
Formalisation des spécifications de bases de données géographiques pour une meilleure compréhension des données

Nils Gesbert
Jeune Chercheur

Laboratoire COGIT, Institut Géographique National,
2/4 avenue Pasteur,
94165 Saint-Mandé Cedex
nils.gesbert@ign.fr

RÉSUMÉ. La conception des bases de données géographiques, notamment de celles produites par l'IGN, s'appuie sur des modèles conceptuels qui sont complétés par des spécifications textuelles. Ces spécifications contiennent de nombreuses informations liées à la sémantique du domaine ainsi qu'à l'expertise et au savoir-faire sur l'acquisition des données. Nous proposons donc d'enrichir les modèles conceptuels des bases à partir de ces spécifications. Notre démarche consiste à réaliser à partir de ces spécifications une ontologie du domaine concerné (entités du monde géographique) afin d'établir ultérieurement des appariements entre ontologie et schémas des bases. L'enrichissement sémantique des modèles conceptuels des bases de données géographiques à l'aide d'une ontologie issue des spécifications nous semble pertinent pour une meilleure interprétation des données. À terme, il s'agit de faciliter l'intégration de plusieurs bases via les ontologies.

ABSTRACT. Conception of a geographical database like those produced by IGN relies on conceptual models which are completed by textual specifications. These specifications contain much information related to the domain's semantics, and to the experts' knowledge about data acquisition. So we wish to extend the conceptual models with these specifications. To this end, we use the specifications to build an ontology of the domain (domain of the geographical entities) and then link this ontology to the databases' schemas. Semantical enrichment of the databases' conceptual models with an ontology built from the specifications seems to us a good way to better data interpretation. Our goal is to eventually use this for the integration of several databases.

MOTS-CLÉS: BD géographiques, spécifications, ontologie, modélisation, terrain conceptualisé

KEYWORDS: Geographical DB, specifications, ontology, modelisation, conceptualised ground.

1. Introduction

Les bases de données d'information topographique produites par l'IGN, telles que la BDCarto ou la BDTopo Pays, ont un contenu qui s'apparente à celui d'une carte topographique : elles ont pour vocation de représenter l'ensemble des entités représentatives du paysage, sur une zone donnée et à une résolution (échelle) donnée. Elles disposent d'un schéma conceptuel orienté objet, dont un exemple est montré figure 1, qui indique les différentes classes d'objets de la base. À l'exception des objets composés (comme ici le cours d'eau nommé), ces classes possèdent toutes un attribut géométrie, qui indique la forme et la position de l'objet. Cet attribut permet la visualisation des données ; les données sont saisies graphiquement et utilisées de même. Ceci a pour conséquence que le schéma conceptuel comprend relativement peu de relations entre objets : beaucoup de relations apparaissent à la visualisation mais ne sont pas inscrites en tant que telles dans la base de données. Ainsi, dans la BDTopo pays, il est fréquent qu'un objet de classe tronçon de cours d'eau et un objet de classe surface d'eau représentent une même entité, le premier permettant de suivre la continuité d'un cours d'eau et le second portant la géométrie détaillée de la portion de cours d'eau correspondante. La relation entre deux tels objets est visible par la superposition géométrique, mais on n'en voit pas trace dans le schéma conceptuel.

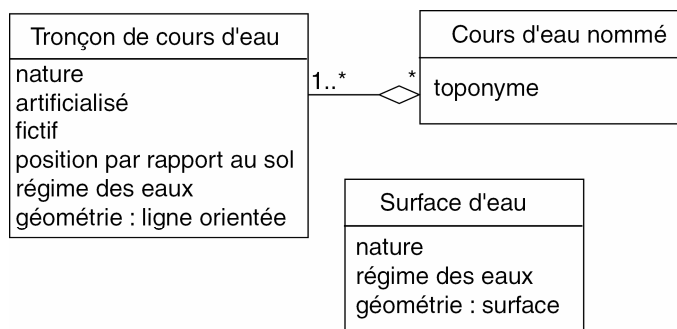


Figure 1. Extrait du schéma conceptuel de la BDTopo Pays concernant le thème hydrographie

Une autre particularité de ces bases de données topographiques est leur non-exhaustivité, qui n'est pas non plus visible dans le schéma conceptuel. En effet, une base de données topographiques qui comprend des classes " bâtiment " et " route " ne contient en général pas des objets correspondant à **tous** les bâtiments et **toutes** les routes de la zone considérée : elle ne contient que les plus importants, les plus représentatifs. La question qui se pose alors tout naturellement est : comment détermine-t-on ce qui est important et ce qui ne l'est pas ? La réponse à cette question dépend de la base considérée, en particulier de sa résolution, et repose en grande partie sur l'expertise du domaine que possèdent les opérateurs qui saisissent

les données. Afin d'éviter autant que possible les ambiguïtés, elle est décrite en détail dans les volumineux documents textuels que sont les spécifications des bases de données.

Pour certaines utilisations des données, ainsi que pour l'intégration de plusieurs bases de données de ce type, seules ces spécifications décrivent la sémantique des objets de la base de données avec une précision suffisante. Mais le format en texte libre est peu adapté à des traitements automatiques, et la structure actuelle des spécifications présente d'autres inconvénients que nous détaillerons. C'est pourquoi nous nous intéressons dans cet article à la formalisation de ces spécifications, avec pour objectif premier de faciliter l'intégration de plusieurs bases de données de ce type.

Il existe des travaux sur la construction de spécifications formelles (Fougères *et al.*, 1999) et des langages de représentation de spécifications tel le langage Z (Spivey, 1992), cependant il s'agit essentiellement pour ces travaux de formaliser des spécifications de *logiciels*, donc de décrire des fonctionnalités ; notre problématique est ici de formaliser des spécifications de bases de données, c'est-à-dire de décrire un contenu statique. Le schéma conceptuel de la base de données est déjà une partie de cette description, mais l'information complémentaire à ce schéma contenue dans les spécifications informelles est particulièrement importante dans notre cas, et aucune étude n'a encore été faite à notre connaissance sur la formalisation de ce type particulier de spécifications.

Nous présenterons dans une première partie la forme de ces spécifications et ce qu'on y trouve ; puis nous proposerons une méthode pour représenter l'information qu'elles contiennent de façon à mettre en valeur l'implicite et les subtilités qui n'étaient pas directement apparentes tout d'abord ; enfin, dans une troisième partie, nous présenterons une formalisation de cette méthode de représentation sous forme d'un métamodèle objet.

2. Présentation des spécifications

Les spécifications des bases de données, tout du moins celles dont nous disposons à l'IGN, sont toutes organisées de la même façon : on a une section pour chacune des classes de la base (le modèle étant exprimé avec le formalisme objet). Cette section, dont un exemple est montré en figure 2, est elle-même divisée en plusieurs parties. Une partie *définition* indique ce que représente, dans le monde réel, un objet de la classe. La seconde partie, *sélection*, qui est souvent la plus importante, tâche quant à elle d'explicitier les critères qui déterminent exactement, en fonction du monde réel, l'extension de la classe, c'est-à-dire l'ensemble de ses instances. Elle donne donc à la fois des informations complémentaires sur la signification des objets de la classe (ainsi “ les talwegs qui ne sont pas marqués par la présence régulière de l'eau sont exclus ” restreint ce que peut représenter un objet tronçon de cours d'eau) et des informations sur l'exhaustivité (ainsi “ **Tous** les

cours d'eau permanents [...] sont inclus. ") Ensuite, une partie *modélisation géométrique* indique plus précisément à quoi correspond la géométrie de l'objet ; une partie *attributs* décrit les significations des attributs autres que la géométrie. L'importance de la géométrie dans le contexte de l'information géographique justifie qu'elle ait une partie dédiée, bien qu'on puisse la considérer comme un simple attribut.

Définition

Portion de cours d'eau, réel ou fictif, permanent ou temporaire, naturel ou artificiel, homogène pour l'ensemble des attributs et des relations qui la concernent, et qui n'inclut pas de confluent.

Sélection

Le réseau hydrographique composé des objets <tronçon de cours d'eau> est décrit de manière continue.

La continuité du réseau n'est toutefois pas toujours assurée dans les cas suivants :

- arrivée d'un cours d'eau en ville
 - infiltration d'un cours d'eau (ex. perte en terrain calcaire)
 - arrivée d'un petit ruisseau temporaire dans une large plaine où son tracé se perd
 - zones de marais où les connexions et interruptions du réseau restent indicatives
- Tous les cours d'eau permanents, naturels ou artificiels, sont inclus.

Les cours d'eau temporaires naturels sont inclus, à l'exception des tronçons de moins de 200 m situés aux extrémités amont du réseau.

Les cours d'eau temporaires artificiels ou artificialisés sont sélectionnés en fonction de leur importance et de l'environnement (les tronçons longeant une voie de communication sont exclus, ainsi que les fossés).

Les talwegs qui ne sont pas marqués par la présence régulière de l'eau sont exclus.

Tous les cours d'eau nommés de plus de 7,5 m de large sont inclus (tronçon de cours d'eau d'attribut <fictif> = " oui " superposé à un objet de classe <surface d'eau>).

Fossé : Les gros fossés de plus de 2 m de large sont inclus lorsqu'ils coulent de manière permanente. Les fossés dont le débit n'est pas permanent sont sélectionnés en fonction de l'environnement. Ils sont généralement exclus lorsqu'ils longent une voie de communication.

Modélisation géométrique

A l'axe et à la surface du cours d'eau (tel qu'il se présente sur les photographies aériennes). L'orientation de l'objet définit le sens d'écoulement. Elle n'est pas significative dans les zones très plates (ex. marais) ni pour les canaux. [...]

Attribut : Nature

Définition : attribut permettant de distinguer les tronçons de cours d'eau libres des obstacles

Type : liste

Valeurs d'attribut : cours d'eau indifférencié / barrage / cascade / écluse [...]

Figure 2. Extrait des spécifications de la classe " tronçon de cours d'eau " de la BD Topo Pays

On constate que ces spécifications donnent l'information dans le sens de la lecture de la base de données, c'est-à-dire qu'il est relativement aisé, étant donné un objet de la base, de savoir à quoi il correspond dans la réalité : on regarde la spécification correspondant à sa classe, qui nous liste toutes les possibilités. Mais si l'on cherche à retrouver dans la base un objet réel, il est nécessaire de consulter l'index des entités représentées (à supposer que celui-ci existe, ce qui n'est pas toujours le cas, et qu'il soit complet) pour trouver la ou les classes concernées, puis de chercher dans les spécifications correspondantes si et comment notre objet peut être représenté ; autrement dit l'information n'est pas directement accessible.

Cette structure présente de gros inconvénients, notamment si l'on dispose de plusieurs bases de données indépendantes couvrant le même territoire, qu'on souhaite les intégrer ou même seulement les comparer pour choisir la plus adaptée à un usage donné. En effet, les regroupements en classes ne sont en général pas les mêmes dans chacune des bases. Par exemple, dans la BDCarto, l'une des bases de données produites par l'IGN, la classe "Tronçon de cours d'eau" comprend les aqueducs, mais dans la BDTopo Pays, une autre de ces bases, ceux-ci sont regroupés dans la classe "canalisation" avec les oléoducs et les gazoducs. On ne peut donc pas (en général) simplement comparer les spécifications d'une classe dans une base avec celles d'une classe qui lui correspondrait dans une autre base et, sauf pour des questions vraiment triviales, on a très rapidement besoin de la totalité des spécifications des deux bases pour être sûr de ne pas oublier de cas particulier.

D'autre part, le format en texte libre est peu adapté à un traitement automatique, or beaucoup d'informations qu'on ne trouve que dans les spécifications sont indispensables à une interprétation correcte des données, notamment tout ce qui est sélection, surtout si l'on veut faire interpréter lesdites données par un programme informatique ne disposant pas des connaissances implicites d'un utilisateur humain qui lui éviteraient les plus grosses erreurs, et auxquelles le programmeur n'aura peut-être même pas pensé puisqu'elles sont implicites.

Aussi nous proposons de représenter les spécifications des bases de données d'une façon plus formalisée et en les structurant différemment. Dans la mesure où une base de données géographiques correspond à une certaine vue, partielle, d'un terrain, il est logique de considérer les spécifications comme la description d'une fonction qui à un terrain donné associe une représentation. Il n'est bien sûr pas nécessaire, pour seulement décrire la signification du contenu de la base, de s'intéresser au véritable processus de saisie des données : on peut décrire cette fonction de façon plus simple et plus abstraite.

3. Comment représenter les spécifications ?

Nous proposons donc de considérer les spécifications comme la description d'une fonction allant du monde réel vers la base de données.

Afin d'éviter les ambiguïtés, nous reprenons le vocabulaire de (Brodaric *et al.*, 2002), qui propose un schéma abstrait pour la conception de bases de données géographiques, faisant intervenir entre autres les notions de concept, de classe, d'instance et d'occurrence. Dans ce schéma, un *concept* est décrit par son intension, qui comprend un mécanisme abstrait de classification permettant de distinguer les objets qui sont des instances de ce concept de ceux qui n'en sont pas. En général, ce mécanisme représente l'expérience et la cognition humaines, qu'on ne cherche pas à analyser dans notre contexte. Ainsi, nous savons reconnaître un cours d'eau ; cela nous permet de parler du concept de " cours d'eau " sans décrire le processus exact permettant cette reconnaissance. Un concept possède également des *propriétés*, par exemple on peut parler de la largeur ou de la navigabilité d'un cours d'eau. Une *classe* est l'extension d'un concept, c'est-à-dire l'ensemble des objets qui sont reconnus comme appartenant à ce concept ; ainsi la classe " cours d'eau " est l'ensemble de tous les cours d'eau. On appelle *occurrence* une entité susceptible d'être placée dans une classe (par exemple une ligne sombre sur une photo aérienne). On appelle *instance* une entité effectivement placée dans une classe (la même ligne une fois qu'elle a été reconnue comme un cours d'eau). Les auteurs distinguent également l'*instanciation* de la *classification*. L'*instanciation* consiste à créer une nouvelle occurrence et à la mettre dans une classe, ce qui correspond à la création d'un objet dans une base de données, donc à la phase de saisie. La *classification* au contraire consiste à placer dans une classe une occurrence préexistante, ce qui correspond à la reconnaissance d'une entité du monde réel comme appartenant à un concept donné, donc à la phase d'interprétation. D'autre part, nous utiliserons de préférence le terme d'" entité " en parlant du monde réel et celui d'" objet " en parlant de la base de données.

Notre fonction qui déduit du monde réel la base de données peut se décomposer en trois étapes : conceptualisation, sélection et modélisation (figure 3) .

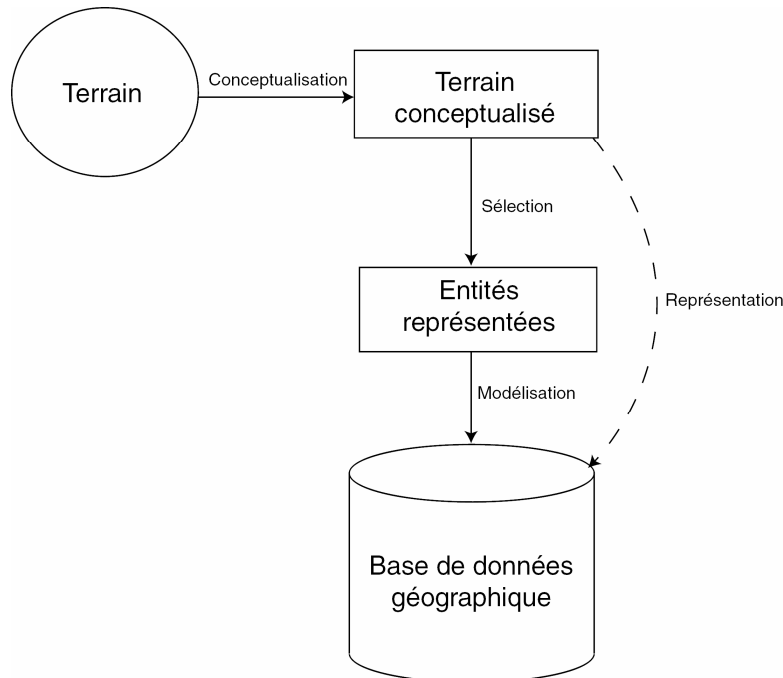


Figure 3. Les différentes étapes menant du terrain à la base de données

Seules les deux dernières étapes, sélection et modélisation, sont décrites dans les spécifications. La première étape correspond à la conceptualisation du terrain. En effet, les spécifications indiquent comment sont représentés les éléments du terrain ; or ces éléments, qui correspondent aux occurrences au sens de (Brodaric *et al.*, 2002), sont issus d'une conceptualisation qui n'est pas triviale. (Smith *et al.*, 1998) mentionne un certain nombre de spécificités qui distinguent les concepts géographiques des autres types de concepts : tout d'abord, un objet géographique est indissociable par nature de l'espace où il se trouve et l'on ne peut séparer le “quoi” du “où” ; les propriétés géométriques et topologiques, en particulier la notion de frontière, sont fondamentales. D'autre part, l'existence de certains objets, tels que les baies ou les péninsules, est le fruit de la cognition humaine. Les auteurs décrivent de tels concepts comme “des ombres projetées par le raisonnement et le langage humains sur l'espace géographique”.

On peut remarquer ensuite que cette conceptualisation dépend du point de vue, et en particulier de l'échelle, ou plutôt, dans le cas des bases de données, de la résolution. Certains concepts ne peuvent exister qu'à certaines résolutions, par exemple il est impossible de définir la limite d'une forêt à un mètre près ou d'en individualiser les arbres à une résolution de 20 m. D'autres peuvent exister à toutes

les résolutions considérées, mais leurs instances changent de nature, ainsi un fleuve qui se divise en plusieurs bras quand on le regarde à 100 m près peut ne plus le faire (dans le terrain conceptualisé, s'entend) lorsqu'on le regarde avec une résolution d'un kilomètre. En quelque sorte, la conceptualisation du terrain à différentes résolutions correspond à ce qu'on voit depuis différentes distances : face à un mur, on voit des briques ; depuis le sommet d'une colline, on peut apercevoir une ville. Toutefois, on peut définir un “ terrain conceptualisé ” abstrait dont ces diverses conceptualisations seraient des vues à une résolution donnée. C'est notamment intéressant dans le cas d'une base de données multi-échelles, ainsi que dans un contexte où plusieurs bases de données coexistent pour représenter une même partie du monde réel : il est plus pratique de considérer que ces bases représentent différentes vues d'une même entité abstraite issue d'une conceptualisation commune du monde réel plutôt que différentes conceptualisations de ce monde.

Les deux autres étapes correspondent à la subdivision des spécifications que nous avons vue en figure 2 : la sélection a une partie dédiée, et la modélisation est composée de la partie modélisation géométrique et de la partie attributs (qu'on pourrait appeler “ modélisation non géométrique ”). Mais, tandis que les spécifications, sous leur forme actuelle, ne font cette subdivision que classe par classe, nous proposons de la faire au niveau global, ce qui nous permettra d'utiliser une classification des concepts du terrain ne correspondant pas classe par classe au schéma conceptuel de la base de données. En effet, on a vu plus haut que les regroupements en classes diffèrent d'une base de données à l'autre, s'affranchir de cette structure est donc nécessaire pour l'intégration.

Pour résumer, les spécifications correspondent à la fonction de représentation (sélection puis modélisation) de la figure 3. Pour les représenter, nous avons adopté dans un premier temps le formalisme objet UML et nous avons défini un profil UML permettant de représenter les concepts géographiques, les classes de la base de données et les spécifications.

4. Description du métamodèle

La démarche proposée va consister à exprimer la fonction de représentation sous forme de liens entre le terrain conceptualisé et la base de données. Nous aurons donc trois parties : l'une correspondant au schéma actuel de la base de données, la seconde au terrain conceptualisé et la troisième au processus de sélection/modélisation. On peut noter que, selon la définition de (Gruber, 1993), la deuxième partie de notre modèle, qui spécifie la conceptualisation du terrain, représente une ontologie : la spécification d'une conceptualisation. (Voir (Guarino *et al.*, 1995) pour les différentes acceptions du terme “ ontologie ”).

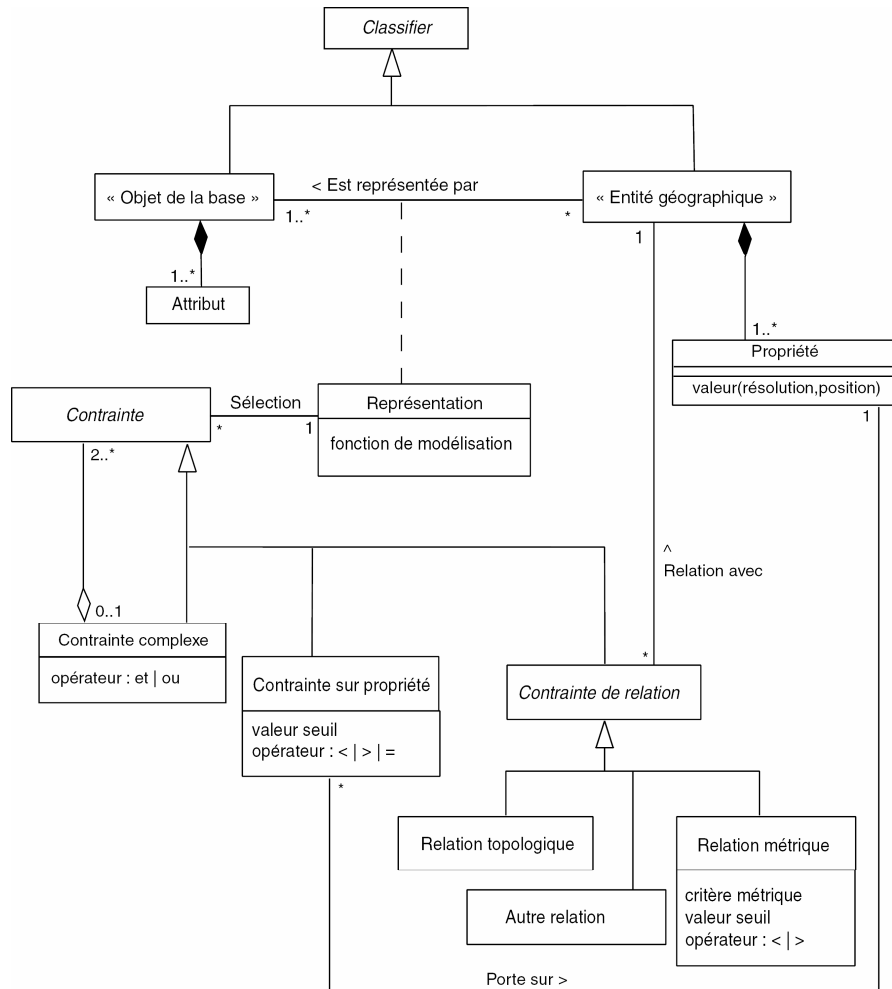


Figure 4. Extrait du profil UML utilisé pour représenter les spécifications

Le profil UML (extension du méta-modèle UML) dont l'essentiel est représenté figure 4 permet de structurer la représentation des spécifications. “Objet de la base” et “entité géographique” sont des spécialisations de *Classifier* du méta-modèle UML (OMG, 2003) et en tant que telles peuvent participer à tous les liens classiques des classes (généralisation/spécialisation, composition, etc.). Cependant ces deux méta-classes ont des statuts différents : les instances d’“Objet de la base” sont des classes au sens habituel des bases de données, éléments du schéma conceptuel de la base. Des instances de ces classes sont créées lors de la saisie de la base, modifiées lors de la mise à jour, etc.. Au contraire, une instance d’“Entité géographique”

représente une classe au sens de Brodaric et Gahegan, c'est-à-dire l'ensemble des entités du terrain (occurrences) correspondant à un concept donné.

Ainsi l'ensemble des instances d'“ Objet de la base ” forme le schéma conceptuel de la base et l'ensemble des instances d'“ Entité géographique ” représente l'ontologie associée (ensemble des concepts géographiques mentionnés dans les spécifications). Les spécifications proprement dites, modélisation et sélection, sont représentées dans la classe d'association *Représentation*, dont les instances indiquent les liens d'appariement entre concepts de l'ontologie et classes de la base. Nous allons détailler ces trois parties.

4.1. Schémas conceptuels des bases et modèle du terrain conceptualisé

L'instanciation de la métaclasse “ objet de la base ” donne l'ensemble des classes des schémas conceptuels des bases concernées. Ces classes possèdent un certain nombre d'attributs, parmi lesquels, en général, une géométrie ainsi que divers attributs thématiques.

La partie modèle du terrain conceptualisé, quant à elle, est constituée d'une hiérarchie de classes, instances d'“ Entité géographique ”, représentant les concepts utilisés dans les spécifications. Les concepts choisis doivent autant que possible être des concepts communs à plusieurs bases de données ; le but est de n'avoir qu'une même ontologie qu'on utilisera pour les spécifications de différentes bases (en l'étendant le cas échéant).

Ces classes possèdent des propriétés dont la valeur peut éventuellement dépendre de la position considérée à l'intérieur de l'instance (par exemple, pour une rivière, la largeur ou la navigabilité varient tout au long du cours), comme il est proposé dans le modèle MADS (Parent *et al.*, 1997). Comme dit plus haut, on ne suppose pas les propriétés des entités du terrain conceptualisé directement accessibles : elles ne le sont qu'à une certaine résolution. Ici la résolution intervient comme une précision (par exemple, on prend la largeur d'une rivière à 3 m près) et peut également indiquer un niveau de généralisation, par exemple la représentation des lacets d'une route nécessite un traitement particulier à petite échelle. La largeur de la rivière avec une précision infinie ou le parcours exact de la route sont des abstractions qu'on ne peut utiliser telles quelles pour créer les objets de la base. D'un point de vue orienté-objet, ces propriétés correspondent donc à des méthodes plus qu'à des attributs, et la résolution est un paramètre de la méthode.

4.2. Liens d'appariement

La classe d'association *Représentation* indique tout d'abord les liens d'appariement qui existent entre classes du modèle conceptuel (“objet de la base ”) et concepts de l'ontologie (“ entité géographique ”). Elle précise ensuite la manière dont on déduit des instances des concepts du terrain conceptualisé les instances de la

base de données, c'est-à-dire : sous quelles conditions on les crée, ce qui correspond à la sélection, et, le cas échéant, comment on remplit les attributs des objets (y compris la géométrie) en fonction des propriétés de l'entité qu'ils représentent, ce qui correspond à la modélisation proprement dite. Pour cette dernière, il sera nécessaire de définir un certain nombre de primitives représentant en quelque sorte l'expertise de l'opérateur. On supposera par exemple l'existence de fonctions relativement simples, telles que “obtenir le contour de l'objet à la résolution [paramètre]”, ou plus compliquées comme “extraire l'axe du réseau à la résolution [param1] en supprimant les culs-de-sac de longueur inférieure à [param2]”.

Chaque instance de la classe d'association *Représentation* relie un concept de l'ontologie (instance d'“entité géographique”) à une instance d'“objet de la base”. Cette instance de *Représentation* est liée par l'association *sélection* à une contrainte que l'entité géographique à représenter doit vérifier.

La classe *Contrainte* permet de représenter les différents types de critères de sélection (Mustière *et al.*, 2003). Plusieurs contraintes élémentaires peuvent être combinées à l'aide d'opérateurs logiques pour constituer une contrainte complexe. Nous distinguons deux types de contraintes élémentaires : tout d'abord, la contrainte sur propriété, qui porte sur la valeur d'une propriété de l'entité elle-même. L'association *Porte sur* permet d'indiquer précisément de quelle propriété il s'agit, étant donné que l'entité en possède a priori plusieurs. Un cas particulier en est la contrainte géométrique, la plus rencontrée, notamment pour spécifier des tailles minimales. Par exemple, “seuls les bâtiments de plus de 20 m² sont inclus dans la classe “bâtiment surfacique” ” indique que la propriété superficie d'une entité géographique “bâtiment” doit être supérieure à la valeur seuil de 20 m² pour que ce bâtiment soit représenté par un objet de la base “bâtiment surfacique”. Cette spécification est représentée par une contrainte sur propriété liée à l'instance de *Représentation* qui relie “bâtiment” et “bâtiment surfacique”.

Ensuite, la contrainte de relation, qui permet de spécifier que l'entité considérée doit ou ne doit pas entretenir une relation avec une entité d'une autre classe (ou éventuellement une autre entité de la même classe). L'association *Relation avec* indique pour ce type de contrainte quelle est l'autre classe d'entités en question. Ainsi rencontre-t-on dans l'exemple de la figure 2 “les tronçons [de cours d'eau] longeant une voie de communication sont exclus”. Il s'agit d'une contrainte sur la représentation des cours d'eau (entité géographique) par des tronçons (objet de la base), donc liée à la relation *Représentation* reliant ces deux classes, et la classe avec laquelle la relation est considérée (*Relation avec*) est l'entité géographique “voie de communication”.

Typiquement, il peut s'agir d'une relation métrique (condition sur la distance entre les deux entités) ou topologique ; mais les spécifications étant souvent relativement floues dans ce type de cas pour laisser libre cours à l'interprétation de l'opérateur qui saisit les données, il est nécessaire d'ajouter un troisième type de relation, “autre”, par exemple pour les relations telles que “mener à”. Il est à noter

que les relations topologiques, comme le contact, doivent être comprises à la résolution près.

Lorsque l'entité géographique considérée respecte les contraintes de sélection, on sait qu'elle est (sauf erreur dans la base) représentée par des objets dans la base de données. Pour indiquer la façon dont le contenu de ces objets est déduit des propriétés de l'entité, la classe *Représentation* contient une fonction de modélisation. Cette fonction sera exprimée dans un langage formel encore à déterminer ; elle pourra recourir aux méthodes de la classe de l'ontologie et aux primitives mentionnées ci-dessus. Il est à noter que, comme les propriétés des entités peuvent dépendre de la position, il peut a priori en aller de même pour les résultats des fonctions donnant les valeurs des attributs des objets de la base, par exemple pour la propriété "navigable" d'un cours d'eau et l'attribut correspondant. Les spécifications précisent comment agir dans ces cas, par exemple prendre la valeur moyenne, ou découper en plusieurs objets pour les différentes valeurs prises par l'attribut. Ainsi un cours d'eau sera découpé en tronçons navigables et en tronçons non navigables. Les quelques opérations possibles (découpage, valeur moyenne, valeur maximale...) seront représentées par des primitives.

4.4. Exemples d'utilisation

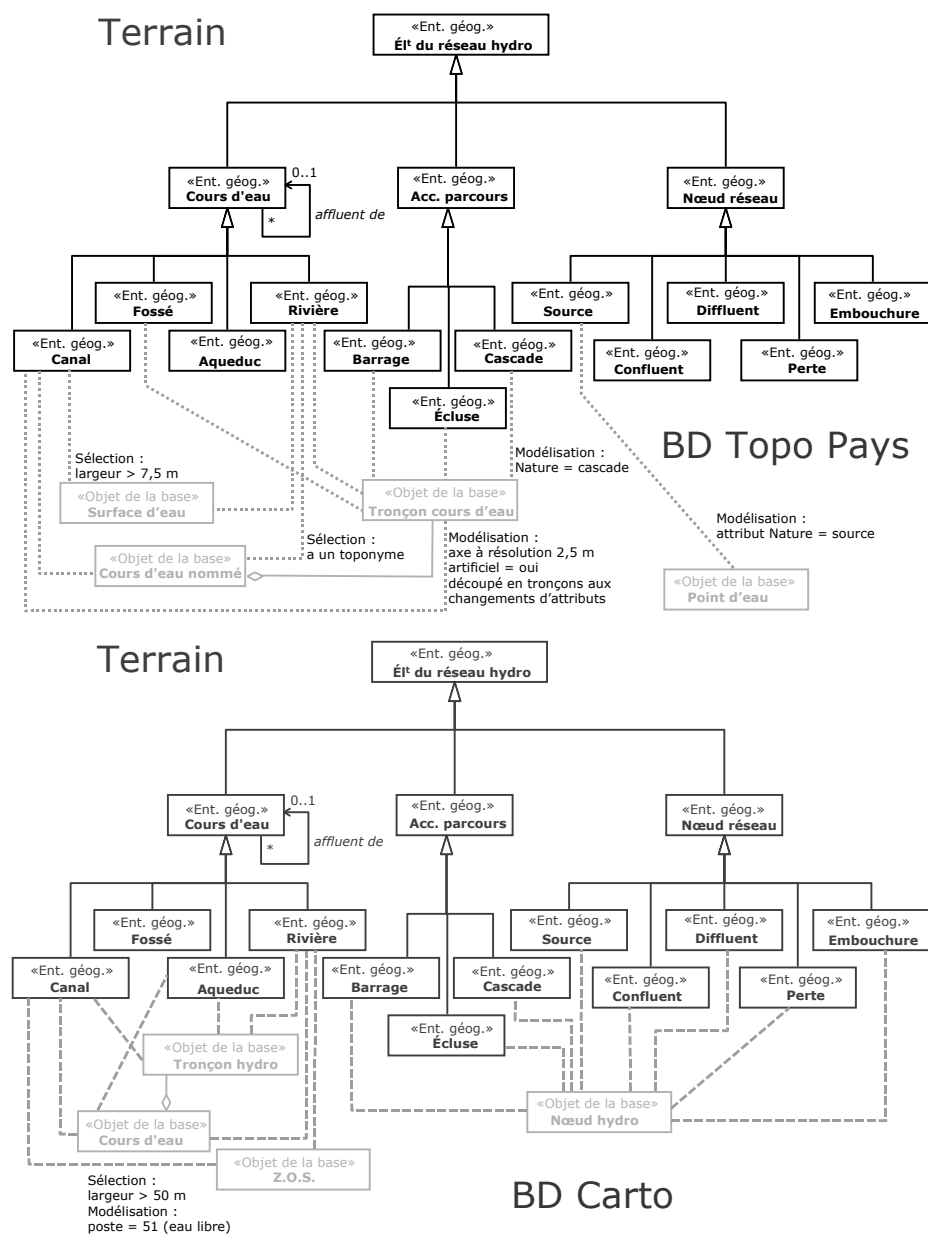


Figure 5. Exemples d'instanciations du métamodèle pour le réseau hydrographique de la BDTopo Pays et de la BD Carto

La figure 5 montre deux exemples d'instanciation de notre métamodèle, pour le réseau hydrographique, respectivement dans les bases de données BDTopo Pays et BDCarto. La partie du haut, correspondant à l'ontologie, a été tout d'abord réalisée pour la BDTopo Pays — les concepts ayant été simplement déterminés par la lecture des spécifications à la recherche de mots-clefs — puis étendue et légèrement modifiée pour pouvoir servir également à la BDCarto. Les parties inférieures correspondent respectivement aux classes des deux bases pour le réseau hydrographique, et les liens en pointillés indiquent les instances de la classe d'association *Représentation*. Le contenu de ces instances (contrainte de sélection et fonction de modélisation) n'a pas été indiqué intégralement mais l'est partiellement sur certains liens, à titre d'exemple.

Seules des modifications mineures ont été nécessaires pour adapter l'ontologie de façon à pouvoir l'utiliser pour les deux bases, ce qui met bien en valeur le fait qu'au-delà des différences de modélisation, elles représentent une même réalité et utilisent les mêmes concepts. Ainsi, par exemple, l'absence de classe “ nœud hydrographique ” dans la BDTopo se traduit par le fait que les accidents de parcours y sont représentés par des *tronçons*, ce qui provoque une différence apparemment importante entre les structures des deux spécifications. Grâce à notre modèle, on peut voir immédiatement que cette différence n'est que superficielle et que les deux bases représentent bien les trois mêmes types d'accidents de parcours.

Cette représentation des spécifications a également l'avantage de faire clairement apparaître la multi-représentation de certaines entités, ainsi le fait qu'une rivière soit dans certains cas représentée simultanément par des tronçons et des surfaces est directement apparent sur le schéma.

5. Conclusion

Nous avons proposé dans cet article un modèle pour la représentation des spécifications de bases de données géographiques et leur intégration avec les modèles conceptuels des bases correspondantes.

Nous espérons qu'un tel modèle permettra une meilleure compréhension des données, et en particulier qu'il facilitera la comparaison de plusieurs jeux de spécifications. L'application la plus importante en serait l'intégration de plusieurs bases de données géographiques à différentes échelles (Devogele *et al.*, 1998), qui pourrait mener à une base de données multi-représentation (Vangenot, 2001). Ce problème a déjà été abordé (Devogele, 1997), mais directement au niveau des données elles-mêmes, c'est-à-dire en termes d'appariement géométrique. Or l'intégration doit également traiter les schémas de données. Cela peut être fait par exemple via la définition d'un schéma “ médiateur ” (vision globale et unifiée) et de correspondances entre schéma médiateur et schémas locaux des diverses bases, ou au moins, si l'on renonce à la vision globale, de correspondances entre les schémas locaux. Il s'agit en quelque sorte d'un appariement sémantique, et l'ontologie joue un

rôle important dans cette étape (Partridge, 2002) ; le but principal de notre modèle est de la faciliter. Il servira également à déterminer les incohérences entre les différentes bases ainsi que les erreurs d'appariement (Sheeren, 2002), en effet seules les spécifications peuvent nous permettre de savoir si une différence de représentation est normale et due à la différence de point de vue ou s'il s'agit d'une erreur ou d'une différence d'actualité des données.

Références

- Brodaric B., Gahegan M., “ Distinguishing Instances and Evidences of Geographical Concepts for Geospatial Database Design ”, Egenhofer M. J., Mark D. M., Eds., *GIScience 2002*, n° 2478 Lecture Notes in Computer Science, Springer-Verlag, 2002, p. 22–37.
- Devogele T., Processus d'intégration et d'appariement de bases de données géographiques ; application à une base de données routières multi-échelles, Thèse de doctorat, Université de Versailles, déc 1997.
- Devogele T., Parent C., Spaccapietra S., “ On spatial database integration ”, *International Journal of Geographical Information Science*, vol. 12, n° 4, 1998, p. 335–352.
- Fougères A.-J., Trigano P., “ Construction de spécifications formelles à partir des spécifications rédigées en langage naturel ”, *Document numérique*, vol. 3, n° 3-4, 1999, p. 215–239
- Gruber T. R., “ Toward Principles for the Design of Ontologies Used for Knowledge Sharing ”, Guarino N., Poli R., Eds., *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer Academic Publishers, 1993.
- Guarino N., Giaretta P., “ Ontologies and Knowledge Bases : Towards a Terminological Clarification ”, Mars N. J., Ed., *Towards Very Large Knowledge Bases*, IOS Press, Amsterdam, 1995.
- Mustière S., Gesbert N., Sheeren D., “ A Formal Model for the Specifications of Geographical Databases ”, Levachkine S., Serra J., Egenhofer M., Eds., *Semantic Processing of Spatial Data, proceedings of workshop GeoPro 2003*, 2003, p. 152–159.
- OMG (Object Management Group), Unified Modeling Language Specification, version 1.5, <http://www.omg.org/technology/documents/formal/uml.htm>, 2003.
- Parent C., Spaccapietra S., Zimanyi E., Donini P., Plazanet C., Vangenot C., Rognon N., Pouliot J., Crausaz P.-A., “ MADS : un modèle conceptuel pour des applications spatio-temporelles ”, *Revue internationale de géomatique*, vol. 7, n° 3-4, 1997.
- Partridge C., The Role of Ontology in Integrating Semantically Heterogeneous Databases, rapport technique n° 05/02, juin 2002, LADSEB-CNR, Padoue.
- Sheeren D., “ L'appariement pour la constitution de bases de données géographiques multi-résolutions : vers une interprétation des différences de représentations ”, *Revue internationale de géomatique*, vol. 12, n° 2/2002, 2002, p. 151–168.

- Smith B., Mark D. M., “ Ontology and Geographic Kinds ”, Poiker, Chrisman, Eds., *Proceedings of the Eighth International Symposium on Spatial Data Handling*, International Geographical Union, Geographic Information Science Study Group, 1998, p. 308–320.
- Spivey J.-M., *The Z Notation: A Reference Manual, Second Edition*, Prentice Hall International, 1992.
- BDTopo Pays/Agglo, spécifications de contenu version 1.2, 2002.
- Vangenot C., Multi-représentation dans les bases de données géographiques, Thèse de doctorat, École polytechnique fédérale de Lausanne, 2001.