Approche de modélisation multidimensionnelle des données complexes : Application aux données médicales

Sid Ahmed Djallal Midouni, Jérôme Darmont, Fadila Bentayeb

Approche de modélisation multidimensionnelle des données complexes : Application aux données médicales

Sid Ahmed Djallal Midouni*, Jérôme Darmont**, Fadila Bentayeb**

* Département d'informatique
Université Abou Bekr Belkaid - Tlemcen
B.P.230- Tlemcen 13000, Algérie
djmidouni@hotmail.com
** Équipe BDD, Laboratoire ERIC
Université Lumière – Lyon 2
5 avenue Pierre Mendès-France
69676 Bron Cedex, France
{jdarmont | bentayeb}@ eric.univ -lyon2.fr

Résumé. La vocation d'un entrepôt de données est l'analyse de données pour l'aide à la décision dans les entreprises. La modélisation multidimensionnelle est la base des entrepôts de données et de l'analyse en ligne (OLAP). Ces techniques sont efficaces pour traiter les données simples numériques, mais elles ne sont pas adaptées aux données variées et hétérogènes provenant de différentes sources, appelés communément données complexes. Dans cet article, nous abordons le problème de la modélisation multidimensionnelle des données complexes à travers le cas des données médicales du projet MAP (Médecine d'Anticipation Personnalisée). Nous proposons un métamodèle multidimensionnel étendu pour les données médicales en généralisant le modèle cardiovasculaire du projet MAP. Enfin, nous avons spécifié et réalisé un outil d'aide à la conception d'entrepôt de données médicales.
Mots Clés : Modélisation des entrepôts de données, modèle multidimensionnel, analyse en ligne, données complexes.

1 Introduction

L'intérêt pour l'analyse de données s'est développé énormément ces dernières années. Les entreprises se sont rendues compte de l'efficacité de la technologie OLAP (*On-line Analytical Processing*) dans l'analyse et l'exploration des données. Cette technologie est utilisée dans les systèmes d'aide à la décision. Le plus souvent, ces systèmes sont basés sur des techniques d'entreposage de données pour exploiter la grande masse d'informations disponibles dans les entreprises à des fins d'analyse et d'aide à la décision.

La manière la plus appropriée pour faciliter cette analyse OLAP est la modélisation multidimensionnelle des données. Cette dernière représente les données comme des points dans un espace multidimensionnel, Kimball (1996), Vassiliadis et Sellis (1999).

La modélisation multidimensionnelle est donc une technique qui vise à organiser les données de telle sorte que les applications OLAP soient performantes et efficaces. Cependant, cette technique n'est pas adaptée à un certain type de données, dites complexes.

Depuis quelques années, la nécessité de gérer et de traiter ce type de données n'a cessé de s'accentuer à cause de leur variété (texte, image, son, vidéo, etc.). Cette variété de données met clairement en évidence la nécessité de créer de nouveaux modèles multidimensionnels pour ces nouveaux types de données qui sont qualifiées de complexes. C'est dans ce contexte que doit être repensée la modélisation multidimensionnelle.

Les modèles existants offrent un cadre agréable pour mener la modélisation multidimensionnelle des données simples, mais ils ne sont pas adaptés aux données complexes. En effet, les données complexes comportent des mesures non additives, non agrégeables et qui ont des niveaux de granularité différents, ce qui rend leur intégration dans des structures multidimensionnelles plus difficile.

Le présent travail vise à apporter des solutions au problème de la modélisation multidimensionnelle de données complexes, en l'occurrence les données médicales du projet MAP (Médecine d'Anticipation Personnalisée[1]). Notre objectif est de proposer un modèle multidimensionnel pour ces données biomédicales, plus particulièrement pour les données du magasin cardiovasculaire et de généraliser ce modèle vers un métamodèle pour entrepôts de données médicales. Le rôle de cet entrepôt est d'intégrer et de stocker toute information utile aux médecins MAP et de conserver l'historique des données médicales pour permettre les analyses nécessaires aux prises de décision.

Outre cette introduction, nous présentons dans la section 2 une définition des données complexes suivie des principaux travaux traitant la modélisation multidimensionnelle des données, plus précisément des données complexes. La section 3 est relative à notre contribution par la proposition du modèle multidimensionnel du module cardiovasculaire qui sera généralisé par la suite vers un métamodèle permettant de prendre en charge tous les types de données du projet MAP. La section 4 décrit une implémentation possible de ce métamodèle dans une base de données relationnelle ainsi que la manière de l'instancier pour définir les autres magasins de données du projet MAP. La dernière section conclut ce travail et présente quelques perspectives d'utilisation et de recherche ouvertes par ce métamodèle.

2 Modélisation multidimensionnelle des données complexes

La description des données complexes nécessite une certaine précision et un espace de représentation adapté. A ce jour, il n'existe pas de modèle universel pour toutes les formes de données complexes. Les données sont qualifiées de complexes si elles sont, Darmont et al. (2005):

- multiformats : l'information est représentée sous différents formats (base de données, données numériques, symboliques, textes, images, sons, vidéos...) ; et/ou
- multistructures : les données peuvent être structurées, non structurées ou semi-structurées (bases de données relationnelles, collection de documents XML...); et/ou
- multisources : les données proviennent de différentes origines (bases de données réparties, Web...) ; et/ou
- multimodales : un même phénomène est décrit par plusieurs canaux ou points de vue (radiographies et diagnostic audio d'un médecin pour évaluer l'état de santé

[1] Projet CREALYS de création d'entreprise porté par le Dr Jean-Marcel Ferret et cofinancé par la Région Rhône-Alpes et l'Université Lumière Lyon 2

d'un patient, données exprimées dans des échelles ou des langues différentes...) ; et/ou

- multiversions : les données sont évolutives en termes de définition ou de valeur (bases de données temporelles, recensements périodiques dont les critères évoluent...).

Le modèle de données multidimensionnel est le cœur d'un système décisionnel, il est l'objet de plusieurs travaux. Certains proposent des langages algébriques pour faciliter l'interrogation et la manipulation des données de l'entrepôt, Agrawal et al. (1995), Cabibbo et Torlone (1998), Pedersen et Jensen (1999), Pokorny et Sokolowsky (1999), Ravat et al. (2001), Teste (2000).

Ces différentes propositions sont parfaitement adaptées aux applications de données classiques, mais ne répondent pas complètement aux exigences des applications à base de données complexes telles que les applications médicales. La majorité de ces travaux ne prennent pas en compte les objets de structure complexe. Cependant, Olivier Teste a spécifié des modèles de représentation et des langages de manipulation qui sont dédiés aux entrepôts et magasins de données complexes et évolutives et qui sont basés sur le paradigme objet, Teste (2000). Il a intégré par ailleurs dans son modèle la dimension temporelle afin de conserver l'évolution des données de manière pertinente.

L'intégration et la structuration des données complexes dans une base de données classique ont déjà été réalisées, Darmont et al. (2002). Ces structures permettent la gestion et la consultation des données mais elles ne sont pas appropriées à l'analyse des données. Le plus souvent, les données complexes sont stockées dans les bases de données pour qu'elles soient retrouvées plus facilement.

Tanasescu et al. ont conçu un modèle UML générique basé sur un modèle général pour mieux identifier et représenter tous les types des données complexes afin qu'elles soient prêtes au processus de modélisation multidimensionnelle, Tanasescu (2003), Darmont et al. (2002). Dans le même article, les auteurs ont proposé l'utilisation des techniques de fouille de données permettant l'extraction des caractéristiques des données complexes en vue de leur modélisation multidimensionnelle.

Les efforts de modélisation des données spatiales, considérées comme un autre type de données complexes, se concentrent sur la représentation arbitraire des objets géométriques (points, lignes, polygones, etc.) dans un espace multidimensionnel, Guting (1994). La technologie SOLAP est basée sur une structure multidimensionnelle pour supporter l'analyse spatio-temporelle, Rivest et al. (2001). Miquel et al. proposent des solutions pour concevoir ces structures lorsque les sources de données sont hétérogènes des points de vue temporel, spatial et sémantique, Miquel et al. (2001). Ces structures sont ensuite explorées dans l'environnement SOLAP. D'autres auteurs, comme Zghal et al. se sont intéressés aux problèmes de la modélisation multidimensionnelle des données spatiales en se basant sur le développement d'un entrepôt spatial, Zghal et al. (2003).

Dans le domaine médical, Pederson et Jensen proposent un modèle multidimensionnel intégrant des données temporelles et imprécises pour la gestion des patients d'un hôpital, Pedersen et Jensen (1999). Ils ont résolu les problèmes de validité et d'incertitude des données respectivement par l'ajout au modèle du temps de validité et de probabilité. Les mêmes auteurs proposent des solutions d'intégration des documents XML et des données relationnelles dans une base de données multidimensionnelle en vue de leur analyse OLAP, Jensen et al. (2001).

Ravat et al. s'intéressent à l'aide à la décision dans le domaine médical, Ravat et al. (2001). Ils proposent un modèle de données étendu pour les bases de données multidimensionnelles qui améliore l'analyse, le suivi et le contrôle des dépenses de santé, de l'activité des médecins et du comportement consommateur des patients. Cependant ces études se limitent aux données du dossier du patient et ne traitent pas la complexité des données médicales.

Peu de recherches s'intéressent à la modélisation multidimensionnelle des données médicales. Ces travaux se révèlent inadaptés à notre contexte de travail car ils ne prennent pas en compte le problème de l'hétérogénéité des données médicales.

3 Modélisation multidimensionnelle de l'entrepôt médical MAP

Ce travail s'est déroulé dans le cadre du projet MAP (Médecine d'Anticipation Personnalisée). Il s'agit d'apporter des solutions aux problèmes posées par la modélisation multidimensionnelle des données médicales. Les modèles existants tels que les modèles en étoile ou en constellation sont inadaptés aux données médicales du projet MAP. Pour cela, le but de ce travail est de proposer un modèle multidimensionnel qui permette de traiter et d'analyser ce type de données complexes.

L'entrepôt de données médicales MAP est organisé sous forme d'une collection de magasins de données (*datamarts*), Darmont et Olivier (2006, 2008). Chaque magasin contient les données spécifiques, à une spécialité médicale (par exemple, les données des analyses biologiques, les données biométriques, les données cardiovasculaires, etc.) et est défini par un ensemble de faits et de dimensions partagées avec d'autres magasins de données. Suite aux problèmes rencontrés lors de la modélisation multidimensionnelle du magasin cardiovasculaire et à cause de la variété et de la complexité des données, nous allons nous intéresser plus particulièrement au module cardiovasculaire, tout en essayant de généraliser cette solution aux autres magasins du projet MAP.

Notre démarche de modélisation est incrémentale. Nous construisons un nouveau modèle multidimensionnel, **Cardio-M**. Ensuite, nous créons un métamodèle qui généralise le modèle Cardio-M pour pouvoir modéliser les autres magasins de données du projet MAP. En d'autres termes, l'idée derrière cette démarche est de modéliser le module le plus complexe dans l'entrepôt médical MAP afin d'extraire les différents concepts qui vont permettre de créer un métamodèle générique pour générer les autres modules de l'entrepôt MAP.

3.1 Le modèle multidimensionnel du magasin cardiovasculaire

L'étude du module cardiovasculaire a permis d'observer deux sujets d'analyse importants, les résultats des analyses et la conclusion du médecin sur ces résultats, qui sont étudiés selon plusieurs axes d'analyse : Individu, Type d'Examen, Analyse, Temps, Médecin, Machine ou Document.

Pour modéliser les données cardiovasculaires, nous avons utilisé un schéma en étoile. Nous avons mis les trois mesures (*Conclusion*, *Normale* et *Valeur*) dans une seule table de faits, cette table est liée à toutes les dimensions mentionnées ci-dessus. Cependant, ce modèle pose un problème majeur. Nous avons deux mesures qui ne dépendent pas totalement des mêmes dimensions, ce qui nous a amené à proposer un modèle en constellation. Nous

avons mis les deux mesures dans deux tables de faits séparées et entourées chacune par les dimensions appropriées.

Cette solution n'est pas tout à fait adaptée à nos faits qui dépendent l'un de l'autre puisqu'on ne peut pas analyser l'un sans avoir l'autre. Ceci explique nos motivations pour établir le lien entre les deux tables de faits. Les deux mesures ont un degré de granularité différent qui est exprimé par le lien hiérarchique existant entre les tables de faits.

Les données de la table de faits *Examen* sont analysées selon six dimensions (Individu, Type Examen, Temps, Médecin, Machine et Document), tandis que la table de faits *Résultat Examen* est vue selon trois axes d'analyses (Examen, Temps et Analyse).

La table de faits *Examen* contient deux mesures (*Normale* et *Conclusion*), la mesure Normale peut avoir deux valeurs ("oui" ou "non") qui contiendra la réponse à la question : "le patient a-t-il déjà eu une alerte cardiovasculaire ?" et la mesure *Conclusion* contient la conclusion du médecin sur un examen donné.

La table de faits *Résultat Exam*, qui contient les résultats des analyses passées dans un examen donné, est liée à la première table de faits par un lien hiérarchique. Dans la première table, *Examen*, on trouve les valeurs globales et génériques, et dans la deuxième on a le détail de ces valeurs. En d'autres termes, un enregistrement de la table de fait *Examen* est équivalent à un ensemble d'enregistrements de la table *Résultats Exam*, ce qui explique la relation **un à plusieurs** entre ces deux tables.

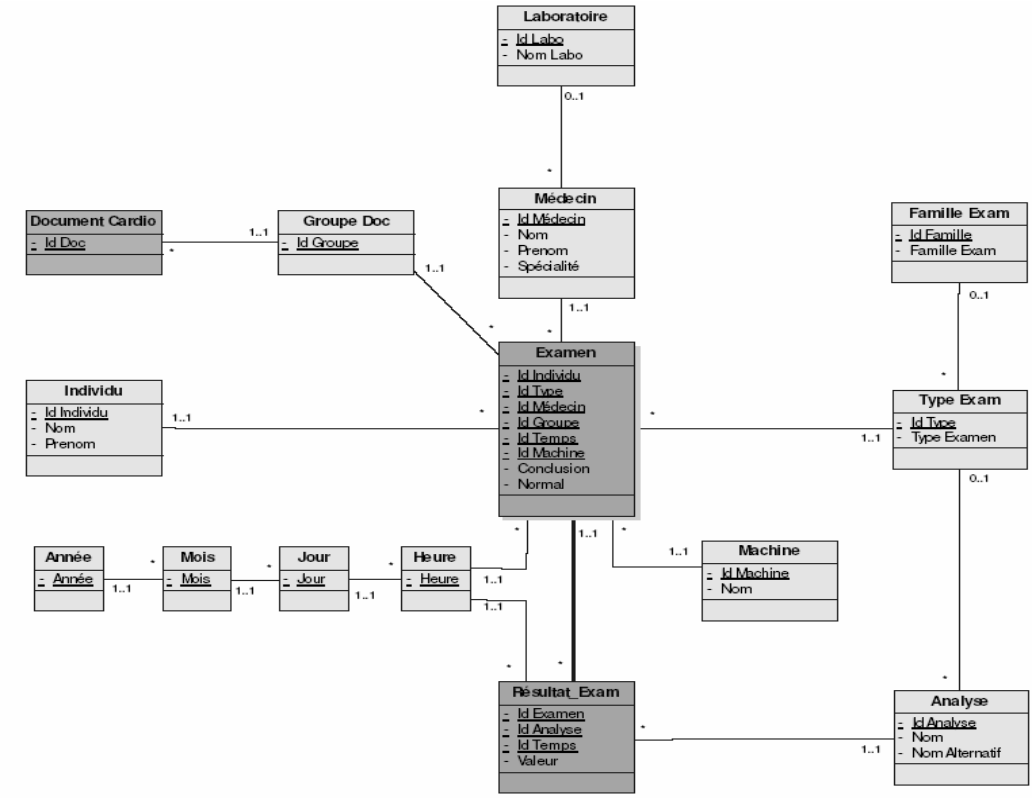

FIG. 1 – Nouveau modèle Cardiovasculaire - Cardio-M.

Pour prendre en compte le lien qui existe entre la mesure *Conclusion* de la table de faits *Examen* et la mesure *Valeur* de la table *Résultat_Exam*, nous avons étendu le modèle en constellation classique de telle façon que ce lien soit représenté. Ainsi, nous avons défini un lien hiérarchique entre la table de faits *Examen* et la table de faits *Résultat_Exam*.

La table de faits *Examen* joue un double rôle dans ce modèle, elle est considérée comme une table de faits par rapport aux dimensions qui sont autour d'elle (Individu, Type Examen, Temps, Médecin, Machine et Document) et elle joue le rôle de dimension par rapport à la table de faits de niveau de granularité plus bas (*Résultat Exam*). Dans notre cas, la table de faits *Examen* contient les résultats agrégés de la table de faits *Résultat Exam*.

Nous constatons dans la partie droite du modèle une hiérarchie liée à la dimension *Analyse* qui est liée à la table de faits *Résultats Exam*. Cette dimension est hiérarchisée en types d'examen et familles d'examen. Cette hiérarchie (*Type_Exam* / *Famille-Exam*) permet de prendre en compte les différents types d'examens du projet MAP (biologiques, biométriques, cardiovasculaires…), elle est commune avec les autres modules du projet MAP.

Décrivons maintenant les dimensions du modèle Cardio-M. Pendant la modélisation du module cardiovasculaire, nous avons constaté qu'il y a un autre type de dimensions, en plus des dimensions classiques et temporelles qui représentent des axes d'analyse, ce sont les **dimensions multimédia** qui vont contenir tous les fichiers sources médicales. On ne peut pas appliquer l'analyse OLAP sur ce type de dimensions mais elles servent d'axe de vérification et de révision pour le médecin en cas de doute. Introduire ce type de dimensions dans les modèles multidimensionnels nécessite des méthodes pour construire un cube de données avec elle. Les trois types de dimension sont détaillés dans la section suivante

3.2 Métamodèle

Afin de modéliser l'entrepôt de données médicales MAP, c'est à dire la collection de magasins de données, nous proposons un métamodèle orienté objet (FIG. 2) permettant de créer les différents magasins de données du projet MAP. Ce métamodèle est une généralisation du modèle cardiovasculaire décrit auparavant avec la prise en compte des faits complexes et les nouveaux concepts définis lors de la modélisation multidimensionnelle du magasin cardiovasculaire.

Peu de travaux proposent des métamodèles multidimensionnels. Leur objectif est de spécifier des métamodèles pour représenter les bases de données multidimensionnelles. Par exemple, Zghal et al. ont spécifié un métamodèle pour la construction d'un entrepôt de données spatiales, Zghal et al. (2003). Un deuxième métamodèle multidimensionnel proposé par Abelló (Abelló, 2002) est basé sur le langage UML pour donner un plus de sémantique en profitant des concepts objets tel que les relations de généralisation et de composition. Un autre métamodèle propose des concepts génériques pour la conception des entrepôts de données tout en essayant de résoudre les problèmes des relations multiples, Harbi et al. (2008). Cependant ces travaux ne sont pas suffisants et ne sont pas adaptés pour représenter les concepts multidimensionnels comme nous souhaitons le faire.

Le standard de l'OMG (*Object Management Group*), CWM[2] (*Common Warehouse MetaModel*), propose un ensemble de métamodèles pour les techniques d'entrepôt de données. Cet ensemble CWM est assez complet pour modéliser un entrepôt de données dans

[2] Voir en annexe : Common Warehouse Metamodel (CWM).

son ensemble. Mais le métamodèle multidimensionnel proposé par CWM représente les aspects multidimensionnels d'une façon générale, il ne prend pas en compte de tous les objets d'une base de données multidimensionnelle, OMG (2003). Il faut le combiner avec d'autres métamodèles du même standard pour avoir une représentation plus complète. Nous avons essayé de prendre en compte, dans un seul métamodèle, tous les composants d'une base de données multidimensionnelle. De plus, le métamodèle CWM ne permet pas de spécifier et de représenter les nouveaux concepts multidimensionnels que nous avons proposés (tables de faits multiples et hiérarchisées, dimensions multimédia).

Notre métamodèle constitue une extension de ces trois derniers travaux qui soit applicable aux données médicales. La figure 2 montre une représentation UML du métamodèle, ce qui nous permet de mieux représenter les concepts multidimensionnels génériques (Dimensions, Faits, Mesures) et les autres concepts extraits de notre étude du module cardiovasculaire (le lien entre les tables de fait, les différents types de dimensions).

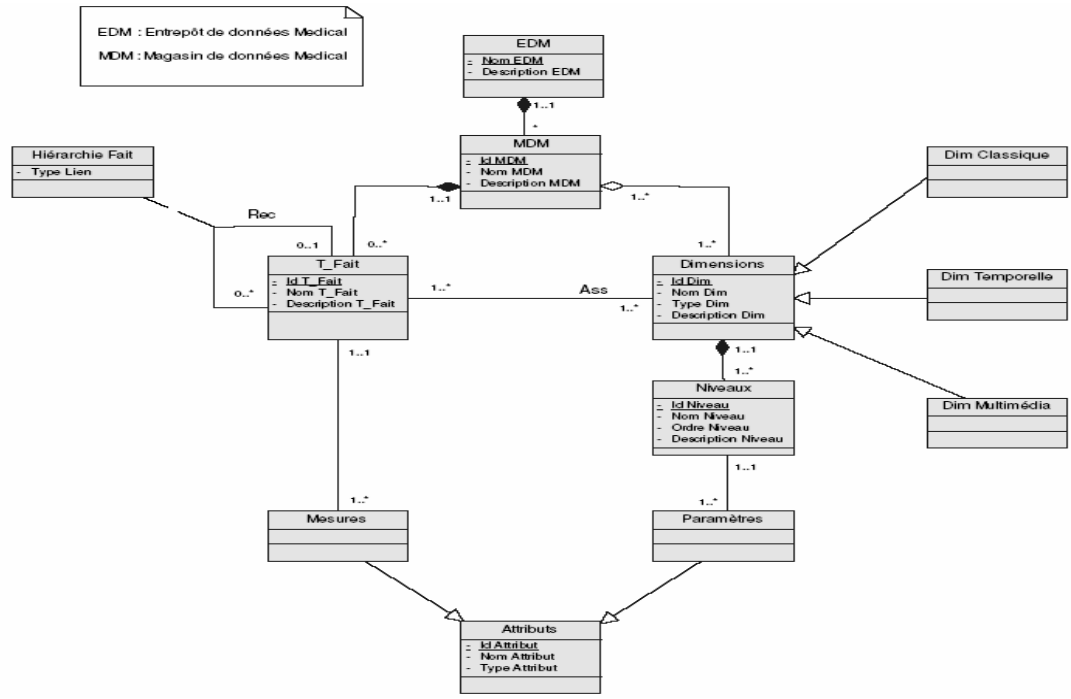

FIG. 2 –Métamodèle MAP.

L'instanciation de ce métamodèle nous permet de créer les différents magasins de données dans l'entrepôt médical du projet MAP. Ce dernier, qui est représenté par la classe EDM, est composé d'un ensemble de magasins de données représenté par la classe MDM. Chaque magasin (MDM) est caractérisé par un ensemble de faits, qui représentent les sujets d'analyses, et un ensemble de dimensions, qui représentent les axes d'analyse. A chaque fait correspond une ou plusieurs mesures et à chaque dimension correspond un ensemble de paramètres. Ces deux derniers concepts (Mesures et Paramètres) héritent de la même classe

Attribut mais ils ont une sémantique différente dans les bases de données multidimensionnelles.

Les faits complexes sont caractérisés par la relation récursive **Rec** qui permet d'associer à chaque table de fait une autre table de faits. Cette relation permet d'exprimer la hiérarchie des tables de faits.

La hiérarchie de dimensions est matérialisée par les deux classes Dimension et Niveaux. On associe à chaque niveau son ordre hiérarchique dans la dimension. Par exemple, l'ordre du niveau Heure de la dimension Temps est égal à zéro et l'ordre de Jour égal à un (FIG. 1).

Nous distinguons trois types de dimensions dans notre modèle : les dimensions classiques, les dimensions temporelles et les dimensions multimédia.

Dimension classique. Ces dimensions servent à enregistrer les valeurs pour lesquelles sont analysées les mesures de l'activité. Une dimension est généralement formée de paramètres (attributs) textuels (pour restreindre la portée des requêtes) et discrets (les valeurs possibles sont bien déterminées et constantes), Kimball (1996).

Dimension temporelle. La dimension temporelle joue un rôle primordial dans les modèles dimensionnels, elle est présente dans tous les magasins de données de notre entrepôt MAP. La dimension temporelle s'ajoute à l'entrepôt pour maintenir l'historique de l'évolution des données médicales dans le temps. Cette dimension est généralement considérée comme une dimension normale.

Dimension multimédia. Ce type de dimension contient les différents types de données multimédias contenues dans notre entrepôt médical. On trouve par exemple : des électrocardiogrammes et des échocardiogrammes. Les données de ce type de dimensions sont difficiles à manipuler par les outils d'analyses actuels, leur but dans notre entrepôt médical est l'archivage pour vérification et contrôle, en cas de doute, des résultats d'analyses. Par exemple, dans le module cardiovasculaire, la dimension Document contient les différents documents multimédias utilisés dans le suivi médical d'un individu donné.

4 Prototype

Afin de valider notre métamodèle (section 3.2), nous avons développé un prototype d'aide à la conception et à la modélisation de notre entrepôt MAP, intitulé **GEDM** (**G**énérateur d'**E**ntrepôt de **D**onnées **M**édicales).

Notre outil facilite la tâche de l'administrateur MAP pour créer et générer des magasins de données afin de construire l'entrepôt global, tout en respectant nos nouveaux concepts définis dans le métamodèle des données médicales. En effet, l'élaboration de l'entrepôt MAP suit un processus de développement à trois niveaux : conceptuel, logique et physique.

La figure 3 décrit ce processus de modélisation. La génération des magasins de données passe généralement par ces trois étapes :

- premièrement, on génère une instance du métamodèle (FIG. 2), qui représente le modèle multidimensionnel du magasin de données en cours de modélisation;
- deuxièmement, on fait la transformation de ce modèle soit vers un fichier XML, soit vers une base de données relationnelle;
- dans la troisième étape, nous choisissons soit un entrepôt XML, soit un entrepôt relationnel, respectivement.

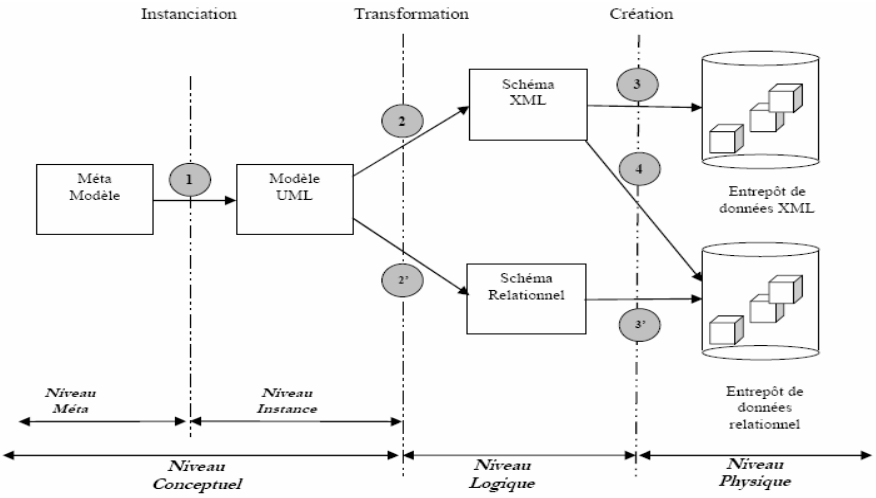

FIG. 3 – Processus de modélisation.

Nous avons opté dans ce travail pour la solution relationnelle. En effet, GEDM est un prototype implanté sur le SGBD Oracle version 10g. Le choix d'un SGBD relationnel est motivé par la grande capacité de stockage ainsi que la performance lors de la manipulation des données. En effet, les systèmes de gestion de bases de données relationnels offrent d'excellentes performances en terme de rapidité d'accès, de volume de stockage et de stabilité des données.

GEDM se base sur une approche incrémentale. L'administrateur MAP élabore l'entrepôt étape par étape en construisant les différents magasins de données du projet MAP. L'architecture de ce prototype, comme le montre la figure 4, est composée essentiellement d'une interface utilisateur et d'un générateur de scripts.

- L'interface utilisateur permet de définir les magasins de données MAP, en introduisant les différents éléments (dimensions, faits…) du schéma dimensionnel.
- Le générateur de scripts est le module responsable de la génération des scripts qui permettent la création du schéma de l'entrepôt de données MAP dans une base de données relationnelle, en s'appuyant sur notre métamodèle.

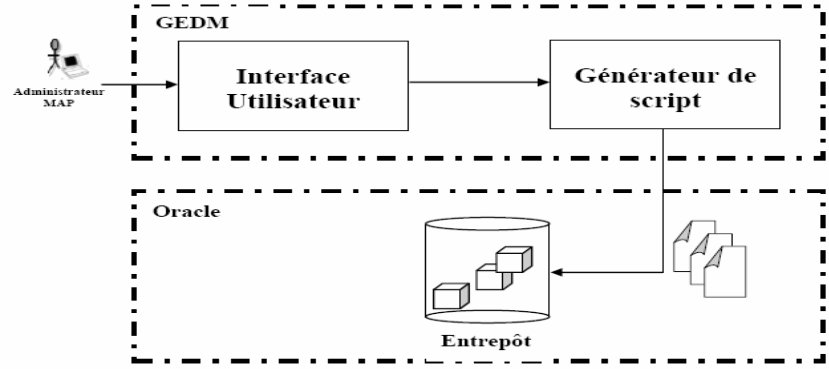

FIG. 4 – Architecture du prototype GEDM.

5 Conclusion et perspectives

Le travail présenté dans cet article traite de la modélisation multidimensionnelle des données complexes. Notre objectif est d'intégrer les données médicales du projet MAP, qui se compose de plusieurs magasins de données, dans une structure multidimensionnelle, pour apporter une aide au processus décisionnel. Pour répondre à cet objectif, nous avons proposé une approche de modélisation et d'implémentation de l'entrepôt médical en nous basant sur un métamodèle que nous avons conçu et développé.

Dans un premier temps, nous avons modélisé le magasin de données le plus complexe du projet MAP, le magasin cardiovasculaire. Pendant cette modélisation, nous avons constaté la difficulté de modéliser et d'intégrer les données médicales telles que les données cardiovasculaires dans une structure multidimensionnelle. Par conséquent, nous avons senti le besoin de proposer de nouveaux concepts qui étendent les modèles existants vers un nouveau type de modèle.

Dans un deuxième temps, nous avons proposé un métamodèle en généralisant le modèle multidimensionnel du module cardiovasculaire. L'apport de notre métamodèle se situe au niveau de la prise en compte des tables de faits multiples et hiérarchisées et des dimensions multimédia. L'instanciation de ce métamodèle permet de spécifier et de définir les différents magasins de données de l'entrepôt MAP indépendamment des plateformes techniques.

Enfin, nous avons développé ce métamodèle en implémentant un prototype GEDM, acronyme de **G**énérateur d'**E**ntrepôt de **D**onnées **M**édicales. Il comporte une interface utilisateur et un module générateur de scripts permettant de créer automatiquement les différents composants de l'entrepôt de données.

Les perspectives que nous envisageons de conduire sont les suivantes.

- Notre approche étant incrémentale, à partir des retours d'usage, nous essayons de faire évoluer le prototype que nous avons réalisé afin de lui permettre une meilleure manipulation de tous les éléments de notre entrepôt.
- Nous travaillons également à la généralisation progressive du métamodèle en ajoutant des nouveaux concepts afin de prendre en compte d'autres types de données complexes (par exemple, définir de nouveaux types de mesures et d'autres types de dimensions). Ainsi l'objectif est l'élaboration de nouveaux modèles de plus haut niveau d'abstraction.
- Nous envisageons également la définition d'une méthodologie de conception et de construction pour les entrepôts de données médicales. A l'heure actuelle, il existe des méthodes de conception des entrepôts de données mais elles ne sont pas adaptées aux données complexes médicales. Nous prévoyons une extension des méthodes existantes afin d'offrir un cadre complet pour concevoir des systèmes décisionnels comportant des données complexes.
- A travers ce travail, nous pensons définir de nouveaux opérateurs OLAP. Il est en effet nécessaire de prévoir l'extension des opérateurs OLAP existants pour prendre en compte les nouveaux concepts définis dans ce travail (les faits multiples hiérarchisés et les dimensions multimédia). Une voie possible, pour les faits multiples, est de s'inspirer des opérateurs OLAP traitant les hiérarchies de dimensions pour définir des nouveaux opérateurs permettant le traitement des faits hiérarchisés.

Références

Abelló A., YAM² (Yet Another Multidimensional Model): A Multidimensional Conceptual Model, PhD Thesis, Universitat Politècnica de Catalunya. Barcelona, April 2002.

Agrawal R., Gupta A., Sarawagi S., Modeling Multidimensional Databases, Research Report, IBM Almaden Research Center, San Jose (California), 1995. ICDE'97 p 232-243.

Cabibbo L., Torlone R., A Logical Approach to Multidimensional Databases. EDBT 1998:183-197.

Darmont J., Boussaid O., Bentayeb F., Rabaseda S., Zellouf Y., Web multiform data structuring for warehousing, In C. Djeraba, ed., Multimedia Mining: A Highway to Intelligent Multimedia Documents; Multimedia Systems and Applications, Vol. 22, Kluwer, 2002, 179-194.

Darmont J., Boussaid O., Ralaivao J., Aouiche K., An Architecture Framework for Complex Data Warehouses, 7th International Conference on Enterprise Information Systems (ICEIS 05), Miami, USA, May 2005.

Darmont J, Olivier E, "A complex data warehouse for personalized, anticipative medicine", 17th Information Resources Management Association International Conference (IRMA 06), Washington, USA, May 2006, 685-687; Idea Group Publishing, Hershey, PA, USA.

Darmont J, "Entreposage de données complexes pour la médecine d'anticipation personnalisée", 9th International Conference on System Science in Health Care (ICSSHC 08), Lyon, France, September 2008.

Darmont J, Olivier E, "Biomedical Data Warehouses", Encyclopaedia of Healthcare Information Systems, IGI Publishing, Hershey, PA, USA, May 2008, 149-156.

Guting R. H., An introduction to spatial database systems, VLDB Journal, 1994.

Harbi N., Meuke-Fante M., Bentayeb F., Boussaïd O., "Un méta modèle multidimensionnel générique pour la conception des entrepôts de données", 3ème Atelier sur les Systèmes Décisionnels (ASD 08), Mohammédia, Maroc, 2008.

Jensen M., Moller T., Pedersen TB., Specifying OLAP cubes on XML data, Journal Of Intelligent Information Systems, 17(2/3):255--280, 2001.

Kimball R., The Data Warehouse Toolkit: Practical techniques for building dimensional data warehouses. John Wiley. 1996.

Miquel M., Bédard Y., Brisebois A., Conception d'entrepôts de données géospatiales à partir de sources hétérogènes : Exemple d'application en foresterie,ISI-NIS, volume X, 2001.

OMG, Common Warehouse Metamodel (CWM) Specification, March 2003, Version 1.1.

Pedersen TB., Jensen CS., Multidimensional Data Modeling for Complex Data, In Proceedings of ICDE, pp. 336--345, 1999.

Pokorny J., Sokolowsky P., A Conceptuel Modeling Perspective for Data Warehouses, Electronic Business Engineering / 4. Internationale Tagung Wirtschaftsinformatik 1999.

Ravat F., Teste O., Zurfluh G., Modélisation multidimensionnelle des systèmes décisionnels, In Actes des 1ères Journées Francophones d'Extraction et de Gestion des Connaissances - EGC 2001, Janvier 2001, Nantes.

Rivest,S., Bédard, Y. & Marchand P., 2001, Towards better support for spatial decision-making: Defining the characteris, SOLAP, Geomatica, 2001.

Tanasescu A., Modélisation multidimensionnelle de données complexes, EDIIS DEA ECD, Université Lumière Lyon2, 2003.

Teste O., Elaboration d'entrepôts de données complexes, INFORSID - mai 2000, Lyon.

Vassiliadis P., Sellis T., A Survey on Logical Models for OLAP Databases, SIGMOD Record 28(4): 64-69, 1999.

Zghal H., Faiz S., Ben Ghézala H., CASME: A CASE Tool for Spatial Data Marts Design and Generation, DMDW 2003.

Annexe: Common Warehouse Metamodel (CWM)

Le CWM est le standard de l'OMG pour les techniques liées aux entrepôts de données. Il couvre le cycle de vie complet de modélisation, construction et gestion des entrepôts de données. Le CWM définit un métamodèle qui représente les méta-données aussi bien métiers que techniques qui sont le plus souvent trouvées dans les entrepôts de données. Il est utilisé à la base des échanges de méta-données entre systèmes hétérogènes.

Le CWM comprend actuellement un certain nombre de méta-modèles concernant les entrepôts de données (représentation des données, analyse, gestion). Les métamodèles de données permettent de modéliser des ressources comme les bases de données relationnelles, les bases de données orientées objets. Une couche d'analyse du CWM définit des métamodèles pour les transformations de données, OLAP, la visualisation, la nomenclature et le data-mining. Une couche de gestion est constituée de métamodèles représentant les processus standard, la journalisation et la planification des activités.

Summary

The main purpose of data warehouses is to support decision-making. Multidimensional data modelling forms the basic issue in data warehousing and On-Line Analytical Processing (OLAP). Although these techniques are very efficient when working on simple numerical data, applying them onto heterogeneous and complex data imported from different sources is a very challenging task.

This paper aims at addressing the issue of multidimensionally modelling complex data, and more precisely medical data from the MAP project (Personalized Anticipative Medicine). We propose an extended multidimensional metamodel that is applied onto medical data by generalizing the complex cardiovascular model of the MAP warehouse. We also present a software tool that we have developed to achieve the computer-aided design of medical data warehouses.

Keywords: Data warehouse modelling, multidimensional model, OLAP, complex data.