Conception d'entrepôts de données géospatiales à partir de sources hétérogènes. Exemple d'application en foresterie.

Maryvonne Miquel*,, Yvan Bédard*, Alexandre Brisebois***

** Centre de Rrecherche en Géomatique Université Laval*
Québec (Canada) G1K 7P4
*** LISI-INSA de Lyon Bât B. Pascal - 7 av. Capelle*
69621 Villeurbanne Cedex France
{maryvonne.miquel, yvan.bedard, alexandre.brisebois}@scg.ulaval.ca

RÉSUMÉ. *Dans cet article, nous abordons les problèmes liés à l'intégration de données spatio-temporelles au sein d'un entrepôt de données. Dans de nombreux cas, les spécifications des jeux de données géospatiales évoluent et les données sont hétérogènes à la fois des points de vue temporel, spatial et sémantique. Afin d'explorer et d'analyser des jeux de données spatio-temporelles dans une application SOLAP (Spatial On-Line Analytical Processing), nous définissons les notions de dimension spatiale géométrique et de dimensions spécifiques ou génériques. A l'aide de ce type de dimensions, nous proposons deux approches pour modéliser des structures multidimensionnelles. L'objectif final est de permettre l'extraction de connaissances géographiques par l'exploration des données détaillées associées à une époque et des études temporelles sur les données intégrées et comparatives. A partir d'un exemple pris dans le domaine de la foresterie, nous évaluons l'implémentation de ces deux modèles.*

ABSTRACT. *This paper presents some problems linked to the integration of data in a spatio-temporal data warehouse. In many cases, the specifications of the data sets have evolved over time and data sources have temporal, spatial and semantic heterogeneity. In order to explore and analyse spatio-temporal data sets in a SOLAP (Spatial On Line Analytical Processing) application, we define geometric spatial dimension and specific or generic thematic dimensions. Using these dimensions, we propose two approaches to model multidimensional structures. The final objective is to support geographic knowledge discovery through data exploration of detailed data for an epoch and of integrated comparable data for time-variant studies. Using a practical example in the field of forestry, we evaluate the implementation of these two models.*

MOTS-CLÉS : *Modèle Multidimensionnel, Entrepôts de données géospatiales, Spatial On-Line Analytical Processing, Dimension Spatiale Géométrique, Intégration de Données, Foresterie*

KEYWORDS: *Multidimensional Model, Geospatial Data Warehouses, Spatial On-Line Analytical Processing, GeometricSpatial Dimension, Data Integration, Forestry*

Miquel, M., Y. Bédard & A. Brisebois, 2002, Conception d'entrepôts de données géospatiales à partir de sources hétérogènes, exemple d'application en foresterie, Ingénierie des Systèmes d'information,Vol. 7, No. 3, p. 89-111

1. Introduction

La technologie des entrepôts de données, les bases de données multidimensionnelles et les outils d'exploration et d'analyse dédiés font émerger de nouvelles solutions pour le recueil et l'utilisation décisionnelle des données géospatiales. L'utilisation des références spatiales, ou références géographiques, lorsqu'elles sont disponibles, devient même une source d'information supplémentaire pour comprendre et analyser les données. La représentation cartographique, parce qu'elle est synthétique, immédiate et facile à interpréter s'intègre parfaitement bien à un environnement décisionnel. Cependant, la mise en place d'un entrepôt de données et l'exploitation de ces données géospatiales doivent tenir compte des spécificités liées à ce type de données et aux besoins relatifs des utilisateurs afin de tirer le meilleur parti de la richesse des informations contenues dans ces sources de données. En effet, classiquement, l'information géospatiale est organisée selon des attributs descriptifs et géométriques. Un entrepôt de données géospatiales a donc pour vocation de collecter, d'intégrer et d'historiser ce type de données afin de les rendre accessibles aux décideurs. Lorsque les données recueillies portent sur la représentation d'un phénomène sur un territoire à différents temps d'observation, les problèmes d'intégration des données sont à la fois sémantiques, temporels et spatiaux surtout si les différentes sources de données n'ont pas été acquises dans l'objectif d'alimenter un entrepôt. En particulier, les problèmes d'intégration spatiale de données cartographiques hétérogènes nécessitent de nouvelles solutions pour minimiser les impacts négatifs souvent invisibles lors de la seule utilisation de noms géographiques comme référence dans la dimension géographique d'un hypercube (approche habituelle). L'objectif du présent article est de présenter quelques problèmes et solutions relatifs à l'exploitation de données géospatiales sur plusieurs décennies, particulièrement dans un contexte où les spécifications et les techniques d'acquisition évoluent au cours des années.

1.1. *La dimension spatiale*

Habituellement, la dimension spatiale est traitée comme les autres dimensions, soit par l'utilisation de noms de lieux hiérarchisés (ex. pays, régions, municipalités). Avec une telle dimension où la référence spatiale est uniquement nominale, les données géométriques servant à représenter cartographiquement les objets du territoire ne sont pas prises en compte et exploitées. Aussi, plus récemment, les notions de dimensions et de mesures spatiales portant sur la géométrie des objets ont été définies et les problèmes liés à leur prise en compte ont été étudiés [BED 01]. Trois types de dimensions spatiales sont distingués :

- *la dimension spatiale non géométrique*, ou traditionnelle, pour laquelle les données considérées sont non géométriques. Par exemple, la dimension «unités administratives» peut être une dimension hiérarchique de type municipalités-régions-pays qui sert à représenter des données nominatives pour localiser un phénomène

dans l'espace. L'utilisation de la technologie des entrepôts de données permet de gérer et d'exploiter ce type de dimensions. Cependant, l'analyse spatio-temporelle des données s'en trouve relativement réduite, allant même jusqu'à ne pas permettre de voir certaines relations entre les phénomènes étudiés.

- *la dimension spatiale géométrique* qui est une dimension dans laquelle tous les niveaux des hierarchies ont une définition géométrique, i.e. sont représentés cartographiquement. Ainsi, tous les niveaux de ce type de dimension sont décrits par des objets géométriques (par exemple des polygones pour les pays et les régions, des points pour les municipalités). Cette dimension doit être navigable tout autant dans sa représentation cartographique que dans les vues tabulaires et graphiques.

- *La dimension spatiale mixte*, représentée par une dimension où seulement certains niveaux sont cartographiés. Par exemple, les niveaux les plus fins peuvent être associés à une géométrie alors que les niveaux supérieurs sont uniquement nominaux. L'inverse est également possible ainsi que toute autre combinaison telle qu'avoir uniquement certains niveaux intermédiaires qui soient cartographiés. Ce choix peut se justifier par la simplification des échelles de mesures car il permet de passer d'un niveau de données quantitatifs à un niveau qualitatif sur le plan géographique.

Pour la suite de ce texte, nous restreindrons l'appellation de dimension spatiale aux deux derniers types de dimensions car ce sont les seules à exploiter pleinement les données géospatiales.

Afin de compléter l'utlilisation des données géospatiales dans les modèles multidimensionnels, un nouveau type de mesures est également introduit. La mesure spatiale est une mesure qui représente géométriquement tous les objets issus du croisement des dimensions et sur lequels portent des opérateurs géométriques. Ainsi, en cours de navigation cette mesure permet de regrouper des objets ayant les mêmes caractéristiques et de les mettre en évidence sur les cartes. Introduire ce type de mesures dans les modèles multidimensionnels nécessite de trouver des méthodes pour construire un cube de données avec ce type de mesures [HAN 98][STE 00].

Le couplage des fonctionnalités des technologies OLAP et des Systèmes d'Information Géographique (SIG) a ouvert la voie à l'émergence d'une nouvelle catégorie d'outils d'aide à la décision plus adaptés à l'exploration et à l'analyse spatio-temporelles de ces données. Ces outils ont été regroupés sous la terminologie SOLAP (Spatial On-Line Analytical Processing) [BED 97]. Par l'association de la représentation cartographique et de la navigation OLAP, L'utilisateur se déplace dans la structure multidimensionnelle et obtient des représentations des données via un affichage cartographique, tabulaire ou en diagramme statistique qui sont fonction des dimensions, des mesures et des niveaux de hiérachie sélectionnés. Les caractéristiques requises sont clairement définies pour la visualisation des données, leur exploration et leur structuration et des solutions technologiques sont proposées [RIV 01]. Le prototype réalisé dans cette étude s'appuie sur ces solutions. Enfin, l'ajout de la représentation cartographique des phénomènes étudiés signifie aussi que l'on doive prendre en compte leur évolution spatiale. Par exemple, la dimension et la

forme des peuplements forestiers vont évoluer avec le temps, l'occupation du sol d'une municipalité va se modifier avec les années, et ainsi de suite. Il est possible de suivre cette évolution s'il existe des mécanismes régulatoires assurant un suivi de cette évolution (ex. émission de permis municipaux). Par contre, il peut également s'avérer impossible de suivre explicitement l'évolution d'un objet (ex. un peuplement forestier) parce qu'il s'agit d'un phénomène délimité par des frontières floues marquant artificiellement une transition graduelle, dont l'évolution est naturelle et non réglementée, et qui est réobservé par une nouvelle production cartographique avec une période donné (ex. tous les 10 années).

1.2. *Evolutions temporelles des spécifications dans les sources de données*

Au cours des deux dernières décennies, l'évolution très rapide des technologies géomatiques pour acquérir et traiter les données, comme par exemple la télédétection, le positionnement par satellite GPS, les systèmes d'information géographique, les lasers aéroportés, les serveurs universels, la photogrammétrie et la vision numérique, les appareils topométriques intelligents, se traduit par une évolution de la nature même des données acquises. Aussi, peut-on considérer que toute organisation a fait évoluer à plusieurs reprises ses spécifications au cours des 30 dernières années. Le remplacement des technologies analogiques des années 70 par les technologies numériques plus récentes a très fortement influencé ce qui est mesuré sur le territoire, le type et la qualité de la donnée recueillie. Lors de l'intégration des données géospatiales et de l'alimentation de l'entrepôt, il est indispensable d'apporter des solutions appropriées pour prendre en compte ces évolutions sans provoquer une perte de qualité et de quantité des informations initiales.

La plupart des modèles multidimensionnels proposés [CAB 98][LEH 98][HUR 99][TES 01][RAV 01] sont basés sur une organisation des données en dimensions et tables de faits [KIM 96] dans lesquelles les dimensions sont statiques et les tables de faits reflètent les aspects dynamiques de l'entrepôt. Dans la réalité, les dimensions elles-mêmes subissent des évolutions au cours du temps. Ce problème, identifié sous l'appellation des « Slowly Changing Dimensions » [Kim96], commence à trouver des réponses. Le plus souvent, le problème est résolu par une intégration des données et des méthodes de mise à jour consistant à ramener les données dans une vision unifiée [HUR 99][EPS 01]. Cette solution, produisant un ensemble de données parfaitement homogènes du point de vue des dimensions, peut cependant se traduire par une perte d'informations. Plus récemment, Pedersen propose un modèle multidimensionnel permettant de représenter des données complexes où le problème des évolutions temporelles des dimensions est étudié [PED 01]. Lorsque les évolutions des données sources sont connues, Eder introduit des matrices de transformation pour passer d'un système de référence à un autre [EDE 01]. Mendelzon expose particulièrement bien le problème des évolutions des dimensions en montrant comment une requête peut donner lieu à diverses interprétations selon la

façon dont l'évolution des structures est gérée [MEN 00]. Il introduit la notion de modèle multidimensionnel temporel et de requêtes OLAP temporelles qu'il utilise par la suite pour produire des données en temps consistant, c'est-à-dire telles qu'elles étaient au moment où elles ont été recueillies. S'appuyant sur les résultats des bases de données temporelles et la notion de temps de validité, il gère à la fois les évolutions de schémas de dimensions et les évolutions des instances. Nous montrerons comment son modèle peut être simplifié et adapté pour répondre aux besoins de notre étude.

Cet article traite des problèmes liés à la constitution de structures multidimensionnelles dont l'alimentation est faite à partir de données géospatiales hétérogènes existantes et présentant des problèmes d'intégration à la fois sémantique, temporelle et spatiale. Ces structures sont ensuite explorées dans un environnement SOLAP. La section 2 présente un exemple d'application qui consiste à créer un entrepôt de données provenant de trois inventaires forestiers décennaux. Cet exemple met en évidence les difficultés rencontrées pour la constitution d'entrepôts de données géospatiales et les besoins des utilisateurs. Il sera utilisé par la suite pour illustrer les solutions proposées. La section 3 donne des solutions pour concevoir une dimension spatiale à partir de données géométriques représentées en mode vectoriel et des dimensions thématiques qui supportent les évolutions temporelles des spécifications des sources de données. Dans la section 4, nous montrons comment différents modèles logiques multidimensionnels peuvent être conçus à partir des dimensions proposées. La section 5 décrit l'utilisation des structures multidimensionnelles dans une application SOLAP réalisée pour l'exploration et l'analyse des données des trois inventaires forestiers. Les avantages et les inconvénients des différentes approches sont ensuite évalués.

2. Exemple d'application et nature des problèmes d'hétérogénéité

La forêt est une ressource naturelle importante au Québec (Canada) de par son étendue (757 900 km^2) et le nombre d'emplois générés (plus de 170 000 emplois). Il est donc normal que ce secteur d'activité s'intéresse aux nouvelles technologies pour améliorer la connaissance des évolutions des zones forestières afin de gérer efficacement cette ressource. Nous avons étudié les apports de la technologie des entrepôts de données géospatiales à ce domaine par l'intégration de plusieurs inventaires forestiers effectués sur la même portion de territoire.

2.1 Inventaires forestiers

Effectuer un inventaire forestier consiste à partitionner la surface de la forêt étudiée en zones (appelées peuplements) qui présentent des caractéristiques forestières homogènes (essence, âge, densité, hauteur, etc.). Les données géométriques fournissent l'information sur la position et la forme des peuplements.

Les structures de données géométriques sont de type vectoriel. Les données descriptives fournissent des informations qualitatives ou quantitatives sur les caractéristiques des peuplements. Les mesures effectuées sur le peuplement sont par exemple la surface occupée ou le volume de bois. Au Québec, la cartographie forestière est établie aux moyens de photo-interprétation à partir de photographies aériennes validées par des échantillonnages sur le terrain, un tel inventaire est effectué environ tous les 10 ans. Compte tenu de la vitessse d'évolution des phénomènes étudiés, cette période d'acquisition est suffisante pour leurs observations. A chaque nouvel inventaire correspond une nouvelle carte forestière dont le jeu de données n'a aucun lien avec le précédent. Pour notre expérimentation, nous avons plus particulièrement travaillé avec trois inventaires de la forêt Montmorency des années 1973, 1984 et 1992. Le dernier inventaire (fin 2001) est en cours de publication et pourra être prochainement intégré à l'étude.

2.2 Problèmes d'hétérogénéité

Les problèmes d'hétérogénéité sont liés à la nature des objets spatiaux considérés lors des inventaires et à l'évolution sémantique des données descriptives. Les objets spatiaux pris en compte par les inventaires forestiers sont les peuplements. Ceux-ci n'étant définis que pour un inventaire donné, ils ne peuvent être utilisés comme objets de référence pour l'étude des évolutions. Il en découle une hétérogénéité des données géométriques tel que le démontre la figure 1. Les données descriptives sont déterminées par la législation, le mode d'acquisition et les spécifications propres à chaque inventaire. L'évolution de ces paramètres rend la tache d'intégration des données particulièrement difficile. La traduction de cette hétérogénéité est regroupée dans le tableau 1.

Année	**Nombre de peuplements forestiers**	**Nombre d'attributs**	**Unités de gestion**
1973	≈ 1 700	5	Peuplement
1984	≈ 2 400	14	Peuplement forestier Compartiment Unité de paysage
1992	≈ 3 800	19	Peuplement écoforestier Polygone écologique Station forestière

Tableau 1. *Caractéristiques des différents inventaires*

Durant les 20 ans recouvrant la période d'étude, l'affinement de la définition de la notion de peuplements associée à l'amélioration de la technologie se traduit par

une augmentation d'un facteur 2 du nombre de peuplements, le nombre d'attributs évolue également et leur sémantique diffère sensiblement. Enfin, les modes de gestion de la forêt ont changé, ce qui se traduit par une définition différente de la notion d'unités de gestion constituées de regroupements de peuplements.

L'hétérogénéité des données se retrouve également dans la variation des domaines de valeurs des attributs et de leur codage. Si les problèmes d'hétérogénéité de codage sont simples à résoudre lors de la constitution puis de l'alimentation de l'entrepôt par des procédures d'équivalence, l'uniformisation sémantique est plus difficile à obtenir. Ainsi, on constate par exemple qu'un attribut peut être organisé dans des domaines de valeurs différents d'un inventaire à l'autre (par exemple, l'âge peut être en domaines de valeurs qualitatives ou quantitatives) ou une valeur d'attribut d'un domaine de valeurs peut être absente dans un autre domaine. L'étude des 3 sources de données constituées des 3 inventaires forestiers fait apparaître que seuls 12 % des types de données sont demeurés identiques sur la période de 20 ans. Réduire l'ensemble des données aux seules valeurs immédiatement comparables ne se justifie pas lorsqu'on approfondit l'analyse à une période donnée car cela entraîne une perte d'information importante. Les besoins des utilisateurs imposent de maintenir dans l'entrepôt à la fois les données détaillées et agrégées et de permettre conjointement leur exploration.

2. Intégration de données géospatiales

L'exemple précédent montre que l'hétérogénéité des données géospatiales doit être traitée au niveau géométrique et descriptif. Pour le premier type de données, cela consiste à trouver une organisation de la surface du territoire en entités spatiales invariantes dans le temps. Ces entités formeront les membres du niveau le plus fin de la dimension spatiale. Le deuxième type de données pose un problème plus classique d'intégration de données hétérogènes avec la contrainte de faire cohabiter dans le même modèle des données détaillées temporellement non comparables et des données agrégées temporellement comparables.

3.2. Données géométriques

L'évolution des frontières d'un découpage de la surface d'un territoire est un problème fréquent lorsqu'on étudie des ensembles de données réparties selon des zones géographiques découlant d'une organisation administrative, politique ou sectorielle. Il est donc indispensable de tenir compte de ces évolutions dans la constitution de l'entrepôt de données géospatiales. Un découpage territorial est généralement défini par un ensemble de polygones (entités géométriques) représentés en mode vectoriel et référencés dans un système cartésien. A chaque entité géométrique est associé un ensemble d'attributs descriptifs valides dans un intervalle de temps inclus dans l'intervalle de validité de l'entité. Pour ramener les

données descriptives à des entités géométriques invariants dans le temps, deux méthodes peuvent être utilisées. La première méthode consiste à partitionner la surface du territoire par la superposition (overlay) de toutes les entités provenant de différents découpages, les entités géométriques d'un même découpage étant disjointes ou adjacentes. Les entités obtenues par cette opération morcelleent la surface en sous entités qui héritent à chaque instant des attributs descriptifs des entités dont elles sont issues. Ainsi, les sous entités forment une référence spatiale invariante pour l'ensemble des découpages et leurs caractéristiques descriptives sont connues à tout instant. Cette méthode présente un certain nombre d'inconvénients. Tout d'abord, l'introduction de nouvelles données avec un nouveau découpage du territoire entraîne une nouvelle détermination des sous entités géométriques et une réaffectation des attributs descriptifs. De plus, le nombre de sous entités peut croître de façon non maîtrisée. A titre d'exemple, l'opération d'overlay des 3 inventaires forestiers (environ 8 000 peuplements) engendre 50 000 sous peuplements, dont plusieurs de taille trop petite pour être siginificatifs. Enfin, cette opération produit des sous entités sans intérêt car seul leur regroupement en entités géométriques correspondant à un découpage donné a un sens pour l'analyste.

La deuxième méthode repose sur une structure en mosaïque du territoire selon un mode matriciel. La surface du territoire est alors représentée à l'aide de cellules régulières. On obtient ainsi un découpage de référence fixe. A chaque cellule et pour chaque instant sont associés les attributs de l'entité qui le recouvre par une opération d'overlay. Le choix de la dimension de la cellule de base est fonction de la dimension des entités à représenter. Il doit être effectué afin de minimiser la perte d'information en veillant à ce que chaque entité soit représentée par au moins une cellule. L'avantage de cette méthode est d'autoriser l'ajout de nouvelles données selon un nouveau découpage sans modifier l'organisation des données existantes. Les inconvénients proviennent du nombre de cellules à définir a priori pour couvrir convenablement l'ensemble des entités. Dans le cas de la constitution d'un entrepôt de données à partir des trois inventaires forestiers, 162 000 cellules ont été créées, chaque cellule représentant un carré de 20x20 mètres. La redondance descriptive est alors très importante puisqu'on duplique les attributs descriptifs pour chaque cellule alors qu'un seul jeu d'attributs descriptifs suffit à décrire une entité géométrique à un instant donné. Enfin, la représentation matricielle provoque une impression d'imprécision dans le tracé des limites. Cependant, la technologie des entrepôts de données supporte bien la gestion de grands ensembles de données dénormalisées et, dans le cas des peuplements des inventaires forestiers, l'imprécision dans le tracé des frontières des entités géométriques correspond bien à l'imprécision de leur délimitation. Cette méthode a été retenue pour fournir une référence spatiale invariante des peuplements forestiers (figure 1).

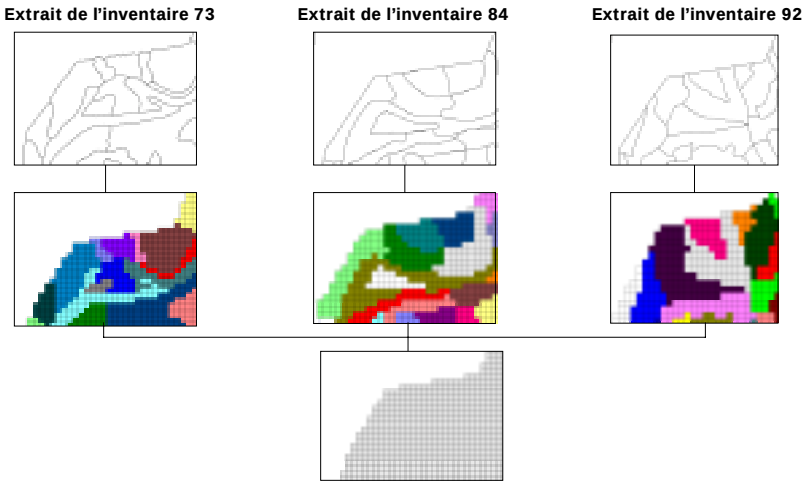

Figure 1. *Transformation des données en mosaique*

La dimension spatiale est organisée en concordance avec les règles de gestion forestière. Elle se compose de 5 niveaux, le niveau le plus bas de la dimension spatiale est la cellule avec ses caractéristiques géométriques, les cellules sont regroupées par peuplements, eux mêmes regroupés par compartiments ou polygones écologiques qui produisent ensuite des unités ou des stations forestières (figure 2).

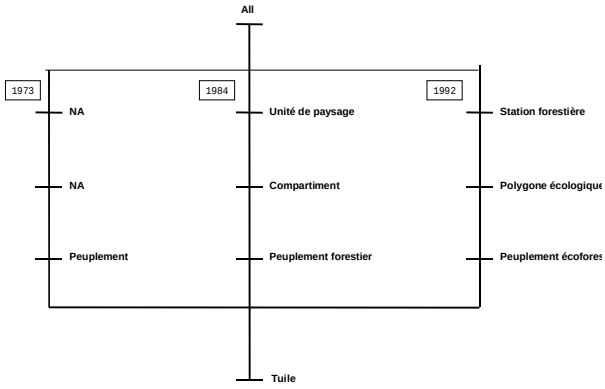

Figure 2. Hiérarchies alternatives de la dimension spatiale pour les inventaires forestiers

Le dernier niveau All est la forêt dans sa totalité. Le regroupement étant dépendant du mode de gestion de chaque inventaire, une dimension composée de plusieurs hiérarchies alternatives a été créée pour permettre une navigation selon ces différents modes. Comme l'opération d'overlay permet d'associer chaque cellule régulière à une instance de n'importe quel niveau géographique, nous avons ici une dimension spatiale géométrique telle que définie à la section 1.1.

3.2 Données descriptives

Afin de prise en compte des données détaillées non comparables sur tout un intervalle d'étude et des données agrégées comparables, nous introduisons la notion de membres ayant une validité temporelle (membres spécifiques) et de membres dont la validité est étendue à toute l'intervalle d'étude (membres génériques). Seul le niveau le plus fin de chaque dimension descriptive (ou thématique) est composé de membres spécifiques afin de représenter la diversité des données; les niveaux supérieurs ayant des membres génériques sur lesquels pourront porter les études comparatives. Les schémas des dimensions décrivent leur organisation hiérarchique (dans notre étude, nous n'utilisons que des dimensions ayant une seule hiérarchie), les instances des dimensions décrivent comment les membres des différents niveaux interagissent. Les liens permettant de lier un membre spécifiques à un autre membre du niveau supérieur ont également une durée de validité (figure 3).

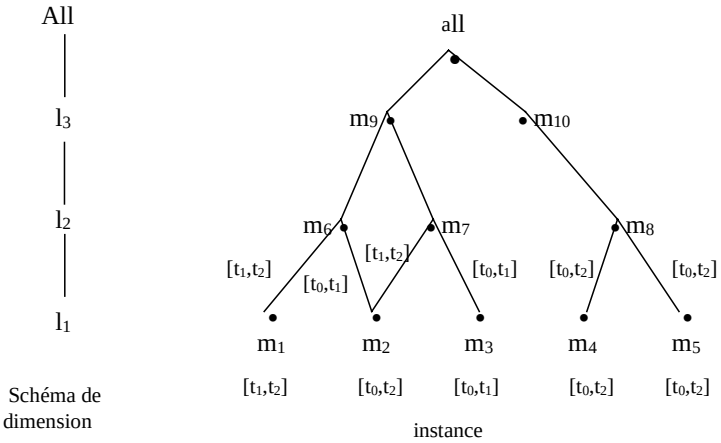

figure 3. Exemple de s*chéma de dimension (a) et instance (b). On a par ex. :*
$m_2 \rightarrow m_6$ est valide pour $t \in [t_0,t_1]$ et
m_2 et m_4 sont valides pour $t \in [t_0,t_2]$.

L'ensemble des dimensions de l'analyse de l'ensemble des données et des axes d'intérêt du décideur. Ainsi, la surface d'une zone forestière peut être étudiée selon les caractéristiques des arbres qui la composent à savoir l'essence, la hauteur, l'âge, les perturbations et la densité. 5 dimensions thématiques sont définies. Une étude des données sources permet également d'organiser les caractéristiques décrivant les données en niveaux hiérarchiques.

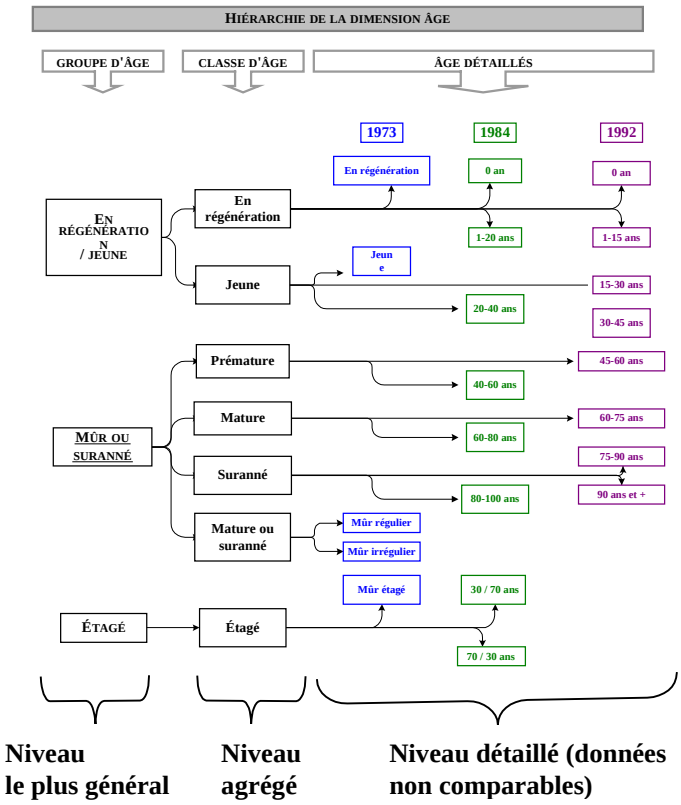

Figure 4. *Membres de la dimension Age [REB 98]*

Pour la dimension Age Par exemple, les classes d'âge utilisées pour l'inventaire de 1973 sont qualitatives, elles sont quantitatives et segmentées en intervalle de 20 ans pour l'inventaire de 1984 et en intervalle de 15 ans pour l'inventaire de 1992 (figure 4). L'ensemble de ces catégories constitue les membres spécifiques du niveau le plus fin de la dimension Age. La dimension Age est organisée selon les niveaux suivants : « All », « groupe d'âge » (3 membres), « classe d'âge » (7 membres), « âge

détaillé » (21 membres dont 5 pour l'inventaire de 1973, 8 pour celui de 1984 et 8 pour celui de 1992). Le membre « *âge détaillé 20-40* » n'est valide que pour l'inventaire de 1984 ainsi que le lien qui l'unit au membre « *jeune*». Ce dernier membre est générique et les mesures associées à ce membre sont comparables au cours du temps. On pourra donc analyser la variation des surfaces d'arbres jeunes pour la période couverte par les trois inventaires.

La figure 5 présente les autres dimensions thématiques qui émergent de l'étude des inventaires forestiers. Les dimensions Essence, Hauteur, Perturbation suivent le modèle précédemment introduit avec des membres spécifiques pour le niveau le plus fin. Les labels des niveaux hiérarchiques ont été choisis en conformité avec le vocabulaire du domaine forestier. On peut noter que la dimension Densité est une dimension plate avec un seul niveau. Comme la classification de la densité des peuplements n'a pas varié d'un inventaire à l'autre, les membres de ce niveau sont génériques.

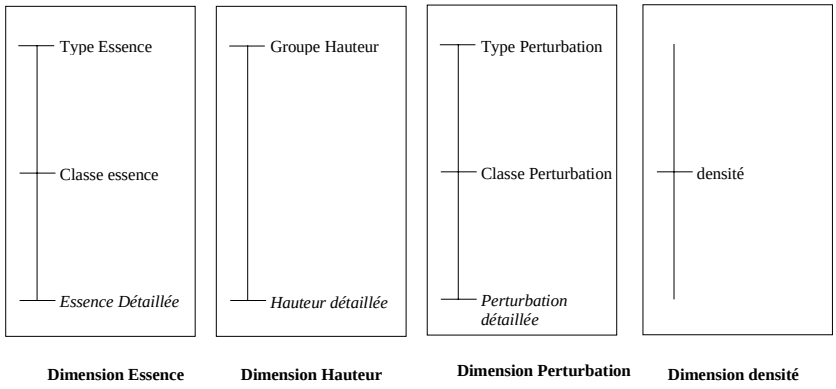

Figure 5. *Autres dimensions thématiques pour les inventaires forestiers, les niveaux ayant des membres génériques sont en italiques*

4.Modèles multidimensionnels

Dans un entrepôt de données géospatiales, trois types de dimensions cohabitent : la dimension temporelle, la dimension spatiale et les dimensions thématiques. La dimension temporelle décrit l'organisation du temps d'étude en conformité avec les besoins de l'utilisateur. Cette dimension peut être une hiérarchie simple ou complexe mais son instance ne comprend pas de membres spécifiques. La dimension spatiale décrit la représentation de la surface du territoire, elle pourrait comporter des membres spécifiques mais cela restreindrait la représentation cartographique des données à un instant donné selon le seul découpage territorial valide à cet instant. Or, on peut imaginer que l'utilisateur veille projeter les données d'une période de

temps dans un découpage associé à une autre période de temps. En ramenant l'ensemble des découpages territoriaux à un système de référence spatiale fixe, on permet ce type d'analyse. Nous avons choisi d'utiliser la notion de membres spécifiques uniquement pour les dimensions thématiques. Plusieurs modèles logiques peuvent être conçus à partir ces dimensions thématiques. Nous présentons les deux modèles qui ont été utilisés pour constituer un entrepôt de données provenant des inventaires forestiers.

4.1 Tables de faits spécifiques et table de faits générique

Le premier modèle est composé de plusieurs tables de faits pouvant partager des dimensions. Une solution pour exploiter les dimensions thématiques que nous avons défini est d'extraire des instances valides (appelées sous dimensions) sur un intervalle de temps donné. Une table de faits est définie à partir d'un jeu de dimensions ayant la même plage de validité. Il y aura donc autant de tables de faits spécifiques qu'il y a d'intervalles de temps considérés. Une table de faits supplémentaire (table de faits générique) est définie à partir des dimensions thématiques réduites au niveau supérieur, c'est-à-dire dont les niveaux ne comprennent que des membres génériques.

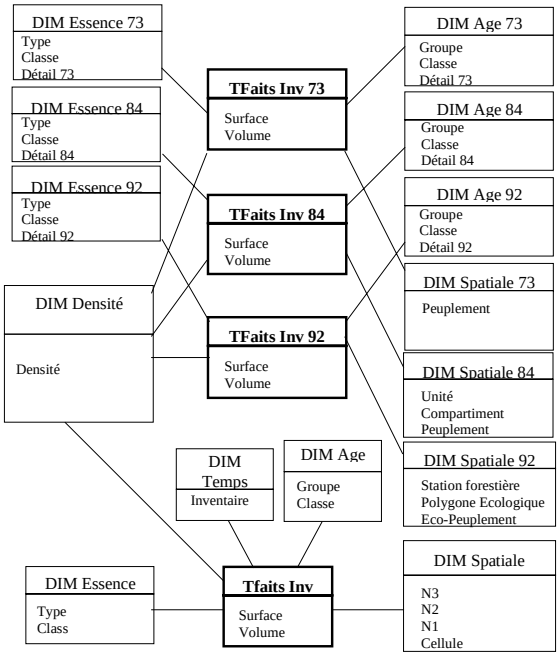

Figure 6. *Modèle en constellation pour les inventaires forestiers*

La figure 6 présente le modèle en constellation appliqué aux inventaires forestiers. A partir des dimensions thématiques précédemment définies, nous avons extrait des sous dimensions pour chaque inventaire, ce qui revient à ne conserver au niveau le plus fin que les membres spécifiques valides pour l'inventaire considéré et à conserver les niveaux génériques supérieurs. La dimension spatiale est traitée différemment. En effet, les cellules ont été introduites pour normaliser la représentation géométrique des différents inventaires. En choisissant de créer une table de faits par inventaire, il n'y a plus d'intérêt à utiliser ce grain particulièrement fin car l'ensemble des cellules d'un même peuplement présente les mêmes caractéristiques qui n'évoluent pas puisque la table de faits ne porte que sur un seul inventaire. La dimension spatiale retenue pour chaque inventaire est donc le peuplement avec sa géométrie comme niveau fin et il y aura une dimension spatiale par inventaire. Ainsi chaque inventaire donne lieu à la constitution d'une table de faits temporellement et spatialement spécifique. Dans cet exemple, la dimension temporelle pour chaque inventaire est omise car le niveau le plus fin est l'année, correspondant à l'intervalle de validité des tables de faits spécifiques. Par contre, la cellule est utilisée comme niveau le plus fin de la dimension spatiale dans la table de faits temporellement générique et une dimension temporelle est définie. Ses membres sont les années d'inventaires.

4.2 Table de faits unique

Le deuxième modèle exploite pleinement les dimensions que nous avons introduites et la notion de temps consistant. Une seule table de faits est définie et les dimensions thématiques associées sont formées des membres spécifiques et des membres génériques. Le modèle est un modèle en étoile classique comme le montre sa représentation dans le cas des inventaires forestiers (figure 7).

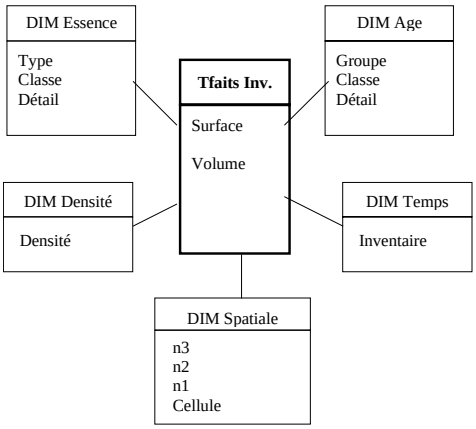

Figure 7. *Modèle en étoile pour les inventaires forestiers*

Dans ce modèle, les dimensions thématiques ont un niveau fin appelé *Détail* composé de membres spécifiques. Cette représentation ne permet pas de décrire plus particulièrement l'organisation des dimensions qui n'apparaît que dans les instances de schémas de dimensions. Chaque mesure de la table de faits se rapporte à une cellule géométrique issue de la représentation matricielle du territoire.

5. Prototype et évaluation

Avant d'être intégrées dans un modèle multidimensionnel, les données brutes subissent un certain nombre de transformation. Elles sont nettoyées afin d'éliminer les erreurs de codage puis recodifiées en fonction des nouvelles classifications définies par les niveaux hiérarchiques des dimensions. Les entités géométriques décrivant les peuplements sont rapportées à une couverture régulière en mosaique du territoire. Afin de conserver des valeurs correctes des mesures au niveau agrégé, les mesures (ici surfaces et volumes) sont pondérées. Enfin, les applications SOLAP étant conçues pour manipuler des objets géométriques en mode vectoriel, la représentation matricielle est convertie en représentation vectorielle en représentant chaque cellule par un polygone. Les différentes solutions sont ensuite implémentées sous MS SQL Server et Analysis Services. Compte tenu de ce choix d'outils, nous avons adapté les modèles logiques afin de tenir compte des contraintes d'implémentation. Ainsi, le premier modèle aboutit à la génération de n+1 structures multidimensionnelles exploitées en fonction des requêtes de l'utilisateur, avec n structures multidimensionnelles temporellement et spatialement spécifiques qui regroupent, pour chacune d'elles, les données comparables à un niveau fin sur un intervalle de temps et une structure temporellement et spatialement générique regroupant l'ensemble des données comparables à un niveau plus grossier. Dans le cas des inventaires forestiers, 4 structures multidimensionnelles ont été créées pour le premier modèle. Le deuxième modèle produit une seule structure multidimensionnelle. Ces différentes structures sont stockées dans une base de données multidimensionnelle propriétaire (MOLAP).

Les deux applications SOLAP, conçus pour supporter l'exploitation spatio-temporelle des structures multidimensionnelles et correspondant aux deux solutions présentées, ont été développés en utilisant les composants logiciels de Proclarity et Intergraph Geomedia. Pour chaque application, une application Visual Basic s'appuie sur les fonctionnalités de Proclarity pour la navigation sur les données descriptives avec tous les opérateurs classiques d'une application OLAP et les fonctionnalités de Geomedia pour la navigation spatiale et la représentation cartographique. La figure 8 donne un exemple de l'interface de l'application SOLAP. Elle est composée de 2 fenêtres. A droite, la fenêtre représente les éléments d'une application OLAP classique avec la sélection des dimensions d'analyse, les opérateurs de navigation et la représentations des données descriptives (ici un tableau et un histogramme). A gauche, la fenêtre spatiale représente la carte des

données sélectionnées. La même sémiologie (couleurs, formes) est utilisée dans les deux fenêtres.

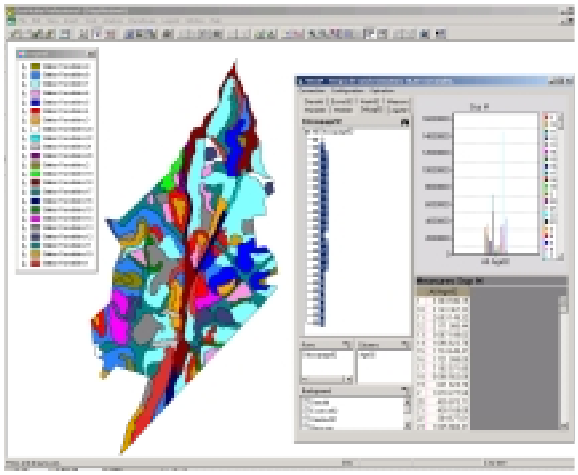

Figure 8. *interface de l'application SOLAP*

Le nombre d'éléments de la table de faits dépend de l'organisation des données. La taille du cube dépend des paramètres introduits lors de sa création car nous utilisons une fonctionnalité de SQL-Server Analyse Services qui détermine le nombre d'agrégations à créer en fonction du taux d'agrégation voulu ou de la taille maximale du cube. L'évaluation qualitative de l'implémentation des deux solutions met en évidence certaines caractéristiques (tableau 2). Les capacités d'analyse de chaque structure sont évaluées du point de vue spatial, temporel et de l'accès aux données détaillées.

La première solution minimise la taille des données stockées dans les stuctures spécifiques car elles sont organisées par peuplement. Ce qui explique que le nombre de faits dans les tables varient. En revanche, les données sont dupliquées entre les structures associées à chaque inventaire et la structure générique. L'ajout d'un nouvel inventaire ne modifie pas les structures spécifiques, seule la structure générique sera à recalculer. Cette solution nécessite le développement d'une interface particulière afin de sélectionner la structure multidimensionnelle correspondant à la requête de l'utilisateur. La navigation temporelle n'est possible que sur la stucture générique alors que les données détaillées sont accessibles par les structures spécifiques. La navigation selon des niveaux plus agrégés est possible dans tous les cas. On peut noter cependant que si pour les inventaires forestiers, la première solution est acceptable, elle devient vite inopérante lorsque trop de changements interviennent, entraînant une mutiplicité des tables de faits spécifiques.

	Analyse spatiale	Analyse temporelle	Données détaillées
Première solution			
3 structures temporellement et spatialement spécifiques	Niveau peuplement	Non	Oui
1 structure générique	Niveau cellule	Oui	Non
Deuxième Solution			
1 structure multidimensionnelle	Niveau cellule	Oui	Oui

Tableau 2. *Comparaison des deux solutions*

La deuxième solution fournit une vue plus intégrée des données. Dans l'exemple des inventaires forestiers, les temps de validité des liens et des membres spécifiques sont confondus, ce qui simplifie l'implémentation des intervalles de temps de validité. Le cube résultant est très creux puisque beaucoup de données au niveau détaillé ne sont pas disponibles du fait même de l'introduction des membres spécifiques. Cette particularité mériterait d'être mieux traitée lors de l'implémentation. L'avantage de cette structure est de représenter simplement les données telles qu'elles ont été acquises (notion de temps consistant), de plus les métadonnées donnent des informations supplémentaires sur l'organisation des données détaillées proches de celles des sources de données. Enfin, dans cette solution, les données détaillées de tous les inventaires sont accessibles, l'utilisateur peut alors étudier les zones du territoire ayant des caractéristiques définies sur plusieurs années comme par exemple, analyser les zones composées de bouleaux blancs en 84 (membre spécifique de la dimension essence en 84) et dont l'âge est de plus de 90 ans (membre spécifique de la dimension âge en 92). Les capacités d'exploration, d'analyse et de mise en corrélation sont donc augmentées.

6. Conclusion

Dans cet article, nous proposons des solutions pour concevoir des structures multidimensionnelles lorsque les sources de données sont hétérogènes du point de

vue temporel, spatial et sémantique. Pour traiter l'hétérogénéité spatiale, les données en mode vectoriel sont ramenées dans un référentiel matriciel fixe composé de cellules régulières. Puis, de nouvelles classifications des données descriptives sont introduites en groupant les attributs descriptifs initiaux en attributs génériques, i.e invariant temporellement. Les dimensions thématiques sont définies à partir de cette nouvelle organisation et des besoins des utilisateurs. Enfin, plusieurs modèles multidimensionnels sont proposés, puis implémentés et évalués avec une application SOLAP. La première solution consiste à créer des cubes pour chaque tranche de temps invariante et un cube pour les niveaux d'agrégation comparables. Cette solution limite les capacités d'analyse et ne convient plus si les sources de données comportent trop de variations dans l'organisation des données descriptives. La deuxième solution consiste à intégrer les données détaillées et agrégées dans le même cube, en ajoutant des temps de validité aux membres des dimensions descriptives et à leurs liens. Ainsi, les évolutions temporelles des dimensions thématiques sont conservées.

Remerciements

Ce travail a été réalisé grâce au le soutien financier du Réseau de CentresD'Excellence GEOIDE, pour le projet DEC#2 : Conceptualisation des fondations technologiques pour la prise de décision à l'aide du World Wide Web. Maryvonne Miquel remercie le Centre de Recherche en Géomatique (CRG) de l'université Laval et plus particulièrement Yvan Bédard et son équipe pour leur accueil durant son année de congé pour recherche.

Bibliographie

[BED 97] Bédard, Y., "Spatial OLAP", Vidéoconférence, *2ème Forum annuel sur la R-D, Géomatique VI: Un monde accessible*, 13-14 novembre 1997, Montréal (Canada)

[BED 01] Bédard, Y., T. Merrett & J. Han,. "Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery". Chapter of the book *Geographic Data Mining and Knowledge Discovery* edited by H. Miller and J. Han, Research Monographs in GIS series edited by Peter Fisher and Jonathan Raper, Taylor & Francis, 2001

[CAB 98] Cabibbo L., Torlone R., "A logical approach to multidimensional databases", *Conference on Extending Database Technology (EDBT) '98*, Valencia (Espagne), Mars 1998

[EDE 01] Eder J., Koncilia C., "Changes of dimension in Temporal Data", *3rd International Conference on Data Warehousing and Knowledge Discovery* - Munich, Sept 3-7 2001

[EPS 01] Espil M, Vaisman A., "A. Efficient Intentional Redefinition of Aggregation Hierarchies in Multidimensional Databases", *ACM International Workshop on Data Warehousing and OLAP (DOLAP)* 2001

[HAN 98] Han J., Stefanovic N., and Koperski K., "Selective Materialization: An Efficient Method for Spatial Data Cube Construction", Proc. *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Melbourne, Australia, 1998, pp. 144-158.

[HUR 99] Hurtado C., Mendelzon A., and Vaisman A., "Maintaining Data Cubes under Dimension Updates" IEEE *International Conference on Data Engineering*, 1999

[KIM 96] Kimball R., *The Data Warehouse Toolkit*. J. Wiley and Sons, 1996

[LEH 98] Lehner W., "Modeling Large Scale OLAP Scenarios", *Conference on Extending Database Technology (EDBT)'98*, Valence (Espagne), Mars 1998

[MEN 00] Mendelzon A., Vaisman A.: "Temporal Queries in OLAP". *VLDB 2000*: 242-253, 2000

[PED 01] Pedersen T.B., Jensen C., Dyreson C., 2001 "A foundation for capturing and querying complex multidimensional data*", Information Systems* 26 383-423

[RAV 01] Ravat F., Teste O., Zurfluh G., "Modélisation multidimensionnelle des systèmes décisionnels", *Extraction des Connaissances et Apprentissage* (ECA), Volume 1 - n°1-2/2001 - EGC 2001, Hermès (ed.), 17-19 Janvier 2001, Nantes, Loire-Atlantique, France) - ISBN 2-7462-0216-6, pp.201-212

[Riv01] Rivest, S., Bédard, Y. & Marchand P., "Towards better support for spatial decision-making: Defining the characteristics of Spatial On-Line Analytical Processing (SOLAP) ", *Geomatica* pp 539 - 555.

[REB 98] Rebout, C., Adaptation d'une base de données pour une application SOLAP pour l'aide à l'aménagement intégré des ressources forestières, Rapport de DESS, Université Joseph Fourier, Grenoble, France

[STE 00] Stefanovic N., Han J., and Koperski K., "Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes", *IEEE Transactions on Knowledge and Data Engineering*, 12(6): 938-958, 2000.

[TES 01] Teste O., "Towards Conceptual Multidimensional Design in Decision Support Systems", Proceedings of the *5th East-European Conference on Advances in Databases and Information Systems*, Research Communications Vol. 1, A. Caplinskas, J. Eder (eds.) - ISBN 9986-05-449-4, September 25-28 2001, pp77-88, Vilnius (Lithuania).