p.

ISO 19101: *Geographic information - Reference model.* International Organization for Standardization (TC 211), **2002** (19101)

ISO 19107: *Geographic information - Spatial Schema*. International Organization for Standardization (TC 211), **2003** (19107)

ISO 19108: *Geographic information - Temporal Schema. International* Organization for Standardization (TC 211), **2002** (19108)

ISO 19109: *Geographic information - Rules for application schema*. International Organization for Standardization (TC 211), **2005** (19109)

ISO 19110: *Geographic information - Methodology for feature cataloguing.* International Organization for Standardization (TC 211), **2005** (19110)

ISO 19111: *Geographic information - Spatial referencing by coordinates*. International Organization for Standardization (TC 211), **2007** (19111)

ISO 19112: *Geographic information - Spatial referencing by geographic identifiers.* International Organization for Standardization (TC 211), **2003** (19112)

ISO 19115: *Geographic information – Metadata.* International Organization for Standardization (TC 211), **2003** (19115)

ISO 19117: *Geographic information – Portrayal.* International Organization for Standardization (TC 211), **2005** (19117)

ISO 19119: *Geographic information – Services.* International Organization for Standardization (TC 211), **2005** (19119)

ISO 19131: *Geographic information* - Draft international standard. International Organization for Standardization (TC 211), **2007** (19131)

ISO 19136: *Geographic information - Geography Markup Language (GML).* International Organization for Standardization (TC 211), **2007** (19136)

ISO 19139: *Geographic information - Metadata, Implementation specification.* International Organization for Standardization (TC 211), **2007** (19139)

Kashyap, V. & Sheth, A. *Semantic and schematic similarities between database objects: A context-based approach.* VLDB Journal, **1996**, Vol. 5, pp. 276-304

Kavouras, M. & Kokla, M. *Theoris of Geographic Concepts - Ontological Approaches to Semantic Integration.* CRC Press - Taylor & Francis Group, **2008**, pp. 319

Klien, E. *Semantic Annotation of Geographic Information.* Thèse de doctorat. Institute for Geoinformatics, University of Muenster, **2008**

Kovacs, K., Dolbear, C., Hart, G., Goodwin, J. & Mizen, H. *A Methodology for Building Conceptual Domain Ontologies.* **2006**

Kuhn, W. *Geospatial Semantics: Why, of What, and How?* Journal on Data Semantics. Special Issue on Semantic-based Geographical Information Systems, LNCS, **2005**, pp. 1-24

Kuhn, W. *Semantic Reference Systems.* International Journal of Geographical Information Science, **2003**, Vol. 17(5), pp. 405-409

Kuhn, W. & Raubal, M. *Implementing Semantic Reference Systems.* M. Gould, R. L. & Coulondre, S. *(ed.).* 6th AGILE Conference on Geographic Information Science. **2003**, pp. 63 - 72

Lassoued, Y., Wright, D., Bermudez, L. & Boucelma, O. *Ontology-Based Mediation of OGC Catalogue Service for the Web: A Virtual Solution for Integrating Coastal Web Atlases.* Proceedings of the 3rd International Conference on Data and Software Engineering ICSOFT' 2008. Springer, **2008**

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

Lutz, M., Christ, I., Witte, J., Klien, E. & Hübner, S. *Overcoming Semantic Heterogeneity in Spatial Data Infrastructures.* Stroink, D. L. *(ed.).* Geotechnologien Science Report - Information Systems in Earth Management. Koordinierungsbüro GEOTECHNOLOGIEN, **2006**(8), pp. 12-31

Manoah, S., Boucelma, O. & Lassoued, Y. *Schema Matching in GIS.* Proceedings of the Artificial Intelligence: Methodology, Systems, and Applications, 11th International Conference, AIMSA. Lecture Notes in Computer Science, **2004**, Vol. 3192/2004, pp. 500-509

Mark, D., Smith, B., Egenhofer, M. & Hirtle, S. *Ontological Foundations for Geographic Information Science.* McMaster, R. & Usery, L. *(ed.).* Research Challenges in Geographic Information Science. CRC Press - Taylor & Francis Group, **2004**, pp. 335 - 350

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. *WonderWeb Deliverable D18 - Ontology Library (final).* Laboratory For Applied Ontology - ISTC-CNR, **2003**(18)

Mechouche A., Abadie N., Mustière S., *Alignment Based Measure of the Distance between Potentially Common Parts of Lightweight Ontologies*, 5th International Workshop on Ontology Matching, 9th International Semantic Web Conference (ISWC'10), 7 Novembre, Shanghai (China), **2010**

Mechouche, A., Abadie, N., Prouteau, E. & Mustière, S. *Advances in Knowledge Discovery and Management (AKDM-3), A refereed Book of Chapters: post-proceedings of EGC 2011.* Editors: F. Guillet, B. Pinaud, G. R. & Zighed, D. *(ed.).* Chapter Ontology-based Formal Specifications for User-Friendly Geospatial Data Discovery. **2012**, Vol. 3

Minard, A.-L. *Etat de l'art des ontologies d'objets géographiques.* Rapport de stage de Master 1 TAL. Université de Lille 3, **2008**

Nambiar, U., Ludaescher, B., Lin, K. & Baru, C. *The GEON Portal: Accelerating Knowledge Discovery in the Geosciences.* Proceedings of the 8th annual ACM international workshop on Web information and data management. ACM, **2006**, pp. 83-90

Nebert, D.D. *Developing Spatial Data Infrastructures: The SDI cookbook.* Douglas D. Nebert, Technical Working Group Chair, G. S. D. I. *(ed.).* **2004**, pp. 171

Noy, N.F. & McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology''.* Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics, **2001**(Technical Report SMI-2001-0880)

Olteanu, A.-M. *Appariement de données spatiales par prise en compte de connaissances imprécises*. Thèse de doctorat. Université de Marne-La-Vallée, **2008**

OGC. OGC 03-040. *OGC ReferenceModel*. Open Geospatial Consortium, **2003**

OGC. OGC 07-006r1. *Catalogue Services Specification*. Open Geospatial Consortium, **2007**

OGC. OGC 09-025r. *Web Feature Service 2.0 Interface Standard.* Open Geospatial Consortium, **2010**

OGC. OGC 07-036. *OpenGIS Geography Markup Language (GML) Encoding Standard*. Version 3.2.1. Open Geospatial Consortium, **2007**

Parent, C. & Spaccapietra, S. *Database Integration: the Key to Data Interoperability.* In Advances in Object-Oriented Data Modeling, M. P. Papazoglou, S. Spaccapietra, Z. Tari (Eds.), The MIT Press, **2000**

Parlement Européen et Conseil de l'Union Européenne. Directive 2007/2/CE du Parlement Européen et du Conseil du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE) [en ligne]. Journal officiel de l'Union Européenne du 25 avril 2007, L108/1, Disponible sur : http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:FR:PDF (Consulté le 08.07.2011). **2007**

Partridge, C. *The Role of Ontology in Integrating Semantically Heterogeneous Databases.* National Research Council, Institute of Systems Theory and Biomedical Engineering (LADSEB-CNR), Group of Conceptual Modeling and Knowledge Engineering, **2002**

Paul, M. & Ghosh, S.K. *An Approach for Service Oriented Discovery and Retrieval of Spatial Data.* Proceedings of the 2006 international workshop on Service-Oriented Software Engineering. ACM, **2006**, pp. 88-94

Peuquet, D., Smith, B. & Brogaard, B. *The Ontology of Fields.* Compte-rendu de réunion d'experts réalisée dans le cadre du projet Varenius : Panel on Computational Implementations of Geographic Concepts. National Center for Geographic Information and Analysis, Bar Harbor, Maine, 11-13 juin **1998**

Picard V., *Instanciation automatique des liens entre ontologies et schémas de bases de données géographiques à partir des spécifications en langage naturel*. Rapport de stage, Master 2 Documents Electroniques et Flux d'Information, Université de Paris 10 – Nanterre, **2007**, 83 p.

Poveda-Villalón, M., Suárez-Figueroa, M.C. & Gómez-Pérez, A. *A Double Classification of Common Pitfalls in Ontologies.* Nathalie Aussenac-Gilles, J. T. F.-B. & Stevens, R. *(ed.).* EKAW 2010 - Workshop W4 - Ontology Quality. **2010**

Probst, F. *Semantic Reference Systems for Observation and Measurement.* Thèse de doctorat. Institute for Geoinformatics, University of Muenster, **2007**

Rahm, E. & Bernstein, P.A. *A survey of approaches to automatic schema matching.* The VLDB Journal, Springer-Verlag New York, Inc., **2001**, Vol. 10, pp. 334-350

Reed, C. *(ed.).The Spatial Data Infrastructures cookbook.* Chapitre Standards Suites for Spatial Data Infrastructure. **2009**

Reitz, T. *A Mismatch Description Language for Conceptual Schema Mapping and Its Cartographic Representation.* Sara Irina Fabrikant, Tumasch Reichenbacher, M. v. K. C. S. *(ed.).* Geographic Information Science - 6th International Conference, GIScience 2010. Springer, **2010**, pp. 204-218

Reitz, T. & Kuijper, A. *Applying Instance Visualisation and Conceptual Schema Mapping for Geodata Harmonisation.* Monika Sester, L. B. & Paelke, V. *(ed.).* Advances in GIScience - Proceedings of the

12th AGILE Conference. Springer, Lecture Notes in Geoinformation and Cartography, **2009**, pp. 173-194

Reitz, T., de Vries, M. & Fitzner, D. *A5.2-D3 [3.3] Conceptual Schema Specification and Mapping.* Humboldt project, **2009**(002 - Final)

Rodriguez, A., Egenhofer, M. & Rugg, R. *Assessing Semantic Similarities among Geospatial Feature Class Definitions.* Vckovski, A., Brassel, K. & Schek, H.-J. *(ed.).* Interoperating Geographic Information Systems. Chapter 16. Interoperating Geographic Information Systems, Springer Berlin / Heidelberg, **1999**, Vol. 1580, pp. 189-202

Ruas, A. *Modèle de généralisation de données géographiques à base de contraintes et d'autonomie.* Thèse de doctorat. Université de Marne-la-Vallée, Laboratoire COGIT, Institut Géographique National, **1999**

Ruas, A. & Bianchin, A. *Généralisation et représentation multiple.* Ruas, A. *(ed.).* Chapter Échelle et niveau de détail. Hermes Lavoisier, **2002**, pp. 25-44

Schade, S. *Computer-Tractable Translation of Geospatial Data.* International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission), **2010**, Vol. 5. Disponible sur: http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/154/181

Sheth, A.P. & Larson, J.A. *Federated database systems for managing distributed, heterogeneous, and autonomous databases.* ACM Computing Surveys, **1990**, Vol. 22(3), pp. 183-236

Shvaiko, P. & Euzenat, J. *Ontology Matching: State of the Art and Future Challenges.* IEEE Transactions on Knowledge and Data Engineering, **2011**(99)

Shvaiko, P. & Euzenat, J. *A Survey of Schema-based Matching Approaches.* Journal on Data Semantics (JoDS) IV, **2005**, Vol. 3730, pp. 146-171

Simperl, E., Mochol, M. & Bürger, T. *Achieving Maturity: the State of Practice in Ontology Engineering in 2009.* International Journal of Computer Science and Applications, **2010**, Vol. 7(1), pp. 45 - 65

Smith, B. & Mark, D. *Ontology and Geographic Kinds.* 8th International Symposium on Spatial Data Handling (SDH'98). **1998**, pp. 308-320

Sure, Y., Staab, S. & Studer, R. *Ontology Engineering Methodology.* et R. Studer, S. S. *(ed.).* Handbook on Ontologies. Springer, **2009**, Vol. 2, pp. 135 - 152

Suárez-Figueroa, M.C. & Gómez-Pérez, A. *The NeOn Methodology Handbook - NeOn Methodology in a Nutshell.* Chapitre 10 - Reusing Domain Ontologies as a Whole. **2009.** Disponible sur: http://www.neon-project.org/web-content/media/book-chapters/Chapter-10.pdf

Uitermark, H. *Ontology-based geographic data set integration.*Thèse de doctorat. Université de Twente, **2001**

Uschold, M. & Grüninger, M. *Ontologies: principles, methods, and applications.* Knowledge Engineering Review, **1996**, Vol. 11(2), pp. 93-155

Uschold, M. & King, M. *Towards a Methodology for Building Ontologies.* Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95. **1995**

Usländer, T. *Reference Model for the ORCHESTRA Architecture (RM-OA).* ORCHESTRA consortium - Fraunhofer IITB, **2007**(D3.2.3 RM-OA Version 2)

Vandenbroucke, D., Crompvoets, J., Janssen, K. & Biliouris, D. *INSPIRE & NSDI SoP. D4.1 - Summary report regarding the results of the European Assessment of 34 NSDI (first year) - September 2010.* Spatial Applications Division K.U. Leuven Research & Development, **2010**

Vatant, B., Rozat, L., Vandenbussche, P.-Y., Bucher, B. & Abadie, N. *Méthodes et indicateurs pour la sélection d'ontologies fiables et utilisables.* Livrable du projet DataLift, **2011**(D2.1)

Vieu, L. *Spatial Representation and Reasoning in Artificial Intelligence.* Stock, O. *(ed.).* Spatial and Temporal Reasoning. Spatial and Temporal Reasoning, Kluwer, **1997**, pp. 3-41

Volz, S. *Data-Driven Matching of Geospatial Semantics.* et D.M. Mark, A. C. *(ed.).* Spatial Information Theory – COSIT. Springer, **2005**(3693), pp. 115-132

de Vries, M. & Reitz, T. *Conceptual schema matching with the Ontology Mapping Language: requirements and evaluation.* International Conference on Geographic Information Science (AGILE'09), Pre-Conference Workshop "Challenges in Spatial Data Harmonisation". **2009**

W3C. *OWL Web Ontology Language Overview.* McGuinness, D.L. & van Harmelen, F. (eds.). W3C Recommandation. **2004**

W3C. *OWL 2 Web Ontology Language: New Features and Rationale* W3C Recommandation. **2009**

Wache, H. *Semantische Mediation für heterogene Informationsquellen.* Thèse de doctorat. Berlin: Akademische Verlagsgesellschaft Aka, **2003**

Wiederhold, G. *Mediators in the architecture of future information systems.* IEEE COMPUTER, **1992**, Vol. 25(3), pp. 38-49

Winter, S. *Ontology: buzzword or paradigm shift in GI science?* International Journal of Geographical Information Science, Taylor & Francis, **2001**, Vol. 15(7), pp. 587-590

Wu, Z. & Palmer, M. *Verbs semantics and lexical selection.* Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, **1994**, pp. 133-138

Yujian, L. & Bo, L. *A Normalized Levenshtein Distance Metric.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **2007**, Vol. 29(6), pp. 1091-1095

Zhao, T., Zhang, C., Wei, M. & Peng, Z.-R. *Ontology-Based Geospatial Data Query and Integration.* Proceedings of the 5th international conference on Geographic Information Science. Springer-Verlag, **2008**, pp. 370-392

Annexes

Annexe 1. Processus d'alignement automatique proposé par (Hamdi et al., 2008) et mis en œuvre dans la phase d'ancrage du processus d'appariement de schémas présenté au chapitre 4.2

Le processus d'alignement d'ontologies que nous avons mis en œuvre lors de la phase d'ancrage de notre approche d'appariement de schémas est orienté ; il vise à détecter des relations de correspondance entre les concepts d'une ontologie source (O_{source}) et ceux d'une ontologie cible (O_{cible}). Pour chaque paire de concepts sources et cibles des relations d'équivalence (*isEq*), de généralisation-spécialisation (*isGeneral* ou *isA*) ou de proximité sémantique (*isClose*) peuvent être détectées. En outre, à chaque paire de concepts alignés est attribué un score de similarité permettant d'évaluer la fiabilité de la relation de correspondance établie par le système. Ce score correspond à la valeur obtenue par calcul d'une mesure de similarité entre les chaînes de caractères des labels désignant les concepts comparés. Ainsi pour chaque paire de concepts (C_{cible}, C_{source}), de labels respectifs c_c et c_s dont on note la plus longue sous-chaîne de caractères commune x, la similarité des labels est évaluée à l'aide de la mesure suivante :

$$s(c_c, c_s) = \frac{2 * |x|}{(|c_s| + |c_c|)}$$

Lorsqu'au sein d'une paire de concepts dont on cherche à évaluer la similarité, les concepts comparés possèdent plusieurs labels, la similarité des chaînes de caractères est évaluée pour toutes les combinaisons possibles de paires de labels. Au final, seul le score le plus élevé est conservé.

Les relations de correspondance entre concepts des ontologies source et cible sont déduites à l'aide des valeurs de similarités ainsi calculées. Les paires de concepts obtenant une valeur de similarité élevée, supérieure à un seuil déterminé S_{eq}, sont alignées et les concepts les composant sont considérés comme équivalents. En dessous de ce seuil, des relations de correspondances peuvent également être établies. Dans ce cas, les paires de concepts concernées doivent vérifier divers critères :

- **Critère d'inclusion de label** : Soit c_c, le label du concept C_{cible} possédant la valeur de similarité la plus élevée avec le label c_s d'un concept C_{source} de l'ontologie source. Si la valeur de similarité de (c_c , c_s) est supérieure à un seuil S_{incl} donné, et si c_c est inclus dans c_s, alors on suppose que c_s désigne un concept plus précis que c_c, et une relation de correspondance de type (C_{source} *isA* C_{cible}) est générée. Inversement, si c_s est inclus dans c_c, alors c'est une relation de type (C_{source} *isGeneral* C_{cible}) qui est créée. Supposons que l'on compare deux concepts de labels respectifs c_c, « caserne de pompiers », et c_s, « caserne ». Le label « caserne » est inclus dans « caserne de pompiers ». Ce dernier label désigne un concept plus précis que le premier. Une relation de correspondance de type (« caserne de pompiers » i*sA* « caserne ») est donc créée.

- **Critère de similarité relative** : Soient c_{cMax1} et c_{cMax2}, les deux labels de l'ontologie cible possédant les valeurs de similarité les plus élevées avec le label c_s d'un concept C_{source} de l'ontologie source. On appelle « similarité relative » le quotient de la valeur de similarité de c_{cMax2} par celle de c_{cMax1}. Si ce rapport est inférieur à un seuil S_{simrel} donné, alors :

- Si la valeur de similarité de (c_{cMax1} , c_s) est supérieure à un seuil donné S_1, et si c_s est inclus dans c_{cMax1}, alors une relation de correspondance de type (C_{source} *isClose* C_{cMax1}) est générée.

- Si la valeur de similarité de (c_{cMax1} , c_s) est supérieure à un seuil donné S_2, avec $S_2 < S_1$, alors une relation de correspondance de type (C_{source} *isClose* C_{cMax1}) est générée.

- Si la valeur de similarité de (c_{cMax1} , c_s) est supérieure à un seuil donné S_3, alors une relation de correspondance de type (C_{source} *isA* $C_{pere}(C_{cMax1})$) est générée, où $C_{pere}(C_{cMax1})$ est le concept père de C_{cMax1}.

Ce critère de similarité relative permet de départager deux concepts cibles n'ayant pas de valeurs de similarité de labels avec un même concept source suffisants pour être considérés comme équivalents à ce concept, mais néanmoins toujours candidats à l'alignement en raison de la supériorité de leurs scores de similarité par rapport à ceux des autres concepts de l'ontologie cible.

- **Critère de structure** : Soient c_{cMax1}, c_{cMax2} et c_{cMax3}, les trois labels de l'ontologie cible possédant les valeurs de similarité les plus élevées avec le label c_s d'un concept C_{source} de l'ontologie source. Si ces trois labels possèdent des valeurs de similarité avec c_s supérieures à un seuil donné S_{pere}, et que leurs concepts respectifs possèdent un concept père commun, alors une relation de correspondance de type (C_{source} *isA* C_{pere}) est générée, où C_{pere} est le concept père de ces trois concepts. Ce critère permet de détecter des relations de correspondance pour des concepts cibles dont aucun label ne vérifie les critères précédents, mais dont la supériorité du score de similarité par rapport à ceux des autres concepts de l'ontologie cible, et l'existence de concepts frères possédant également des scores de similarité élevés autorise à supposer l'existence relation de type *isA* entre C_s et le concept père de ces concepts cibles.

A chacune des relations de correspondance créées est associé un score égal à la valeur de similarité des labels des concepts concernés. Celui-ci permet de disposer d'une estimation de la fiabilité des relations de correspondance calculées.