Cohérence logique dans les systèmes OLAP spatiaux : un état de l'art

S. Bimonte, K. Boulil, François Pinet

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

Cohérence logique dans les systèmes OLAP spatiaux : un état de l'art

Sandro Bimonte[2], Kamal Boulil[1], François Pinet[2]

1. *Université de Strasbourg, France*

boulil@unistra.fr

2. *Irstea, TSCF, Aubière, France*

prenom.nom@irstea.fr

RESUME. Les systèmes d'Entrepôts de Données et « OnLine Analytical Process » spatiaux (EDS et SOLAP) sont des technologies d'aide à la décision permettant l'agrégation et l'analyse multidimensionnelle de gros volumes de données spatiales. En réponse à des actions utilisateurs sur l'interface cliente (exploration), les systèmes SOLAP agrègent les données de l'EDS le long de différentes hiérarchies des dimensions pour calculer des indicateurs d'analyse à différents niveaux de détails. La qualité des indicateurs d'analyse dépend donc de trois facteurs : la qualité des données entreposées, la qualité des agrégations et la qualité de l'exploration des données.

La qualité des données entreposées dépend de critères comme la précision, l'exhaustivité et la cohérence logique. La cohérence logique des données est généralement contrôlée par les contraintes d'intégrité qui définissent les conditions que les données doivent satisfaire. La qualité d'agrégation peut être ramenée à la cohérence logique entre les natures des éléments qui sont impliqués dans l'agrégation SOLAP (par ex., mesure, fonction d'agrégation). Cette cohérence d'agrégation est affectée par des problèmes structurels et de problèmes sémantiques . La qualité d'exploration dépend essentiellement de la consistance des requêtes utilisateur (par ex. quelles ont été les valeurs de température en URSS en 2010 ?). Dans cette article nous étendons la notion de cohérence logique des données aux deux autres composantes fondamentales des systèmes SOLAP, à savoir l'agrégation et la requête. Nous présentons un état de l'art et des travaux sur la définition des contraintes d'intégrité pour garantir la cohérence logique au niveau des trois composantes (données, agrégations, et requêtes), ainsi qu' une évaluation de ces travaux et de notre proposition par rapport à un ensemble de critères que nous avons définis. Cette évaluation montre que notre proposition satisfait tous les critères contrairement aux autres.

MOTS-CLES : Cohérence logique, OLAP Spatial, Contraintes d'intégrité

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

1. Introduction

L'informatique décisionnelle apporte des solutions pour la modélisation, l'interrogation et la visualisation de données dans un objectif d'aide à la décision. Les entrepôts de données (ED) et les outils d'analyse « On-Line Analytical Processing » (OLAP), représentent une solution efficace pour l'informatique décisionnelle (Immon, 1996), car ils permettent de produire des indicateurs agrégés à partir de grandes quantités de données alphanumériques. Cependant, malgré l'explosion de la disponibilité de données spatiales, obtenues grâce à de nouvelles techniques d'acquisition comme les satellites, les capteurs, etc., les systèmes OLAP ne présentent aucun instrument pour la gestion et l'analyse de telles données. Des solutions, connues sous le terme de Spatial OLAP (SOLAP), qui visent à intégrer la donnée spatiale dans les processus OLAP, ont donc été développées. Les systèmes SOLAP sont des outils d'aide à la décision qui intègrent les techniques d'analyse des processus OLAP et des Systèmes d'Information Géographique (SIG). Le concept de SOLAP a été défini par Yvan Bédard en 1997 comme : « *Une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographiques, graphiques et tabulaires* » (Bédard, 1997). Les processus SOLAP impliquent une redéfinition des concepts d'OLAP et d'entrepôts de données d'un point de vue formel et sur le plan de l'implémentation (Bédard, 2009).

La technologie SOLAP a été déjà employée avec succès dans différents domaines d'application comme la santé, le marketing, l'environnement, l'agriculture (Pinet *et al.*, 2010).

La qualité est une notion très importante qui est recherchée dans toute production de biens ou de services et aussi dans les systèmes informatiques et logiciels (ISO, 2000; ISO/IEC, 2001; Daniel, 2005). Il s'agit d'un facteur capital qui conditionne la réussite et le bon fonctionnement du système ou du produit. La norme ISO 9000 (ISO, 2000) définit la qualité comme : « l'ensemble des propriétés ou caractéristiques d'un produit ou service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites des utilisateurs ». La qualité est donc une notion relative qui dépend des besoins des utilisateurs : un même produit peut donner lieu à des évaluations de qualité différentes selon des besoins distincts. A un niveau général, la qualité d'analyse dans les systèmes (S)OLAP est une problématique de recherche très importante car ces systèmes sont utilisés pour aider le processus de prise de décision au sein de divers types d'organisations (Rizzi *et al.*, 2006; Salehi, 2009). La qualité d'analyse (S)OLAP telle que nous la définissons dans (Boulil *et al.*, 2012c ; 2012b) est déclinée en trois types : (a) la qualité de données, (b) la qualité d'agrégation et (c) la qualité d'exploration des données. Selon la norme ISO 9113:2002, la qualité de données (spatiales) dépend principalement de trois facteurs: la cohérence logique (absence de contradictions dans les données en rapport aux règles logiques définies dans leur spécification), la précision (données présentant le moins d'écarts possibles avec la réalité) et l'exhaustivité (présence de toutes les données nécessaires à la prise de décision). La qualité d'agrégation, traitée dans la

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

littérature sous le terme d'« agrégeabilité » (traduction du terme anglais Summarizability), définit des conditions pour garantir une agrégation correcte et sensée des mesures le long des hiérarchies de dimensions (Lenz et Shoshani, 1997). La qualité d'exploration vise à éviter les problèmes d'interprétation liés à des requêtes insensées ou invalides (Levesque et al., 2007 ; Boulil et al., 2012c ; 2012b).

La prise en compte des problèmes de qualité dans les systèmes SOLAP est donc un point crucial. Motivé par l'absence d'un cadre global qui résume les avancés dans ce domaine, nous proposons dans ce papier un état de l'art des travaux concernant la *cohérence logique* des données spatiales dans le processus d'entreposage et d'analyse en ligne. Une solution très utilisée pour vérifier la cohérence logique dans le domaine des bases de données et des entrepôts de données sont les Contraintes d'Integrité (CI). Les CI sont définis comme des assertions qui expriment les conditions qui sont censées être satisfaites par les données (Goertzen et Stausberg, 2007). Un exemple de CI est un trigger SQL sur une clé primaire qui évite l'insertion d'un tuple dans une base de données relationnelle si la clé primaire existe déjà.

L'article est structuré de façon suivante : la section 2 introduit les concepts principaux de OLAP et Spatial OLAP ; la section 3 introduit les problématiques de qualité dans le SOLAP ; l'évaluation des travaux existants est faite dans la section 4.

2. Concepts principaux

2.1. Entrepôts de données et OLAP

Un entrepôt de données(ED) est « *Une collection de données, intégrées, non volatiles et historiques pour la prise de décision* » (Kimball, 1996). Dans un entrepôt de données, les données sont organisées en accord avec le modèle multidimensionnel (Inmon, 1996). Ce modèle définit les concepts de faits et de dimensions pour décrire l'information décisionnelle conformément au point de vue des décideurs. Les faits représentent les sujets d'analyse et sont décrits par des attributs généralement numériques appelés mesures. Une instance de faits représente l'ensemble des valeurs mesures et des membres de dimensions.

Les dimensions représentent les axes d'analyse des mesures; elles sont organisées en hiérarchies. Une hiérarchie de dimension décrit une organisation hiérarchique de niveaux (ou niveaux d'agrégation). Une instance d'un niveau est appelée membre. Ces niveaux d'agrégation correspondent à différent niveaux de détail auxquels les indicateurs d'analyse peuvent être observés. Nous appelons instance de hiérarchie la hiérarchie formée par les membres des niveaux d'une dimension.

Un exemple de modèle multidimensionnel est montré en Figure 1. Il concerne l'analyse des ventes (le fait) qui présente une mesure (la quantité), et il est décrit par trois dimensions. Par exemple la dimension Temps présente trois niveaux (Jour, Mois, Année). Un membre du niveau Année est par exemple l'année 1990.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

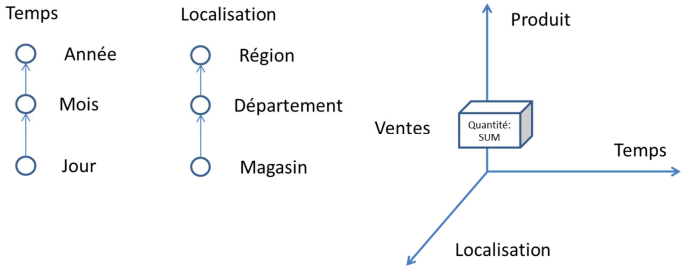

Figure 1. Exemple de modèle multidimensionnel.

Une instance d'un fait par exemple est : <19-9-1990, Carrefooor, Wiii, 150>.

Les méthodes courantes pour explorer les données d'un ED sont les processus OLAP. Ces systèmes permettent une exploration interactive et intuitive des données. Ils définissent pour cela un ensemble d'opérateurs qui permettent de calculer et de manipuler des cubes de données. Structurellement, les cubes de données (ou aussi hypercubes) correspondent à des vues du modèle multidimensionnel (Pinet et Schneider, 2010). Les opérateurs OLAP permettent d'avoir une information plus ou moins détaillée en agrégeant les mesures par des fonctions d'agrégation (par exemple Sum, Max, etc.). Les opérateurs classiques sont les opérateurs de forage vers le haut (Roll-up) et vers le bas (Drill-down) qui permettent respectivement de monter/descendre dans les hiérarchies de dimension (Inmon, 1996). Par exemple, l'opération de Roll-Up permet de passer d'une agrégation faite par ville à une agrégation faite par région. L'opération de Drill-down permet de faire l'inverse.

Les systèmes OLAP visualisent les données à travers un client OLAP qui offre une interface utilisateur avec des outils de rendu, d'analyse interactive, et parfois de fouille de données. Le paradigme de visualisation le plus adopté par les clients OLAP est le tableau croisé (Pinet et Schneider, 2010). Il s'agit d'un tableau multidimensionnel auquel sont associés des totaux et des sous-totaux. Les interactions de l'utilisateur avec les différentes composantes (par ex. un clic sur un membre d'un niveau de dimension) déclenchent les opérateurs OLAP. Les tableaux croisés sont souvent couplés avec des affichages graphiques, par exemple en barres, en secteurs, en bulles etc., qui permettent d'avoir une vue synthétique des données.

2.2. Entrepôts de données spatiales et SOLAP

2.2.1. Concepts principaux

La nécessité croissante d'intégrer l'information spatiale dans l'analyse multidimensionnelle OLAP a conduit à l'introduction des concepts d'Entrepôts de Données Spatiales (EDS) et du Spatial OLAP (SOLAP). Les Entrepôts de Données Spatiales (EDS) permettent d'intégrer et d'historiser de très gros volumes de données géoréférencées provenant de sources diverses pour supporter les processus de prise de décision (Stefanovic et al., 2000). Les systèmes SOLAP permettent une exploration rapide et interactive suivant une approche multidimensionnelle et à plusieurs niveaux de granularité du contenu de l'EDS (Bédard et al.,2007). Ces

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

systèmes enrichissent les capacités d'analyse des systèmes OLAP, en combinant les méthodes d'analyse multidimensionnelles OLAP avec des navigations et visualisations cartographiques issues des SIG.

Les systèmes d'EDS et SOLAP se basent sur le modèle spatio-multidimensionnel qui étend le modèle multidimensionnel OLAP et permet de prendre en compte la composante spatiale de l'information géographique en la représentant en axes d'analyse (dimension spatiale) et/ou en mesures (mesure spatiale) (Malinowski et Zimányi, 2008). Une dimension spatiale présente un ou plusieurs niveaux ayant un attribut spatial (une géométrie). Une mesure spatiale représente soit la géométrie du phénomène analysé (par ex. une zone d'épandage agricole) ou le résultat d'opérations spatiales topologiques ou métriques comme le résultat de l'opérateur topologique d'intersection, la surface d'une région spatiale, etc.

En plus de la visualisation cartographique des résultats des requêtes décisionnelles, la représentation des données spatiales dans les niveaux d'agrégation permet principalement d'utiliser des prédicats spatiaux pour la sélection et le regroupement des données lors de l'application des opérateurs SOLAP comme par exemple le Roll-up spatial; la représentation des données spatiales en mesures rend possible leur agrégation à l'aide de fonctions d'agrégation spatiale (par ex. union spatiale).

2.2.2. Architecture

Les systèmes SOLAP sont généralement implémentés suivant une architecture relationnelle (basée sur un Système de Gestion de Bases de Données Relationnelles, SGBDR) constituée de quatre couches logicielles (cf. Figure 2) : ETL (*Extract, Transform and Load tool*) spatial, EDS, Serveur SOLAP et Client SOLAP (Rivest et al., 2003). Le rôle de l'ETL spatial est l'intégration et le chargement périodique au sein de l'EDS de toutes les données qui sont nécessaires à l'analyse SOLAP. La couche de stockage ou EDS est responsable de l'historisation et de l'organisation des données intégrées en vue de leur analyse. Elle est généralement constituée d'un entrepôt de données qu'on appelle primaire qui stocke généralement au niveau le plus fin toutes les données nécessaires à la prise de décision au sein de l'organisation, et éventuellement de petits EDS focalisés sur des besoins particuliers (par ex. relatifs à un groupe d'utilisateurs ou répondant à certaines exigences d'optimisation), appelées magasins de données (Data marts) et qui sont soit calculés à partir de l'EDS ou à partir des sources de données. Cette couche est gérée par le SGBDR spatial (par ex. Oracle Spatial, PostGis, etc.) qui utilise des techniques d'optimisation de stockage et de requêtes adaptées aux données spatiales (Sampaio et al., 2006; Malinowski et Zimányi, 2008; da Silva et al., 2010). Le serveur SOLAP (par ex. GeoMondrian) permet l'exécution rapide des requêtes d'analyse multidimensionnelles sur les données entreposées en se basant sur un schéma d'analyse appelé schéma OLAP qui définit les hypercubes de données spatiales, leurs dimensions et indicateurs et indique le mapping de ces structures OLAP vers les structures relationnelles de l'EDS (tables de fait et de dimension). Pour cette exécution rapide des requêtes, il implémente un ensemble d'opérateurs SOLAP (par ex. Roll-up spatial) pour le calcul et la manipulation des hypercubes de données

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

spatiales. Enfin, le client SOLAP (par ex. Map4Decision) met à disposition des utilisateurs une série d'interfaces graphiques intuitives qui déclenchent les opérateurs SOLAP et permettent l'exploration interactive et la visualisation des hypercubes de données en utilisant différents formats d'affichage interactifs: tableaux croisés, diagrammes statistiques, cartes, etc. Les affichages tabulaires sont synchronisés avec les autres types de visualisation.

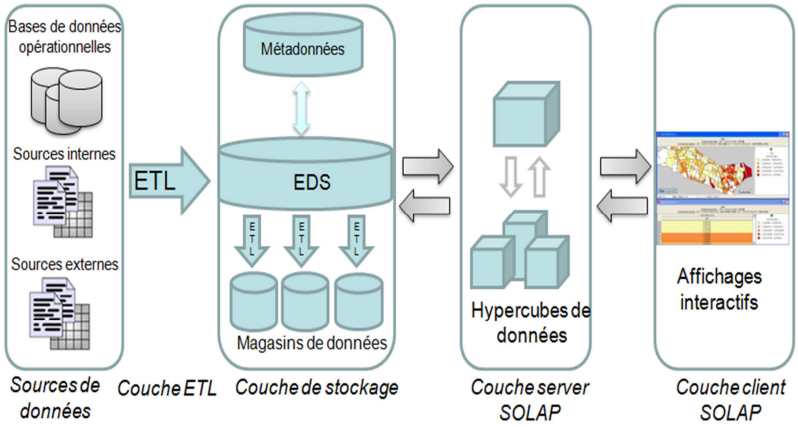

Figure 2. Architecture SOLAP.

3. SOLAP et Qualité

La qualité d'analyse (S)OLAP dépend de trois types de qualité : la qualité des données (spatiales) entreposées, la façon dont les mesures (spatiales) sont agrégées et la façon dont les décideurs explorent les données agrégées. Dans cette section nous introduisons trois grandes classes de type de qualité liées : (a) aux données, (b) à l'agrégation, et (c) à l'exploration. Cette classification sera par la suite utilisée pour décrire les travaux existants.

La qualité des données (spatiales) est une notion relative et complexe qui nécessite le recours à plusieurs composantes (ou critères) pour être évaluée (ISO/TC 211, 2002 ; Devillers et Jeansoulin, 2005 ; Devillers *et al.*, 2007). Divers travaux de recherche ont proposé divers critères quantitatifs et qualitatifs pour déterminer la qualité d'un jeu de données (spatiales). Ces travaux proposent deux définitions pour la qualité des données en considérant deux points de vue, celui des producteurs de données et celui des utilisateurs de données :

(a) Qualité interne : La qualité interne désigne l'absence d'erreurs/imperfections dans les données en rapport à des règles précises de production (ISO/IEC, 2001 ; 2003). Cette qualité qualifiée de qualité intrinsèque est mesurée au travers de critères quantitatifs et qualitatifs comme la précision, l'exhaustivité, la cohérence logique, la généalogie, etc.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

(b) Qualité externe : La qualité externe correspond au niveau d'adéquation existant entre les données et les besoins des utilisateurs dans un contexte donné (ISO/IEC, 2001 ; 2004). Elle exprime la capacité des données à répondre à un usage particulier. C'est donc une qualité dépendante des besoins des utilisateurs qui peut être qualifiée de qualité d'usage relative.

Dans la suite nous présentons le concept de cohérence logique selon trois axes : les données, l'agrégation et l'exploration.

3.1 Données

La cohérence logique décrit le degré d'adhésion (conformité) des données aux règles logiques des structures de données, des attributs et des relations (la structure de données peut être conceptuelle, logique, physique) (ISO/TC 211 ; 2002). Autrement dit, la cohérence logique désigne l'absence de contradictions dans le jeu de données en rapport aux règles logiques présentes dans sa spécification. Ses sous éléments sont :

a. Cohérence conceptuelle : adhésion des données aux règles du schéma conceptuel;

b. Cohérence de domaine : adhésion des valeurs à leurs domaines de valeurs;

c. Cohérence de format : le degré auquel les données sont stockées en accord avec la structure physique du jeu de données;

d. Cohérence topologique : justesse des caractéristiques topologiques encodées explicitement du jeu de données.

Dans la littérature, la cohérence logique des données dans le domaine des ED et EDS a été traitée par différents travaux (Carpani et Ruggia, 2001; Ghozzi et al., 2003a; Malinowski et Zimányi, 2008; Pinet et Schneider, 2009; Salehi, 2009; Pinet et Schneider, 2010). Ces travaux proposent la définition de Contrainte d'Intégrité (CI) pour interdire l'entreposage de données erronées. D'autres approches ont été proposées pour traiter les problèmes de qualité de données lors de l'exploration par exemple par la reformulation des requêtes (Pedersen et al., 2001). Ces travaux sont décrits en Section 4.3.1.

3.2. Agrégation

Un autre aspect fondamental de la qualité d'analyse (S)OLAP concerne la qualité d'agrégation des données. En effet, avoir des données entreposées de bonne qualité ne garantit pas des indicateurs d'analyse exacts et qui ont du sens pour l'application considérée. La qualité des indicateurs obtenus suite à l'agrégation des données, en plus de la qualité des données dépend à la fois (a) de conditions de schéma sur les structures multidimensionnelles et (b) de conditions sémantiques sur les définitions des indicateurs d'analyse.

Nous proposons de regrouper et qualifier ces deux types de conditions sous deux composantes, la cohérence structurelle d'agrégation et la cohérence sémantique d'agrégation. La cohérence structurelle d'agrégation se rapporte essentiellement à

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

des contraintes fixant les cardinalités des relations fait-dimension et entre niveaux d'agrégation. Par exemple, une ville n'est associée qu'à une seule région et une région n'est associée qu'à un seul pays. Ces contraintes permettent d'éviter les problèmes bien connus du comptage en double des mesures et des agrégats incomplets qui peuvent être induits par des hiérarchies et relations fait-dimension irrégulières/non agrégeables, c'est-à-dire qui ne satisfont pas ces contraintes. Les relations non strictes peuvent engendrer le comptage en double. La cohérence sémantique concerne essentiellement la compatibilité entre les natures des trois éléments fondamentaux d'une agrégation (S)OLAP : la fonction d'agrégation, la mesure et la hiérarchie de dimension (par ex. la somme de mesures semi-additives comme la quantité en stock des produits selon le temps induit des résultats incorrects), et également la compatibilité des unités de mesure des données agrégées.

Cette problématique de cohérence d'agrégation (« Agrégeabilité » ou Summarizability), a été introduite pour la première fois dans le domaine des bases de données statistiques par (Rafanelli et Shoshani, 1990). Ce travail définit l'agrégeabilité (ou plus précisément l'agrégeabilité structurelle) comme la possibilité de calculer des agrégats corrects à un niveau de granularité plus élevé en réutilisant des agrégats plus fins. Par exemple, la population des départements Français sert à calculer la population des régions, puis la population des régions permet de calculer la population totale de la France. L'agrégeabilité (structurelle et sémantique) a été ensuite étudiée dans les processus OLAP par le travail de (Lenz et Shoshani, 1997). Ce travail de référence définit trois conditions pour garantir l'agrégeabilité : (a) disjonction (*disjointness*), (b) complétude (*completeness*) et (c) compatibilité de types (*type compatibility*). Les deux premières conditions sont plutôt des conditions structurelles, elles permettent respectivement d'éviter les problèmes de comptage en double des mesures et des agrégats incomplets. La condition de disjonction permet de vérifier que les ensembles de données (instances de fait ou membres des niveaux des dimensions) formés à partir des membres se trouvant aux mêmes niveaux de détail en suivant les liens d'agrégation et les liens fait-dimension, sont disjoints. Par exemple, la hiérarchie spatiale de la Figure 3 est une hiérarchie non stricte qui ne respecte pas cette condition puisque la "parcelle 2" appartient aux deux villes "Clermont-Ferrand" et "Aubière". De ce fait les deux ensembles de membres fils de ces deux villes ne sont pas disjoints.

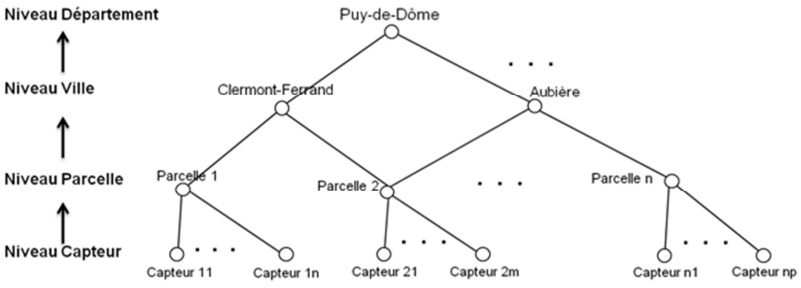

Figure 3. Exemple de hiérarchie non stricte

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

La condition de complétude vérifie que ces ensembles sont complets; en d'autres termes, toutes les données (membres et instances de fait) sont présentes et sont correctement assignées aux différents ensembles; ceci permet d'éviter le problème des agrégats incomplets. Prenons l'exemple d'un entrepôt servant à afficher les sommes des ventes d'une chaîne de magasins. S'il manque des ventes dans certains magasins dans l'entrepôt, alors les agrégats aux niveaux départements, régions (etc.) seront incomplets (ici sous-évalués). La condition de compatibilité de types est une condition sémantique qui vérifie que l'application de la fonction d'agrégation à la mesure le long de la dimension est valide en considérant la nature de la mesure (par ex. stock, flux, ou valeur par unité), le type de dimension (temporelle ou non temporelle) et le type de fonction d'agrégation utilisée (distributive ou pas). Par exemple, la somme des populations annuelles de France selon le temps (par décennie par exemple) n'est pas correct du fait du comptage en double des individus. Par rapport à cette dernière condition de compatibilité de types, plusieurs classifications des mesures, fonctions d'agrégation, et hiérarchies ont été proposées.

Une analyse détaillée des causes d' « inagrégeabilité » est proposée par (Horner et Song, 2005). Cette étude distingue trois familles de problèmes : problèmes de schéma (liés aux différents types de hiérarchies de dimension irrégulières), problèmes de données (imprécisions et incohérences dues à l'emploi de différents instruments de mesure) et problèmes de calcul (agrégation des mesures avec des fonctions inappropriées (incompatibilité de types – dans le sens de (Lenz et Shoshani, 1997)) et agrégation de mesures ayant des unités incompatibles).

Dans le contexte des EDS, (Bimonte et al., 2009) montrent la dépendance qui peut exister dans certaines applications SOLAP entre les agrégations de mesures spatiales et numériques, et comment la topologie des membres spatiaux affecte l'exactitude de l'agrégation des mesures numériques.

Pour éviter ces problèmes d'inagrégeabilité et garantir les conditions définies par (Lenz et Shoshani, 1997) différentes approches sont proposées dans la littérature. Ces approches sont présentées et discutées en Section 4.3.2.

3.3 Exploration

Les outils (S)OLAP sont une catégorie de logiciels axés sur l'exploration interactive et rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation (Bédard, 2009). Les utilisateurs de ces systèmes sont essentiellement des analystes d'affaires et des décideurs, qui sont supposés ignorer les contraintes liant les données multidimensionnelles (Inmon, 2005). Ils peuvent donc formuler assez facilement des requêtes multidimensionnelles inconsistantes en définissant des combinaisons insensées ou invalides de membres et d'indicateurs (Devillers *et al.*, 2007 ; Boulil et al., 2012c). Ces requêtes invalides dépendent de l'application considérée. L'exécution de ces requêtes retourne souvent des valeurs nulles. Par exemple ce sera certainement le cas si on définit des CI de données qui interdisent le stockage de données (instances de fait) définies par ces combinaisons. Un problème se pose au niveau de l'interprétation de la vacuité de ces résultats. L'utilisateur qui a formulé la requête va probablement interpréter ce résultat nul

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

comme une absence de données, au lieu de réaliser que sa requête est insensée (Levesque *et al.*, 2007). Par ex. l'absence de ventes dans les magasins de RDA en 2000 s'explique en fait par la disparition de la RDA. Des interprétations erronées des données peuvent mener le décideur à prendre de mauvaises décisions. Les approches possibles pour éviter ces problèmes d'interprétation sont de définir des méthodes pour informer l'utilisateur de la qualité des résultats de requêtes et/ou interdire l'exécution de ces requêtes incohérentes et/ou les reformuler en requêtes valides, etc. Toutes ces approches nécessitent de vérifier si la requête utilisateur définit ou non une combinaison problématique. Ces travaux sont décrits en Section 4.3.3.

4. La cohérence logique dans les entrepôts de données et les entrepôts de données spatiales

Dans cette section, nous introduisons les Contraintes d'Intégrité (Sec. 4.1), nous définissons un ensemble de critères pour évaluer les travaux existants sur les CIs et les EDS (Sec 4.2) et nous détaillons cette évaluation dans la Section 4.3.

4.1. Contraintes d'Intégrité

Dans le domaine des bases de données, les CI sont définis comme des assertions qui expriment les conditions qui sont censées être satisfaites par les données (Goertzen et Stausberg, 2007). Elles sont utilisées principalement pour vérifier la cohérence logique des bases de données. Elles peuvent être vérifiées/contrôlées lors de l'insertion de nouvelles données dans la base de données; et/ou la modification; et/ou la suppression de données existantes. Le mécanisme qui permet ce contrôle est connu sous le nom de mécanisme de contrôle d'intégrité. Dans le domaine des entrepôts de données (y compris spatiales), les CI sont typiquement utilisées pour garantir la cohérence des données intégrées à partir de plusieurs sources (qualité de données) (Carpani et Ruggia, 2001 ; Ghozzi et al., 2003b ; Salehi, 2009; Pinet et Schneider, 2010) ; spécifier les conditions qui garantissent l'agrégation correcte des mesures le long des hiérarchies de dimension (qualité d'agrégation) (Salehi, 2009 ; Prat et al., 2010) et vérifier la consistance des requêtes multidimensionnelles (qualité d'exploration) (Boulil et al., 2012b). Les CI doivent être définies au niveau conceptuel (avant d'être codées) afin de permettre leur validation par les experts du domaine car ce sont les langages utilisés à ce niveau qui servent de base à la communication entre les utilisateurs et les concepteurs (Torlone, 2003).

4.2. Evaluation des travaux

Proposition d'une classification des CI (C1). Les classifications facilitent l'identification des CI par les analystes. Elles peuvent aussi servir de base à la définition de patrons de conception spécifiques à chaque classe de CI. Ces patrons intégrés dans des Ateliers de Génie Logiciel (AGL) vont faciliter la modélisation conceptuelle et la validation des contraintes et permettre leur implémentation automatique dans différents niveaux de l'architecture SOLAP ; pour chaque patron,

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

il est possible de définir les règles de transformation du niveau conceptuel vers des représentations physiques dans différentes couches de l'architecture SOLAP.

Nombre de types de CI considérés (C2). La qualité de l'analyse SOLAP dépend forcément de toutes les CI pouvant être définies à différents niveaux de l'architecture SOLAP. La valeur (élevé, moyen, réduit) de ce critère concerne le nombre de classes de CI considérées

Niveau d'abstraction (C3). La plupart des travaux sur les CI préconisent leur définition au niveau conceptuel. Ceci facilite leur validation par les experts du domaine.

Utilisation de standards pour spécifier les CI et le modèle multidimensionnel (C4). L'utilisation des langages standards comme UML et OCL (qui sont intégrés dans de nombreux AGL) pour la spécification conceptuelle des CI (S)OLAP et du modèle (spatio)-multidimensionnel rend plus facile leur spécification et permet leur implémentation automatique.

Utilisation de notations visuelles pour exprimer certaines CI (C5). Dans certains cas, les notations visuelles améliorent la lisibilité des modèles conceptuels et facilitent l'interaction avec les décideurs et spécialistes du domaine (Papajorgji *et al.*, 2010).

Implémentation automatisée des CI (C6). Il est nécessaire pour une implémentation automatique des CI (S)OLAP d'avoir une formalisation (basée sur des langages standards) des règles de mapping entre le modèle conceptuel et les représentations physiques utilisable dans différentes plateformes. Il faut aussi permettre une implémentation au sein de générateurs de code ou bien l'utilisation de générateurs de code existants (comme préconisé par l'approche MDA (OMG, 2003) par exemple). L'évaluation d'une méthode selon ce critère est considérée satisfaisante si cette méthode formalise les mappings en se basant sur des standards implémentés ou/et utilise des générateurs de code pour une implémentation automatique des CI.

Dans la suite de cette section nous allons présenter et discuter les limites et avantages des différents langages de modélisation conceptuelle de CI dans le domaine des bases de données spatiales et entrepôts de données (**Utilisation de standards pour spécifier les CI et le modèle multidimensionnel (C4)**).

Les langages de modélisation conceptuelle des CI, dans les domaines des bases de données spatiales et des EDS, sont classées en quatre grandes familles (Salehi *et al.*, 2007) : (a) langages naturels, (b) langages visuels, (c) langages logiques et (d) langages hybrides. Ces catégories ainsi que leurs sous catégories sont décrites dans ce qui suit. La description faite par (Salehi *et al.*, 2007) de ces catégories est résumée ici.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

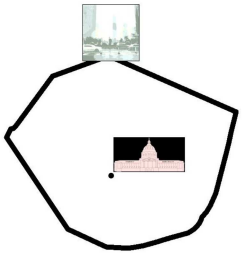

Figure 4. Exemple d'expression d'une CI avec un langage visuel

Les *langages naturels* sont issus des langages de communication de tous les jours entre les humains (comme l'Anglais et le Français). Comparés à d'autres types de langages, ces langages permettent une spécification plus facile des CI et plus de lisibilité (les utilisateurs comprennent plus facilement les spécifications édictées en langues naturelles). Ces langages sont classés en deux sous-catégories : (a) langages naturels libres et (b) langages naturels contrôlés. Un *langage naturel libre* correspond exactement à un langage naturel (par ex. le Français) sans aucune restriction sur le vocabulaire, la syntaxe et la sémantique du langage. La syntaxe décrit la façon dont les éléments du langage doivent être combinés pour former des expressions (ou phrases) correctes. La sémantique décrit le sens des éléments et des expressions correctes du langage. Ces langages offrent des vocabulaires riches et sont les plus faciles à utiliser. En revanche, ils présentent plusieurs ambiguïtés (par ex. un même mot peut avoir différents sens selon le contexte de son utilisation) et sont donc difficilement interprétables automatiquement par la machine. Malgré les progrès important en traitement automatique du langage naturel, les langages naturels restent encore aujourd'hui difficilement exploitables dans le domaine de la représentation des CI en vue d'une exploitation informatique automatisée.

Les *langages naturels contrôlés* sont des sous-ensembles de langages naturels dont la sémantique et la syntaxe ont été réduites. Ces langages ont été proposés dans le but de réduire l'ambiguïté des langages naturels tout en gardant leur facilité d'expression. Par exemple, ce type de langage est proposé par (Ubeda et Egenhofer, 1997). Suivant ce langage, une CI spatiale définit une relation spatiale topologique du modèle 9IM (Egenhofer et Herring, 1994) (par ex. inside) entre les objets de deux classes (Classe 1 et Classe 2), avec un qualificatif indiquant le nombre de fois que la relation doit être vraie (jamais, exactement *n* fois, etc.). La syntaxe générale est la suivante :

```
(Classe 1, Relation, Classe 2, Qualificatif)
```

Les *langages visuels* se basent uniquement sur des notations visuelles et graphiques pour la spécification des CI (Pizano et al., 1989 ; Salehi et al., 2007). Ces notations servent à décrire une scène spatialement dans le but de représenter les contraintes entre les objets. Des pictogrammes ou des icônes peuvent aussi être utilisés pour schématiser les objets. Ces langages présentent en revanche

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

d'importantes limites qui sont liées à : leur manque d'expressivité (un langage visuel ne peut exprimer qu'un nombre réduit de CI), la taille des expressions (les langages visuels nécessitent un grand espace - une à plusieurs pages) pour exprimer les contraintes contrairement aux autres langages (une à deux lignes), l'ambiguïté liée à différents contextes culturels et la difficulté d'apprentissage. La Figure 4 montre une CI avec un langage visuel – on cherche à exprimer ici qu'une mairie est dans les limites d'une ville. Les notations visuelles ici présentés sont proches de ce que l'on peut trouver dans les langages de requête visuel (Bonhomme et al., 1999).

Les *langages logiques* se basent sur des logiques formelles telles que la logique du premier ordre pour la spécification des CI (Hadzilacos et Tryfona, 1992). Par conséquent, ces langages fournissent une sémantique et syntaxe assez précises permettant d'éviter de mauvaises interprétations et des ambiguïtés des/dans les spécifications de CI. Ces langages présentent également un bon niveau d'expressivité relativement aux autres. L'inconvénient majeur de ces langages est leur faible lisibilité et leur difficulté d'écriture puisqu'ils nécessitent un certain niveau de connaissances mathématiques de la part des concepteurs et utilisateurs qui n'ont pas forcément cette connaissance.

Les langages hybrides sont des combinaisons des langages précédents (langages naturels, et/ou visuels et/ou logiques). Ces langages sont classés selon la composante dominante du langage (notations visuelles ou texte) en deux sous-catégories : langages hybrides visuels et langages hybrides naturels.

Dans *les langages hybrides visuels*, la partie dominante est constituée par des notations visuelles, essentiellement des symboles visuels. Ces langages combinent la lisibilité des langages visuels avec la richesse des langages naturels. Dans le domaine des bases de données spatiales, les exemples de cette catégorie sont les modèles conceptuels spatio-temporels MADS (Parent et al., 1999) et PVL (Perceptory) (Bédard et al., 2004).

Bien que ces langages/modèles améliorent la lisibilité des spécifications vis-à-vis des utilisateurs, ils ne permettent d'exprimer qu'un très petit nombre de contraintes. De plus dans le cas d'applications complexes comme par exemple celles du SOLAP, qui présentent beaucoup de contraintes, l'ajout des spécifications de contraintes au modèle de données dégrade sa lisibilité.

Les langages hybrides naturels combinent un langage naturel qui constitue la partie dominante avec des pictogrammes ou des symboles.

Un *langage hybride naturel avec pictogrammes* est un langage naturel enrichi avec un ensemble de pictogrammes. Les pictogrammes sont des symboles visuels intuitifs (c'est-à-dire un pictogramme se caractérise par sa ressemblance picturale à la notion du monde réel qu'il représente), concis et "standardisés" utilisés pour faciliter la modélisation des données et des CI.

Langages hybrides naturels avec symboles combinent un langage naturel contrôlé (partie dominante) avec des symboles visuels spécifiques (autres que des pictogrammes). Les exemples de ces langages sont le langage standard de

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

contraintes objet OCL développé par l'OMG, les langages ad hoc proposés par (Ghozzi *et al.*, 2003a) et par (Hurtado et Mendelzon, 2001 ; Hurtado *et al.*, 2005).

OCL (Object Constraint Language) est un langage formel pour l'expression de contraintes et de requêtes sur des diagrammes UML (diagrammes de classes, états-transitions, etc.) (OMG, 2006). Ce standard adopté par l'OMG présente de nombreux avantages (OMG, 2006 ; Chimiak-Opoka et Lenz, 2007). C'est un langage qui combine la facilité d'expression et la lisibilité des langages naturels avec la non ambiguïté amenée par les langages formels. Une extension de OCL pour les données spatiales a été proposée dans (Pinet, 2007).

4.3 Etude des des CI dans les EDS

4.3.1 Données

La problématique d'incohérence logique des données dans les systèmes d'ED(S) est due principalement à la phase d'intégration des sources données (Salehi, 2009). En effet, en raison des différentes hétérogénéités qui peuvent exister entre les sources de données, de nombreuses incohérences peuvent être rencontrées lors de cette phase importante du processus d'entreposage. Par exemple, dans une application d'analyse des ventes, après intégration des sources de données, on peut avoir des géométries de villes qui se chevauchent; des géométries qui ne sont pas dans leurs domaines de définition, etc.

Quelques travaux traitant de l'incohérence logique ont été menés concernant la définition des contraintes d'intégrité dans les ED (Carpani et Ruggia, 2001 ; Ghozzi et al., 2003a ; Pinet et Schneider, 2009 ; Turki et al., 2010) et EDS (Malinowski et Zimányi, 2008 ; Salehi, 2009 ; Pinet et Schneider, 2010) (Song et al., 2001).

Dans le domaine des EDS, (Malinowski et Zimányi, 2008) proposent le modèle conceptuel spatio-multidimensionnel ER, Spatial MultiDimER. Ce modèle utilise les pictogrammes du modèle MADS pour la spécification de contraintes d'intégrité spatiale simples. Plus exactement, les pictogrammes du modèle MADS sont utilisés pour (a) représenter la géométrie d'un niveau d'agrégation spatial ou d'une mesure spatiale (CI de domaine de valeurs), (b) exprimer une relation spatiale topologique (par ex. coveredBy) entre les géométries de deux niveaux d'agrégation spatiaux reliés par une relation d'agrégation dans une hiérarchie de dimension spatiale et (c) enfin spécifier une relation spatiale topologique entre des niveaux d'agrégation spatiaux feuille de plusieurs dimensions spatiales afin de définir un fait spatial. Les relations spatiales topologiques considérées sont les relations définies par le modèle 9IM.

Les auteurs proposent également un mapping de ce modèle vers le modèle logique Objet-Relationnel et une implémentation manuelle des contraintes sous forme de triggers, vues et fonctions SQL dans le SGBD spatial Oracle Spatial 10g. Cependant, en plus des CI d'agrégation sémantiques (condition de compatibilité de types (Lenz et Shoshani, 1997)) et d'exploration, beaucoup de types de CI de données ne sont pas considérés, par exemple les CI qui portent sur les attributs thématiques, les relations entre attributs thématiques et spatiaux, les instances de

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

fait, etc. De plus, le langage visuel proposé (ensemble de pictogrammes) ne peut spécifier que les types de CI spatiales décrites précédemment.

(Salehi, 2009) proposent un modèle conceptuel ad hoc pour les hypercubes de données spatiales, une classification détaillée des CI des hypercubes spatiaux et un langage hybride naturel avec pictogrammes pour la spécification de ces CI au niveau conceptuel. La catégorisation de CI proposée considère les CI de données et d'agrégeabilité. Les contraintes de données définies peuvent contraindre les données dimensionnelles (CI traditionnelles) ou les données factuelles (CI de faits). Les CI sur les dimensions considérées peuvent porter sur un ou plusieurs attributs d'un membre de dimension (CI Inter 0 et CI Inter 1 respectivement), plusieurs membres (CI Inter 2), plusieurs niveaux d'agrégation (CI Inter 3), ou plusieurs dimensions (CI Inter 4). Les CI de faits peuvent contraindre une instance de fait (CI F-Inter 0), plusieurs instances de fait d'une seule hypercellule[1] (CI F-Inter 1), ou plusieurs instances de fait de plusieurs hypercellules (CI F-Inter 2). Ces différents types de CI sont modélisés au niveau conceptuel en utilisant un langage hybride naturel avec pictogrammes.

Ce travail comme les précédents ne propose pas d'implémentation pour les CI spécifiées. De plus, la catégorisation de CI proposée ne prend pas en compte les CI de données impliquant plusieurs hypercubes et les CI d'exploration. Finalement, le langage de spécification de CI proposé n'est pas basé sur des langages standards, ce qui rend l'implémentation difficile. A travers quelques exemples et en se basant sur le modèle multidimensionnel UML qu'ils proposent, (Pinet et Schneider, 2009 ; Pinet et Schneider, 2010) montrent comment les langages de spécification de CI OCL et Spatial OCL peuvent être utilisés pour exprimer des CI d'ED et d'EDS au niveau conceptuel. Ces contraintes sont définies pour garantir la cohérence logique des données intégrées dans l'entrepôt à partir de plusieurs sources hétérogènes. Différents détails sur les cadres d'implémentation des CI sous la forme de requêtes ou triggers sont discutés dans (Pinet, 2010)

Les travaux de (Pinet et Schneider, 2009 ; Pinet et Schneider, 2010) ne proposent ni classification de CI, ni implémentation pour les contraintes spécifiées. De plus, les auteurs ne considèrent pas les CI d'agrégation et d'exploration. Tous les travaux ci-dessus utilisent ou proposent des langages de spécification de CI en plus des langages de définition des structures de données multidimensionnelles; ceci permet d'exprimer un plus grand nombre de types de contraintes et d'éviter la complexification du modèle de données en séparant la spécification des CI de la spécification des structures de données. Contrairement aux travaux ci-dessus, (Pestana et al., 2005 ; Glorio et Trujillo, 2008) se basent seulement sur les modèles de données pour spécifier les CI. Ces modèles n'expriment que des contraintes d'intégrité simples, c'est-à-dire des contraintes de domaine de valeurs des attributs (par ex. géométries des niveaux d'agrégation spatiaux et mesures spatiales) et des contraintes de multiplicité des relations d'agrégation et de fait-dimension.

[1] Une hypercellule définit le schéma commun à un ensemble d'instances de fait agrégées ou detaillée.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

(Boulil *et al.*, 2013) s'appuyant sur un le profil UML ICSOLAP pour la définition des contraintes sur les données en utilisant par exemple le modèle de la Figure 5, les auteurs permettent d'exprimer des contraintes sur les hiérarchies. Le modèle conceptuel de la Figure 5 permet l'analyse spatio-multidimensionnelle de valeurs d'humidité et de précipitation. En particulier, le fait est représenté par la classe « EnvFactors » stéréotypé « Fact » qui contient deux mesures numériques: « Humidity » (humidité) et « Precipitation » (pluviométrie). Ces mesures sont analysées selon deux dimensions : une dimension spatiale « Nodes » représentée avec un package stéréotypé « SpatialDimension » qui groupe les capteurs (« Node ») par parcelle (« Plot »; et une dimension temporelle « Time » avec les niveaux heure et jour.

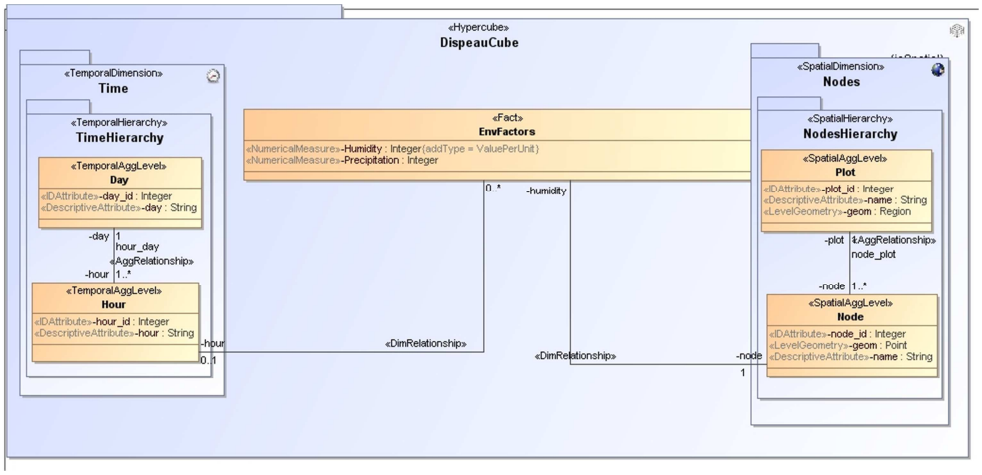

Figure 5. Modèle multidimensionnel d'analyse de l'humidité et de pluviométrie spécifié avec notre profil UML.

En utilisant Spatial OCL et le profil UML, nous pouvons facilement exprimer les CI de données hiérarchiques, comme par exemple « La géométrie d'un capteur doit être contenue dans celle d'une parcelle »

```
context Node inv DataIC1:
self.geom.isInside(self.Plot.geom) or
self.geom.coveredBy(self.Plot.geom)
```

Le tableau 1 résume les travaux existants pour les ED spatiaux en les évaluant selon les critères décrits précédemment.

Référence	**C1**	**C2**	**C3**	**C4**		**C5**	**C6**
				CI	**Modèle**		
(Glorio et Trujillo, 2008)	Non	Réduit	Conc	Oui (UML)	Oui (UML)	Oui	Non
(Malinowski et Zimányi,	Non	Réduit	Conc	Non (LHV)	Oui (E/R)	Oui	Non

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

2008)							
(Pinet et Schneider, 2010) (Pinet et Schneider, 2009)	Non	Moyen	Conc	Oui (OCL)	Oui (UML)	Non	Non
(Salehi, 2009)	Oui	Elevé	Conc	Non (LHN, LNC)	Non	Oui	Non
(Bouil et al, 2012)	Oui	Elevé	Conc	Oui (Spatial OCL, UML)	Oui (UML)	Oui	Oui
(Pestana et al., 2005)	Non	Réduit	Conc	Oui (UML)	Oui (UML)	Oui	Non

Tableau 1. Evaluation des travaux existants sur les CI pour la qualité de données dans les ED et les EDS. Conc : Conceptuel; Méta : Métamodèle; LN : Langage Naturel; LNC : Langage Naturel Controlé; LHN : Langage Hybride Naturel; LL : Langage Logique; LHV : Langage Hybride Visuel.

4.3.2 Agrégation

Les conditions d'agrégeabilité définies dans (Lenz et Shoshani, 1997) peuvent être regroupées en deux catégories : (a) des conditions structurelles (c'est-à-dire disjonction et de complétude, cf. Section 3.2) et des conditions sémantiques (condition de compatibilité de types, cf. Section 3.2).

4.3.2.1 Conditions structurelles d'agrégeabilité

Les conditions structurelles d'agrégeabilité (disjonction et complétude) dépendent (Mazón et al., 2009) :

- des types de hiérarchies de dimension (par ex. les relations non strictes peuvent engendrer le comptage en double des mesures);

- et des types de relations fait-dimension liant les faits aux dimensions (par ex. les relations fait-dimension incomplètes peuvent engendrer des agrégats incomplets).

(A) Classification des hiérarchies

Dans la littérature plusieurs typologies de hiérarchies ont été proposées (Mazón et al., 2009). Mais de façon général, par rapport à leur agrégeabilité structurelle, les hiérarchies sont classées en deux catégories (Mazón et al., 2009) : régulières et irrégulières.

Les hiérarchies régulières satisfont les deux premières conditions d'agrégeabilité à savoir la complétude et la disjonction. Ces hiérarchies sont strictes, onto et couvrantes.

Les hiérarchies strictes spécifient pour chaque membre feuille un seul chemin d'agrégation vers le membre racine de la hiérarchie. Dans les hiérarchies comportant un seul chemin conceptuel (par ex. la hiérarchie spatiale

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

Magasin<Ville<Département) cette condition se traduit par la présence d'exactement un membre père pour chaque membre non racine de la hiérarchie (une ville pour chaque magasin et un département pour chaque ville). Les hiérarchies strictes garantissent la première condition d'agrégeabilité (c'est-à-dire la disjonction). Un exemple de hiérarchie non-stricte est illustrée précédemment en Figure 3, Capteur<Parcelle<Commune<Département, où chaque parcelle peut être rattachée à plus d'une commune.

Dans *les hiérarchies onto*, tout membre non feuille possède au moins un membre fils et tous les membres feuille se trouvent au même niveau qui est le niveau d'agrégation feuille de la hiérarchie. Ces hiérarchies sont appelées également hiérarchies équilibrées (Malinowski et Zimányi, 2008). Par exemple, la hiérarchie spatiale précédente Magasin<Ville<Département serait également une hiérarchie onto si pour chaque ville on a au moins un magasin.

Dans *les hiérarchies couvrantes,* chaque membre non racine a au moins un membre père, qui de plus ne saute pas des niveaux d'agrégation, autrement dit, il appartient au niveau d'agrégation qui est juste au-dessus de celui de ses fils. Dans la littérature ces hiérarchies sont également connues sous les appellations non ragged et non levels-skipped (Mazón et al., 2009). La hiérarchie précédente Magasin<Ville<Département, serait couvrante si pour chaque magasin on a une ville.

Les hiérarchies irrégulières ne satisfont pas l'une ou les deux conditions d'agrégeabilité décrites précédemment (disjonction et complétude), induisant les problèmes de comptage en double de mesures et/ou d'agrégats incomplets. Il s'agit donc de hiérarchies : non-strictes, et/ou non-couvrantes et/ou non-onto.

(B) Classification des relations fait-dimension

Relativement à leur agrégeabilité structurelle et similairement aux hiérarchies, les relations fait-dimension sont classées également en deux catégories, régulières et irrégulières.

Une relation fait-dimension régulière est une relation *stricte* (chaque instance de fait est liée à au maximum un seule membre de la dimension) et *complète* (chaque instance de fait est liée à au minimum un seule membre de la dimension). Au niveau du schéma conceptuel, elle peut être représentée dans le formalisme UML par une association un à plusieurs entre la classe de fait et la dimension avec une cardinalité un à un du côté de la dimension (la cardinalité de participation des instances de fait).

A l'inverse, une relation *fait-dimension irrégulière* est une relation non-stricte et/ou non complète.

(C) Catégorisation des approches de résolution des problèmes d'agrégeabilité structurelle

Concernant les conditions d'agrégeabilité structurelles, les approches de modélisation multidimensionnelle conceptuelle peuvent être classées en 4 catégories :

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

Catégorie A. Ces approches garantissent l'agrégeabilité en interdisant la présence des structures multidimensionnelles complexes problématiques pour l'agrégeabilité dans les modèles conceptuels. L'inconvénient de ce type d'approches est que de nombreuses situations réelles (par ex. hiérarchies complexes) sont ignorées dans la modélisation. La conséquence de cela est que des analyses pertinentes peuvent être perdues.

Dans cette catégorie, nous retrouvons par exemple le modèle ad hoc DFM (Golfarelli *et al.*, 1998) et les travaux qui utilisent les Dépendances Fonctionnelles pour garantir ces conditions structurelles d'agrégeabilité (Lehner *et al.*, 1998 ; Niemi *et al.*, 2001 ; Lechtenbörger et Vossen, 2003). Ces travaux étendent les DF et Formes Normales des bases de données relationnelles aux bases de données multidimensionnelles.

Catégorie B. Ces approches proposent des modèles conceptuels qui représentent toutes ou quelques-unes des structures multidimensionnelles complexes mais ne fournissent ni solutions pour déterminer l'agrégeabilité de ces structures (c'est-à-dire déterminer si ces structures respectent ou non les conditions structurelles d'agrégeabilité), ni solutions pour résoudre les problèmes d'agrégeabilité qui peuvent être induits par ces structures.

Les travaux de cette catégorie sont par exemple les modèles multidimensionnels conceptuels proposés dans (Abelló *et al.*, 2006 ; Lujan-Mora *et al.*, 2006) et les modèles spatio-multidimensionnels conceptuels proposés dans (Glorio et Trujillo, 2008 ; Pinet *et al.*, 2010 ; Salehi *et al.*, 2010; Boulil *et al.*, 2011). Dans (Boulil *et al.*, 2011) nous exprimons les conditions structurelles d'agrégeabilité sous la forme de contraintes de multiplicité UML et de contraintes OCL d'une façon similaire à (Mazón *et al.*, 2009 ; Pinet et Schneider, 2010), mais leur implémentation n'est pas traitée.

Catégorie C. Ces travaux proposent des modèles qui permettent la représentation conceptuelle des structures multidimensionnelles complexes et proposent des algorithmes pour déterminer l'agrégeabilité de ces structures. Ces algorithmes en revanche ne permettent pas de rendre ces structures agrégeables.

Les principaux travaux de cette catégorie sont (Hurtado et Mendelzon, 2001 ; Hurtado et al., 2005 ; Pinet et Schneider, 2009 ; Pinet et Schneider, 2010); ils sont décrits ci-dessous.

Dans les EDS, (Pinet et Schneider, 2009; Pinet et Schneider, 2010) expriment avec OCL et UML différents types de contraintes d'agrégeabilité structurelles. Les contraintes exprimées peuvent porter sur une ou plusieurs relations d'agrégation et même sur plusieurs chemins d'agrégation. Ces contraintes sont aussi utilisées pour caractériser différents types de hiérarchies (strictes ou non strictes, couvrantes ou non couvrantes, etc.). Ce travail propose un algorithme pour tester l'agrégeabilité d'une hiérarchie de dimension en présence de ces contraintes.

Catégorie D. Les travaux de cette catégorie proposent des modèles qui tolèrent la présence des structures multidimensionnelles irrégulières et des algorithmes et mappings pour transformer ces structures en structures agrégeables. Les principaux

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

travaux dans cette catégorie sont décrits comme suit. Dans les EDS, (Malinowski et Zimányi, 2008) propose une classification qui définit différents types de hiérarchies de dimension complexes. Ces hiérarchies sont représentées au niveau conceptuel en utilisant le modèle (Spatial) MultiDimER. Les auteurs définissent également de manière informelle un ensemble de mappings de ces hiérarchies conceptuelles vers les modèles logiques Relationnel et Objet-Relationnel permettant leur implémentation dans les SGBD existants en garantissant les conditions structurelles d'agrégeabilité. Contrairement à (Pedersen *et al.*, 1999), ce travail en revanche ne définit pas l'ordre d'application de ces mappings dans le cas où la hiérarchie de dimension présente plusieurs irrégularités structurelles (par ex. non stricte et non onto).

Dans le domaine des EDS, (Pedersen et Tryfona, 2001) analysent l'influence des relations spatiales topologiques du modèle 9IM sur l'additivité des mesures métriques (par ex. "surface" de la zone épandue (cf. Section 3.1)) et montrent que celle-ci dépend de la disjonction des objets spatiaux. Cette étude propose une approche pour transformer l'EDS et garantir l'agrégeabilité dans le cas où cette condition n'est pas vérifiée (les objets ne sont pas disjoints). Cette approche consiste à calculer les parties disjointes avec les valeurs de mesure associées et à normaliser les hiérarchies de dimension en appliquant les transformations présentées (Pedersen *et al.*, 1999). (Jensen *et al.*, 2004) proposent une méthode pour évaluer l'imprécision des chemins d'agrégation engendrée par les relations d'inclusion partielle entre membres de dimension spatiaux. Cette méthode permet de choisir le chemin le plus précis parmi plusieurs alternatives lors de l'exécution des requêtes. Les auteurs adaptent également les algorithmes de transformation des hiérarchies irrégulières (non strictes, non onto et non couvrantes) proposés dans (Pedersen et Tryfona, 2001) pour supporter ces relations d'inclusion partielle.

Tous les travaux ci-dessus dans cette catégorie se focalisent sur les hiérarchies de dimension irrégulières. Pour résoudre les problèmes d'agrégeabilité dus aux relations fait-dimension irrégulières (non strictes et incomplètes), (Mazón *et al.*, 2008) propose des transformations pour convertir un modèle multidimensionnel conceptuel non agrégeable (c'est-à-dire qui contient ces relations fait-dimension irrégulières) en un modèle agrégeable. Ces transformations sont exprimées en utilisant le langage QVT et implémentées par un module sous Eclipse. Enfin, tous ces algorithmes et techniques de transformation de structures multidimensionnelles qui sont proposées par les travaux de cette catégorie engendrent une modification et complexification de l'analyse OLAP (Mazón *et al.*, 2009) notamment par l'insertion de valeurs artificielles qui n'ont pas de sens pour les décideurs.

4.3.2.2 Conditions sémantiques d'agrégeabilité

Les conditions sémantiques d'agrégeabilité se rapportent essentiellement à la compatibilité entre les natures sémantiques et types de la mesure, de la fonction d'agrégation et de la hiérarchie de dimension. Dans les processus SOLAP, cette compatibilité dépend de plusieurs paramètres : (a) l'additivité de la mesure (stock, flux ou valeur par unité), (b) la nature de la hiérarchie (son type de structuration (par ex. stricte ou non stricte) décrit précédemment en Section 3.2.1 et son caractère

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

temporel ou pas), (c) le type de fonction d'agrégation (applicable à des données pouvant être additionnées, etc.) (Pedersen et al., 2001), (d) la distributivité de la fonction d'agrégation, et (e) les relations spatiales topologiques entre les membres spatiaux (pour une hiérarchie de dimension spatiale)(Pedersen et Tryfona, 2001).

Concernant cette agrégeabilité sémantique, certains travaux identifient seulement les différentes typologies de mesures, de fonctions d'agrégation et de dimensions et les incompatibilités entre ces typologies sans proposer de solutions pour les représenter au niveau conceptuel ou les implémenter ((Lenz et Shoshani, 1997 ; Shekhar et al., 2001)). Ainsi par rapport à leur additivité[2] selon le temps, les mesures numériques ont été classées dans (Lenz et Shoshani, 1997) en trois types : (a) mesures de type flux, qui peuvent être sommées selon toutes les dimensions; (b) mesures de type stock, qui ne peuvent pas être sommées selon le temps (par ex. la "population"), et (c) mesures de type valeur par unité, qui ne peuvent pas être sommées selon toutes les dimensions (par ex. les prix unitaires de vente). Dans (Pedersen et al., 2001), les fonctions d'agrégation numériques ont été classées en fonction des catégories de données auxquelles elles peuvent s'appliquer en trois types également: applicables aux données pouvant être additionnées (Sum, Count, Avg, Min, et Max), applicables aux données pouvant être moyennées (Count, Avg, Min, Max), et applicables aux données pouvant être uniquement comptées (Count). Par rapport aux types de données auxquels les fonctions d'agrégation SOLAP peuvent être appliquées, les catégories suivantes ont été identifiées dans la littérature :

- Fonctions d'agrégation numérique qui s'appliquent seulement aux mesures de type numérique (par ex. Sum et Avg).
- Fonctions d'agrégation spatiale qui s'appliquent seulement aux données de type géométrique (Point, Polygone, etc.), par ex. l'union spatiale, le centroïde, et l'équipartition (Shekhar et al., 2001;Silva et al., 2008; Ruiz et Times,2009).
- Fonctions d'agrégation booléenne qui s'appliquent seulement aux données booléennes (par ex. Or et And) (Golfarelli et al., 1998).
- Fonctions d'agrégation textuelle qui s'appliquent seulement aux données textuelles (par ex. Topic et Top Keywords (Ravat et al., 2008).
- Fonctions d'agrégation temporelle qui s'appliquent aux données de type temporel (par ex. Instant ou Intervalle de dates), par exemple l'union temporel.
- Fonctions d'agrégation génériques qui s'appliquent à plusieurs types de données (Ravat et al., 2008).

[2] Le mot additivité est particulièrement utilisé pour désigner la possibilité d'agréger la mesure par la fonction Sum [Horner et al., 2004).

Suivant leur distributivité, les fonctions d'agrégation SOLAP (numériques, spatiales ou autres) sont classées de manière formelle en trois catégories (Shekhar et al., 2001) : (a) distributives (par ex. Sum, Union spatiale, etc.), (b) algébriques (par ex. Avg, Centroid, etc.) et (c) holistiques (par ex. Mode, Equipartition, etc.). Les fonctions distributives permettent de réutiliser les agrégats d'un niveau d'agrégation pour calculer des agrégats corrects à un niveau d'agrégation supérieur. Les fonctions algébriques sont des expressions algébriques finies de fonctions distributives. Ces fonctions requièrent donc une manipulation supplémentaire pour pouvoir réutiliser correctement des agrégats. Finalement, les fonctions holistiques qui ne sont ni distributives ni algébriques, nécessitent un recalcul total en utilisant les données les plus fines de l'EDS.

D'autres travaux proposent des solutions pour exprimer ces incompatibilités au moyen de contraintes dans les modèles conceptuels. Ces travaux utilisent différents langages. Mais en règle générale, ces travaux ne fournissent pas d'implémentations et ne donnent aucune indication sur la façon dont ces contraintes peuvent être intégrées dans les outils (S)OLAP pour contrôler l'agrégation. Ces travaux sont décrits comme suit. Certains modèles conceptuels d'ED et d'EDS (Tryfona *et al.*, 1999 ; Malinowski et Zimányi, 2008) ne représentent que le type additif de la mesure. Dans le modèle multidimensionnel ER de (Tryfona *et al.*, 1999), le type additif de la mesure (Flux, Stock ou Valeur par unité) est indiqué par un langage naturel contrôlé (cf. Section 2.3) : une lettre F, S ou V respectivement. Le modèle Spatial MultiDimER de (Malinowski et Zimányi, 2008) représente le type de mesure par des symboles visuels (langage visuel). Les modèles proposés par (Nguyen *et al.*, 2000 ; Ravat *et al.*, 2008) permettent d'exprimer seulement des incompatibilités entre les mesures et les fonctions d'agrégation sans considérer les dimensions. Pour ce faire ces modèles intègrent la fonction d'agrégation dans la définition de la mesure. Par exemple, (Nguyen *et al.*, 2000) propose un modèle conceptuel UML où la mesure est définie par son nom et sa fonction d'agrégation. Les travaux (Golfarelli *et al.*, 1998 ; Abelló *et al.*, 2006 ; Lujan-Mora *et al.*, 2006) spécifient seulement des contraintes d'agrégation qui définissent la ou les fonctions d'agrégation qui peuvent être appliquées à une mesure ou un ensemble de mesures le long d'une ou plusieurs dimensions sans considérer les hiérarchies de dimension. Dans les ED, le modèle DFM proposé par (Golfarelli *et al.*, 1998) représente les fonctions d'agrégation qui peuvent être appliquées à une mesure le long d'une dimension en utilisant un langage hybride visuel (cf. Section 3.2.1). Le modèle YAM2 (Abelló *et al.*, 2006) modélise les règles d'agrégation de mesures le long des dimensions sous forme d'opérations UML stéréotypées « Summarization ». Chaque règle précise une fonction d'agrégation pour une mesure et un ensemble de dimensions. YAM2 définit également des contraintes qui spécifient qu'un niveau d'agrégation est une source invalide pour l'agrégation (les agrégats à ce niveau ne peuvent être réutilisés pour calculer les agrégats des niveaux supérieurs dans la hiérarchie de dimension). (Lujan-Mora *et al.*, 2006) modélise des contraintes sur l'additivité des mesures par des notes UML. Ces contraintes spécifient seulement les dimensions le long desquelles les mesures ne peuvent être sommées (fonction Sum).

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

(Bimonte *et al.*, 2009) proposent un modèle spatio-multidimensionnel qui prend en compte les dépendances entre fonctions d'agrégation spatiales et alphanumériques lors de l'agrégation des mesures spatiales. Pour contrôler la qualité de l'agrégation alphanumérique, ce modèle formalise deux contraintes qui spécifient pour chaque mesure alphanumérique le type de fonctions d'agrégation possible : la contrainte d'agrégation verticale considère le type de la mesure (par ex. additive ou pas) et de la dimension utilisée (temporelle ou pas); et la contrainte d'agrégation horizontale considère le type de mesure et la fonction d'agrégation spatiale utilisée pour agréger la géométrie. Ces contraintes sont spécifiées par des fonctions au niveau métamodèle.

Les travaux de (Prat *et al.*, 2010) et (Salehi, 2009) proposés dans le contexte des ED et EDS respectivement sont les travaux qui représentent le plus de contraintes d'agrégation en relation avec l'agrégeabilité sémantique. En se basant sur un modèle multidimensionnel UML, (Prat *et al.*, 2010) spécifient différentes contraintes d'agrégation sémantiques en utilisant le standard OMG de représentation des règles de production PPR (Production Rule Representation language) pour garantir des agrégations correctes des mesures. Ces règles expriment entre autres les fonctions d'agrégation numériques applicables aux mesures le long des dimensions, l'ordre d'application des fonctions et leur distributivité. Ce travail ne propose pas d'implémentation à ces contraintes.

(Salehi, 2009) définit six catégories de CI d'agrégeabilité. Chaque catégorie est utilisée pour spécifier une interdiction d'un type de combinaisons lors d'une agrégation. Par exemple, les CI de catégorie M-D-AF IC interdisent l'agrégation d'une mesure avec une fonction d'agrégation le long d'une dimension.

(Boulil et al., 2013) permet la définition des contraintes d'agrégation en utilisant le profil UML. Par exemple en utilisant l'application présentée en Figure 5, il est possible définir une contrainte d'agrégation comme montré en Figure 6. Ce modèle définit un indicateur d'analyse, « Avghumidity » (moyenne de l'humidité) ; qui est composé d'une seule règle d'agrégation (stéréotype «AggRule»). Cette règle indique que cet indicateur est calculé en agrégeant la mesure « Humidity » (valeur marquée aggregatedAttribute de l'indicateur) avec la fonction Avg (valeur marquée aggregator de la règle d'agrégation).

Le type d'additivité de la mesure « Humidity » est « valeur par unité » ce qui signifie que son agrégation avec la somme selon toutes les dimensions n'as pas de sens et par conséquent interdite par les contraintes d'agrégation de notre profil.

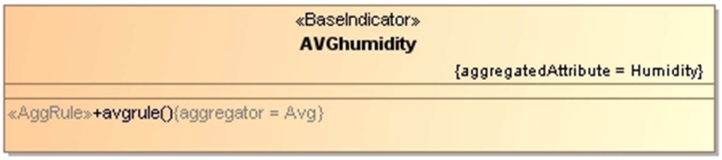

Figure 4. Exemple d'indicateur d'analyse spécifié avec notre profil UML.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

Les CI d'agrégation sont implémentées au niveau métamodèle comme des contraintes OCL du profil UML et permettent la définition des règles d'agrégation (« AggRule ») correctes. Par exemple pour éviter de sommer des mesures de type valeur par unité, la contrainte OCL suivante a été implémentée :

```
context AggRule inv notSumValuePerUnitMeasure :
if
(baseIndicator.aggregatedAttribute.OclIsKindOf(Measure)
and baseIndicator.aggregatedAttribute.addType =
'ValuePerUnit')
then aggregator.name <> 'Sum'
```

Le tableau 2 résume les travaux existants les ED spatiaux en les évaluant selon les critères décrits précédemment. Nous montrons dans la colonne Contrainte si les travaux s'intéressent aux contraintes structurelles (Struc) et/ou sémantique (Sem).

Référence	**Contrainte**	**C1**	**C2**	**C3**	**C4**		**C5**	**C6**
					CI	**Modèle**		
(Bimonte et al., 2009)	Sem	Non	Réduit	Méta	Non (LN)	Non	Non	Non
(Salehi, 2009)	Oui	Oui	Elevé	Conc	Non (LHN, LNC)	Non	Oui	Non
(Bouil et al, 2012)	Struc Sem	Oui	Elevé	Conc	Oui (Spatial OCL, UML)	Oui (UML)	Oui	Oui
(Glorio et Trujillo, 2008)	Struc Sem	Non	Réduit	Conc	Oui (UML)	Oui (UML)	Oui	Non
(Pinet et Schneider e al., 2010)	Struc	Non	Moyenne	Conc	Oui (UML)	Oui (UML)	Oui	Non
(Malinowsky et Zimanyi, 2008)	Struc Sem	Non	Réduit		Non (LHV)	Oui (E/R)	Oui	Non
(Pedersen et Tryfona, 2001)	Struc Sem	Non	Réduit	Méta	Non	Non	Non	Non

Tableau 2. Evaluation des travaux existants sur les CI pour la qualité d'agregation dans les ED et les EDS. Conc : Conceptuel; Méta : Métamodèle; LN : Langage Naturel; LNC : Langage Naturel Controlé; LHN : Langage Hybride Naturel; LL : Langage Logique; LHV : Langage Hybride Visuel.

4.3.3 Exploration

La qualité d'exploration a été principalement traitée dans le domaine des processus OLAP par les travaux de (Sapia, 1999 ; Böhnlein *et al.*, 2002) et dans le domaine du SOLAP par (Levesque *et al.*, 2007). De manière générale ces travaux ne proposent ni langages pour la spécification conceptuelle des contraintes d'exploration, ni de langages pour leur implémentation.

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

Dans (Boulil et al., 2013), les concepteurs peuvent aussi exprimer des CI sur les requêtes SOLAP en utilisant la partie méta modèle des CI sur les requêtes SOLAP du profil. En règle générale, une requête SOLAP est une combinaison de mesures et de membres de différentes dimensions. Ainsi, notre méta modèle de CI sur les requêtes peut être utilisé par exemple pour interdire la définition des combinaisons non valides. Dans les contraintes, ces ensembles de membres sont spécifiés par des attributs stéréotypés «MemberSet». Le domaine de valeurs d'un attribut stéréotypé «MemberSet» est un sous-ensemble des membres d'un niveau de dimension. Ce domaine peut être précisé en utilisant une expression OCL qui sélectionne les membres du niveau impliqué dans la contrainte de requêtes SOLAP (valeur marquée condition).

Un exemple de notre représentation UML des contraintes sur les requêtes est montré à la figure 7. Dans cette contrainte, on précise que cela n'a pas de sens de combiner dans une requête SOLAP, le capteur « node1 » et toutes les heures du 21 Mars 2012 (OCL de la balise Hourset20120321 : self.hour = '21/03/2012 00H' or self.hour = '21/03/2012 01H' ... or self.hour = '21/03/2012 24H') car ce capteur était en panne a ce moment.

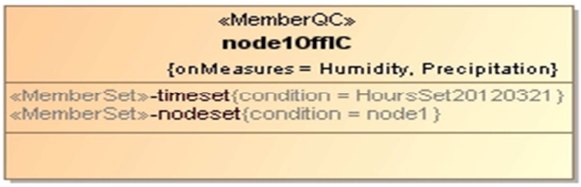

Figure 5. Exemple de CI de requête spécifié avec notre profil UML.

Finalement, ce cadre conceptuel de spécification des CI SOLAP est implémenté dans un outil qui prend en entrée le fichier XMI du modèle conceptuel UML et celui des CI Spatial OCL et automatiquement génère des requêtes et triggers SQL pour les contraintes de données, et des formules MDX pour les contraintes de requêtes (les contraintes d'agrégation sont gérées dans l'atelier de génie logiciel et donc au niveau de la conception du modèle spatio-multidimensionnel).

(Levesque et al., 2007) présentent un framework pour la visualisation des risques associés à la mauvais interprétation des résultats de requêtés SOLAP. Ils classent ces risques dans différentes classes.

Le tableau 3 résume les travaux existants les ED spatiaux en les évaluant selon les critères décrits précédemment.

Référence	**C1**	**C2**	**C3**	**C4**		**C5**	**C6**
				CI	**Modèle**		
(Levesque et al., 2007)	Non	Réduit	LHV	non	non	Oui	non
(Bouil et al, 2012)	Oui	Elevé	Conc	Oui (Spatial OCL,	Oui (UML)	Oui	Oui

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

				UML)			

Tableau 3. Evaluation des travaux existants sur les CI pour la qualité d'exploration dans les ED et les EDS. QD : Qualité de Données; QA : Qualité d'Agrégation; QE : Qualité d'Exploration; Conc : Conceptuel; Méta : Métamodèle; LN : Langage Naturel; LNC : Langage Naturel Controlé; LHN : Langage Hybride Naturel; LL : Langage Logique; LHV : Langage Hybride Visuel.

4.3.4 Discussion

Comme le montre les tableaux 1, 2 et 3, à l'exception de notre proposition, aucun travail ne considère tous les trois types de qualité : qualité de données (QD), qualité d'agrégation (QA) et qualité d'exploration (QE). L'évaluation du troisième critère (C2) montre que (Salehi, 2009) est le travail qui considère le plus grand nombre de types de contraintes concernant les qualités de données et d'agrégation. La catégorisation de CI proposée néanmoins nécessite d'être étendue par d'autres types de CI comme les CI de données sur plusieurs hypercubes, les CI d'exploration et les CI de métadonnées. A l'exception de (Sapia, 1999 ; Böhnlein *et al.*, 2002 ; Bimonte *et al.*, 2009 ; Prat *et al.*, 2010), tous les travaux définissent les CI au niveau conceptuel (C3), c'est-à-dire sur des modèles conceptuels. La plupart des travaux expriment les contraintes en utilisant des langages non standards (logiques, naturels, visuels ou hybrides) qui sont difficiles à implémenter (C4/CI). Dans le domaine des EDS, seul (Pinet et Schneider, 2010) utilise un langage standard de spécification de CI au niveau conceptuel (OCL) mais n'exprime que quelques types de CI de données (C2). Dans les ED, seulement (Prat *et al.*, 2010) utilisent un langage standard, mais pour exprimer que certains types de CI au niveau métamodèle (C2 et C3). Les travaux de (Abelló *et al.*, 2006 ; Glorio et Trujillo, 2008) se basent sur des profils UML et n'exploitent pas des langages dédiés à la spécification des CI; par conséquent ils n'expriment que très peu de contraintes (C2). (Abelló *et al.*, 2006) ne modélise qu'un type de CI d'agrégation (les fonctions d'agrégation applicables aux mesures le long des dimension) et (Glorio et Trujillo, 2008) spécifie des CI de données très simples (CI de domaine de valeurs des attributs et de multiplicité). Enfin, aucun travail ne propose une implémentation automatisée des CI (S)OLAP (C5).

Notre proposition représentée en dernière ligne du tableau satisfait tous les critères retenus :

Nous considérons les trois types de qualité et nous proposons une classification des CI SOLAP (critères C1 et C2) suivant les quatre composantes fondamentales des systèmes SOLAP (métadonnée, donnée, agrégation, exploration ou requête). Notre classification définit quatre grandes familles de CI : (a) de métadonnées, (b) de données, (c) d'agrégation et (d) d'exploration qui vérifient respectivement la cohérence logique des métadonnées des différentes sources de données intégrées, des données entreposées, des agrégations des mesures le long des hiérarchies de dimension, et des requêtes SOLAP. Elle réorganise et étend la classification de (Salehi, 2009) essentiellement par les CI de requêtes, de métadonnées et données impliquant plusieurs hypercubes. Nous proposons des définitions conceptuelles (C3) des CI et du modèle multidimensionnel basées sur les langages standards UML et

Author-produced version of the article published in Revue Internationale de Géomatique, 2016, 26 (1), 97-131.
The original publication is available at http://www.sciencedirect.com/
doi : 10.3166/RIG.26.97-131

OCL (C4). Pour les contraintes de requêtes nous proposons une représentation sous forme de diagrammes de classes UML (C5), avec des contraintes Spatial OCL pour exprimer des sélections complexes de membres. En effet, les langages visuels offrent plus de lisibilité mais en revanche nécessitent beaucoup d'espace pour exprimer les contraintes. Enfin, nous proposons une implémentation automatique (C6) du modèle multidimensionnel et des CI en termes de langages d'implémentation standards (SQL et MDX) qui est basée sur des générateurs de code. Voir les travaux de (Pinet, 2010) et (Duboisset, 2007) pour plus de détails sur l'évaluation des stratégies de génération de code.

5. Conclusions et perspectives

Les systèmes d'Entrepôts de Données et OLAP spatiaux (EDS et SOLAP) sont des technologies d'aide à la décision permettant l'analyse multidimensionnelle de gros volumes de données spatiales. En réponse à des actions utilisateurs sur l'interface cliente (exploration), les systèmes SOLAP agrègent les données de l'EDS le long de différentes hiérarchies de dimension pour calculer des indicateurs d'analyse à différents niveaux de détails. La qualité des indicateurs d'analyse dépend donc de trois facteurs : la qualité des données entreposées, la qualité des agrégations et la qualité de l'exploration des données.

La qualité des données entreposées dépend de critères comme la précision, l'exhaustivité et la cohérence logique. La cohérence logique des données est généralement contrôlée par les contraintes d'intégrité qui définissent les conditions que les données doivent satisfaire. La qualité d'agrégation peut être ramenée à la cohérence logique entre les natures des éléments qui sont impliqués dans l'agrégation SOLAP (par ex., mesure, fonction d'agrégation). Cette cohérence d'agrégation est affectée par des problèmes structurels (par ex. les hiérarchies non strictes qui peuvent engendrer le comptage en double des mesures) et de problèmes sémantiques (par ex. sommer les valeurs d'humidité n'as pas de sens). La qualité d'exploration dépend essentiellement de la consistance des requêtes utilisateur (par ex. quelles ont été les valeurs de température en URSS en 2010 ?). Dans cette article nous avons étudié la notion de cohérence logique des données aux deux autres composantes fondamentales des systèmes SOLAP, à savoir l'agrégation et la requête. Nous avons présenté un état de l'art et des travaux sur la définition des CI pour garantir la cohérence logique au niveau des trois composantes (données, agrégations, et requêtes), ainsi qu'une évaluation de ces travaux et de notre proposition par rapport à un ensemble de critères que nous avons définis. Cette évaluation montre que notre proposition satisfait tous les critères contrairement aux autres.

Différentes pistes de recherche restent ouvertes.

Les travaux présentés dans cet article portent sur les problèmes d'incohérence logique qui peuvent se poser au niveau des données, des agrégations et des requêtes. Afin de mettre en place un système de gestion de qualité SOLAP qui tient compte de toutes les composantes de la qualité de données définies par exemple dans le standard ISO 19113, il serait intéressant d'élargir ces travaux aux autres types de

problèmes de qualité qui sont la non exhaustivité et l'imprécision. Cette perspective de recherche soulève d'abord des questionnements quant aux représentations conceptuelle et physique des données présentant ces types de problèmes (données imprécises et/ou non exhaustives), à leur agrégation (est-il par exemple nécessaire de redéfinir les fonctions d'agrégation SOLAP ?) et à leur interrogation (définition de langages de requêtes intégrant des prédicats d'imprécision et d'exhaustivité). Ensuite la question de la définition de langages pour la spécification et l'implémentation des contraintes sur ce type de données, de requêtes (e.g. requêtes vagues) et d'agrégations peut être investiguée. Pour l'imprécision par exemple ceci pourrait se faire en étendant des travaux effectués dans les bases de données spatiales pour la représentation des données spatiales vagues et des travaux dans les ED. Finalement, comme pour l'incohérence logique, il peut être intéressant d'étudier l'implémentation des contraintes de précision et d'exhaustivité tout au long du processus d'entreposage et d'analyse des données.

Dans cet article nous avons étudié les CI au niveau des couches EDS et Serveur SOLAP. Il serait intéressant d'étendre cette implémentation au niveau de l'ETL. Cette implémentation ETL nécessitera certainement la définition de modèles qui représentent les sources de données; ensuite des techniques de propagation de contraintes vers les premiers niveaux possibles de leur vérification dans le processus ETL pourraient être employées. Aussi, une étude plus détaillée pour déterminer quelles sont les couches SOLAP les plus adéquates pour l'implémentation de chaque type de CI serait une perspective intéressante. De la même façon, la prise en compte de ces contraintes au niveau du client SOLAP, nous semble une solution intéressante.

References

Abelló A., Samos J., Saltor F. (2006). YAM2: a multidimensional conceptual model extending UML. *Information Systems*, vol. 31, n°6, p. 541-567.

Akehurst D. H., Bordbar, B. (2001). On Querying UML Data Models with OCL. *Actes de 4th International Conference on The Unified Modeling Language, Modeling Languages, Concepts, and Tools*, p. 91-103.

Batini C. Scannapieca M. (2006). *Data quality: Concepts, methodologies and techniques*, Springer-Verlag, New York.

Bédard Y. (2009). site Web "Spatial OLAP." <http://www.spatialbi.com/>.

Bédard Y., Larrivee S., Proulx M. J., Nadeau M. (2004). Modeling geospatial databases with plug-ins for visual languages: A pragmatic approach and the impacts of 16 years of research and experimentations on perceptory. *Lecture Notes in Computer Science*, vol. 3289, p. 17-30.

Bimonte S., Boulil K., Pinet F., Kang M.A.: Design of Complex Spatio-multidimensional Models with the ICSOLAP UML Profile - An Implementation in MagicDraw. ICEIS (1) 2013: 310-315

Bimonte S., Villanova-Oliver M., Gensel J. (2009). A Multidimensional Model for Correct Aggregation of Geographic Measures. *Evolving Application Domains of Data Warehousing and Mining: Trends and Solutions*, IGI global, p.162-183

Böhnlein M., Plaha M., Ulbrich-vom Ende A. (2002). Visual Specification of Multidimensional Queries based on a Semantic Data Model. *Vom Data Warehouse zum Corporate Knowledge Center (DW)*, p. 379-397.

Bonhomme C., Trépied C., Laurini R. (1999). A Visual Language for Querying Spatio-Temporal Databases. ACM GIS'99.

Boulil K., Bimonte S., Pinet F., Carluer N., Lauvernet C., Cheviron B., Miralles A. Chanet J.-P. (2013). Guaranteeing the Quality of Multidimensional Analysis in Data Warehouses of Simulation Results: Application to Pesticide Transfer Data Produced by the MACRO Model. *Ecological Informatics*, vol. 16, p. 41-52.

Boulil K., Bimonte S., Mahboubi H., Pinet F. (2010). Vers la définition des contraintes d'intégrité d'entrepôts de données spatiales avec OCL. Revue des Nouvelles Technologies de l'Information B6, p. 121-136.

Boulil K., Bimonte S., Pinet F. (2011). Un modèle UML et des contraintes OCL pour les entrepôts de données spatiales. De la représentation conceptuelle à l'implémentation. *Ingénierie des Systèmes d'Information*, vol. 16, n°6, p. 11-39.

Boulil K., Bimonte S., Pinet F. (2012b). A UML & Spatial OCL based approach for handling quality issues in SOLAP systems. *Proceedings of the14th International Conference on Enterprise Information Systems (ICEIS 2012)*, Wroclaw, Poland.

Boulil K., Bimonte S., Pinet F. (2012c). Un cadre conceptuel basé sur UML et Spatial OCL pour la définition des contraintes d'intégrité dans les systèmes SOLAP. *Actes des 8èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 2012)*, Bordeaux, France.

Carpani F., Ruggia R. (2001). An Integrity Constraints Language for a Conceptual Multidimensional Data Model.

Chimiak-Opoka J. D., Lenz C. (2007). Use of OCL in a Model Assessment Framework: An experience report. *Electronic Communications of the EASST*, vol. 5.

Daniel, L (2005). Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering*, vol. 55, n° 3, p. 243-276.

Devillers R., Bedard Y., Jeansoulin R. and Moulin B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science*, vol. 21, n°3, p. 261-282.

Devillers R., Jeansoulin, R. (2005). *Qualité de l'information géographique*, Hermès Science Publications.

Duboisset M. (2007). *Un Système de Contraintes d'Intégrité OCL pour les Bases de Données Spatiales: Application à un Système d'Information pour l'Epandage Agricole*, Université Blaise Pascal.

Duboisset M., Pinet F., Kang M. A. and Schneider M. (2007). A general framework to implement topological relations on composite regions. *Lecture notes in computer science*, vol. 4653, p. 823-833.

Egenhofer M., Herring J. (1994). *Categorizing binary topological relations between regions, lines, and points in geographic databases*.

Ghozzi F., Ravat F., Teste O., Zurfluh G. (2003a). *Constraints and Multidimensional Databases*.

Ghozzi F., Teste O., Zurfluh G. (2003). Modèle multidimensionnel à contraintes. *Revue d'Intelligence Artificielle*, vol. 17, n°1-3, p. 43-55.

Glorio O., Trujillo J. (2008). *An MDA Approach for the Development of Spatial Data Warehouses. Data Warehousing and Knowledge Discovery*. I.-Y. Song, J. Eder and T. Nguyen, Springer Berlin / Heidelberg, p. 23-32.

Goertzen R., Stausberg J. (2007). A grammar of integrity constraints in medical documentation systems. *Computer Methods and Programs in Biomedicine*, vol. 86, n°1, p. 93-102.

Golfarelli M., Maio D., Rizzi S. (1998). Conceptual Design of Data Warehouses from E/R Schema. *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences*, vol. 7, IEEE Computer Society.

Hadzilacos T., Tryfona N. (1992). A Model for Expressing topological Integrity Constraints in Geographic Databases. *Proceedings of the International Conference GIS - From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning*.

Horner J., Song I.-Y. (2005). A taxonomy of inaccurate summaries and their management in OLAP systems. *Proceedings of the 24th international conference on Conceptual Modeling*, p. 433-448.

Hurtado C. A., Gutierrez C., Mendelzon A. O. (2005). Capturing summarizability with integrity constraints in OLAP. *ACM Transactions on Database Systems*, vol. 30, n°3, p. 854- 886.

Hurtado C. A. Mendelzon A. O. (2001). Reasoning about Summarizability in Heterogeneous Multidimensional Schemas. *Proceedings of the 8th International Conference on Database Theory*,p. 375-389.

Inmon W. H. (2005). *Building the Data Warehouse, 4th Edition*, John Wiley & Sons.

ISO (2000). *ISO 9000: Quality Management Systems: Fundamentals and Vocabulary*.

ISO/IEC (2001). *ISO/IEC 9126-1:2001 -- Software engineering -- Product quality -- Part 1: Quality model.*

ISO/IEC (2003). *ISO/IEC TR 9126-3:2003 -- Software engineering -- Product quality – Part 3: Internal metrics.*

ISO/IEC (2004). *ISO/IEC TR 9126-4:2004 -- Software engineering -- Product quality – Part 4: Quality in use metrics.*

ISO/TC 211 (2002). *ISO 19113:2002: Geographic information -- Quality principles*

Jensen C. S., Kligys A., Pedersen T. B., Timko I. (2004). Multidimensional data modeling for location-based services. *VLDB Journal*, vol. 13, n°1, p. 1-21.

Klasse Objecten (2009). *OCL Tools and Services Web Site*. <http://www.klasse.nl/ocl>.

Lechtenbörger J., Vossen G. (2003). Multidimensional normal forms for data warehouse design, *Information Systems*, vol. 28, n° 5, p. 415-434.

Lehner W., Albrecht J., Wedekind H. (1998). Normal Forms for Multidimensional Databases. *Proceedings of the 10th International Conference on Scientific and Statistical Database Management*, IEEE Computer Society, p. 63-72.

Lenz H.-J., Shoshani A. (1997). *Summarizability in OLAP and statistical data bases*, IEEE.

Levesque M., Bédard Y., Gervais M., Devillers R. (2007). *Towards managing the risks of data misuse for spatial datacubes.*

Lujan-Mora S., Trujillo J., Song I.-Y. (2006). A UML profile for multidimensional modeling in data warehouses. *Data & Knowledge Engineering*, vol. 59, n° 3, p. 725-769.

Ma H., Schewe K.-D. Thalheim B. (2009). Geometrically Enhanced Conceptual Modelling. *Conceptual Modeling - ER 2009*. Laender A., Castano S., Dayal U., Casati F., Oliveira J. De,. p. 219-233.

Malinowski E., Zimányi E. (2008). Advanced Data Warehouse Design. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications, Data-Centric Systems and Applications* Berlin Heidelberg, 2008.

Mandel L., Cengarle M. (1999). On the Expressive Power of OCL FM'99 — Formal Methods. p. 713-713.

Mansmann S., Neumuth T., Scholl M. (2007). Multidimensional Data Modeling for Business Process Analysis. *ER 2007*, p. 23-38

Mansmann S., Scholl M. (2006). Extending Visual OLAP for Handling Irregular Dimensional Hierarchies. *DaWaK 2006*, p. 95-105

Mazón J.-N., Lechtenbörger J., Trujillo J. (2008). Solving summarizability problems in fact-dimension relationships for multidimensional models. *Proceedings of the ACM 11th international workshop on Data warehousing and OLAP*. p. 57-64.

Mazón J.-N., Lechtenbörger J., Trujillo J. (2009). A survey on summarizability issues in multidimensional modeling. *Data & Knowledge Engineering*, vol. 68, n°12, p. 1452-1469.

Mazón J.-N., Trujillo J., Lechtenbörger, J. (2006). A set of QVT relations to assure the correctness of data warehouses by using multidimensional normal forms.

Proceedings of the 25th international conference on Conceptual Modeling. p. 385-398.

Nguyen T. B., Tjoa A. M., Wagner R. (2000). An Object Oriented Multidimensional Data Model for OLAP. *Proceedings of the First International Conference on Web- Age Information Management*, p. 69-82.

Niemi T., Nummenmaa J., Thanisch P. (2001). Logical Multidimensional Database Design for Ragged and Unbalanced Aggregation Hierarchies. *Proceedings of 3rd International Workshop on Design and Management of Data Warehouses*.

Olivé A. (2006). A method for the definition of integrity constraints in object-oriented conceptual modeling languages. Data and Knowledge Engineering, vol. 59, n° 3, p. 559-575.

OMG (2003). *MDA Guide, Version 1.0.1, Object Management Group*.

OMG (2006). *Object Constraint Language OMG, Available Specification, Version 2.0, Object Management Group*.

OMG (2007). *MOF 2.0/XMI Mapping, Version 2.1.1, Object Management Group.*

Papajorgji P., Pinet F., Miralles A., Jallas E. and Pardalos P. M. (2010). Modeling: A Central Activity for Flexible Information Systems Development in Agriculture and Environment, IJAEIS, vol. 1, n°1, p. 1-25.

Parent C., Spaccapietra S., Zimányi E. (1999). Spatio-temporal conceptual models: data structures + space + time. *Proceedings of the 7th ACM international symposium on Advances in geographic information systems*. p. 26-33.

Pedersen T. B., Jensen C. S., Dyreson C. E. (1999). Extending Practical Pre-Aggregation in On-Line Analytical Processing. *Proceedings of the 25th International Conference on Very Large Data Bases*, p. 663-674.

Pedersen T. B., Tryfona N. (2001). Pre-aggregation in Spatial Data Warehouses. *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, p. 460-480.

Pestana G., da Silva M. M., Bedard Y. (2005). Spatial OLAP modeling: an overview base on spatial objects changing over time. *IEEE 3rd International Conference on Computational Cybernetics*. 2005.

Pinet F. (2010). *HDR - Modélisation des contraintes d'intégrité dans les systèmes d'information environnementaux*, 112 p.

Pinet F., Miralles A., Bimonte S., Vernier F., Carluer N., Gouy V., Bernard S. (2010). The use of UML to design agricultural data warehouses. *International Conference on Agricultural Engineering, AGENG 2010*.

Pinet F., Schneider M. (2009). A Unified Object Constraint Model for Designing and Implementing Multidimensional Systems. *Journal on Data Semantics*, vol. 13, p. 37-71.

Pinet F., Schneider M. (2010). Precise design of environmental data warehouses. *Operational Research*, vol. 10, n°3, p. 349-369.

Pizano A., Klinger A., Cardenas A. (1989). Specification of spatial integrity constraints in pictorial databases. *Computer*, vol. 22, n° 12, p. 59-71.

Prat N., Wattiau I., Akoka J. (2010). Representation of aggregation knowledge in OLAP systems. *Actes de 18th European Conference on Information Systems*, paper 71.

Rafanelli M., Shoshani A. (1990). STORM: a statistical object representation model. *Proceedings of the fifth international conference on Statistical and scientific database management*, Charlotte, North Carolina, United States, Springer-Verlag New York, p. 14-29.

Rizzi S., Abelló A., Lechtenbörger J., Trujillo J. (2006). Research in data warehouse modeling and design: dead or alive?. *Actes de 9th ACM international workshop on Data warehousing and OLAP*, Arlington, Virginia, USA, ACM Press, p. 3-10.

Salehi M. (2009). *Developing a Model and a Language to Identify and Specify the Integrity Constraints in Spatial Datacubes*. Univ. Laval, Faculté des études supérieures de l'Université Laval: 2002.

Salehi M., Bédard Y., Mostafavi M., Brodeur J. (2007). On Languages for the Specification of Integrity Constraints in Spatial Conceptual Models. *Advances in Conceptual Modeling – Foundations and Applications*. J.-L. Hainaut, E. Rundensteiner, M. Kirchberget al, Springer Berlin / Heidelberg, vol. 4802, p. 388-397.

Salehi M., Bédard Y., Rivest S. (2010). Formal Conceptual Model and Definition Framework for Spatial Datacubes. *Geomatica*, vol. 3, n° 64, p. 313-326.

Sapia C. (1999). On modeling and predicting query behavior in OLAP systems. *Actes de International Workshop on Design and Management of Data Warehouses*, Heidelberg, Germany, p. 1-15.

Shekhar S., Lu C. T., Chawla S., Vatsavai R. R. (2001). Map Cube: A Visualization Tool for Spatial Data Warehouses. *Geographic Data Mining and Knowledge Discovery*, CRC Press, p. 74-109.

Song I.Y., Rowen W., Medsker C., Ewen E. (2001). An Analysis of Many-to-Many Relationships Between Fact and Dimension Tables in Dimensional Modeling. *Actes de International Workshop on Design and Management of Data Warehouses*, Interlaken, Switzerland, p. 1-13.

Torlone R. (2003). Conceptual multidimensional models. *Multidimensional databases*, IGI Global, p. 69-90.

Tryfona N., Busborg F., Christiansen J. G. B. (1999). starER: a conceptual model for data warehouse design, *Actes de 2nd ACM international workshop on Data warehousing and OLAP*, Kansas City, United States, ACM Press, p. 3-8.

Turki I. Z., Jedidi F. G., Bouaziz R. (2010). Multiversion data warehouse constraints. Proceedings of the ACM 13th international workshop on Data warehousing and OLAP, Toronto, Canada, ACM Press, p. 11-18.

Ubeda T., Egenhofer M. J. (1997). Topological Error Correcting in GIS. *Actes de 5th International Symposium on Advances in Spatial Databases*, Springer-Verlag, p. 283-297.