

Une approche basée sur la qualité pour faciliter l'intégration de modèles de cubes de données spatiales

Tarek Sboui*, Mehrdad Salehi**, Yvan Bédard***, Sonia Rivest****

Chaire industrielle CRSNG en bases de données géospatiales décisionnelles
Centre de recherche en géomatique, 0611 Pavillon Casault, Département des Sciences
Géomatiques

Faculté de Foresterie et de Géomatique
Université Laval, Québec, Canada,
G1V 0A6

*Tarek.Sboui.1@ulaval.ca

**Mehrdad.Salehi.1@ulaval.ca

***Yvan.Bedard@scg.ulaval.ca

****Sonia.Rivest@scg.ulaval.ca

<http://sirs.scg.ulaval.ca/YvanBedard>

Résumé. L'intégration de cubes de données spatiales permet de faciliter l'accès et la réutilisation des données qui proviennent de différents cubes afin de répondre à des besoins d'analyse stratégique. Cette intégration fait face à des problèmes complexes d'hétérogénéité des cubes de données spatiales. Malgré des travaux intéressants dans le domaine de l'intégration des données, ces problèmes particuliers aux cubes de données spatiales n'ont pas été traités. Cet article explique l'intérêt de l'intégration des cubes de données spatiales, présente une catégorisation des problèmes d'hétérogénéité liés aux modèles des cubes de données spatiales, et propose une approche pour aider à prendre les décisions appropriées concernant l'intégration des cubes de données spatiales. L'approche consiste en un cadre général qui se base sur une structure descendante et une méthode qui propose et évalue un ensemble d'indicateurs de qualité des modèles de cubes à intégrer. L'approche est illustrée par un exemple d'application.

1 Introduction

Les données spatiales sont très utiles pour décrire, visualiser, et analyser efficacement les phénomènes réels qui peuvent être localisés sur ou sous la surface de la terre (ex. pays, routes et accidents de la route). Selon Franklin (1992), jusqu'à 80% des données d'une organisation ont une composante spatiale. Les données spatiales sont de plus en plus nombreuses grâce à l'évolution des outils d'acquisition de données (ex. GPS, images satellitaires et photographies aériennes) et des méthodes de structuration (ex. matriciel et vectoriel) et de représentation (ex. représentations 2D, 3D). De plus, des outils et des méthodes de représentation des données spatiales (ex. des outils de visualisation

cartographique) ont été développés pour mettre en évidence les caractéristiques spatiales des données (ex. position, forme, taille et orientation) et les relations qui existent entre elles (ex. intersection et adjacence) afin de faciliter leur interprétation. Par exemple, ces outils de représentation fournissent une vue cartographique des phénomènes et aident à mieux les comprendre et les analyser (Bédard et al., 2007). Ces caractéristiques font en sorte que les données spatiales sont considérées comme des données complexes.

En outre, des innovations importantes ont vu le jour dans le domaine des technologies de l'information, particulièrement dans les domaines des technologies de bases de données et des systèmes d'aide à la décision (SAD). Les SAD sont des systèmes d'information qui permettent aux analystes d'effectuer des analyses complexes. En effet, ces outils fournissent des techniques, des données et des solutions qui aident les usagers à identifier et à résoudre les problèmes liés à la prise de décisions stratégiques (Turban et Aronson, 2000). Les SAD utilisent généralement les entrepôts de données qui permettent l'exploration de données actuelles et historiques à différents niveaux d'agrégation (Yougworth, 1995; Rivest et al., 2005). Les entrepôts de données se basent généralement sur une structure multidimensionnelle, ils contiennent alors ce qu'on appelle des « cubes de données ». Les cubes de données facilitent la navigation rapide des données selon différents thèmes (i.e., dimensions) et selon différents niveaux de granularité (ex. d'un niveau détaillé à un niveau plus général). Lorsque les cubes de données contiennent des données spatiales (nous les appelons « cubes de données spatiales »), ils permettent à la fois de profiter des avantages de la structure multidimensionnelle et de la représentation cartographique des données spatiales.

Par ailleurs, dans quelques situations, on peut avoir besoin d'utiliser plus d'un cube de données spatiales à la fois pour répondre à un besoin particulier. Par exemple, supposons que l'on ait besoin d'analyser le risque d'infection humaine par le virus du Nil occidental, et que l'on dispose de deux cubes de données spatiales : un cube contenant la propagation du virus du Nil occidental, et un autre contenant la densité de population dans des régions spécifiques. Une méthode efficace d'utiliser les cubes de données spatiales d'une façon homogène est d'intégrer ces cubes. L'intégration de données permet de combiner des données provenant de différentes sources en créant un schéma commun de ces sources ou en générant une base de données plus complète afin de répondre à des besoins spécifiques (Ziegler et Dittrich, 2004). L'intégration des cubes de données spatiales supporte efficacement les outils exploitant ces cubes, tels que les outils OLAP (*On-Line Analytical Processing*) ou SOLAP (*OLAP Spatial*) et les outils de fouille de données (*Data Mining*) ou fouille de données spatiales (*Spatial Data Mining*), permettant ainsi l'analyse décisionnelle stratégique (Bédard et al., 2001; Boussaid et al., 2003; Rivest et al., 2005).

Cependant, l'intégration de cubes de données spatiales fait face à des problèmes complexes d'hétérogénéité de ces cubes de données. Nonobstant l'hétérogénéité spatiale *per se*, l'hétérogénéité ici traitée résulte d'une différence au niveau de la représentation ou de la qualité de deux éléments sémantiquement liés¹ appartenant aux cubes de données spatiales à intégrer. Par exemple, deux systèmes de référence spatiale différents peuvent être utilisés pour les données spatiales de deux niveaux de différents cubes. L'hétérogénéité représente un problème majeur pour l'intégration des données spatiales et, si elle n'est pas traitée, peut engendrer une mauvaise utilisation des données et augmenter le risque de mauvaises

¹ Deux éléments de deux cubes de données spatiales (i.e. deux dimensions, deux niveaux ou deux mesures) sont sémantiquement liés lorsqu'ils représentent différemment les mêmes phénomènes spatiaux.

décisions stratégiques. L'hétérogénéité des cubes de données spatiales peut se situer au niveau des modèles, incluant schémas et métadonnées (ex. différentes significations et différents systèmes de références spatiales) ou bien au niveau des données, incluant les membres des dimensions et les valeurs des mesures (ex. différentes représentations géométriques d'un même membre existant dans des cubes différents).

Plusieurs travaux ont été proposés afin de mesurer la liaison sémantique. La plupart de ces travaux se basent sur les ontologies (Nedas et Egenhofer, 2008). Schwering (2008) présente une revue des méthodes de mesure de la similarité sémantique dans le domaine spatial. D'autres techniques peuvent également être utilisées pour mesurer les liens sémantiques tels que la proximité géosémantique basée sur une matrice à neuf intersections qui est largement utilisé dans le domaine spatial (Brodeur, 2004). Selon notre approche, le choix d'une méthode pour mesurer la liaison sémantique est la tâche d'un intervenant humain ou d'un agent système. Comme notre approche se veut flexible et supporte la méthode choisie par l'intervenant, la suite de cet article utilise un résultat générique à cet égard.

Certaines études ont tenté d'intégrer des bases de données spatiales (Batini et al., 1986; Devogele et al., 1998; Harvey et al., 1999; Ziegler et Dittrich, 2004) et des cubes de données (Mangisengi et al., 2001; Tseng et Chen, 2005). Boussaid et al. (2003) ont étudié l'intégration des données complexes (ex. textes, images et sons). Par contre, ils ne se sont pas intéressés aux problèmes spécifiques aux cubes de données spatiales. Ainsi, a notre connaissance, aucune recherche n'a adressé les problèmes d'intégration des cubes de données spatiales.

Le développement d'une approche pour adresser ces problèmes s'avère nécessaire pour supporter l'analyse décisionnelle stratégique. Une telle approche nécessite la détermination des problèmes d'intégration des cubes, ainsi que le développement d'une approche pour adresser ces problèmes à différents niveaux d'agrégation (cube, dimensions, hiérarchies, niveaux et mesures). Dans ce travail, nous expliquons l'intérêt de l'intégration des cubes de données spatiales, nous présentons une catégorisation des problèmes d'hétérogénéité des modèles de ces cubes, et nous proposons une approche pour aider à l'intégration de modèles de cubes de données spatiales. Cette approche assiste l'intervenant en imitant son modèle mental de données. Elle consiste en un cadre général et une approche qui permettent de définir et évaluer un ensemble d'indicateurs de qualité des modèles à intégrer. Ces indicateurs sont présentés d'une façon intuitive à l'intervenant pour l'aider à prendre les décisions appropriées à propos de l'intégration des cubes de données spatiales. La catégorisation et le cadre proposés constituent une base pour d'autres travaux futurs en lien avec l'intégration des cubes de données spatiales.

La prochaine section présente les concepts liés aux cubes de données spatiales et explique le besoin de l'intégration des cubes de données spatiales et le principe d'intégration de ces cubes de données. La section 3 présente une revue de la littérature traitant de l'intégration de bases de données. La section 4 propose une catégorisation des problèmes d'hétérogénéité qui peuvent se présenter lors de l'intégration de différents cubes de données spatiales. La section 5 propose une approche pour aider à l'intégration de modèles de cubes de données spatiales. Cette approche est illustrée par un exemple d'application. Enfin, la section 6 conclut ce travail et présente quelques perspectives.

2 Les cubes de données spatiales et leur intégration

2.1 Les cubes de données spatiales

Un cube de données contient souvent des données collectées à partir de différentes sources (transactionnelles ou décisionnelles). Ces données sont structurées en mesures agrégées selon des dimensions, ou thèmes d'analyse (Thomsen et al., 1999). Une dimension contient une ou plusieurs hiérarchies qui comprennent des niveaux de détails organisés selon un ordre. Les agrégations des mesures sont généralement pré-calculées (en tout ou en partie) à partir des combinaisons possibles de membres des dimensions et sont optimisées afin de faciliter une recherche rapide d'informations et faciliter le processus de prise de décisions stratégiques en fournissant des réponses très rapides aux requêtes.

La structure des cubes de données (i.e. structure multidimensionnelle) est en accord avec le modèle mental de hiérarchies croisées de l'utilisateur et, par conséquent, elle est appropriée pour l'exploration de données et pour la prise de décisions stratégiques. Quelques travaux ont démontré que la structure multidimensionnelle est plus adaptée pour l'analyse stratégique que la structure transactionnelle (Yougworth, 1995; Bédard et al., 2001; Rivest et al., 2005). Basés sur la structure multidimensionnelle, les outils exploitant les cubes de données, tels que les outils OLAP et SOLAP, fournissent des fonctionnalités intuitives pour interroger des données. Les outils SOLAP utilisent généralement les cubes de données spatiales et appliquent des opérateurs spatiaux de navigation tels que le forage spatial, le remontage spatial et le forage latéral spatial. Le forage spatial permet à l'utilisateur de naviguer d'un niveau général à un niveau plus détaillé dans une dimension spatiale géométrique (ex. visualiser une province d'intérêt à l'intérieur d'un pays affiché à l'écran). Le remontage spatial permet de naviguer d'un niveau détaillé des données à un niveau plus général dans une dimension spatiale géométrique (ex. visualiser le pays auquel appartient la province affichée à l'écran). Le forage latéral spatial permet de visualiser d'autres membres du même niveau de détail dans une dimension spatiale géométrique (ex. visualiser les provinces canadiennes à partir de la province de Québec (Bédard et al., 2005).

Dans le contexte des cubes de données spatiales, trois types de dimensions spatiales peuvent être distinguées : les dimensions spatiales non-géométriques, les dimensions spatiales géométriques et les dimensions spatiales mixtes (Rivest et al., 2003). Dans la dimension spatiale non-géométrique, on considère uniquement les données nominales (ex. les noms des provinces). Ce type de dimension spatiale ne contient aucune information sur les aspects géométriques et, par conséquent, n'est pas suffisant pour supporter la cartographie et une pleine analyse spatio-temporelle des données.

Dans une dimension spatiale géométrique, les membres contiennent des formes géométriques à tous les niveaux de détail (ex. des polygones qui représentent les comtés de la province de Québec). Ces formes géométriques sont géographiquement référencées afin de permettre leur visualisation cartographique et leur analyse. Finalement, la dimension spatiale mixte combine des données géométriques et nominales. La figure 1 présente un exemple des trois types de dimensions spatiales.

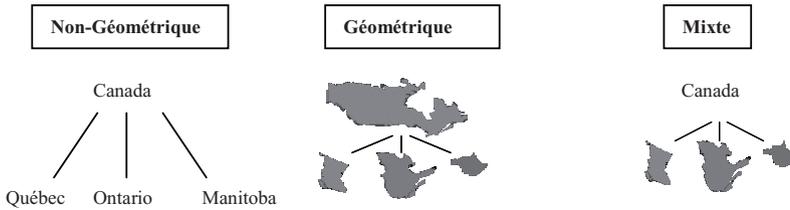


FIG. 1 – Les trois types de dimensions spatiales (Rivest et al., 2003).

Également, deux types de mesures spatiales peuvent être distingués : les mesures spatiales géométriques et les mesures spatiales non-géométriques (Bédard et al., 2007). La mesure spatiale géométrique est le résultat de l'application d'un opérateur spatial qui retourne une géométrie (ex. la projection des lacs situés à moins de 100 mètres des routes peut être représentée par des lignes sur les routes). La mesure spatiale non-géométrique (ex. nombre de kilomètres de route par comté, présence ou non de lac à moins de 100 mètres d'une route) résulte d'une opération spatiale métrique ou topologique retournant une valeur plutôt qu'une géométrie.

Les cubes de données spatiales visent à analyser un sujet en se basant sur une représentation graphique. Par conséquent, un cube de données spatiales est considéré comme étant un cube qui contient au moins une dimension géométrique ou mixte et/ou une mesure spatiale géométrique.

2.2 L'intégration de cubes de données spatiales

L'intégration de données est un processus qui vise à utiliser des données provenant de différentes sources en créant un schéma commun de ces sources ou une base de données plus complète afin de répondre à des besoins spécifiques (Ziegler et Dittrich, 2004).

L'intégration de bases de données spatiales est généralement plus complexe que celle de bases de données non-spatiales. Cette dernière repose sur des techniques fiables et bien documentées (Gervais, 2004). Par contre, l'intégration de bases de données spatiales nécessite des traitements spécifiques et complexes (Faiz, 1999; Gervais, 2004) (ex. l'appariement des géométries des objets spatiaux), et fait face à plusieurs problèmes complexes d'hétérogénéité (ex. les problèmes liés aux différences d'échelles et de généralisation cartographique, de méthodes de géo-référencement, et de précision spatiale). L'intégration de cubes de données spatiales fait face à des complexités supplémentaires par rapport aux deux premières (ex. les problèmes liés aux différences de méthodes d'agrégation).

Tout d'abord, nous définissons la notion d'intégration de cubes de données spatiales (le « quoi »), et nous discutons les besoins qui expliquent l'intérêt d'une telle intégration (le « pourquoi »).

2.2.1 Qu'est-ce que l'intégration des cubes de données spatiales ?

L'intégration des cubes de données vise à créer soit 1) un nouveau schéma qui permet l'accès et la réutilisation des données qui existent dans différents cubes, soit 2) un cube de données spatiales qui combine les données de cubes existants. Comme ce deuxième type d'intégration consiste à peupler davantage le cube par le processus ETL (i.e. faire des ajouts de membres), nous décrivons ci-après uniquement le premier type d'intégration.

Le schéma ou le nouveau cube de données spatiales doit être approprié pour l'analyse et la prise de décisions stratégiques dans le cas d'une problématique particulière. Généralement, l'intégration des cubes de données spatiales peut faire référence à :

1. intégrer les mesures. On peut distinguer deux types d'intégration de mesures. Premièrement, on peut ajouter une nouvelle mesure dans un cube à partir d'un autre cube (ex. ajouter dans un cube *Élection municipale* une mesure *Âge* provenant d'un cube *Élection nationale* grâce aux dimensions et membres communs). Deuxièmement, on peut créer une nouvelle mesure unifiée (ex. *Densité de population* qui provient de *Superficie* du cube A et *Nombre de personnes* du cube B, grâce aux dimensions et membres communs).
2. intégrer les dimensions de ces cubes. On peut distinguer trois types d'intégration de dimensions. Le premier type consiste à intégrer les dimensions appartenant à des cubes différents afin d'en créer une nouvelle, seule et unique (ex. fusionner les dimensions *Province* et *État* des cubes Canada, USA et Mexique pour faire une seule dimension du cube Amérique du Nord). Un cas particulier consiste à sélectionner une des dimensions sémantiquement liées de deux cubes pour en créer une nouvelle (ex. sélectionner une des deux dimensions *Découpage électoral* et *Découpage municipal* qui appartiennent à deux cubes différents). Le deuxième consiste à ajouter une ou plusieurs dimensions d'un cube à un autre (ex. ajouter à un cube existant une dimension « temps » provenant d'un autre cube). Dans ce cas, le résultat d'intégration des modèles de cubes peut être un modèle en constellation. Le troisième consiste à modifier une dimension d'un cube en utilisant des éléments (niveaux, membres) de dimension provenant d'un autre cube (ex. ajouter un niveau *Multi-pays* et transformer une hiérarchie *Territoire* qui était simple pour un cube d'un seul pays en une hiérarchie multiple tenant compte des particularités des divisions territoriales de chaque pays).

Dans tous ces cas, il s'agit d'intégrer les dimensions ou les mesures en tenant compte des nouvelles informations requises par l'utilisateur. Le choix entre les différentes possibilités ci-haut n'est pas toujours évident. Ce choix demande une connaissance avancée de la part de l'intervenant ainsi que des outils pour faciliter et supporter son intervention. La présente approche se situe dans ce contexte et vise à développer un cadre et une méthode afin de faciliter ce travail.

2.2.2 Quels sont les besoins qui expliquent l'intérêt de l'intégration des cubes de données spatiales?

Dans le domaine de la prise de décision stratégique, on peut avoir besoin d'utiliser des données qui existent dans différents cubes de données spatiales. Nous regroupons ces besoins en trois groupes :

a) Navigation simultanée et rapide de différents cubes de données spatiales : Les utilisateurs de différentes disciplines peuvent avoir le besoin d'accéder et naviguer simultanément à travers plusieurs cubes de données spatiales. Par exemple, des situations d'urgence, telles que la propagation d'un feu de forêt d'une province à une autre, nécessitent l'interrogation simultanée des cubes développés dans ces provinces. Plus spécifiquement, nous pouvons intégrer les données à partir de différents cubes de données spatiales, développés dans ces provinces, afin d'obtenir les informations appropriées et intervenir rapidement à différents niveaux (local et national) ou dans différents domaines (ex. géographique et économique). Dans une telle situation, l'intégration de cubes de données spatiales permet de réduire les pertes.

b) Insertion des données spatiales dans un cube : Les données stockées dans des cubes ne sont pas uniquement collectées à partir de sources transactionnelles; elles peuvent également être collectées à partir d'autres cubes (Bédard et al., 2001). L'intégration qui étaient établis la collecte de données à partir d'autres cubes. Par exemple, un cube de données sur la construction des ponts peut être chargé à partir d'un autre cube qui contient des données décrivant la circulation routière.

c) Analyse interactive de l'évolution des phénomènes : afin d'analyser l'évolution des phénomènes (ex. les peuplements forestiers), il est souvent indispensable de comparer des données décrivant ces phénomènes à différentes époques. Dans certains cas, cette comparaison se fait en utilisant des cubes de données spatiales qui étaient établis à différentes époques pour analyser un phénomène donné. Par exemple, l'évolution des peuplements forestiers de la forêt Montmorency peut être étudiée en intégrant des cubes de données spatiales d'inventaires décennaux qui ont été construits à différentes époques (Miquel et al., 2002).

Il est possible, par ailleurs, de demander pourquoi ne pas intégrer les données directement des sources à partir desquelles on a créé les cubes de données spatiales ? On peut distinguer trois raisons d'intégrer les cubes de données spatiales plutôt que les sources de données :

a) Nous n'avons probablement plus accès aux sources de données à partir desquelles les cubes de données ont été créés à cause de différentes raisons (ex. l'entreprise ne donne plus accès aux sources de données, contraintes de confidentialité, transformations trop coûteuses).

b) Nous avons besoin des données historiques qui existent généralement uniquement dans les cubes de données spatiales. En effet, dans les sources de données transactionnelles, les données historiques sont généralement modifiées ou remplacées par de nouvelles données avant d'être détruites ou archivées, tandis que les cubes conservent les données historiques pour des fins d'analyse stratégique (Bédard et al., 2001).

c) Dans le contexte de la prise de décision, intégrer des cubes de données spatiales est plus efficace que d'intégrer les sources de données. En effet, dans un cube de données spatiales, contrairement aux sources de données, les agrégations possibles des mesures, pour toutes les combinaisons possibles des membres, peuvent être pré-calculées en utilisant différents opérateurs (ex. opérateurs spatiaux tels que l'intersection). Ces agrégations exigent généralement un travail laborieux qui consiste en, par exemple, la définition de procédures d'agrégation ou la définition d'une nouvelle couche spatiale comme agrégation d'autres couches. Parfois, ces agrégations ne peuvent pas être automatisées et nécessitent l'intervention d'un opérateur humain. La réutilisation des cubes de données nous évite de redéfinir et de reconstruire ces agrégations, et ainsi de réduire fortement le temps et le coût de développement des cubes de données spatiales.

3 État de l'art sur l'intégration de bases de données

De nombreuses approches ont été proposées pour résoudre les problèmes d'intégration des bases traditionnelles de données. Les approches peuvent être classifiées selon différents critères tels que le *degré d'importation* de données à partir de différentes bases (les entrepôts de données et les multibases), le *degré d'intégration* (ex. architectures faiblement couplées et architectures fortement couplées), le *type de données* à intégrer (ex. fichiers et bases de données relationnelles), et les *caractéristiques* des bases à intégrer (ex. distribution et autonomie) (Sheeren, 2005). Le degré d'importation et le degré d'intégration sont les critères les plus souvent utilisés dans la littérature (Busse et al., 1999; Rahm et Bernstein, 2001; Sheeren, 2005). Selon ces derniers critères, l'intégration peut être matérialisée ou non-matérialisée.

1. *Intégration matérialisée*. Consiste à copier l'ensemble de données de différentes bases et à les rassembler dans une base unique (ex. les entrepôts de données (Inmon, 1996; Ranjan et Khalil, 2008)).
2. *Intégration non-matérialisée*. Consiste à définir une technique d'utilisation des bases de données sans les intégrer physiquement (i.e., les données sont gardées au niveau des bases d'origine). Différentes architectures peuvent être distinguées dans cette catégorie : les *architectures faiblement couplées* qui correspondent aux *architectures multibases*, les *architectures fortement couplées* qui correspondent aux *architectures fédérées* et les *architectures de médiation* (Sheeren, 2005):
 - (a) *Architecture multibase* : Elle utilise les schémas des bases d'origine (sans définir un schéma unifié des données). Les bases d'origine sont donc faiblement intégrées et gardent une grande autonomie. L'interrogation des bases se fait à l'aide d'un langage commun. Cependant, la cohérence entre ces bases n'est pas assurée (Kim et al., 2007).
 - (b) *Architecture fédérée* : Elle consiste à définir un schéma unifié appelé *schéma fédéré* qui constitue l'interface d'accès aux bases à intégrer. L'intégration se fait au niveau des schémas. À l'inverse des architectures multibase, les architectures fédérées sont dites fortement couplées (Sheth et Larson, 1990; Shankar et al., 2007).
 - (c) *Architecture de médiation* : Elle consiste à définir une couche logicielle (appelée « médiateur ») permettant d'accéder et interroger différentes bases hétérogènes. Afin d'assurer l'intégration, le médiateur utilise des informations à propos des données des différentes bases (métadonnées). Contrairement à l'architecture fédérée, celle-ci est principalement conçue pour l'interrogation et non pour la mise à jour des bases d'origine (Wiederhold, 1992; Boucelma et al., 2002; Lutza et al., 2008).

À ces dernières, nous devons ajouter les entrepôts virtuels de données qui effectuent le processus ETL (Extract, Transform, Load) à la volée, incluant l'intégration, dès l'ouverture d'une séance de travail pour ensuite ne jamais matérialiser le résultat de cette intégration.

Il faut noter qu'il n'existe pas un consensus ni sur la taxonomie des approches de l'intégration des bases de données, ni sur la terminologie utilisée. Par exemple, certains auteurs utilisent le terme « multibase » pour indiquer toute intégration de bases de données (Sheth et Larson 1990; Sheeren, 2005).

Au cours des dernières années, plusieurs travaux se sont focalisés sur la résolution des conflits liés à l'intégration de bases de données complexes telles que les données spatiales

(Wicaksana, 2007). Ces conflits sont généralement causés par la représentation de la même réalité en utilisant différents objets spatiaux (ex. deux cartes décrites représentées différemment mais indiquant les mêmes routes).

Les approches d'intégration de bases de données spatiales ont tenté (1) d'adapter les approches d'intégration des bases de données non-spatiales et (2) d'utiliser des outils spécifiques pour comparer et intégrer les données spatiales représentant les mêmes phénomènes réels (Devoegele et al., 1998; Laurini, 1998; Harvey et al., 1999; Boucelma et al., 2002; Xiong et Sperling, 2004, Kunapo et al., 2007). Deux approches principales peuvent être distinguées (Devoegele, 1997; Sheeren, 2005) :

- Une approche mono-représentation : consiste à fusionner les représentations utiles des données de deux bases.
- Une approche multi-représentations : consiste à préserver les différentes représentations en tolérant les redondances (ex. le concept *vuel* proposé par Bédard et Bernier (2002)).

En plus, certains comités nationaux et internationaux ont proposé des normes qui visent à unifier la description des données et des métadonnées spatiales en fournissant un ensemble commun de terminologies et de définitions. Le fait d'unifier la description des données spatiales permet d'éviter ou de minimiser les problèmes d'hétérogénéité, facilitant ainsi leur intégration. Des exemples de normes sont : (1) Geographic extensible Markup Language (GML), élaborée par l'Open GIS Consortium (OGC) (Portele, 2007) pour encoder les données spatiales, et (2) ISO 19115:2003. Geographic information – metadata (ISO, 2003), élaborée par ISO pour définir les métadonnées des informations géographiques.

Concernant les cubes de données, certains travaux d'intégration se sont focalisés sur l'appariement des structures des cubes en se basant sur une représentation commune en utilisant un langage tel que XML (Boussaid et al., 2003; Pérez et al., 2008). Quelques travaux se sont basés sur l'approche fédérée (Bruckner et al., 2001; Mangisengi et al., 2001; Nguyen et al., 2001; Tseng et Chen, 2005). D'autres se sont basés sur l'architecture *Grid*² (Niemi et al., 2003).

Malgré qu'il existe une multitude d'approches pour intégrer les bases de données spatiales et que ceci fasse l'objet de tout un domaine d'études (i.e. la géomatique), nous n'avons trouvé aucun travail portant sur l'intégration des cubes de données spatiales et la présente recherche constitue une première en ce sens. Bien que les approches présentées ci-dessus peuvent être utilisées pour faire une partie de l'intégration des cubes spatiaux, une telle intégration fait face à d'autres problèmes particuliers au niveau de la structure multidimensionnelle (mesures spatiales, dimensions spatiales, niveaux spatiaux et membres spatiaux), au niveau des agrégations spatiales, au niveau des croisements des membres spatiaux et au niveau des autres métadonnées à propos des données spatiales (ex. système de référence spatiale, couverture spatiale, méthode d'acquisition de données spatiales). Ces problèmes existent également avec les cubes non-spatiaux, mais leur complexité est moindre. Une étude approfondie des nombreuses facettes de l'intégration des données spatiales dépasse largement le cadre d'un seul article puisqu'il y a un grand nombre d'alternatives, que le tout est grandement contextuel et qu'il est typiquement impossible (ou risqué) d'avoir une procédure complètement automatique. Par conséquent, l'intégration des cubes de données

² L'infrastructure *Grid* est utilisée pour lier plusieurs ressources informatiques telles que des personnes et des données (Foster et Kesselman, 1999).

spatiales nécessite une étude approfondie et une approche spécifique permettant à l'intervenant de juger, au fur et à mesure, de la pertinence du résultat possible (souvent à l'aide d'indicateurs explicites ou de vérifications échantillonnées).

4 Catégorisation des problèmes d'intégration de modèles de cubes de données spatiales

L'intégration de cubes de données spatiales consiste à intégrer les modèles (schémas et métadonnées) ainsi que les données (les membres des dimensions et les valeurs des mesures) de ces cubes de données. L'intégration de modèles fait face à des problèmes d'hétérogénéité au niveau des schémas, au niveau des métadonnées ainsi qu'au niveau de la qualité de ces éléments, alors que l'intégration des données fait face à des problèmes d'hétérogénéité des données *per se* et de leur qualité.

Dans cet article, nous considérons que l'*hétérogénéité des schémas* (différence au niveau de la structure) et ainsi que l'*hétérogénéité des métadonnées* (différence au niveau de tous les autres éléments du modèle d'un cube tels que système de référence spatiale, contraintes d'intégrité, etc.) font partie de l'*hétérogénéité au niveau de la représentation* des modèles. Un exemple d'hétérogénéité des schémas est la différence au niveau de la définition de la hiérarchie d'une même dimension dans deux cubes différents. Par exemple, la hiérarchie de la dimension *Région Administrative* peut avoir les niveaux suivants dans un cube : *Pays, Province, Comté, Municipalité*; et les niveaux suivants dans un autre cube : *Pays, Région, Municipalité*. Un exemple d'hétérogénéité des métadonnées est la différence au niveau de la précision de la localisation cartographique des membres spatiaux. Un exemple d'hétérogénéité des données spatiales est la représentation différente d'une même route sur deux cartes; elle est représentée avec une ligne double symbolique simplifiée sur une carte, alors qu'elle est représentée avec une seule ligne centrale telle que mesurée sur une autre carte. Dans cet article, nous nous intéressons aux problèmes d'intégration de modèles de cubes de données spatiales, et nous proposons une catégorisation de ces problèmes. Pour introduire cette catégorisation, nous présentons un exemple d'utilisation de deux cubes de données spatiales (voir figure 2). Pour déterminer le risque d'infection par le virus du Nil occidental sur la population, on intègre deux cubes de données spatiales *C1* et *C2* (modélisés respectivement dans la figure 2(a) et la figure 2(b)). Le premier cube de données spatiales *C1* est utilisé pour déterminer la densité de la population dans des régions spécifiques et pour des périodes données. Le deuxième cube de données spatiales *C2* est utilisé pour suivre la propagation du virus du Nil occidental. L'exemple proposé est extrait d'un cas réel qui consiste à déterminer le risque d'infection par le virus du Nil occidental sur la population canadienne.

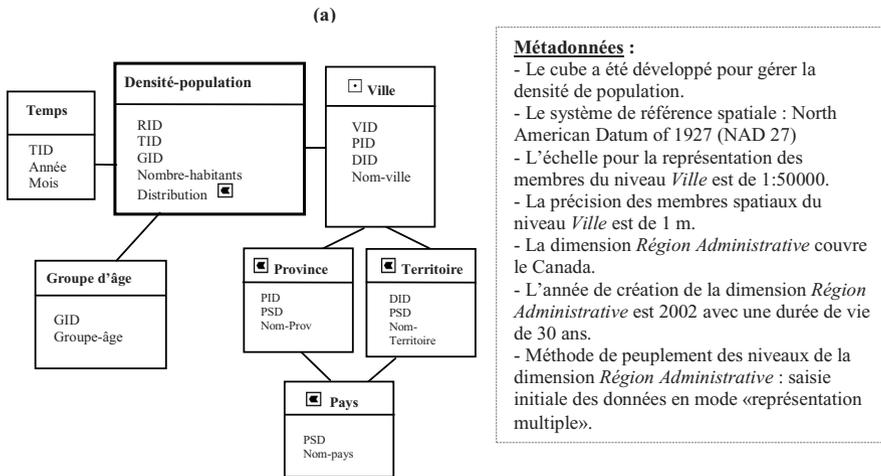
C1 contient trois dimensions (*Temps, Région Administrative, et Groupe d'âge*) et une table de faits (*Densité-population*) avec une mesure spatiale géométrique (*Distribution*), une mesure numérique (*Nombre-habitants*) et les clés étrangères des dimensions. La dimension *Région Administrative* de *C1* contient une hiérarchie qui contient quatre niveaux qui sont ordonnés comme suit : *Pays, Province, Territoire* et *Ville*. La dimension *Temps* de *C1* contient une hiérarchie avec deux niveaux : *Année* et *Mois*. La dimension *Groupe d'âge* de *C1* contient une hiérarchie avec un seul niveau.

C2 contient quatre dimensions (*Période, Région, Famille de virus, et Forêt*) et une table de faits (*Propagation virus du Nil occidental*) qui permet de déterminer les endroits affectés

par le virus. Cette table contient une mesure spatiale géométrique (*Zone-affectée*) qui indique les types de zones forestières potentiellement affectées par les différentes familles de virus du Nil occidental pour chaque période. La dimension *Période* de *C2* contient une hiérarchie avec deux niveaux : *Année* et *Mois*. La dimension *Région* de *C2* contient deux hiérarchies : une hiérarchie avec quatre niveaux : *Pays*, *Province*, *Territoire* et *Cité*, et une autre avec trois niveaux : *Pays*, *État* et *Cité*. Les dimensions *Famille de virus* et *Forêt* contiennent chacune une hiérarchie qui contient un seul niveau.

Les niveaux spatiaux géométriques ainsi que les mesures spatiales géométriques sont modélisés à l'aide de pictogrammes spatiaux. Un pictogramme est un symbole qui facilite la représentation des géométries dans le modèle d'une base de données spatiale (Bédard et Larrivée, 2007). Dans notre exemple, nous utilisons les pictogrammes développés dans l'outil de modélisation spatio-temporelle *Perceptory*³ où le pictogramme «□» représente une géométrie de type point, le pictogramme «▣» représente une géométrie de type ligne, et le pictogramme «■» représente une géométrie de type polygone. Par exemple, dans le modèle de la figure 2(a), les géométries des niveaux de la dimension *Région Administrative* de *C1* (*Ville*, *Province*, *Territoire*, *Pays*) sont modélisées en utilisant respectivement le pictogramme «□» et trois pictogrammes «▣». Par ailleurs, la mesure spatiale *Distribution* est modélisée en utilisant le pictogramme «■».

Dans notre exemple, les métadonnées des cubes de données spatiales contiennent des informations à propos du système de référence spatiale, de l'échelle et de la précision des membres de quelques niveaux spatiaux, la couverture spatiale et l'année de création de la dimension de quelques dimensions, ainsi que la méthode de peuplement de quelques niveaux. Il faut noter que les métadonnées peuvent contenir d'autres informations selon le besoin.



³ Site web de *Perceptory* : <http://sirs.scg.ulaval.ca/perceptory>

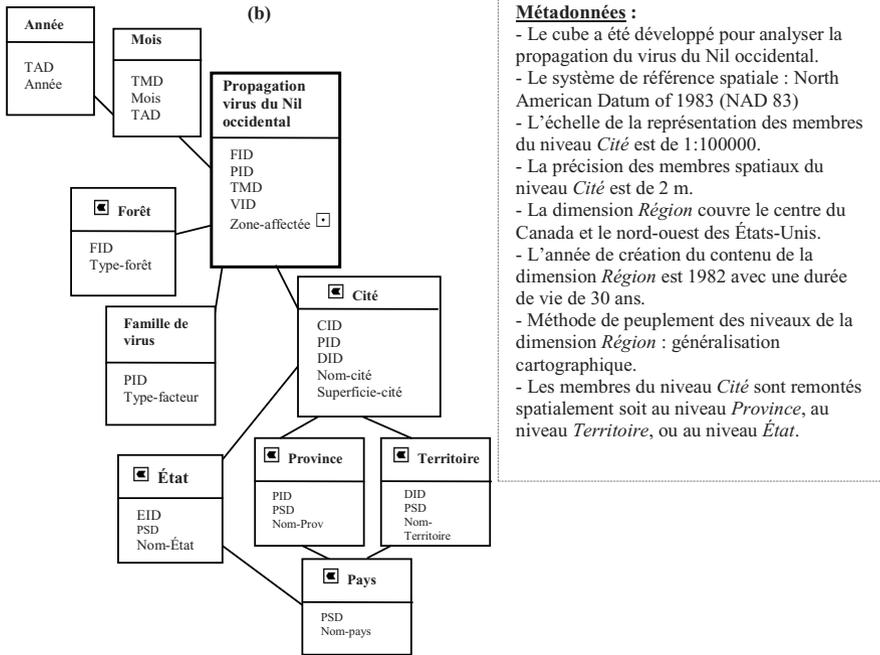


FIG. 2 (a et b) – Deux modèles de cubes de données (C1 et C2).

Les problèmes d'hétérogénéité des modèles de cubes de données spatiales peuvent exister à cinq niveaux différents: cubes, dimensions, hiérarchies, niveaux de dimensions et mesures. Dans chaque niveau, l'hétérogénéité peut être observée au niveau des schémas ou au niveau des métadonnées des cubes de données spatiales. Afin de représenter les problèmes d'hétérogénéité d'une façon rigoureuse, nous proposons une formalisation des schémas des éléments de cube. Dans cette formalisation, *C* indique un cube, *D* indique une dimension, *H* indique une hiérarchie, *N* indique un niveau, *a* indique un attribut.

1. $N = \{a_1, a_2, \dots, a_n\}$: un niveau est un ensemble de n attributs. n étant la cardinalité de l'ensemble. Par exemple, le niveau *Ville* = {nom_ville, position_ville}.
2. $H = (\{N_1, N_2, \dots, N_n\}, <)$: une hiérarchie est un ensemble de h niveaux et une relation d'ordre entre ces niveaux, h étant la cardinalité de l'ensemble. La relation d'ordre est définie comme suit : $\forall N_1, N_2 \in H, N_1 < N_2$ si N_2 fore vers le haut N_1 (roll-up). Par exemple, la hiérarchie (*Ville, Province, Territoire, Pays*) de la dimension *Région Administrative* = (*Ville < Province < Pays, Ville < Territoire < Pays*).
3. $D = \{H_1, H_2, \dots, H_d\}$: une dimension est un ensemble de d hiérarchies, d étant la cardinalité de l'ensemble et étant habituellement égal à 1. Par exemple, la

dimension *Région Administrative* = { *Ville* < *Province* < *Pays* , *Ville* < *Territoire* < *Pays* } a une cardinalité de 1.

4. Un mesure est un attribut *a* qui décrit le sujet d'analyse.
5. $C = (\{D_1, D_2, \dots, D_c\}, \{m_1, m_2, \dots, m_q\})$: un cube est un ensemble de *c* dimensions et de *q* mesures.

Nous proposons cinq catégories d'hétérogénéités qui peuvent exister entre les modèles de cubes de données spatiales :

1. Hétérogénéité Cube-à-Cube

- *Hétérogénéité au niveau du schéma*. Elle apparaît quand deux cubes ont des mesures qui sont sémantiquement liées et qui sont représentées selon différents nombres de dimensions. Par exemple, les deux mesures *intersection routière* et *points d'intersection* peuvent être analysées selon respectivement deux dimensions (*Route* et *Région*) et trois dimensions (*Route*, *Région* et *Temps*).
- *Hétérogénéité au niveau des métadonnées* :
 - *Différence de la date de création de cube*. Elle apparaît lorsque deux cubes ont été créés à deux dates différentes.

2. Hétérogénéité Mesure-à-Mesure

- *Hétérogénéité au niveau du schéma* :
 - *Différence de primitives géométriques*. Elle apparaît lorsque deux mesures géométriques sémantiquement liées ont différentes primitives géométriques (ex. ligne vs polygone).
- *Hétérogénéité au niveau des métadonnées* :
 - *Hétérogénéité au niveau de l'agrégation*. Apparaît lorsque différentes fonctions ont été utilisées pour agréger des mesures spatiales sémantiquement liées. Par exemple, la fonction utilisée pour l'agrégation d'une mesure est l'union géométrique (i.e. la distribution du niveau supérieur est calculée en fonction de l'union des distributions des niveaux inférieurs), alors que la fonction utilisée pour l'agrégation d'une autre est le centre de gravité (i.e. la *Zone-affectée* du niveau supérieur est calculée en fonction du centre de gravité des zones des niveaux inférieurs).

3. Hétérogénéité Dimension-à-Dimension

- *Hétérogénéité au niveau du schéma* :
 - *Inégalité du nombre de hiérarchies*. Elle apparaît quand les cardinalités des dimensions sémantiquement liées sont différentes. Plus précisément, supposons que n_1 est la cardinalité de dimension D_1 , n_2 est la cardinalité de dimension D_2 . si $n_1 \neq n_2$ alors il y a une hétérogénéité Dimension-à-Dimension au niveau schéma. Dans notre exemple, la dimension *Région Administrative* contient une seule hiérarchie : (*Pays*, *Province*, et *Ville*), alors que la dimension *Région* contient deux hiérarchies : (*Pays*, *Province*, *Territoire*, et *Cité*) et (*Pays*, *État*, et *Cité*).
- *Hétérogénéité au niveau des métadonnées* :
 - *Non-correspondance des contraintes de dimensions*. Elle apparaît quand des contraintes de deux dimensions⁴ sémantiquement liées sont

⁴ Hurtado et al. (2005) ont introduit la notion de contrainte de dimension.

incohérentes. Par exemple, la contrainte de la dimension *Région Administrative* indique que tous les membres du niveau *Ville* sont remontés spatialement au niveau *Province*, tandis que la contrainte de la dimension *Région* indique que les membres du niveau *Cité* sont remontés spatialement soit au niveau *Province*, au niveau *Territoire*, ou au niveau *État*.

4. Hétérogénéité Hiérarchie-à-Hiérarchie

- *Hétérogénéité au niveau du schéma* :
 - *Inégalité du nombre de niveaux*. Elle apparaît quand les cardinalités des hiérarchies de dimensions qui sont sémantiquement liées sont différentes. Plus précisément, supposons que n_1 est la cardinalité de la hiérarchie H_1 , n_2 est la cardinalité de la hiérarchie H_2 . si $n_1 \neq n_2$ alors il y a une hétérogénéité Hiérarchie-à-Hiérarchie au niveau du schéma. Par exemple, la hiérarchie spatiale (*Pays*, *Province*, *Territoire* et *Ville*) de la dimension *Région Administrative* contient quatre niveaux, alors que la hiérarchie spatiale (*Pays*, *État*, et *Cité*) de la dimension *Région* contient trois niveaux.
 - *Inégalité d'ordre de niveau*. Elle apparaît quand des hiérarchies de dimensions qui sont sémantiquement liées ont différents ordres de niveaux. Plus précisément, pour chaque combinaison de couples des niveaux qui sont sémantiquement liés $((n_1, n_2), (n_1', n_2'))$, si $n_1 < n_2$ et $\neg (n_1' < n_2')$, alors il y a une hétérogénéité Hiérarchie-à-Hiérarchie au niveau du schéma. Par exemple, dans un cube de données, l'ordre des niveaux d'une hiérarchie spatiale peut être comme suit : *Ville*, *Comté*, et *Province*. Alors que, dans un autre cube, les mêmes niveaux sont ordonnés comme suit : *Ville*, *Comté*, et *Province* d'un côté et *Ville* et *Province* de l'autre côté.
- *Hétérogénéité au niveau des métadonnées* :
 - *Hétérogénéité au niveau de la couverture spatiale*. Elle apparaît quand les membres des hiérarchies de deux dimensions sémantiquement liées ont des couvertures territoriales différentes. Par exemple, la hiérarchie de la dimension *Région Administrative* du cube de données spatiales *C1* (*Pays*, *Province*, *Territoire* et *Ville*) couvre le Canada, tandis que dans le deuxième cube de données *C2*, la hiérarchie de la dimension *Région* (*Pays*, *Province*, *Territoire* et *Cité*) couvre le centre du Canada et le nord-ouest des États-Unis.
 - *Hétérogénéité au niveau de la méthode de peuplement des niveaux des hiérarchies*. Elle apparaît quand les niveaux de deux hiérarchies appartenant à deux dimensions sémantiquement liées sont peuplés en utilisant deux méthodes différentes. Par exemple, les niveaux de la hiérarchie (*Pays*, *Province*, *Territoire* et *Ville*) de la dimension *Région Administrative* du cube de données spatiales *C1* sont peuplés en utilisant une saisie initiale de données en mode « représentation multiple ». Alors que, dans le deuxième cube de données *C2*, les niveaux de la hiérarchie (*Pays*, *Province*, *Territoire* et *Cité*) de la dimension *Région* sont peuplés en utilisant la généralisation cartographique.

5. Hétérogénéité Niveau-à-Niveau

- *Hétérogénéité au niveau du schéma* :
 - *Inégalité du nombre d'attributs*. Elle apparaît quand les cardinalités des niveaux de hiérarchies qui sont sémantiquement liées sont différents. Plus précisément, supposons que n_1 est la cardinalité du niveau N_1 , n_2 est la cardinalité du niveau N_2 , si $n_1 \neq n_2$ alors il y a une hétérogénéité Niveau-à-Niveau au niveau du schéma. Par exemple, le niveau spatial ville dans *C1* est décrit en utilisant un seul attribut : `nom_ville`, alors que le niveau cité dans *C2* est décrit en utilisant deux attributs : `nom_cité` et `superficie_cité`.
 - *Différence de primitives géométriques*. Elle apparaît lorsque, dans deux cubes de données spatiales, des niveaux sémantiquement liés ont différentes primitives géométriques. Par exemple, dans le cube de données *C1*, chaque membre du niveau *Ville* est représenté par un point, alors que dans le cube de données *C2*, chaque membre du niveau *Cité* est représenté par un polygone.
- *Hétérogénéité au niveau des métadonnées* :
 - *Différence de systèmes de référence spatiale*. Il y a différents systèmes de référence spatiale qui peuvent être utilisés pour déterminer la position des objets spatiaux (ex. le système de coordonnées ellipsoïdal global (latitude-longitude-hauteur standard) et le système de coordonnées x,y,z). Dans notre exemple, les niveaux de la dimension *Région Administrative* du cube *C1* sont basés sur le système *North American Datum of 1927 (NAD 27)*, tandis que ceux de la dimension *Région* du cube *C2* sont basés sur le système *North American Datum of 1983 (NAD 83)*.
 - *Hétérogénéité au niveau de la résolution de représentation*. Elle apparaît lorsque des niveaux appartenant à des dimensions sémantiquement liées sont représentés avec un niveau de détail géométrique correspondant traditionnellement à différentes échelles cartographiques. Dans notre exemple, la résolution de représentation des détails cartographiques du niveau *Province* du cube de données *C1* est typique d'une carte topographique à l'échelle 1:50000, alors que la résolution géométrique du niveau *Province* du cube de données *C2* équivaut aux détails d'une carte topographique 1:100000.
 - *Hétérogénéité au niveau de la précision*. Elle apparaît lorsque des niveaux appartenant à des dimensions sémantiquement liées des cubes de données sont représentés par des géométries localisées avec des précisions différentes. Dans notre exemple, les niveaux *Ville* de *C1* et *Cité* de *C2* dans les deux cubes n'ont pas la même précision (précision de 1 m dans le cube de données *C1* et précision de 2 m dans le cube de données *C2*), ce qui entraîne inévitablement des problèmes de chevauchement de frontières.
 - *Hétérogénéité temporelle*. Elle apparaît lorsque des membres des niveaux appartenant à des dimensions sémantiquement liées des cubes de données ont été collectés à des périodes différentes. Par exemple, la figure 3 montre que les *municipalités* d'une même couverture spatiale sont représentées différemment suite à la fusion administrative des deux villes A et C en 2006.

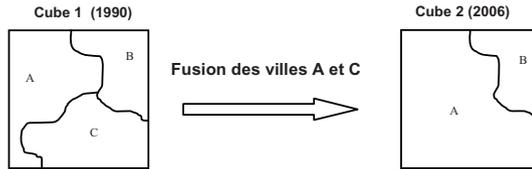


FIG. 3 – Hétérogénéité temporelle : fusion des municipalités A et C en 2006

Il faut noter ici que les hétérogénéités au niveau des schémas sont de type structurel alors que les hétérogénéités au niveau des métadonnées ne font qu'indiquer des problèmes affectant directement le contenu, i.e. les données elles-mêmes. Dans le contexte spatial, l'identification des problèmes liés aux métadonnées peut être très difficile car ces problèmes peuvent aussi concerner des informations subjectives qui n'ont pas été mentionnées explicitement. Seuls quelques exemples ont été mentionnés ci-avant mais les sources d'hétérogénéité peuvent se compter par plusieurs centaines : systèmes de référence spatiale, méthodes de positionnement dans ces systèmes, instrumentations d'acquisition, méthodes d'observations, algorithmes de pré-traitement, structures propriétaires des logiciels traitant des données géométriques, structuration applicative des bases de données spatiales, algorithmes de transformation et de traitement des données, modes de représentation des données, généralisation cartographique, etc.

Dans la section suivante nous proposons une approche pour aider à traiter les problèmes d'hétérogénéité.

5 Une approche d'aide à l'intégration de modèles de cubes de données spatiales

Dans certains cas, les problèmes d'hétérogénéité au niveau des modèles peuvent être faciles à résoudre grâce aux logiciels spécialisés en géomatique (ex. changement de datum de référence). Dans d'autres cas, l'hétérogénéité est très difficile, voire même impossible à résoudre automatiquement, et fait appel à tout un domaine d'expertises spécialisées, soit la géomatique (incluant la géodésie et le GPS, la photogrammétrie, la télédétection, la topométrie, l'hydrographie, la cartographie et les SIG (Systèmes d'Information Géographique)). Le but de notre travail est de proposer des outils et informations utiles pour aider l'intervenant à prendre la décision d'intégrer les modèles des cubes de données spatiales. Pour cela nous proposons 1) un cadre général qui peut être utilisé par les intervenants humains et 2) une approche qui permet de définir et évaluer un ensemble d'indicateurs qui montrent la qualité des modèles et des métadonnées à intégrer. Ces indicateurs sont présentés d'une façon intuitive à l'intervenant pour l'aider à prendre une décision à propos de l'intégration des cubes de données spatiales.

5.1 Cadre général pour résoudre les problèmes d'intégration de modèles de cubes de données spatiales

Le cadre général proposé dans ce travail vise à guider l'intervenant pour l'identification des problèmes d'hétérogénéité liés aux modèles et pour la prise de décisions vis-à-vis de ces problèmes.

Le cadre est composé principalement de cinq phases successives d'identification et de résolution des problèmes d'hétérogénéité. Ces phases correspondent à cinq niveaux conceptuels différents allant du niveau le plus général au niveau le plus détaillé : niveau *cubes*, niveau *mesures*, niveau *dimensions*, niveau *hiérarchies* et niveau *niveaux*. À chacun de ces niveaux, l'intervenant analyse les métadonnées et les structures pour tirer une conclusion sur le niveau d'hétérogénéité de chacun de ces deux aspects. L'intervenant commence par examiner le cube dans son ensemble, puis il continue en allant plus en détails dans les autres niveaux (approche descendante). À l'issue de chaque phase, l'intervenant peut prendre l'une des décisions suivantes (voir figure 4) :

- Suspendre l'intégration des cubes de données spatiales dont les problèmes liés aux modèles présentent un risque élevé qui peut conduire à des conséquences néfastes pour l'intégration. Dans ce cas, l'intervenant n'a pas besoin d'identifier et traiter les problèmes d'hétérogénéité dans les niveaux restants.
- Poursuivre l'intégration des cubes de données spatiales s'il n'y a pas de problèmes d'intégration (ex. les éléments ne sont pas sémantiquement liés) ou si les problèmes liés aux modèles ne présentent pas un risque élevé. Deux décisions peuvent alors être prises :
 - Résoudre les problèmes d'hétérogénéité. Par exemple, résoudre les problèmes de différence entre des systèmes de référence spatiale en choisissant un système de référence commun.
 - Endurer les problèmes d'hétérogénéité de métadonnées dans le cas où ils ne risquent pas d'affecter de façon jugée significative l'utilisation des données spatiales pour des besoins d'analyse. Par exemple, une différence de précision de 1 mètre pour une intégration de données spatiales qui seront utilisées pour une application touristique.

À l'issue de chaque phase, selon les décisions prises (suspendre l'intégration, résoudre ou endurer les problèmes), l'intervenant devrait rédiger un rapport expliquant (1) les raisons de la suspension, (2) comment le problème a été résolu et les avertissements signalés, ou (3) les raisons pour lesquelles l'intervenant endure les problèmes.

L'identification et le traitement de ces problèmes sont effectués selon une démarche hiérarchique descendante qui présente deux avantages :

- La démarche permet à l'intervenant de prendre des décisions pertinentes à un stade précoce du processus de l'intégration afin de gérer les risques d'intégration des cubes de données (Machlis et Rosa, 1990; Morgan, 1990; Renn, 1998). Cela permet un gain considérable d'efficacité et de temps. En effet, à l'issue de chaque phase, l'intervenant peut suspendre l'intégration avant même d'entrer trop dans les détails et réduire ainsi ses coûts d'analyse et d'intégration. Il peut également poursuivre l'intégration en tenant compte de ses observations au niveau général pour mieux agir au niveau détaillé.

- Cette démarche correspond bien au modèle mental de l'intervenant humain (Yougworth, 1995; Bédard et al., 2001; Rivest et al., 2005). En effet, le cadre est basé sur une structure hiérarchique qui est un des principes essentiels de la cognition humaine (Edwards, 2001; Marchand et al., 2004). Ce principe stipule que les humains regroupent les données et les métadonnées dans des catégories selon leur propre connaissance (Mennis et al., 2000). Ces catégories sont organisées en hiérarchies afin de permettre la réutilisation maximale des données avec le minimum d'effort (Rosch, 1978).

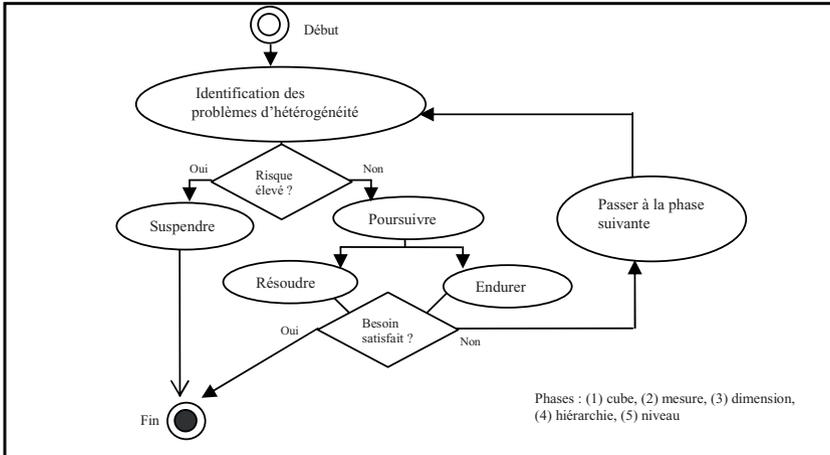


FIG. 4 – Cadre général d'intégration des cubes de données spatiales.

La figure 4 illustre la démarche proposée. L'intervenant commence par identifier les problèmes d'hétérogénéité au niveau le plus général (i.e., niveau *cube*) afin de prendre des décisions pour résoudre ces problèmes. Par exemple, identifiant une différence de systèmes de référence spatiale, l'intervenant peut proposer l'utilisation d'un système de référence commun pour les deux cubes de données. Ensuite, l'intervenant identifie les conflits d'hétérogénéité au niveau *mesure* et traite ces conflits. Par exemple, l'intervenant peut identifier un conflit potentiel lié à la différence des méthodes d'agrégation (ex. union géométrique versus centre de gravité), et décider d'utiliser une seule de ces méthodes pour intégrer les cubes. Après cela, l'intervenant identifie les problèmes d'hétérogénéité au niveau *dimension* et prend des décisions pour résoudre ces problèmes. Par exemple, l'intervenant peut décider d'abandonner l'intégration des cubes s'il identifie un conflit lié à l'hétérogénéité au niveau de la couverture spatiale (ex. cube couvrant la province de Québec versus cube couvrant la province de Manitoba). Il faut noter qu'il est important de traiter les problèmes d'hétérogénéité du niveau *mesure* avant ceux du niveau *dimension* puisque les premiers sont le sujet d'analyse des cubes tandis que les derniers représentent le contexte de cette analyse. Ensuite, l'intervenant identifie les problèmes d'hétérogénéité au niveau hiérarchie et prend des décisions pour résoudre ces problèmes. Par exemple, l'intervenant peut décider d'abandonner l'intégration des cubes lorsque deux dimensions sémantiquement liées

provenant de cubes différents présentent un nombre de niveaux différents et incompatibles. Finalement, l'intervenant identifie les problèmes d'hétérogénéité au niveau *niveau*, et prend une décision vis-à-vis de ces problèmes. Par exemple, l'intervenant peut décider d'ignorer l'hétérogénéité des échelles de représentation cartographique (ex. 1:125000 et 1:120000).

5.2 Définition d'indicateurs pour aider à résoudre les problèmes d'intégration de modèles de cubes de données spatiales

Dans cette section, nous proposons une méthode pour définir et évaluer des indicateurs de qualité des éléments sémantiquement liés de modèles (schéma et métadonnées) de cubes à intégrer en se basant sur les besoins de l'utilisateur final, eux aussi exprimés sous forme de modèle de cube (voir figure 5). Ces indicateurs ont deux objectifs principaux :

1. Premièrement, aider l'intervenant à prendre une décision appropriée à chaque phase (suspendre l'intégration, résoudre ou endurer les problèmes). Par exemple, si les deux éléments hétérogènes sont cruciaux pour l'intégration mais qu'ils ont une qualité médiocre, l'intervenant peut être conseillé de suspendre l'intégration des deux éléments.
2. Deuxièmement, aider l'intervenant à résoudre les problèmes à chaque niveau du cadre proposé (cube-à-cube, mesure-à-mesure, dimension-à-dimension, hiérarchie-à-hiérarchie, niveau-à-niveau). En effet, il s'agit de fournir des informations à propos de la qualité qui peuvent être utiles à la prise de décision. Par exemple, en se basant sur de telles informations, l'intervenant peut être conseillé de :
 - (a) ne pas utiliser un des éléments hétérogènes (celui qui a une qualité inférieure par rapport à l'autre).
 - (b) considérer un élément qui a une excellente qualité en totalité.
 - (c) se servir d'un élément qui a une qualité assez bonne pour en créer un nouveau.
 - (d) ne pas considérer les deux éléments s'ils ont tous les deux une qualité médiocre ou une qualité risquant de produire un résultat médiocre.

Nous distinguons deux catégories d'indicateurs : une catégorie qui contient ceux qui sont liés au schéma et une autre qui contient ceux qui sont liés aux métadonnées. Pour chaque indicateur, nous associons une valeur appartenant à l'intervalle $[0, 1]$ où 1 indique une qualité parfaite tandis que 0 indique une qualité très médiocre. Cette valeur est définie de manière pragmatique et mathématique tel qu'expliqué dans les prochains paragraphes. Cette valeur représente en fait une échelle de mesure ordinale malgré l'utilisation de valeurs quantitatives. Conséquemment, nous pouvons y appliquer les opérateurs $=$, $>$, $<$ mais pas les autres, sauf de façon indicatrice et en connaissance de cause. En d'autres termes, nous pouvons dire qu'une qualité 0,8 est meilleure qu'une qualité 0,4, mais nous ne pouvons pas dire qu'elle est précisément deux fois meilleure, malgré que nous puissions dire qu'elle est *passablement* meilleure (i.e. une comparaison qualitative inspirée d'un calcul quantitatif dont on connaît la nature indicatrice).

Pour mieux illustrer notre méthode, nous proposons donc d'utiliser un modèle pour exprimer les besoins de l'utilisateur final comme celui de la figure 5.

Approche basée sur la qualité pour l'intégration de modèles de cubes

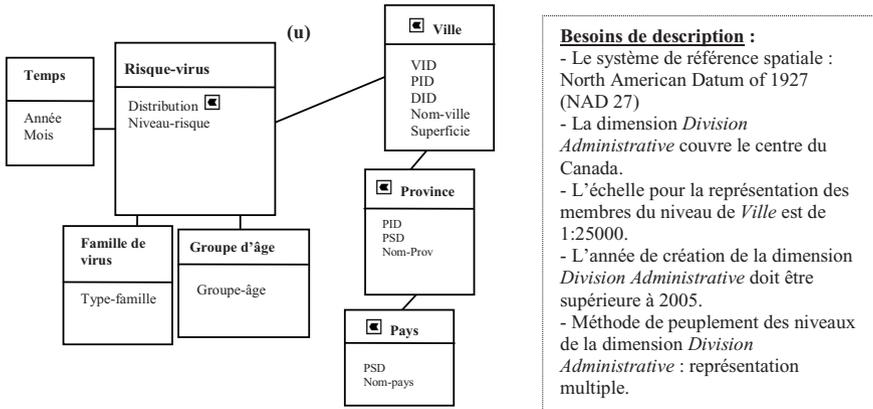


FIG. 5 – Modèle décrivant les besoins de l'utilisateur final.

5.2.1 Indicateurs de qualité de schéma

Afin d'évaluer la qualité des schémas, trois indicateurs ont été définis : pertinence de la primitive géométrique, pertinence de la structure et pertinence de l'ordre de la hiérarchie.

1- *Pertinence de la primitive géométrique* : cet indicateur indique la pertinence de la primitive géométrique utilisée pour représenter un élément du modèle de cube par rapport à la primitive géométrique requise par l'utilisateur final. Cet indicateur peut être évalué en utilisant un tableau qui prédéfinit la correspondance entre les deux primitives. Le tableau 1 représente un exemple de correspondance défini. Les valeurs de ce tableau ont été définies en se basant aussi bien sur les règles de généralisation cartographique reconnues que sur la pratique des cartographes. Ainsi nous avons accordé les valeurs 0, 0.25, 0.5, 0.75, et 1 respectivement lorsque 1) les primitives du modèle ne sont pas suffisantes pour généraliser les primitives requises (ex. un point ne doit pas être utilisé pour généraliser une ligne), 2) les primitives du modèle peuvent être utilisées pour généraliser les primitives requises, cette utilisation n'est cependant pas conseillée car elle ne garantit pas une représentation fidèle de la variété (ex. un point (0D) et une ligne (3D)), 3) les primitives du modèle peuvent être utilisées pour généraliser les primitives requises, mais le résultat de généralisation ne reflétera pas bien la réalité (ex. un point (0D) et une ligne (1D)), 4) les primitives du modèle sont suffisantes pour généraliser les primitives requises, ce résultat reflétera assez bien la réalité (ex. une ligne (1D) et un point (0D)), et 5) les primitives du modèle sont semblables à celles des besoins (ex. une ligne est pertinente pour représenter une ligne).

Cet indicateur est évalué pour chacun des niveaux suivants : mesure-à-mesure et niveau-à-niveau puisque ce sont les niveaux qui peuvent contenir des primitives géométriques dans le schéma.

		Éléments du modèle des besoins			
		0D	1D	2D	3D
Élément des modèles à intégrer	0D	1	0	0.5	0.75
	1D	0	1	0	0
	2D	0.75	0	1	0.5
	3D	0.5	0	0.75	1

TAB. 1 – Évaluation de la pertinence de la primitive géométrique.

2- *Pertinence de structure* : cet indicateur indique la pertinence d'un élément du schéma du modèle par rapport à un élément de schéma requis par l'utilisateur final. L'indicateur est évalué pour chacun des éléments qui sont sémantiquement liés (selon la méthode choisie par l'intervenant) de différents cubes de données spatiales. Il indique le ratio des composants de l'élément du modèle qui sont sémantiquement liés avec celles de l'élément requis. La pertinence de structure P_s est calculée en utilisant la formule suivante :

$$P_s(E, R) = \begin{cases} \frac{N_{CE}}{N_{CR}} ; & \text{si } N_{CE} \leq N_{CR} \\ 1 ; & \text{sinon} \end{cases} \quad (1)$$

Où N_{CR} est le nombre total de composants requis par l'utilisateur, et N_{CE} est le nombre de composants de l'élément qui sont sémantiquement liés avec les composants requis.

3- *Pertinence de l'ordre de la hiérarchie* : cet indicateur est évalué seulement pour le niveau hiérarchie-à-hiérarchie. Il indique la pertinence de l'ordre de chacune des hiérarchies hétérogènes par rapport à la structure de la hiérarchie requise par l'utilisateur final.

Supposons que n_1 et n_2 sont deux niveaux appartenant à une hiérarchie H d'un cube donné et n_1' et n_2' sont deux niveaux appartenant à une hiérarchie B d'un cube exprimant les besoins de l'intégration. La pertinence de l'ordre (O_s) est la moyenne de toutes les pertinences élémentaires des paires de niveaux qui appartiennent à la hiérarchie de chacun des cubes à intégrer (soit o_s). La pertinence élémentaire o_s est déterminée à l'aide de l'expression suivante :

$$\forall n_1, n_2 \in H, \forall n_1', n_2' \in B : [sem_lié(n_1, n_1') \wedge sem_lié(n_2, n_2') \wedge n_1 < n_2] \Rightarrow n_1' < n_2' \quad (2)$$

Où *sem_lié* est une fonction qui vérifie si les niveaux sont sémantiquement liés ou non. La valeur 1 (ou 0) est assignée à la pertinence élémentaire d'ordre (o_s) entre chaque paire de niveaux lorsque l'expression (2) est vraie (ou fausse).

Ainsi, si par exemple la qualité de la structure d'un élément 1 est plus grande que celle d'un élément 2, on peut dire que la structure de l'élément 1 est plus pertinente que celle de l'élément 2. Par conséquent, l'intervenant peut être conseillé de ne pas considérer la structure de l'élément 2.

5.2.2 Indicateurs de qualité externe de métadonnées

La qualité externe (*fitness for use*) de métadonnées correspond à la pertinence d'utiliser les métadonnées dans une application spécifique. Nous n'avons trouvé aucun travail de recherche sur la qualité externe des métadonnées spatiales ni pour les bases de données transactionnelles ni pour les cubes de données spatiales. Dans cette section, nous proposons un ensemble d'indicateurs et une approche quantitative pour évaluer la qualité externe des métadonnées de cubes de données spatiales. Les indicateurs proposés sont : la pertinence des métadonnées par rapport aux besoins et l'actualité des métadonnées.

1- *Pertinence des métadonnées.* Cet indicateur indique le degré de pertinence des métadonnées par rapport aux besoins de l'utilisateur final. La pertinence peut être évaluée à différents niveaux : thématique, spatial et temporel. L'indicateur est estimé en se basant sur le ratio du nombre des éléments de métadonnées qui sont sémantiquement liés aux éléments requis par rapport au nombre total d'éléments requis. La pertinence des métadonnées P_m est évaluée selon la formule suivante :

$$P_m = \begin{cases} w \times \frac{N_{Elt}}{N_{ReqEltm}} & ; \text{si } N_{Elt} \leq N_{ReqEltm} \\ 1 & ; \text{sinon} \end{cases} \quad (3)$$

Où N_{Elt} est le nombre d'éléments thématiques, spatiaux, ou temporels de métadonnées sémantiquement liés aux éléments requis par l'utilisateur, $N_{ReqEltm}$ est le nombre total d'éléments de métadonnées requis, et w une valeur prédéfinie entre 0 et 1 qui indique le niveau d'importance de chaque type (i.e., thématique, spatial, et temporel) pour l'utilisateur final. Si N_{Elt} est égal ou plus grand que $N_{ReqEltm}$, les métadonnées sont parfaitement pertinentes et sa valeur est mise à 1.

2- *Actualité des métadonnées.* Cet indicateur indique le degré d'actualité des métadonnées des cubes de données spatiales sources. Il est évalué en fonction de l'âge des métadonnées par rapport à leur durée de vie. L'âge des métadonnées est le temps écoulé depuis la date de définition des métadonnées (T_{def}) jusqu'à la date d'actualité désirée des cubes (T_{requis}). La durée de vie est le nombre d'années après lequel les métadonnées ne seront plus valides. L'actualité des métadonnées A_m est évaluée en se basant sur la formule suivante :

$$A_m = \begin{cases} 1 - \frac{|T_{requis} - T_{def}|}{DV} & ; \text{si } |T_{requis} - T_{def}| < DV \\ 0 & ; \text{sinon} \end{cases} \quad (4)$$

Où DV est la durée de vie des métadonnées. Une faible valeur de l'actualité diminue la qualité externe des métadonnées. Cette valeur diminue lorsque son âge augmente. La durée de vie et la date de définition des métadonnées peuvent être fournies par le producteur de métadonnées.

Ce qu'il faut rappeler ici pour l'ensemble des indicateurs ci-avant, c'est qu'ils ne visent pas à être exhaustifs ou précis mais plutôt à orienter l'intervenant dans sa prise de décision à propos de l'intégration des modèles de cubes données spatiales. Une telle méthode est fréquemment utilisée dans plusieurs domaines impliquant des facteurs qui sont difficiles, voire même impossible, à évaluer d'une façon exhaustive et précise tels que les domaines de l'épidémiologie, de l'écologie, des sciences économiques, etc. Une telle méthode a été utilisée pour évaluer la qualité de données spatiales par Devillers et al. (2007).

La façon de présenter ces indicateurs à l'intervenant a une grande importance sur la prise de décision. Généralement, les applications d'aide à la décision utilisent un nombre restreint d'indicateurs (Few, 2006). En ce sens, nous suggérons de présenter seulement deux indicateurs à l'intervenant à chaque niveau du cadre proposé : un pour indiquer la qualité du schéma et l'autre pour indiquer la qualité des métadonnées. Ces valeurs sont calculées comme suit :

$$Q_s = \frac{\sum (a \times I_s)}{n} \quad (5)$$

$$Q_m = \frac{\sum (b \times I_m)}{m} \quad (6)$$

Où a et b sont des valeurs prédéfinies entre 0 et 1 qui indiquent l'importance accordée à chaque indicateur de qualité de schéma et de métadonnées respectivement. Les variables n et m sont les nombres d'indicateurs au niveau du schéma et des métadonnées respectivement.

5.2.3 Représentation symbolique des indicateurs de qualité

Les valeurs de ces indicateurs peuvent être représentées en utilisant différents symboles d'avertissement afin d'aider l'intervenant à comprendre le résultat de l'évaluation de ces indicateurs à chaque niveau et ainsi à prendre les décisions appropriées d'une façon intuitive. Nous utilisons un ensemble de symboles d'avertissements montrant la qualité des éléments du modèle des cubes à intégrer. Ces symboles (Danger, Warning, Caution et Notice) sont basés sur la norme ANSI Z535.6 (2006) et utilisés auparavant dans le domaine spatial par Levesque et al. (2007). Ces symboles sont enrichis pour répondre aux besoins d'intégration des cubes de données spatiales. Le tableau 2 montre un exemple de la façon dont des symboles peuvent être prédéfinis en se basant sur les valeurs quantitatives de la qualité des métadonnées. Par exemple, le symbole  indique à l'intervenant la qualité médiocre des dimensions candidates à l'intégration. Ainsi, l'intervenant pourra prendre la décision d'annuler cette intégration qui représente un risque considérable. En procédant ainsi, nous revenons donc à une échelle de mesures ordinale et évitons d'attribuer une valeur trop quantitative aux résultats.

Qualité de modèle	Représentation des indicateurs de qualité	Signification du symbole
$Q = 0$		L'intervenant est conseillé d'arrêter le processus d'intégration.
$0 < Q \leq 0.25$		Il y a un risque élevé si l'intervenant décide de continuer l'intégration.
$0.25 < Q \leq 0.5$		L'intervenant est averti de l'existence de risques potentiels.
$0.5 < Q \leq 0.75$		L'intervenant est invité à faire attention s'il décide de continuer l'intégration.
$0.75 < Q < 1$		Une information sera affichée à l'intervenant.
$Q = 1$		L'intervenant est invité à continuer à un niveau plus détaillé.

TAB. 2 – Une définition des symboles selon la valeur de qualité.

À la liste de symboles présentée ci-dessus, on ajoute : (1) le symbole  qui invite l'intervenant à continuer l'évaluation d'autres d'éléments sémantiquement liés et le symbole  qui indique une absence de métadonnées des éléments sémantiquement liés, et ainsi l'impossibilité d'évaluer la qualité.

5.3 Exemple d'application

Dans les points suivants, nous présentons une évaluation des indicateurs à chacun des niveaux (cube-à-cube, mesure-à-mesure, dimension-à-dimension, hiérarchie-à-hiérarchie, niveau-à-niveau) pour l'exemple déjà présenté (cf. figures 2 et 4). Pour simplifier le calcul, on accorde la valeur 1 pour a et b et cela pour tous les niveaux.

1. Hétérogénéité Cube-à-Cube :

Après avoir identifié les cubes pouvant potentiellement fournir de l'information comblant les besoins exprimés (grâce à la même approche légèrement adaptée et appliquée entre le *cube besoin* et les *cubes sources potentiels*), les cubes sources retenus sont analysés. Lorsque deux cubes sources retenus sont sémantiquement liés (ex. par leur objectif, leur nom ou leurs mesures contextualisées, i.e. en tenant compte des dimensions), alors leur structure et leurs métadonnées sont analysées, deux indicateurs d'hétérogénéité sont produits et une décision est prise quant à la poursuite de l'intégration.

Par contre, comme c'est le cas dans l'exemple présenté dans cet article, si les cubes ou les mesures ne semblent à prime abord pas traiter du même sujet, il est tout de même possible que des éléments de ces cubes (qui rempliraient partiellement les besoins exprimés) puissent être intégrables. Alors, le symbole  sera affiché.

2. Hétérogénéité Mesure-à-Mesure :

Lorsque deux mesures sont sémantiquement liées (ex. par leur nom, leur définition, leur unité de mesure, leur type de donnée, leur nature descriptive ou géométrique)

indépendamment de leur contexte (i.e. sans tenir compte des dimensions), alors leur structure (ex. type de donnée, domaine de valeur, primitive géométrique si mesure spatiale) et leurs métadonnées sont analysées, deux indicateurs d'hétérogénéité sont produits et une décision est prise quant à la poursuite de l'intégration.

Par contre, comme c'est le cas dans l'exemple présenté dans cet article, les mesures des deux cubes ne sont pas sémantiquement liées. Par contre, il est tout de même possible que d'autres éléments de ces cubes (qui rempliraient partiellement les besoins exprimés) puissent être intégrables. Alors, l'intervenant est invité à continuer au niveau suivant (soit le niveau dimension-à-dimension) et le symbole  sera affiché.

3. Hétérogénéité Dimension-à-Dimension :

Si les dimensions des deux cubes ne sont pas sémantiquement liées, l'intervenant ne devra pas aller plus en détails pour ces dimensions, le symbole  sera affiché et l'intervenant pourra passer aux autres combinaisons possibles de dimensions. C'est le cas pour les dimensions *Groupe d'âge* du cube *C1* et *Région* du cube *C2*.

S'il y a un lien sémantique entre deux dimensions, comme dans notre exemple avec les dimensions *Région Administrative* du cube *C1* et *Région* du cube *C2*, alors on évalue l'hétérogénéité entre ces deux dimensions tant pour leur structure que pour leurs métadonnées.

- Hétérogénéité au niveau du schéma :

➤ Pertinence du nombre de hiérarchies

La dimension *Région Administrative* du cube *C1* contient une seule hiérarchie : (*Pays, Province, Territoire* et *Ville*), alors que la dimension *Région* du cube *C2* contient deux hiérarchies : (*Pays, Province, Territoire*, et *Cité*) et (*Pays, État*, et *Cité*). Étant donné que l'utilisateur a besoin d'une seule hiérarchie (*Pays, Province* et *Ville*) puisqu'il ne couvre que le Canada, l'indicateur de qualité « Pertinence de structure » (« Pertinence du nombre de hiérarchies ») est évalué comme suit selon la formule (1):

Pour *C1*, $P_s(\text{Région Administrative, Division Administrative}) = 1/1 = 1$

Pour *C2*, puisque le nombre de hiérarchies de la dimension *Région* est plus grand que celui de la dimension sémantiquement liée du cube de besoins *C3* (*Division Administrative*), donc :

$$P_s(\text{Région, Division Administrative}) = 1$$

Selon la formule (5) la qualité au niveau du schéma :

$$Q_s(\text{Région Administrative, Division Administrative}) = 1$$

$$Q_s(\text{Région, Division Administrative}) = 1$$

Donc, le symbole  sera affiché pour les deux dimensions *Région Administrative* et *Région*. Dans les étapes suivantes (Hiérarchie-à-Hiérarchie, Niveau-à-Niveau), cette analyse sera raffinée pour finalement conduire à un choix parmi les possibilités d'intégration des dimensions sources visant à combler les besoins de *C3*. Les deux schémas de *C1* et *C2* peuvent alors être pris en compte.

– Hétérogénéité au niveau des métadonnées :

➤ Pertinence des métadonnées

Les métadonnées associées à la dimension *Région Administrative* de *C1* contiennent deux éléments qui sont sémantiquement liés aux éléments requis par l'utilisateur (la couverture spatiale de la dimension et l'année de création de la dimension). En revanche, les métadonnées associées à la dimension *Région* de *C2* contiennent un seul élément qui est sémantiquement lié à un élément requis par l'utilisateur (la couverture spatiale de la dimension). Si on considère que le niveau d'importance de l'information spatiale est = 1, alors selon la formule (3) :

$$P_m(\text{Région administrative, Division Administrative}) = 2/2 = 1$$

$$P_m(\text{Région, Division Administrative}) = 1/2 = 0.5.$$

➤ Actualité des métadonnées

Les métadonnées des dimensions *Région Administrative* de *C1* et *Région* de *C2* ont été créées respectivement en 2002 et 1982. De plus, ces deux dimensions ont la même durée de vie, soit 30 ans. Par conséquent selon la formule (4):

$$A_m(\text{Région Administrative, Division Administrative}) = 1 - (2005-2002/30) = 0.9$$

$$A_m(\text{Région, Division Administrative}) = 1 - (2005-1982/30) = 0.23$$

Ainsi, au niveau des métadonnées selon la formule (6) :

$$Q_m(\text{Région Administrative, Division Administrative}) = (1 + 0.9)/2 = 0.95$$

$$Q_m(\text{Région, Division Administrative}) = (1 + 0.23)/2 = 0.61$$

Donc, le symbole  sera affiché pour la dimension *Région Administrative*. Par contre, le symbole  sera affiché pour la dimension *Région*. L'intervenant est alors invité à faire attention en considérant la dimension *Région* dans le processus d'intégration. De plus, l'intervenant est invité à évaluer la qualité des niveaux les plus détaillés de ces deux dimensions (i.e., hiérarchies et niveaux) en tenant compte de cette différence de qualité. Il faut noter ici que, si la qualité de l'une des deux dimensions avait été médiocre, alors le symbole  aurait été affiché et l'intervenant aurait été conseillé de ne plus considérer les niveaux les plus détaillés de cette dimension.

4. Hétérogénéité Hiérarchie-à-Hiérarchie :

Si les hiérarchies des deux dimensions ne sont pas sémantiquement liées, l'intervenant ne devra pas aller plus en détail pour ces hiérarchies, le symbole  sera affiché et l'intervenant pourra passer aux autres combinaisons possibles de hiérarchies des deux dimensions. Par contre, s'il y a un lien sémantique entre deux hiérarchies, comme dans notre exemple avec les hiérarchies (*H1* : *Ville, Province, Territoire et Pays*) de la dimension *Région Administrative* du cube *C1* et (*H2* : *Cité, Province, Territoire et Pays*) de la dimension *Région* du cube *C2*, alors on évalue l'hétérogénéité entre ces deux hiérarchies tant pour leur structure que pour leurs métadonnées.

– Hétérogénéité au niveau du schéma :

➤ Pertinence du nombre de niveaux

Chacune des deux hiérarchies (*H1*) et (*H2*) contient 4 niveaux. Étant donné que la hiérarchie (*H3* : *Ville*, *Province*, *Territoire* et *Pays*) de la dimension *Division Administrative* de *C3* contient 4 niveaux, selon la formule (1) on a :

$$P_s(H1, H3) = 4/4 = 1$$

$$P_s(H1, H3) = 4/4 = 1.$$

➤ *L'ordre de niveaux*

La hiérarchie de la dimension *Région Administrative* du cube *C1* a l'ordre suivant : (*Ville* < *Province*), (*Ville* < *Territoire*), (*Province* < *Pays*) et (*Territoire* < *Pays*). La hiérarchie de la dimension *Région* du cube *C2* a l'ordre suivant : (*Cité* < *Province*), (*Cité* < *Territoire*), (*Province* < *Pays*) et (*Territoire* < *Pays*). La hiérarchie de la dimension *Division Administrative* du cube *C3* a l'ordre suivant : (*Ville* < *Province*) et (*Province* < *Pays*).

Pour le cube *C1*, à l'aide de l'expression (2) :

Les niveaux *Ville*, *Province* (ou *Territoire*) et *Pays* (*C1*) sont, respectivement, sémantiquement liés aux niveaux *Ville*, *Province* et *Pays* (*C3*). De plus, les ordres *Ville* < *Province* (dans *C1*) et *Ville* < *Province* (dans *C3*), donc la pertinence élémentaire de l'ordre (*Ville* < *Province*) $o_s = 1$. Également, $o_s(\text{Ville} < \text{Territoire}) = o_s(\text{Province} < \text{Pays}) = o_s(\text{Territoire} < \text{Pays}) = 1$. Donc, la pertinence de la structure :

$$O_s(H1, H3) = (1+1+1+1)/4 = 1$$

De même, la pertinence de la hiérarchie (*Cité*, *Province*, *Territoire* et *Pays*) du cube *C2* est calculée à l'aide de l'expression (2) :

$$O_s(H2, H3) = (1+1+1+1)/4 = 1$$

Finalement, selon la formule (5) :

$$Q_s(H1, H3) = 1$$

$$Q_s(H2, H3) = 1$$

Donc, la qualité de schéma des deux hiérarchies est bien satisfaisante. En effet, le symbole  sera affiché et l'évaluation du niveau suivant continue.

– *Hétérogénéité au niveau des métadonnées :*

➤ *Pertinence des métadonnées*

Les métadonnées associées à la hiérarchie (*Ville*, *Province*, *Territoire* et *Pays*) du cube *C1* contiennent un élément qui est sémantiquement lié à l'élément requis par l'utilisateur (utilisation de la « représentation multiple » pour le peuplement des niveaux spatiaux). Par contre, les métadonnées associées à la hiérarchie (*Cité*, *Province*, *Territoire* et *Pays*) de *C2* ne contiennent aucun élément sémantiquement lié à l'élément requis par l'utilisateur. Si on considère que le niveau d'importance de l'information spatiale est = 1, alors selon la formule (3) :

$$P_m(H1, H3) = 1/1 = 1$$

$$P_m(H2, H3) = 0/1 = 0$$

Finalement, selon la formule (6) :

$$Q_m(H1, H3) = 1/1 = 1$$

$$Q_m(H2, H3) = 0/1 = 0$$

Donc, le symbole  sera affiché pour la hiérarchie (*Ville, Province, Territoire* et *Pays*). Le symbole  sera affiché pour la hiérarchie (*Cité, Province, Territoire,* et *Pays*). Donc, la considération de la deuxième hiérarchie risque d'endommager l'intégration des cubes de données spatiales. Ainsi, l'intervenant est invité à continuer à évaluer uniquement les niveaux de la première hiérarchie (*Ville, Province, Territoire* et *Pays*).

5. Hétérogénéité Niveau-à-Niveau :

– *Hétérogénéité au niveau du schéma* (Niveau *Ville*) :

➤ *Pertinence du nombre d'attributs.*

Le nombre d'attributs du niveau *Ville* du cube $CI = 1$ (nom-ville). Étant donné que l'utilisateur a besoin de 2 attributs pour ce niveau (non-ville et superficie), alors selon la formule (1):

$$P_s(Ville, Ville) = 1/2 = 0.5$$

➤ *Pertinence de la primitive géométrique :*

Dans le cube CI , chaque membre du niveau *Ville* est représenté par un point (0D), alors que l'utilisateur a besoin d'un polygone (2D) pour représenter les membres du niveau *Ville*. En se basant sur le tableau 1 :

$$P_p(Ville, Ville) = 0.5$$

Par conséquent, selon la formule (5), la qualité de la structure du niveau *Ville* :

$$Q_s(Ville, Ville) = (0.5+0.5)/2 = 0.5$$

Donc, la qualité est moyennement satisfaisante et le symbole  sera affiché à l'intervenant pour l'avertir de l'existence de risques en considérant ce niveau. En se basant sur le cadre proposé, l'utilisateur peut décider soit de résoudre les problèmes liés à ce niveau ou d'endurer les conséquences potentielles de ces problèmes.

– *Hétérogénéité au niveau des métadonnées* (Niveau *Ville*) :

➤ *Pertinence des métadonnées*

Les métadonnées du cube CI contiennent deux éléments qui sont sémantiquement liés aux éléments requis par l'utilisateur (le système de référence spatiale et l'échelle de représentation des membres). Si on considère que la valeur de l'importance de l'information spatiale est = 1, alors selon la formule (3) :

$$P_m(Ville, Ville) = 2/2 = 1$$

Selon la formule (6), la qualité des métadonnées du niveau *Ville* :

$$Q_m(Ville, Ville) = (1+1)/2 = 1$$

Donc, la qualité des deux hiérarchies au niveau des métadonnées est bien satisfaisante. Le symbole  sera affiché à l'intervenant.

De même, la qualité des niveaux *Province, Territoire* et *Pays* du cube CI est évaluée en se basant sur les formules (5) et (6):

$$Q_s(Province, Province) = 1$$

$$Q_m(Pays, Pays) = 1$$

Enfin, dans le cas où un élément d'un des cubes sources est unique (i.e. sémantiquement non lié à aucun autre élément d'un autre cube) et qu'il répond au besoin de l'utilisateur, alors cet élément (mesure, hiérarchie, dimension ou niveau) est intégré dans le nouveau cube et les agrégations nécessaires sont calculées. C'est le cas des dimensions *Groupe d'âge* et *Famille de virus* de notre exemple.

La figure 6 montre un exemple de modèle qui pourrait être obtenu en intégrant les modèles des cubes *C1* et *C2* suivant les différents niveaux du cadre général et utilisant les indicateurs de qualités proposés. Il est important de rappeler ici que l'approche proposée ne vise pas à trouver une solution spécifique pour l'intégration des cubes, mais plutôt à aider l'intervenant à 1) analyser la faisabilité de l'intégration des cubes en se basant sur une approche descendante, et à 2) trouver une solution possible en se basant sur un ensemble d'indicateurs de qualité des différents éléments de cubes. En d'autres termes, ces mêmes indicateurs sont aussi utiles puisqu'ils aident à considérer ou non certains éléments des cubes sources en se basant sur leur qualité aussi bien au niveau du schéma qu'au niveau des métadonnées.

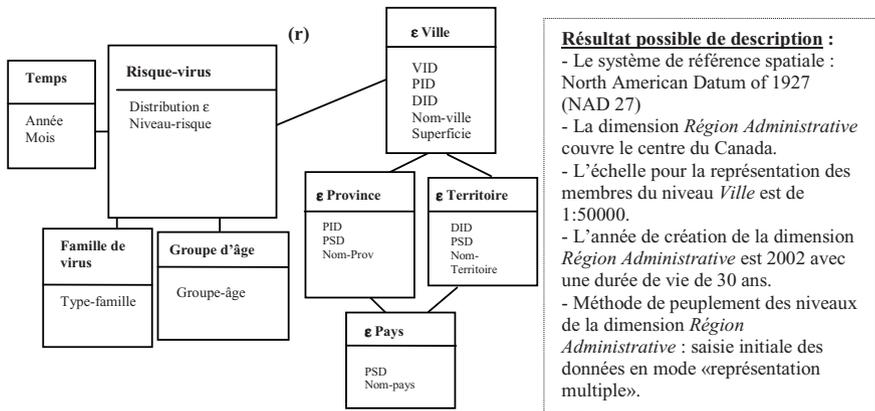


FIG. 6 – Exemple de solution finale d'intégration de modèles.

6 Conclusion

Les données spatiales sont des données complexes qui, une fois stockées dans des cubes, peuvent guider le processus d'analyse et facilitent la prise de décisions stratégiques au sein des organisations. Les cubes de données spatiales sont de plus en plus utilisés par les systèmes d'aide à la décision. Dans des situations spécifiques on peut avoir besoin d'intégrer plusieurs cubes. Cependant, ces cubes sont soit modélisés différemment d'une organisation à une autre ou d'un concepteur à un autre, soit qu'ils couvrent des régions, thématiques ou périodes différentes ou se chevauchant, ou soit qu'ils visent des objectifs légèrement ou fortement différents, qu'ils émanent ou non d'un même entrepôt de données. En de tels cas,

l'hétérogénéité représente un problème complexe pour l'intégration des cubes de données spatiales. Dans ce travail, nous avons présenté l'intérêt d'intégrer des cubes de données spatiales. Ensuite, nous avons proposé une catégorisation des problèmes d'hétérogénéité en tenant compte des différents éléments des cubes de données spatiales. Ainsi, nous avons défini cinq catégories : l'hétérogénéité *Cube-à-Cube*, l'hétérogénéité *Mesure-à-Mesure*, l'hétérogénéité *Dimension-à-Dimension*, l'hétérogénéité *Hiérarchie-à-Hiérarchie* et l'hétérogénéité *Niveau-à-Niveau*. Pour chaque catégorie, nous avons considéré les différentes composantes des modèles de cubes de données spatiales, notamment leur structure et leurs métadonnées. Enfin, nous avons proposé une approche qui consiste en un cadre général pouvant être utilisé par les intervenants impliqués et une méthode d'évaluation de la qualité des éléments des cubes à intégrer.

Le cadre ainsi proposé se base sur une structure hiérarchique allant du niveau le plus général au niveau le plus détaillé. Cette structure correspond au modèle mental de l'intervenant qui peut prendre des décisions appropriées à chaque niveau de la hiérarchie en tenant compte de ses observations au niveau général avant même d'étudier les détails s'ils sont jugés nécessaires.

La méthode propose et évalue un ensemble d'indicateurs qui montrent la qualité du schéma et des métadonnées à intégrer. Les résultats d'évaluation de ces indicateurs sont représentés qualitativement par des symboles qui ont pour but d'aider l'intervenant à prendre les décisions d'une façon intuitive. L'ensemble des indicateurs proposés ne vise pas à être exhaustif ou précis mais plutôt à orienter l'intervenant dans sa prise de décision. De même, la méthode proposée ne se veut pas absolue, mais suffisamment indicatrice pour l'intervenant.

L'exemple présenté a montré comment le cadre ainsi que les indicateurs de qualité proposés peuvent aider l'intervenant à prendre des décisions concernant la suspension ou la continuation de l'intégration. La catégorisation des problèmes d'hétérogénéité ainsi que l'approche proposée constituent une base pour d'autres travaux en lien avec l'intégration des cubes de données spatiales.

Remerciements

Les auteurs tiennent à remercier les organisations suivantes pour le financement de la Chaire de recherche industrielle en bases de données géospatiales décisionnelles : Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG), Recherche et Développement Défense Canada, Hydro-Québec, DVP, Intélec Géomatique, Holonics, KHEOPS Technologies, Syntell, Ressources Naturelles Canada, Transports Québec et l'Université Laval. Les auteurs tiennent à remercier également les rapporteurs anonymes pour leurs judicieux commentaires à l'égard de ce travail.

Références

- ANSI Z535.6 (2006). American national standard for product safety signs and labels.
- Batini, C., M. Lenzerini, et S. B. Navathe (1986). A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18, 323-364.

- Bédard, Y. et E. Bernier (2002). Supporting Multiple Representations with Spatial View Management and the Concept of "VUEL". Joint Workshop on Multi-Scale Representations of Spatial Data, ISPRS WG IV/3, ICA Com. on Map Generalization.
- Bédard, Y., M.-J. Proulx, et S. Rivest (2005). Enrichissement du OLAP pour l'analyse géographique : exemples de réalisation et différentes possibilités technologiques. *Revue des Nouvelles Technologies de l'Information, Cepaduès-Éditions*, 1-20.
- Bédard, Y. et S. Larrivée (2007). Spatial Databases Modeling with Pictogrammic Languages. *Encyclopedia of Geographic Information Sciences*. New York: Springer Verlag, 716-725.
- Bédard, Y., S. Rivest, et M.-J. Proulx (2007). Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective. *Data Warehouses and OLAP: Concepts, Architecture, and Solutions*. R. Wrembel and C. Koncilia, London, UK, Idea Group Publishing, 298-319.
- Bédard, Y., T. Merrett, et J. Han (2001). Fundamentals of spatial data warehousing for geographic knowledge discovery. *Geographic Data Mining and Knowledge Discovery*. Miller, H.J., Han, J. (Eds.).
- Boussaid, O., F. Bentayeb, J. Darmont, et S. Rabaséda (2003). Vers l'entreposage des données complexes : structuration, intégration et analyse. *Ingénierie des Systèmes d'Information*, 8(5-6), 79-107.
- Boucelma, O., Z. Lacroix, et M. Essid (2002). A WFS-Based Mediation System for GIS Interoperability. *ACM International Symposium on Advances in Geographic Information Systems*, 23-28.
- Brodeur, J. (2004) *Interopérabilité des données géospatiales: élaboration du concept de proximité géosémantique*. Thèse de doctorat, Université Laval.
- Bruckner, R. M., T. M. Ling, O. Mangisengi, et A. M. Tjoa (2001). A Framework for a Multidimensional OLAP Model Using Topic Maps. *Web Information Systems Eng. (WISE '01)*, 109-118.
- Busse, S., R.-D. Kutsche, U. Leser et H. Weber (1999). Federated Information Systems: Concepts, Terminology and Architectures. *Technical Report no 99-9*. Technical University of Berlin, 38 p.
- Devillers R., Y. Bédard, R. Jeansoulin. et B. Moulin. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data". *International Journal of Geographical Information Science*, 21(3):261-282.
- Devoegele, T. (1997) *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.
- Devoegele, T., C. Parent, et S. Spaccapietra (1998). On spatial database integration. *International Journal of Geographical Information Science*, 12(4):335-352.
- Edwards, G. (2001). A virtual test bed in support of cognitively-aware geomatics technologies. *Conférence on Spatial Information Theory (COSIT)*. Lecture Notes in Computer Science 2205, Springer, 140-155.

- Faïz, S.O. (1999). Systèmes d'Information Géographique : Information Qualité et Data Mining. *Les Éditions C.L.E.* Tunis, 362 p.
- Few, S. (2006) *Information Dashboard Design: The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.
- Foster, I., et C. Kesselman (1999). *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA: Morgan Kaufmann Publishers.
- Franklin, C. (1992). An introduction to geographic information systems: linking maps to databases. *Database*, 15(2), 13-21.
- Gervais, M. (2004) *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*. Thèse de doctorat, Université Laval et Université Marne-La-Vallée.
- Harvey, F., W. Kuhn, H. Pundt, Y. Bishr, et C. Riedemann (1999). Semantic Interoperability: A Central Issue for Sharing Geographic Information. *Annals of Regional Science*. Special Issue on Geo-spatial Data Sharing and Standardization, 213-232.
- Hurtado C. A., G. Claudio, Mendelzon A. O. (2005). Capturing summarizability with integrity constraints in OLAP. *ACM Transactions on Database Systems*, 30(3), 854-886.
- Inmon, W. H. (1996). *Building the Data Warehouse*. New York: John Wiley & Sons, Second Edition, 401 p.
- ISO: International Standards Organization (2003). Geographic information - metadata. International Organization for Standardization. *ISO 19115:2003*. Geneva, Switzerland.
- Kim, M. -Y., J. -M Seo, et C. -J. Moon (2007). SQL Extension for Multidatabase System. *International Computational and its Applications (ICCSA)*, 283-289.
- Kunapo, J., J. Peterson, et S. Chandra (2007). Spatial data integration for accurate parcel level sub-catchment delineation for hydrological modelling. *Spatial Science Institute Biennial International Conference (SSC2007)*, Hobart, Tasmania, Australia, 824-833.
- Laurini, R. (1998). Spatial multidatabase topological continuity and indexing: a step towards seamless GIS data interoperability. *International Journal of Geographical Information Sciences*, 12(4):373-402.
- Levesque, M.-A., Y. Bédard, M. Gervais, et R. Devillers (2007). Towards managing the risks of data misuse for geospatial cubes de données. *5th International Symposium on Spatial Data Quality (ISSDQ 2007)*, Enchede, Pays-Bas.
- Lutza, M., J. Spradob, E. Klienc, C. Schubertd, et I. Christ (2008). Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences*.
- Machlis, G. E. et E. A. Rosa (1990). Desired risk: Broadening the social amplification of risk framework. *Risk Analysis*, 10:161-68.
- Mangisengi, O., J. Huber, C. Hawel, et W. Essmayr (2001). A Framework for Supporting Interoperability of Data Warehouse Islands Using XML. *Data Warehousing and Knowledge Discovery (DaWaK '01)*, 328-338.

- Marchand, P., A. Brisebois, Y. Bédard, et G. Edwards (2004). Implementation and evaluation of a hypercube-based method for spatio-temporal exploration and analysis. *Journal of the International Society of Photogrammetry and Remote Sensing (ISPRS)*, 59(1-2):6-20.
- Mennis, J. L., D. J. Peuquet, et L. Qian (2000). A conceptual framework for incorporating cognitive principles into geographical database representation. *International Journal of Geographic Information Science*, 14(6):501-520.
- Miquel, M., Y. Bédard et A. Brisebois (2002). Conception d'entrepôts de données géospatiales à partir de sources hétérogènes, exemple d'application en foresterie. *Ingénierie des Systèmes d'Information*, 7(3):89-111.
- Morgan, M. G. (1990). Choosing and Managing Technology-Induced Risks., *Readings in Risk*. T.S. Glickman and M. Gough (Eds), Washington: Resources for the Future.
- Nedas, K.A., et M.J. Egenhofer (2008). Spatial-Scene Similarity Queries. *Transactions in GIS*, 12(6).
- Nguyen, T.B., A.M. Tjoa, et O. Mangisengi (2001). MetaCube-X: An XML Metadata Foundation of Interoperability Search among Web Data Warehouses. *Third International Workshop Design and Management of Data Warehouses (DMDW '01)*, 8.1-8.8.
- Niemi, T., M. Niinimäki, J. Nummenmaa, et P. Thanisch (2003). Applying Grid Technologies to XML Based OLAP Cube Construction. *Fifth International Workshop Design and Management of Data Warehouses (DMDW '03)*, 4.1-4.13.
- Pérez, J. M., R. Berlanga, M. J. Aramburu, et T. B. Pedersen (2008). Integrating Data Warehouses with Web Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 20(7):940-955.
- Portele, C. (2007). OpenGIS Geography Markup Language Encoding Standard (version 3.2.1). *Document OGC*, 7-36.
- Rahm, E. et P. A. Bernstein (2001). A survey of approaches to automatic schema matching. *The International Journal on Very Large Data Bases*, 10:334-350.
- Ranjan, J. et S. Khalil (2008). Building Data Warehouse at life insurance corporation of India: a case study. *International Journal of Business Innovation and Research*, 2(3): 241-261.
- Renn, O. (1998). Three decades of risk research: accomplishments and new challenges. *Journal of Risk Research*, 1(1):49-71.
- Rivest, S., Y. Bédard, M.J. Proulx, et M. Nadeau (2003). SOLAP: a new type of user interface to support spatio-temporal multidimensional data exploration and analysis. *ISPRS Joint Workshop on Spatial, Temporal and Multi-Dimensional Data Modelling and Analysis*, Quebec, Canada.
- Rivest, S., Y. Bédard, M.-J. Proulx, M. Nadeau, F. Hubert, et J. Pastor (2005). SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *ISPRS Journal of Photogrammetry & Remote Sensing*, 60:17-33.

- Rizzi, S., A. Abelló, J. Lechtenböcker, J. Trujillo (2006) Research in data warehouse modeling and design: dead or alive? DOLAP. *Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP*, 3-10.
- Rosch, E. (1978) *Principles of categorization*. In: Rosch, E., Lloyd, B. (Eds.), *Cognition and Categorization*, 27-77.
- Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data - a survey. *Transactions in GIS*, 12(1).
- Shankar, M., A. Sorokine, B. Bhaduri, D. Resseguie, S. Shekhar et J. S. Yoo (2007). Spatio-temporal conceptual Schema Development for Wide-Area Sensor Networks. *International Proceedings of International Conference on Geospatial Semantics(GeoS)*, Mexico City, Mexico.
- Sheeren, D. (2005) *Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales – Une approche combinant l'utilisation de métadonnées et l'apprentissage automatique*. Thèse de doctorat, Université Paris 6.
- Sheth, A. et J. Larson (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183-230.
- Thomsen, E., G. Spofford, et D. Chase (1999). *Microsoft OLAP Solutions*. New York: John Wiley & Sons, 495 p.
- Tseng F. et C. Chen (2005). Integrating Heterogeneous Data Warehouses Using XML Technologies. *Journal of Information Science*, 31(3):209-229.
- Turban, E. et J. E. Aronson (2000). *Decision support systems and intelligent systems (6th ed.)*. New Jersey: Prentice Hall.
- Wicaksana I. W. S. (2007). *Interopérabilité de Systèmes d'Information dans un environnement Pair-à-Pair : une Approche basée sur les Agréments Sémantiques*. Thèse de doctorat, Université de Bourgogne.
- Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38-49.
- Xiong, D., et J. Sperling, (2004). Semiautomated matching for network database integration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59, 35-46.
- Yougworth, P. (1995). OLAP Spells Success For Users and Developers. *Data Based Advisor*, 38-49.
- Ziegler, P. et K. R. Dittrich (2004). Three decades of data integration - all problems solved? *18th IFIP World Computer Congress (WCC 2004)*, 12:3-12.

Summary

The integration of spatial data cubes aims at facilitating the access and the reuse of their content for strategic decision analysis. Such integration, however, faces complex problems related to the heterogeneity of spatial data cubes. While, the integration of databases has been the subject of several research works and standards, no research dealing with spatial

data cubes integration has been found. In this paper, (1) we motivate the need for integrating spatial data cubes, (2) we present a categorization of the heterogeneity problems related to spatial data cubes models, and (3) we propose an approach that supports interveners in making decisions about spatial data cubes integration. This approach consists of a global framework that is based on a top-down structure and a method that proposes and evaluates a set of quality indicators of spatial data cubes models. The proposed approach is illustrated with an example of application.

