Extension du modèle M3 aux évolutions temporelles dans les applications SOLAP

Maryvonne Miquel, Anne Tchounikine

LIRIS-INSA de Lyon, Bâtiment 501, 69621 Villeurbanne, France
{miquel, atchouni}@if.insa-lyon.fr

Résumé. La plupart des modèles multidimensionnels considèrent les faits comme la partie dynamique de l'entrepôt tandis que les dimensions sont vues comme des entités statiques. Dans les applications pratiques, les structures d'analyse évoluent avec le temps. Cette problématique devient particulièrement sensible et complexe dans le cadre d'applications SOLAP (Spatial On Line Analytical Processing). En effet, la dimension spatiale lorsqu'elle décrit un découpage territorial est sujette à modification et ces modifications ont une énorme incidence sur les mesures. Dans cet article nous montrons comment le modèle M3 que nous avons proposé permet de construire un hypercube conservant l'ensemble des évolutions et offre une navigation et une exploitation comparative des données dans les différentes versions des dimensions. Notre proposition est illustrée dans le domaine de la foresterie.

1 Problématique

Les systèmes OLAP (On Line Analytical Processing) se composent de tables de faits, et de dimensions formant la structure multidimensionnelle d'analyse. La plupart des modèles actuels considèrent les faits comme la partie dynamique de l'entrepôt tandis que les dimensions sont vues comme des entités statiques. Pourtant, dans les applications pratiques, les structures d'analyse évoluent avec le temps. Ce problème a été soulevé dès 1996 par Kimball (Kimball. 1996) et défini comme les *slowly changing dimensions*. Ces dernières années, ce domaine de recherche a suscité de nombreux travaux (Blaschka et al. 1999 ; Mendelzon, Vaisman. 2000 ; Eder, Koncilia. 2001 ; Pedersen et al. 2001). Dans de précédents travaux, nous avons élaboré un modèle, le modèle M3 (Modèle Multidimensionnel Multiversion) (Body et al, 2003) qui permet de prendre en compte les évolutions temporelles de schéma de dimension et de membres de dimension dans les modèles multidimensionnels.

Cette problématique devient particulièrement sensible et complexe dans le cadre d'applications SOLAP (Spatial On Line Analytical Processing). Les applications SOLAP introduisent les notions de dimensions et de mesures spatiales portant sur la position et la forme des objets. Les problèmes liés à leur prise en compte ont été étudiés dans (Bédard et al., 2001). Trois types de dimensions spatiales sont distingués :

- *la dimension spatiale non géométrique* ou purement textuelle. C'est celle-ci qui est traditionnellement utilisée dans les applications n'exploitant pas la cartographie et pour laquelle les données considérées ne décrivent ni la position ni la forme des objets sur le territoire. Par exemple, la dimension «unités administratives» peut être une dimension hiérarchique de type municipalités-régions-pays qui utilise des données nominatives pour identifier un phénomène et localiser son inclusion dans l'espace d'un autre phénomène plus

global (ex. Lyon, Rhône, France). L'utilisation de la technologie des entrepôts de données permet de gérer et d'exploiter facilement ce type de dimension puisque la référence spatiale y est purement textuelle et traitée comme toute autre dimension,

- *la dimension spatiale géométrique* qui est une dimension dans laquelle tous les niveaux des hiérarchies 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* combine des niveaux qui sont cartographiés ou non. Par exemple, les niveaux les plus fins peuvent être associés à une géométrie (ex. points des municipalités) alors que les niveaux supérieurs sont uniquement nominaux (ex. noms des pays). L'inverse est possible, c'est-à-dire avoir des géométries uniquement pour les niveaux supérieurs (ex. polygones des pays) et des noms de lieux pour les niveaux inférieurs (ex. noms des municipalités), ainsi que toute autre combinaison. Ce choix permet ainsi de passer d'un niveau de données quantitatif à un niveau qualitatif sur le plan géographique.

Une dimension spatiale géométrique ou mixte doit être capable de stocker et de manipuler des types d'objets complexes tels que des points, agrégats de points, lignes, polylignes, polygones simples et agrégés, comme savent le faire les Systèmes d'Information Géographique (SIG) et que ne permettent pas les environnements OLAP classiques. Le couplage des fonctionnalités des technologies OLAP et des 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 (Bédard, 1997).

La dimension spatiale, lorsqu'elle décrit un découpage territorial, est sujette à modification (réaffectation d'une municipalité à un département, création de communautés urbaines, déplacement de frontières, réunification ou morcellement d'états….). Or ces modifications ont une énorme incidence sur les mesures. Les solutions actuellement proposées telles que les mises à jour, aboutissent à une perte ou à une déformation de l'information préjudiciable à l'analyse décisionnelle. Dans cet article nous montrons comment le modèle M3 que nous avons proposé permet de construire un hypercube conservant l'ensemble des évolutions et offre une navigation et une exploitation comparative des données dans les différentes versions des dimensions.

Dans la première partie de cet article, nous exposons une étude de cas prise dans le domaine de la foresterie et nous mettons en avant les différents types d'évolution à prendre en compte. Dans la deuxième partie nous présentons le modèle M3. La troisième partie montre comment le modèle M3 peut être adapté pour développer une application SOLAP temporellement évolutive telle que celle décrite dans l'étude de cas.

2 Etude de cas

La forêt est une ressource naturelle à évolution lente mesurée par des inventaires réguliers. Les experts en foresterie s'intéressent à la mise en place d'entrepôts de données intégrant différents inventaires dans le but de suivre l'évolution de cette ressource et d'en améliorer la gestion. 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. Notre travail a porté sur l'intégration de trois inventaires décennaux (1973, 1984, 1992).

2.1 Inventaires forestiers

Les inventaires forestiers sont généralement effectués tous les 10 ans. Un inventaire consiste à partitionner la surface de la forêt étudiée en zones (appelées peuplements) qui présentent des caractéristiques forestières homogènes par le type des arbres (ou essence) qui les composent, leur âge, leur hauteur … Une zone est décrite par des données géométriques de type vectoriel qui fournissent l'information sur la position et la forme des peuplements. D'autres données qualifiées de 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. Le schéma de l'entrepôt est représenté figure 1. Il fait apparaître une table de faits contenant les mesures surface et volume ainsi que 5 dimensions : Essence, Age, Densité, Temps et Découpage, cette dernière étant la dimension spatiale qui décrit un peuplement.

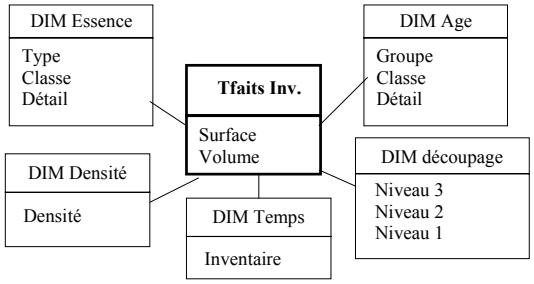

FIG. 1. *Modèle de l'entrepôt de données en foresterie*

2.2 Problèmes d'évolutions

Lorsque l'on analyse les données des différents inventaires à intégrer dans l'entrepôt, on constate que durant les 20 ans recouvrant la période d'étude, l'affinement de la définition de la notion de peuplements associé à l'amélioration de la technologie s'est traduit par une modification sensible de la nature et du nombre de données collectées. Ainsi 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 une période de 20 ans.

Comme le montre la figure 2, la classification des essences a évolué d'un inventaire à l'autre. Par exemple, les 2 premiers inventaires font apparaître la classe des « résineux et bouleaux » alors que, lors du dernier inventaire, les spécialistes en foresterie ont introduit deux sous classes, « sapins et bouleaux blancs » et « sapins et bouleaux jaunes ».

	1973	1984	1992
Résineux et Bouleaux	- Résineux et bouleaux	- Résineux et bouleaux	- Sapins et bouleaux blancs - Sapins et bouleaux jaunes
Résineux et feuillus	- Résineux et feuillus intolérants - Résineux et peupliers	- Résineux et feuillus intolérants - Résineux et peupliers	- Sapins et feuillus intolérants - Sapins et peupliers

FIG 2. *Extrait de la classification des essences selon les inventaires*

Les classes d'âge utilisées ont également évoluées. Elles sont qualitatives pour l'inventaire de 1973, 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 3).

1973	1984	1992
En régénération	0 an	0 ans
Jeune	1-20 ans	1-15 ans
Mûr régulier	20-40 ans	15-30 ans
Mûr irrégulier	40-60 ans	30-45 ans
Mûr étagé	60-80 ans	45-60 ans
	80-100 ans	60-75 ans
	30/70 ans	75-90 ans
	70/30 ans	90 ans et +

FIG 3. *Classes d'âge utilisées selon les inventaires*

Enfin, les données collectées et les modes de gestion de la forêt ayant évolués, la notion même de peuplement change d'un inventaire à l'autre. Ces modifications associées à l'évolution naturelle de la forêt conduisent à des modifications des frontières du découpage en zones. L'organisation elle-même de la dimension est modifiée à mesure des changements de la réglementation et des avancées scientifiques dans le domaine de la foresterie et de l'écologie. Ainsi, en 1973, la dimension spatiale est constituée du seul niveau *Peuplement*, en 1984, et 1992, cette hiérarchie a trois niveaux recouvrant des concepts différents.

L'étude des sources d'information fait donc apparaître plusieurs types d'évolutions :

- pour l'âge et l'essence, l'évolution est au niveau des membres. Dans le cas de l'âge, on passe d'une valuation qualitative à une valeur discrète puis à une modification de cette discrétisation. Dans le cas de l'essence, c'est la classification qui change d'un inventaire à l'autre.

- pour le découpage en zone, il s'agit d'une évolution du schéma de la dimension, {*peuplement}* pour 1973, {*peuplement forestier* < *compartiment* < *unité de paysage}* pour 1984, et {*peuplement écoforestier* < *polygone écologique* < *Station forestière}* pour 1992.

- pour le niveau *Peuplement*, les frontières des zones sont modifiées, il s'agit donc d'une évolution des membres du grain le plus fin.

Fusionner l'ensemble des données dans une structure multidimensionnelle fixe n'intégrant que les formes invariantes des dimensions permettrait de faire des études d'évolution. Cependant, cette solution conduit à ne conserver que les seules valeurs immédiatement comparables. Cette option ne se justifie pas lorsque l'on souhaite approfondir l'analyse à une période donnée puisque les données réelles ont été perdues et elle ne permet pas d'étudier l'évolution spatiale selon des caractéristiques spécifiques à un inventaire (par exemple, le suivi de l'évolution des arbres jeunes sur les 3 inventaires). De même, ramener la représentation cartographique des données sur un seul découpage (par exemple, le plus récent) ne correspond pas aux besoins des utilisateurs. En effet, ils souhaitent pouvoir projeter les données d'une période de temps dans un découpage associé à une autre période de temps pour suivre par exemple l'évolution des essences d'une zone donnée. Il est donc essentiel d'une part, de maintenir l'ensemble des données dans l'entrepôt et d'autre part, de permettre la représentation d'une mesure dans une version choisie par

l'utilisateur. Nous avons défini un modèle appelé M3 (Modèle Multidimensionnel Multiversion) qui permet de gérer les évolutions temporelles de manière explicite (Body et al, 2003). Nous proposons d'adapter ce modèle pour satisfaire les besoins d'évolution dans les applications SOLAP et en particulier en foresterie.

3 Le modèle M3

Notre modèle se fonde classiquement sur la notion de table de faits qui regroupe l'ensemble des mesures représentant les données à analyser et sur les dimensions qui constituent les axes d'analyse (Cabibbo et al, 1998). Dans sa première partie, notre proposition s'appuie sur les travaux complémentaires de Mendelzon (Mendelzon, Vaisman 2000) et d'Eder (Eder, Koncilia, 2001). Les premiers donnent les bases pour prendre en compte les évolutions dans les structures dimensionnelles alors que les seconds fournissent des solutions pour utiliser les connaissances sur les évolutions afin de représenter les données dans une référence donnée. Ainsi, pour prendre en compte les évolutions temporelles, nous avons redéfini les éléments de base d'une structure multidimensionnelle en y ajoutant la notion de temps valide. Une dimension évolutive est donc composée de membres (appelés *versions de membres*) et de liens hiérarchiques (appelés *filiations temporelles*) ayant des temps de validité. Ces éléments permettent de définir la notion de *table de faits temporellement consistante* et de *structure temporelle multidimensionnelle*. Ensuite, nous introduisons de nouveaux concepts pour représenter les évolutions au sein d'une dimension : *les relations de mapping* qui relient les versions de membres ayant un lien historique et *les indices de confiance* qui indiquent la qualité des nouvelles données mappées. L'ensemble de ces éléments permet de construire *la table de faits multiversion* qui regroupe toutes les données selon *les différents modes de représentation* (c'est-à-dire les différentes versions) pour les rendre accessibles à l'utilisateur.

Définition 1 (Version de Membre). Un membre étant un objet ou une entité abstraite d'un intérêt particulier pour l'utilisateur final, une version de membre est définie comme l'état invariant et cohérent d'un membre, sur une tranche temporelle donnée. Une version de membre est représentée par un tuple $<Id, Nom, [A], [Niveau], t_i, t_f>$ où :

Id est un identifiant unique pour cette version du membre.
Nom est le nom du membre.
A est un ensemble d'attributs/valeurs pour cette version de membre.
Niveau est le nom du niveau hiérarchique auquel appartient cette version de membre. Cette propriété est optionnelle mais si elle est définie pour une version de membre, elle doit l'être pour toutes les versions de membres de la dimension. Cela signifie que la dimension est organisée en hiérarchie explicite.
[ti, tf] est la fenêtre temporelle sur laquelle cette version de membre est valide.

Un membre (au sens classique) peut donc avoir plusieurs versions avec éventuellement des périodes de recouvrements.

Définition 2 (Filiation Temporelle). Une filiation temporelle établit un lien hiérarchique entre deux versions de membres. Elle est définie par un tuple $<Id\_from, Id\_to, t_i, t_f>$ où :

Id_from est l'identifiant de la version de membre, enfant dans la relation,
Id_to est l'identifiant du parent dans cette relation,
$[t_i, t_f]$ est la fenêtre temporelle sur laquelle cette relation est valide.

Notons que cette fenêtre temporelle doit être incluse dans l'intersection des fenêtres de validité de chacun des versions de membres intervenant dans la filiation.

Définition 3 (Dimension Evolutive). Une dimension évolutive D est un tuple <*Did, Dname, D, G*> où :

Did est un identifiant unique pour la dimension,
Dname est le nom de la dimension,
D est un ensemble de versions de membres,
G est un ensemble de filiations temporelles définissant la structure hiérarchique et temporelle de la dimension.

Une dimension peut donc être représentée par un graphe orienté où les éléments de *D* sont les nœuds et ceux de *G,* les arêtes orientées. Des contraintes existent sur un tel graphe :

On note, pour un temps *t* donné, D*(t)*=< D*id, D(t),G(t)*> le sous-graphe extrait de D, où $D(t) = \{d \in D \mid t \in [t_i^d ; t_f^d]\}$ est la restriction de *D* à l'ensemble de ses éléments valides au temps *t* et $G(t) = \{g \in D \mid t \in [t_i^g ; t_f^g]\}$ est la restriction de *G* à l'ensemble de ses éléments valides au temps *t*.

Le sous-graphe D*(t)* doit être un graphe orienté et acyclique afin de permettre les agrégations des mesures vers les niveaux hiérarchiques supérieurs.

Par la suite, nous désignerons par *versions de membres-feuilles*, les versions de membres n'ayant pas de fils, pour au moins un instant *t*. En effet, ces versions sont les feuilles du graphe D*(t)*.

Définition 4 (Niveau d'une dimension). Si la propriété *Niveau* est donnée aux versions de membres de la dimension, alors un niveau d'une dimension est défini par l'ensemble des versions de membres ayant la même valeur pour cette propriété. Les niveaux sont simplement les classes d'équivalence sur l'ensemble des versions de membres d'une dimension, pour la relation d'égalité de la valeur *Niveau*. En revanche, si cette valeur n'est pas définie, les niveaux sont constitués des versions de membres de même profondeur, pour un instant *t*, dans le graphe D*(t)* orienté et acyclique associé à la dimension D. On note que l'ensemble des versions de membres constituant un niveau évolue au cours du temps, ce qui découle en fait des évolutions même de la dimension.

Définition 5 (Table de Faits Temporellement Consistante). Etant données *n* dimensions évolutives $D_1 \ldots D_n$, une dimension temporelle *T* et un ensemble de *m* mesures $M=\{m_1,\ldots,m_m\}$, une table de faits temporellement consistante est définie par une fonction de la forme :

$$f : D_1 \times \ldots \times D_n \times T \rightarrow dom(m_1),\ldots,dom(m_m)$$
$$d_1,\ldots,d_n,t \rightarrow v_1,\ldots,v_m$$

où D_i est l'ensemble des versions de membres de la dimension D_i et $dom(m_k)$ est le domaine des valeurs de la mesure m_k. Cette fonction associe à un ensemble de versions de membres-feuilles d_i valides au temps *t*, les valeurs v_k acquises pour la mesure m_k.

Afin de conserver les liens de transition entre les versions de membre, des relations de mapping et des indices de confiance associés sont introduits. Ces notions seront utilisées pour construire par la suite la Table de Faits MultiVersion.

Définition 6 (Indice de Confiance). Un indice de confiance est une valeur décrivant la fiabilité des données. Il permet notamment de distinguer les données sources des données

mappées. Pour ces indices de confiance, une fonction d'agrégation $\otimes_{cf}$ doit être définie pour propager cette notion aux données agrégées. Cette fonction peut être définie soit par une fonction explicite, si les indices de confiance sont définis de façon quantitative, soit par une table de vérité, s'ils sont définis de façon qualitative.

Définition 7 (Relation de Mapping). Une relation de mapping est définie par un tuple $<Id\_from, Id\_to, F, F^{-1}>$ où :

Id_from est l'identifiant de la version de membre-feuille avant évolution,

Id_to est l'identifiant de la version de membre-feuille après évolution,

F est un ensemble de *m* couples $<fm_k, cf_k>$ où fm_k est une fonction de mapping, de $dom(m_k)$ dans lui-même, précisant comment la mesure m_k doit être mappée. cf_k est l'indice de confiance associée à cette fonction fm_k,

F^{-1} est l'ensemble de *m* couples $<fm'_k, cf'_k>$ définissant le passage inverse de la version Id_to vers la version Id_from.

Les relations de mapping ne sont pertinentes que pour les versions de membres apparaissant dans la table de faits temporellement consistante, c'est-à-dire pour les versions de membres-feuilles. En effet, seules ces versions de membres sont associées à des données sources et pourront donc être utilisées par les fonctions de mapping. Pour les versions de membres non feuilles, les valeurs des mesures sont simplement calculées (agrégées) à partir des valeurs de leurs fils.

Définition 8 (Structure Temporelle Multidimensionnelle). Une structure temporelle multidimensionnelle *STM e*st définie par un tuple $<\{D_1, ..., D_n, T\}, RM, f>$ où les D_i sont des dimensions évolutives, *T* est la dimension temporelle, *RM* est un ensemble de relations de mapping et *f,* une table de faits temporellement consistante.

Définition 9 (Version de Structure). Une version de structure V est représentée par un tuple $<VS_{id}, \{D_{1,VSid}, ..., D_{n,VSid}\}, t_i, t_f>$ où Vs_{id} est un unique identifiant de la version de structure et $[t_i, t_f]$ donne la fenêtre temporelle de validité pour cette version. Chaque $D_{i,VSid}$ est la restriction de la dimension D_i à l'ensemble de ses éléments valides (versions de membres et filiations temporelles) pour tout $t \in [t_i, t_f]$. De façon moins formelle, on peut voir les versions de structure comme la représentation des dimensions évolutives valide sur une période où aucun changement ne se produit dans leur structure.

Ces versions forment une partition de l'axe temporel (ou partition historique). Elles peuvent se déduire de l'intersection des fenêtres de validité de chacune des versions de membres et de leurs filiations temporelles.

Par l'introduction du concept de Mode Temporel de Présentation, nous donnons la possibilité à l'utilisateur de choisir entre les différentes représentations possibles.

Définition 10 (Modes Temporels de Présentation). Etant données N versions de structure $V_1,\ldots, V_N$ (déduites d'une structure temporelle multidimensionnelle STM), $MTP = \{mtc, MV_1,..., MV_N\}$ représente l'ensemble des modes temporels de présentation pour les données. *mtc* symbolise le mode temporellement consistant et les MV_i dénotent les modes temporels où les données sont rapportées dans la version de structure V_i.

Définition 11 (Table de Faits MultiVersion). Etant donnés une structure temporelle multidimensionnelle *STM* et l'ensemble *MTP* des modes temporels de présentation qui lui sont associés, la table de faits multiVersion est définie comme une fonction f' telle que:

$$f': D_1 \times ... \times D_n \times T \times MTP \rightarrow dom(m_1) \times ... \times dom(m_m) \times CF^m$$
$$d_1,...,d_n,t,mtp \rightarrow v_1,...,v_m,cf_1,...,cf_m$$

où *CF* est le domaine de valeurs des indices de confiance. Cette fonction associe les valeurs v_k obtenues pour les mesures m_k ainsi que leur indice de confiance cf_k, à un ensemble de versions de membres-feuilles d_i, non forcément valides au temps *t* donné mais en revanche valides pour le mode *mtp* donné (i.e. $d_i \in D_{i,mtp}$), à un instant *t* et à un mode temporel de présentation *mtp*.

Notons que cette table inclut la table de faits temporellement consistante précédemment définie. Nous avons en effet :

$$f'\Big|_{D_1 \times ... \times D_n \times T \times \{mtc\}} = f \times \{ds\}^{Card(M)}$$

où *mtc* est le mode temporel de présentation en temps consistant et *ds* est l'indice de confiance associé aux données sources (sans mapping).

Cette table peut être déduite de la structure temporelle multidimensionnelle telle que précédemment définie, elle se construit en effet à partir de la connaissance des dimensions évolutives, des relations de mapping et de la table de faits temporellement consistante.

Définition 12 (Agrégation des données). Supposons donnés une fonction d'agrégation $\oplus_{m_k}$ pour chaque mesure m_k, $\otimes_{cf}$ la fonction d'agrégation pour les indices de confiance, *mtp* un élément de *MTP* et *t* un instant de l'axe temporel. Les agrégations de données dans le cube peuvent être aisément calculées à partir de la table de faits multiVersion et des filiations temporelles entre versions de membres. Pour établir les données associées à une version de membre non-feuille *d* de la dimension D_I, dont G_I est l'ensemble des filiations temporelles, nous pouvons trouver les enfants (versions de membres-feuille) d_1, …, d_J de *d*. Supposons également que nous avons :

$$\forall j \in [1,J] \quad f'(d^j,d_2...,d_n,t,tmp) = v_1^j,...,v_m^j,cf_1^j,...,cf_m^j$$

Nous pouvons alors obtenir les valeurs pour *d* par :

$$f'(d,d_2,...,d_n,t,mtp) = \bigoplus_{m_1}^{j=1,J} v_1^j,...,\bigoplus_{m_m}^{j=1,J} v_m^j, \bigotimes_{cf}^{j=1,J} cf_1^j,...,\bigotimes_{cf}^{j=1,J} cf_m^j$$

De telles opérations sont réalisées pour construire le cube OLAP.

4 Application du modèle M3 à la foresterie

4.1 Construction du graphe temporel

Comme retenu dans le modèle M3, chaque version de membre a un temps de validité affecté, correspondant à l'inventaire dont il est issu. Ainsi, la version du membre *Résineux et Bouleaux* de la dimension *Essence* est définie par :

< Résineux&bouleaux_Id, «Résineux et Bouleaux »,1,1973,1984>

Cette définition indique que la version du membre *Résineux et Bouleaux* est valide pour [1973,1984]. Les versions des membres *Sapins et Bouleaux Blancs* et *Sapins et Bouleaux Jaunes* de la même dimension, uniquement valides à partir de 1992, sont définies par

< Sapins&bouleaux_blancs_Id, « Sapins et bouleaux_blancs »,1,1992,Now>

< Sapins&bouleaux_jaunes_Id, « Sapins et bouleaux_jaunes »,1,1992,Now>

La dimension est hiérarchisée selon *essence détaillée* < *classe d'essence* < *type d'essence*. Les membres du niveau supérieur ont été choisis de façon à ce que leur temps de validité porte sur les 3 inventaires. Ainsi, la version du membre *Mélangés* du niveau *classe d'essence* est défini par

< Mélangés_Id, « Mélangés »,2,1973,Now>

Les trois versions de membres précédemment décrites ont une filiation temporelle avec la version du membre *Mélangés* du niveau supérieur. Ces liens hiérarchiques héritent du temps de validité de la version de membre fils (figure 4).

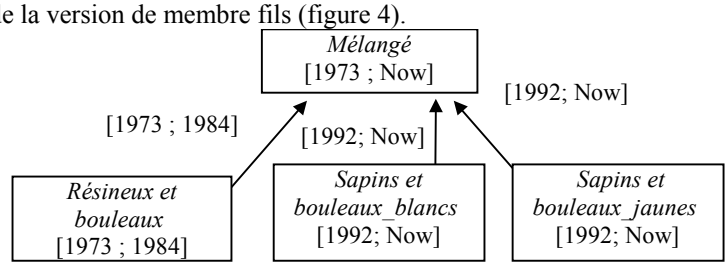

FIG. 4 - *Extrait de la dimension Essence*

La même méthode de constructions de graphes temporels est appliquée aux dimensions Essence, Age, Zone et Densité qui sont également évolutives dans le temps comme nous l'avons vu dans l'étude de cas.

4.2 Relations de mapping et indices de confiance

4.2.1 Relations de mapping pour les dimensions thématiques

Pour passer d'une version de structure à une autre, c'est-à-dire représenter une mesure de l'entrepôt obtenue lors d'un inventaire dans la structure correspondant à un autre inventaire, il est indispensable de définir les relations de mapping d'un membre à l'autre. Ainsi, en reprenant l'exemple des membres de la dimension *Essence*, le membre *Résineux et Bouleaux* est lié aux membres *Sapins et Bouleaux blancs* et *Sapins et Bouleaux Jaunes* par :

<Résineux&Bouleaux_id, Sapin&Bouleaux_jaunes_id, $\{(x \mapsto 0.4x, \text{am})\}$, $\{(x \mapsto x, \text{em})\}$>
<Résineux&Bouleaux_id, Sapin&Bouleaux_blancs_id, $\{(\text{x} \mapsto 0.6\text{x}, \text{am})\}$, $\{(x \mapsto x, \text{em})\}$>

Ces relations de mapping traduisent le fait que, dans la forêt considérée, on estime que les Résineux et Bouleaux se répartissent pour 40% dans la catégorie des Sapins et Bouleaux jaunes et pour 60 % dans la catégorie des Sapins et Bouleaux jaunes. Cette répartition est une approximation comme l'indique l'indice de confiance *am* (Approximate Measure). L'opération inverse (ici le passage de *Sapins et Bouleaux blancs* à *Résineux et Bouleaux*) est par contre un mapping exact *em* (Exacte Measure).

4.2.2 Relations de mapping pour la dimension spatiale

Pour pouvoir extraire les relations de mapping entre les découpages correspondant aux différents inventaires, nous avons choisi de passer par l'intermédiaire d'une référence spatiale invariante dans le temps. Cette méthode repose sur une structuration 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, un peuplement devenant un ensemble de cellules. Ce maillage de la surface permet de déduire la relation de mapping exact reliant chaque zone dans sa version i à sa version i+1. Ainsi, la figure 5-a représente une forêt initialement composée de 2 peuplements P1 et P2. Lors de l'inventaire suivant, trois peuplements sont détectés (figure 5-b) PF1, PF2 et PF3. A partir du maillage représenté, on en déduit les relations de mapping suivantes :

<P1,PF1, $\{(x \mapsto \frac{1}{2}x, em)\}, \{(x \mapsto x, em)\}$ >;<P1,PF2, $\{(x \mapsto \frac{1}{3}x, em)\}\{(x \mapsto \frac{2}{3}x, em)\}$>

<P1,PF3, $\{(x \mapsto \frac{1}{6}x, em)\}\{(x \mapsto \frac{1}{2}x, em)\}$ >;<P2,PF2, $\{(x \mapsto \frac{1}{2}x, em)\}\{(x \mapsto \frac{1}{3}x, em)\}$ >

<P2,PF3, $\{(x \mapsto \frac{1}{2}x, em)\}\{(x \mapsto \frac{1}{2}x, em)\}$ >

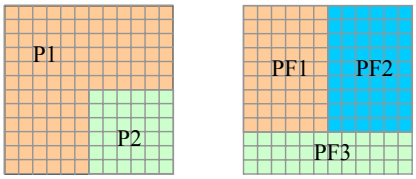

FIG. 5 - *Exemple de l'utilisation du maillage de la surface*

Ces relations de mapping traduisent que l'on peut déduire les mesures associées à PF1, PF2 et PF3 à partir des mesures de P1 et P2 par les relations suivantes :

$$PF1 = \frac{1}{2}\ P1\ ;\ PF2 = \frac{1}{3}\ P1 + \frac{1}{2}\ P2\ ;\ PF3 = \frac{1}{6}\ P1 + \frac{1}{2}\ P2$$

De même, Les mesures associées à PF1, PF2 et PF3 seront représentées selon P1 et P2 en utilisant les fonctions de mapping inverses :

$$P1 = PF1 + \frac{2}{3}\ PF2 + \frac{1}{2}\ PF3\ ;\ P2 = \frac{1}{3}\ PF2 + \frac{1}{2}\ PF3$$

On notera qu'il n'y a pas de relation de mapping entre P2 et PF1, ce qui traduit qu'il n'y a aucun lien entre ces deux versions de membres comme le montre la figure 5.

4.3 Modes Temporels de Présentation

Dans le modèle M3, les modes temporels de présentation regroupent toutes les versions de structure et le mode temporellement consistant. Dans l'exemple de la foresterie, nous avons donc 4 modes temporels de présentation : un mode pour chacun des 3 inventaires

(*MV1973*, *MV1984* et *MV1992*) et le mode temporellement consistant *mtc* correspondant aux données telles quelles ont été acquises.

4.4 Table de faits multiversion

La table de faits multiversion regroupe les différentes mesures selon les dimensions spatiale et descriptives précédemment définies et selon les différents modes temporels de présentation. Les modes temporels de présentation interviennent au niveau de la table de faits multiversion comme une dimension supplémentaire. Les indices de confiance sont des mesures supplémentaires ajoutées à la table de faits. Ainsi, le modèle logique de l'entrepôt multiversion peut être représenté selon la figure 6.

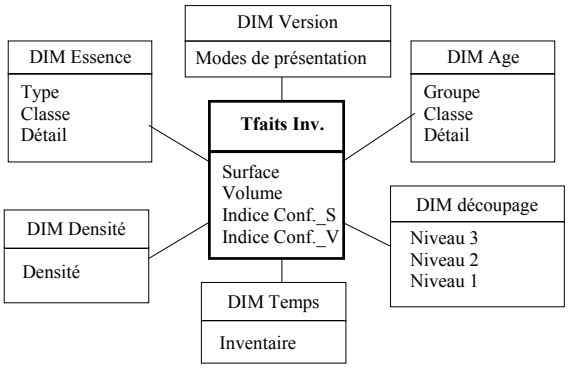

FIG. 6 - *Table de faits multiversion*

Cette solution présente l'avantage de pouvoir naviguer dans les différents modes de présentation comme dans une dimension classique et d'obtenir les indices de confiance des mesures agrégées par des opérations d'aggrégation sur les indices de confiance associés aux mesures non agrégées.

5 Conclusion

L'implémentation de notre modèle a été effectuée avec Microsoft SQL-Server 2000. Nous avons utilisé les dimensions de type parent-enfant qui permettent de représenter des hiérarchies implicites qui sont déduites des liens existant entre les instances. Notre architecture se compose de trois parties : un entrepôt temporel qui contient les données temporellement consistantes et les métadonnées (dont les relations de mapping), l'entrepôt multiversion qui introduit la dimension modes temporels de présentation et implémente la table de faits multiversion et enfin le cube OLAP créé par le serveur OLAP SQL Server 2000 Analysis Services). Une description plus détaillée de cette implémentation pourra être trouvée dans (Body et Al, 2002).

L'introduction des évolutions temporelles dans les modèles multidimensionnels est particulièrement intéressante pour les applications SOLAP où les données varient à la fois selon la dimension spatiale et les dimensions descriptives. La prise en compte des évolutions temporelles accroît les capacités de navigation de l'utilisateur en lui donnant la possibilité de choisir le type de présentation qu'il souhaite, par exemple mapper les données les plus

récentes dans un découpage géographique plus ancien afin de mieux en évaluer les évolutions. Toutes les informations recueillies sont accessibles quelle que soit la représentation choisie, les données étant alors mappées ou d'origine. L'utilisation des indices de confiance nous permet de qualifier la donnée en fonction de sa qualité. Ce type d'approche devrait pouvoir s'appliquer à des évolutions autre que temporelles. Ainsi, nous étudions actuellement l'extension du modèle pour la prise en compte de version de nomenclatures ou de classification et plus généralement, de versions fonctionnelles.

Références

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

Bédard Y., Merrett T., Han J. (2001), Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery, in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS series edited by Peter Fisher and Jonathan Raper, Taylor & Francis, 2001

Blaschka M., Sapia C., Höfling G. (1999), On Schema Evolution in Multidimensional Databases, Proceedings of DaWak'99 Conference, Florence, Italy, 1999.

Body M., Miquel M., Bedard Y., Tchounikine A. (2003), Handling Evolutions in Multidimensional Structure, IEEE International Conference on Data Engineering, ICDE, March 5-8 2003, Inde

Body M., Miquel M., Bedard Y., Tchounikine A (2002), A Multidimensional and Multiversion Structure for OLAP Applications, ACM Fifth International Workshop on Data Warehousing And Olap (DOLAP) in McLean, VA, USA, on November 8, 2002.

Cabibbo L., Torlone R. (1998), A Logical Approach to Multidimensional Databases, Proceedings of the 6th International Conference on Extending Database Technology (EDBT'98), Valencia, Spain.

Eder J., Koncilia C. (2001), Evolution of Dimension Data in Temporal Data Warehouses, Proceedings of the DaWaK'01 Conference, Munich, Germany, 2001.

Kimball R. (1996), The Data Warehouse Toolkit, J.Wiley and Sons, Inc, 1996.

Mendelzon A.O., Vaisman A. (2000), Temporal Queries in OLAP, Proceedings of the 26th VLDB'00 Conference, Cairo, Egypt, 2000.

Pedersen T.B., Jensen C.S., Dyreson C.E. (2001) Foundation for capturing and querying complex multidimensional data, Information Systems Special Issue: Data Warehousing, Vol 26, No 5.

Summary

Most of multidimensional models consider facts as the dynamic part of the data warehouse whereas dimensions are seen as static entities. However, in practical applications, the structures of analysis may evolve with time. These problems become particularly complex with SOLAP (Space One Line Analytical Processing) applications. Indeed, the space dimension, when it describes a territorial division, can be modified, and these modifications affect the measures. In this article we show how the model we propose, named M3, makes it possible to build a hypercube preserving the whole of the evolutions. It also offers navigation and comparative exploration of the data in the various versions of dimensions. Our proposal is illustrated in the field of forestry.