Modélisation de la Structure Complexe des Faits et des Mesures

Estella Annoni, Franck Ravat, Olivier Teste

Modélisation de la Structure Complexe des Faits et des Mesures

Estella Annoni — Franck Ravat — Olivier Teste

IRIT-SIG Institute (UMR 5505)
118 Route de Narbonne, F-31062 Toulouse Cedex 9 France
{annoni, ravat, teste, zurfluh}@irit.fr

RÉSUMÉ. *Les systèmes d'information décisionnels (SID) ont pour objectif de faciliter la prise de décision. Ils reposent principalement sur la modélisation multidimensionnelle où le sujet de l'analyse, appelé fait, avec ses mesures sont représentés au centre d'une étoile dont les branches sont les dimensions de l'analyse avec ses paramètres. De nombreux modèles ont été proposés, mais ils permettent de représenter uniquement les liens entre les dimensions et les paramètres. Certains liens entre les faits ont récemment été étudiés. Les liens entre mesures représentent les corrélations propres à une activité car elles sont les caractéristiques de celle-ci. Mais, il n'existe pas de travaux définissant et modélisant ces liens. En ce sens, nous proposons une modélisation des faits et des mesures qui représente les liens entre ces concepts. Notre modélisation utilise UML car il permet de définir une représentation associée et adaptée à la terminologie des SID et de s'adresser au plus grand nombre de concepteurs décisionnels de part sa notoriété.*

ABSTRACT. *Decision support system (DSS) goal is to help decision-makers. They are modeled mainly in a multdimiensional way, that means activity analysis subjects, called facts, with their measures are represented as the center of a star with dimensions and their parameters around. Several models have been devoted to DSS representation, but they handle only relationships between dimensions and their parameters. Recently, some relationships between facts have been analyzed. Relationships between measures indicate correlations through a given activity because they represent properties of this activity. However, there is no work on these relationships. Therefore, we model facts and measures such as relationships between these concepts. We define a UML-based model because it allows a representation related and adapted to DSS terminology and it is well-known by a large amount of designers.*

MOTS-CLÉS : *systèmes d'information décisionnels, faits, mesures, structure complexe*

KEYWORDS: *decision support systems, facts, measures, complex structure*

1. Introduction

Les systèmes d'information décisionnels (SID) apportent une valeur ajoutée aux organisations pour améliorer leur réactivité et leur compétitivité en facilitant le processus de prise de décision. La modélisation qui est largement reconnue comme étant la plus adaptée aux SID est la modélisation multidimensionnelle (Kimball, 1996) où les activités sont décrites par des sujets d'analyse, appelés faits, ayant des caractéristiques appelés mesures. Elle représente les faits et leurs mesures au centre d'étoiles dont les axes utilisées pour l'analyse des activités sont appelés dimensions. Durant la dernière décennie, de nombreux modèles multidimensionnels ont été proposés reposant sur deux paradigmes connus, Entité-Association (Tryfona *et al.*, 1999, Franconi *et al.*, 2004) et Orienté-Objet (Luján-Mora *et al.*, 2006, Abelló *et al.*, 2006) et sur des paradigmes spécifiques (Golfarelli *et al.*, 1998, Schneider, 2007, Ravat *et al.*, 2008). Ces modèles offrent une analyse détaillée des liens entre les paramètres, l'organisation en hiérarchies des paramètres au sein d'une dimension (Pedersen *et al.*, 2005) et les contraintes entre dimensions (Ghozzi *et al.*, 2003). Ainsi, principalement les spécificités liées aux dimensions et aux paramètres sont étudiées. Cependant, les faits traduisent les tendances des activités à un niveau macroscopique et les mesures à un niveau microscopique. Nous avons constaté que tout besoin d'analyse complexe des données d'une activité porte obligatoirement sur le fait associé et ses mesures. De plus, les listes énonçant les propriétés que doivent vérifier un modèle multidimensionnel incluent le traitement symétrique des dimensions et des faits (Vassiliadis *et al.*, 1999, Rafanelli, 2003), mais il n'est pas vérifié par les modèles existants. Ainsi, face au rôle central des faits et des mesures dans les SID (symbolisé par l'image de l'étoile), comment représenter la complexité de la structure de ces concepts dans son intégralité ?

De récents travaux ont abordé certains liens entre faits (Schneider, 2007) en représentant les relations M-N existantes entre deux faits, mais ils ne traitent pas des liens entre mesures et des liens entre faits et mesures.

Pour répondre à ce besoin en terme de modélisation, nous proposons d'utiliser deux structures différentes pour représenter les faits et les mesures et de matérialiser ces liens sur le schéma conceptuel du SID. Cette modélisation repose sur le paradigme orienté-objet où chaque concept est représenté par une classe UML. L'avantage que procure UML est celui d'associer la terminologie des SID à la représentation et d'étendre les relations entre les classes pour exprimer les spécificités des liens entre les faits et les mesures. De plus, les modèles existants qui permettent de représenter le plus grand nombre de spécificités des SID sont basés sur UML (Luján-Mora *et al.*, 2006, Abelló *et al.*, 2006).

L'article s'organise en six sections. Dans la section 2, nous présentons le contexte des travaux et un état de l'art de la modélisation des faits. Puis, dans la section 3, nous abordons la dynamique des SID. Dans la section 4, nous explicitons la modélisation des faits et des relations entre faits. Dans la section 5, nous décrivons la modélisation

des mesures et des relations entre faits et mesures ainsi qu'entre mesures. Enfin, dans la section 6, nous récapitulons nos propositions pour énoncer nos perspectives.

2. Contexte et état de l'art

L'activité d'une organisation est représentée par le concept de fait à un niveau global, macroscopique. Elle est décrite au niveau microscopique par des éléments mesurables qui sont les mesures. Ces mesures donnent une vue plus précise et plus focalisée de l'activité. Les listes de propriétés requièrent que la complexité de la structure des faits et des mesures soit prise en compte par tout modèle multidimensionnel. Les principaux modèles existants présentés dans le tableau 1 n'abordent que cinq des sept spécificités liées à la structure complexe des faits et des mesures, soient :

1) mesures multiples : propriété d'un fait composé de plusieurs mesures. Cette spécificité est prise en compte par tous les modèles car ils sont postérieurs à 1998, en l'occurrence à celui de (Lehner, 1998) qui définit qu'une seule mesure par fait,

2) faits multiples : propriété d'un schéma multidimensionnel composé de plusieurs faits partageant au moins une dimension. Ce schéma est appelé un « schéma en constellation ». La majorité des modèles vérifient cette spécificité car les grandes organisations ont plus d'une activité et des analyses comparatives de ces activités sont requises par les décideurs,

3) fait dégénéré : un fait est dégénéré par rapport à une dimension si à une instance du fait est associé plusieurs instances de la dimension. Par exemple, la commission d'une vente réalisée par plusieurs commerciaux doit être associée à ces commerciaux. Les commissions ne peuvent pas être une mesure du fait « Ventes » bien qu'elles y soient liées car les autres mesures telles que le chiffre d'affaires « CA » et la « Quantité » sont évaluées au niveau global par rapport à la dimension « Commerciaux ». Ainsi, un fait dégénéré implique un type de liens entre deux faits où une dimension intervient,

4) liens entre faits : les corrélations des activités définies dans un SID. Ces liens ont récemment été étudiés dans les travaux de (Schneider, 2007) car ils décrivent les corrélations d'alimentation et d'analyse des activités à un niveau macroscopique. L'auteur représente uniquement les relations M-N entre faits telle que la relation entre les faits « Ventes » et « Achats » où une instance du fait « Ventes » est associée à plusieurs instances du fait « Achats » (plusieurs produits servent à la fabrication d'un produit vendu) et inversement,

5) mesures dérivées : propriété des mesures qui sont calculées à partir d'autres mesures du SID. Cette spécificité implique que la règle de calcul de chaque mesure utilise d'autres mesures telle que le « CA » qui est calculé à partir de la mesure « Quantité » et d'une donnée issue des systèmes sources, soit le « Prix ».

Les quatre premières spécificités concernent principalement le fait dans sa globalité et les relations qu'il peut établir avec un autre fait et une dimension. Ce constat peut s'expliquer de la façon suivante : auparavant les faits avaient une seule mesure

Spécificités / Paradigmes	Modèles multidimensionnels		Structure complexe des faits et des mesures						
			Mesures multiples	Faits multiples	Fait dégénéré	Mesures dérivées	Liens entre faits	Liens entre mesures	Liens entre faits et mesures
Entité-association	[Tryfona et al., 1999]	StarER	1						
	[Franconi and Kamble, 2004]	GMD	1	1					
Objet	[Lujan-Mora et al., 2006]	GOLD	1	1	1	1	0,5		
	[Abello et al., 2006]	YAM^2	1	1		1			
Spécifiques	[Golfarelli et al., 1998]	DFM	1						
	[Cabibbo and Torlone, 1998]	MD	1	1					
	[Schneider, 2007]		1	1			0,5		

1: modèle représentant complètement la spécificité donnée
0,5: modèle représentant partiellement la spécificité donnée

Tableau 1. *Comparatif des principaux modèles par rapport à la structure complexe du fait.*

et les concepts de fait et de mesure étaient souvent employés de manière indifférente. De plus, les faits peuvent être liés par d'autres relations que les relations M-N car deux faits peuvent être dépendants l'un de l'autre. Par exemple, un fait peut requérir l'alimentation préalable d'un autre fait pour sa propre création et alimentation. Cette information est importante aussi bien pour les concepteurs décisionnels que les décideurs car la cohérence du système et la fiabilité des données en dépendent. De même, les auteurs de (Luján-Mora *et al.*, 2006) représentent uniquement un type de liens, soit celui relatif aux faits dégénérés.

Seule la cinquième spécificité concerne le concept de mesure. Elle indique qu'une mesure peut dériver d'autres mesures, ce qui implique des liens relatifs au calcul. Cependant, ces liens ne sont pas matérialisés sur le schéma. La question suivante se pose : « la propriété de dérivation induit-elle un seul type de liens entre mesures ? ». Nous avons constaté qu'entre une mesure dérivée et une autre mesure, deux types de liens peuvent être définis car les mesures peuvent résulter d'un calcul à partir :

– de données provenant uniquement des sources et qui ne sont pas présentes à l'état brut ou avec format exploitable dans le SID. Dans ce cas de figure, la mesure n'est pas une mesure dérivée,

– de données provenant des sources et de mesures existantes dans le SID ; soit le premier type de mesures dérivées,

– uniquement de mesures du SID ; soit le second type de mesures dérivées.

Les modèles existants proposent tous de représenter les mesures dans la même structure conceptuelle que le fait ou encore dans un autre structure sans des précisions

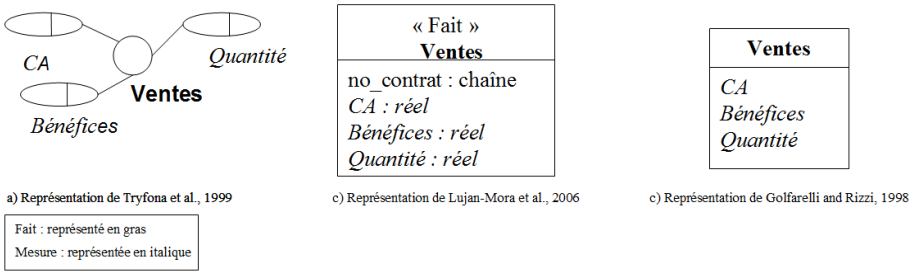

Figure 1. *Représentations du fait « Ventes » suivant des modèles basés sur les trois paradigmes Entité-Association (a), Orienté-Objet (b) et spécifique (c).*

sur les liens et sans la représentation des liens entre mesures (Tryfona *et al.*, 1999). La figure 1 présente des représentations du fait « Ventes » suivant des modèles reposant sur les trois types de paradigmes, Entité-Association, Orienté-Objet et spécifique. Ces représentations permettent uniquement de modéliser qu'une mesure est une caractéristique du fait. Par exemple, le lien entre « CA » et « Bénéfices » n'est pas représenté. Ainsi, « comment modéliser les mesures afin que ces liens soient représentés ? »

Par ailleurs, la propriété de dérivation d'une mesure est aussi liée à la dynamique des SID car elle concerne le calcul de la mesure à partir d'autres mesures ou des données des systèmes sources. Pour représenter cet aspect dynamique de la propriété de dérivation d'une mesure, la modélisation d'une mesure doit aussi représenter les processus liés à la dynamique du SID. Cependant les travaux existants sur la modélisation des processus à l'origine des SID les représentent uniquement après la conception du schéma du SID et dans un autre schéma (Vassiliadis *et al.*, 2002, Luján-Mora *et al.*, 2004). L'évaluation de ces processus lors de l'analyse des besoins et au cours de la conception du schéma n'est pas abordée. De plus, ces processus ne couvrent que l'extraction, la transformation et le chargement des données (ETL - Extraction, Transformation, Load) depuis les systèmes sources ; ceux concernant la préparation des données pour les analyses complexes sont peu abordés. De ce fait, dans cet article, avant de présenter la modélisation des faits, des mesures et des liens associés, nous développons les processus de dérivation et de préparation des données que nous avons notés nécessaires pour le développement de SID fiables. Ainsi, nos contributions par rapport aux travaux existants sur la modélisation multidimensionnelle sont :

– la modélisation des mesures distincte de celle des faits,

– une modélisation représentant aussi bien les données que les processus sur le même schéma conceptuel,

– la représentation des liens entre mesures,

– la représentation des liens entre faits et mesures,

– la représentation complète des liens entre faits.

3. Processus de dérivation et de préparation des données du SID

Un système se caractérise par un aspect statique et un aspect dynamique, en l'occurrence par les données et les processus. Les SID se distinguent des systèmes d'information classiques car ils sont définis à partir de données contenues dans des systèmes sources *via* des processus. Ces processus correspondent aux opérations réalisées sur les données pour créer, alimenter et préparer l'environnement pour faciliter la prise de décision. Comme le mentionne (Inmon, 1997), les traitements ETL représentent 55% du coût de développement d'un SID. Ces traitements concernent principalement la dérivation, la gestion technique des données, mais ils n'abordent pas tous les problèmes liés à la préparation des données. Le processus de préparation des données qui est inclus concerne la transformation, plus précisément le calcul des données. Nous proposons dix processus, comprenant les processus ETL, regroupés en deux sous-ensembles : ceux liés à la dérivation des données et ceux liés à la préparation des données comme indiqué dans la figure 2.

Nous avons constaté que tous les attributs d'une classe UML associée à un concept ne font pas nécessairement l'objet de tous les processus appliqués à cette classe. Nous avons proposé le concept d'informativité introduit dans nos précédents travaux liés à l'analyse des besoins du SID (Annoni *et al.,* 2006). Un concept d'informativité est caractérisé par un symbole associé à un traitement qui indique qu'un attribut fait l'objet de ce traitement (cf. la donnée membre symboleConceptI de la classe ConceptInformativité dans notre métamodèle dans la figure 2). Il se place à côté de la visibilité de l'attribut dans la classe UML. La prise en compte des processus dans le schéma conceptuel du SID a l'avantage de garantir la définition de toutes les données nécessaires pour assurer la dérivation du SID à partir des systèmes sources et la préparation des données pour le contexte décisionnel. Ainsi, la définition des processus met en avant les éventuelles incohérences et les données manquantes.

4. Modélisation du fait et des liens entre faits

Notre modélisation représente les quatre principaux concepts multidimensionnels (fait, dimension, mesure, paramètre) par une classe UML. Elle utilise UML car comme indiqué dans le tableau 1, les modèles basés sur UML sont ceux qui permettent de représenter le plus grand nombre de spécificités. Cet avantage repose sur le fait qu'il est possible d'étendre les classes UML en utilisant des stéréotypes relatifs à la terminologie du domaine et d'étendre les relations entre classes pour modéliser la réalité du SID. Ainsi, la classe UML est dite multidimensionnelle car elle est une extension associée au domaine du décisionnel. Les auteurs de (Luján-Mora *et al.,* 2006) ont introduit ce concept comme étant une classe UML avec un stéréotype associé à la terminologie multidimensionnelle et qui peut être reliée à d'autres classes multidimensionnelles par une relation d'agrégation ou par une extension de l'association UML afin d'exprimer le lien « roll-up » existant entre deux paramètres d'une dimension. Cette définition de la classe multidimensionnelle ne permet pas de représenter toutes les spécificités liées à la structure complexe des faits et des mesures car le lien entre mesures n'est pas un

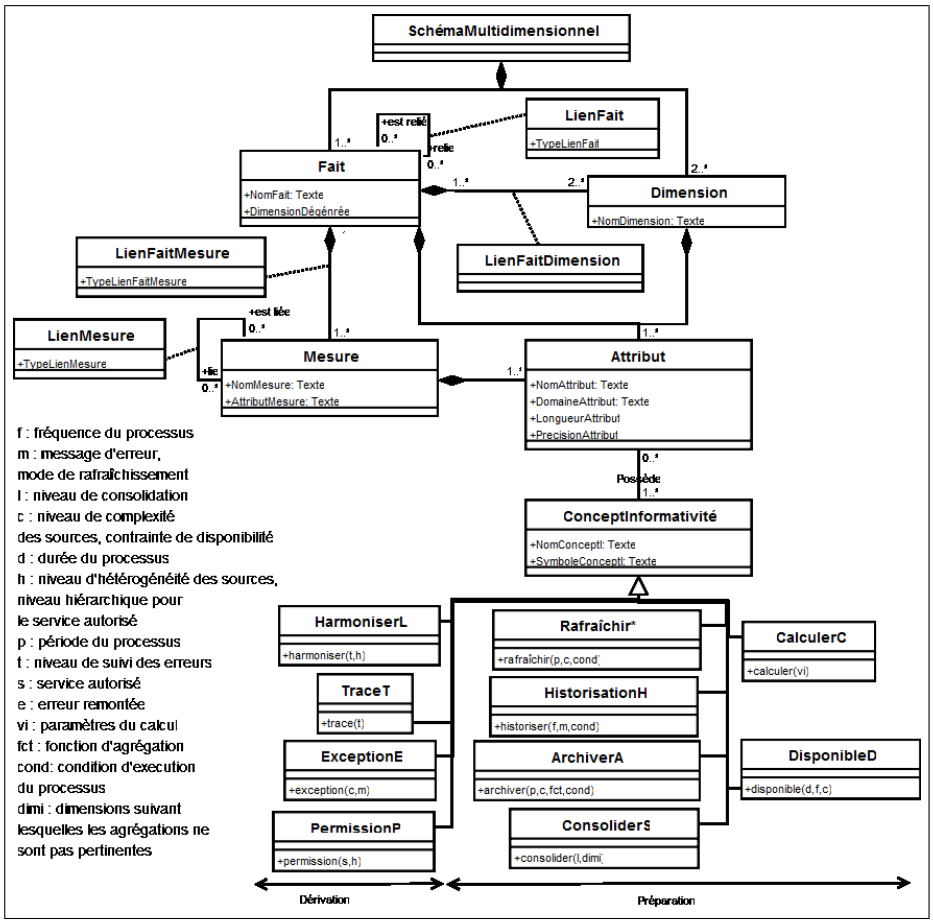

Figure 2. *Métamodèle de notre modélisation des faits et des mesures.*

lien de « roll-up » et les processus liés à ces concepts n'y sont pas représentés. De ce fait, nous étendons la classe multidimensionnelle telle que :

– le compartiment du nom contient le nom de la classe précédé d'un stéréotype correspondant au nom du concept multidimensionnel,

– le compartiment des attributs contient les attributs spécifiques au concept multidimensionnel précédés de symboles associés aux processus de dérivation et de préparation dont ils font l'objet,

– le compartiment des méthodes contient les méthodes associées aux processus de dérivation et de préparation de la classe (cf. section 3),

– les relations entre les classes sont étendues afin d'exprimer les liens entre les concepts relatifs à leur structure complexe.

Dans notre modèle multidimensionnel, nous considérons quatre principales classes multidimensionnelles classes-faits, classes-dimensions, classes-mesures et classes-niveaux associées respectivement aux faits, dimensions, mesures et niveaux (noms donnés aux paramètres organisés dans une dimension et formant ainsi des hiérarchies). Dans les sections suivantes, nous focalisons sur les classes-faits et les classes-mesures. Ainsi, dans notre métamodèle présenté à la figure 2, nous décrivons de manière détaillée uniquement les méta-classes associées Fait et Mesure. La méta-classe Dimension y est partiellement représentée. Par rapport à notre métamodèle, nos contributions sont liées aux méta-classes Fait, Mesure, LienFait, LienFaitMesure et LienMesure. Nous proposons une représentation textuelle et une représentation graphique car un schéma est plus explicite que du texte sachant que les deux représentations expriment chacune toutes les spécificités multidimensionnelles. Nous définissons deux représentations graphiques, globale et détaillée. La représentation détaillée des dimensions permet de montrer ses différents niveaux alors que la représentation globale montre uniquement son nom. Nous utiliserons la représentation détaillée pour les faits et les mesures, mais pour les dimensions nous utilisons la représentation globale.

4.1. *Modélisation d'un fait*

Un fait est le sujet d'analyse lié à un métier d'une organisation. A tout fait est associé une classe multidimensionnelle appelée « classe-fait » dont la modélisation textuelle est présentée dans la définition 1. Dans un schéma multidimensionnel, un valeur de fait est définie pour une combinaison de dimensions, mais une valeur de fait peut aussi être identifiée de manière unique par un attribut appelée dimension dégénéré.

Définition 1 *Soit le fait « F » défini par la classe-fait* $F[< d_1, ...d_m, af_1, ..., af_n >, < md_1, ..., md_p, mp_1, ..., mp_q >, < M_1, ..., M_r >, < F_1, ..., F_s >, < L_1, ..., L_u >]$ *tel que :*

– F : le nom du fait,

– $< d_1, ...d_m >$ *: les dimensions dégénérées du fait,*

– $< af_1, ..., af_n >$ *: les attributs spécifiques au fait,*

– $< md_1, ..., md_p >$ *: les méthodes associées aux processus de dérivation et de préparation des données,*

– $< M_1, ..., M_r >$ *: les mesures liées au fait,*

– $< F_1, ..., F_s >$ *: les autres faits connectés au fait,*

– $< D_i.L_1, ..., D_j.L_u >$ *: les niveaux des dimensions suivant lesquelles le fait est analysé ; ces niveaux peuvent être de granularités différentes.*

La modélisation graphique d'une classe-fait « F » est une classe multidimensionnelle avec le stéréotype « Fait », qui est l'agrégation des classes multidimensionnelles associées aux niveaux $< D_i.L_1, ..., D_j.L_u >$ appartenant aux dimensions suivant les-

quelles le fait est analysé dans le cas d'une représentation détaillée. Cependant, dans le cas d'une représentation globale, seule le nom de la dimension est représenté. La classe-fait est reliée aux autres classes-faits $< F_1, ..., F_s >$ qui lui sont connectées par des relations définies dans la sous-section 4.2 et aux classes associées à ses mesures $< M_1, ..., M_r >$ par des relations définies dans la sous-section 5.2. Le compartiment attribut comprend les attributs correspondant aux dimensions dégénérées $< d_1, ...d_m >$ et ses attributs spécifiques $< af_1, ..., af_n >$. Un fait peut ne pas avoir d'attribut, son compartiment attribut est alors vide. Le compartiment méthode comprend les méthodes de dérivation $< md_1, ..., md_p >$ et de préparation $< mp_1, ..., mp_q >$ des données.

Exemple 1 *Considérons le fait « Ventes » présenté dans la figure 3 sans ses mesures et ses liens avec les autres faits. Sa modélisation textuelle est :*

$Ventes[< no\_contrat, taux\_change >$
$, < disponible(20, heure, jour),\ historiser(annee, 3) >$
$, < CA, Benefices, Quantite >, < >$
$, < Geographie.Ville, Produits.Produit, Commerciaux.Commercial, Temps.Date >$
$]$

Ce fait a un attribut dimension dégénérée appelé « no_contrat » car le numéro de contrat permet d'identifier de manière unique chaque vente. Il a aussi un attribut spécifique nommé « taux_change » qui n'est pas une dimension dégénérée car le taux de change peut être le même pour plusieurs ventes. Il a une méthode de dérivation « disponible(20, heure, jour) » qui signifie que les données des ventes sont disponibles vingt heures par jour. Il a une méthode de préparation « historiser(année, 3) » qui signifie que les données des ventes sont gardées dans le SID durant trois ans. Ce fait a trois mesures « CA, Bénéfices, Quantité » et il est connecté à quatre dimensions « Géographie, Produits, Commerciaux, Temps ».

4.2. *Modélisation des liens entre faits*

Dans les organisations les schémas multidimensionnels sont en constellation car elles gèrent plusieurs activités. Les liens entre les faits d'un SID correspondent aux liens entre les activités. Par exemple, considérons le fait « Ventes » et le fait « Achats » d'une organisation qui achète des matières premières pour fabriquer des produits qu'elle vend. Le chiffre d'affaires de produits vendus peut être analysé par rapport au montant payé pour l'achat des produits nécessaires à leur fabrication. Dans cette organisation, un produit peut servir de matière première à la fabrication de plusieurs produits et plusieurs produits entrent dans la composition d'un produit destiné à la vente. Il y a donc une relation M-N entre le fait « Achats » et le fait « Ventes ». Cependant, dans le cas où plusieurs produits entrent dans la composition d'un produit destiné à la vente, mais qu'un produit peut servir de matière première à la fabrication d'un et d'un seul produit, il y a une relation de type 1-N entre les deux classes-faits. Par

ailleurs, un produit ne peut être vendu que si les produits nécessaires à sa fabrication ont été achetés. Il y a donc un sens dans l'analyse des données liées à ces deux activités, soit une orientation de lecture de la relation entre les faits. Entre deux classes-faits, il n'y a pas que ce type de liens. Comme précisé dans la section 2, il y a un lien relatif à la spécificité de fait dégénéré. Par exemple, le fait « Commissions » est défini par rapport au fait « Ventes » et à la dimension « Commerciaux » car la commission d'une vente peut concerner plusieurs commerciaux. Nous définissons donc deux relations entre des classes-faits.

Définition 2 *Une relation dite de **corrélation** existe entre un fait F_1 et un fait F_2 si les attributs et les mesures de F_1 peuvent être analysés suivant le fait F_2, ses mesures, et ses dimensions. Cette relation n'est pas orientée par défaut, ce qui signifie que la création, l'alimentation et l'analyse des faits F_1 et F_2 sont indépendantes mais que les métiers associés sont liés.*

Cette relation est graphiquement représentée par une simple association UML avec le stéréotype « corrélation » entre les classes-faits F_1 et F_2 pour une relation non-orientée. Cependant, pour une relation orientée, une flèche est ajoutée au niveau de l'extrémité finale dans le sens de la lecture. Elle est textuellement définie dans les deux faits F_1 et F_2 pour une relation non-orientée et pour une relation orientée, le fait indépendant F_2 est défini dans la modélisation du fait dépendant F_1.

Exemple 2 *La classe-fait « Ventes » présentée à la figure 3 est connectée à la classe-fait « Achats » par une relation de corrélation orientée indiquant que la classe-fait « Achats » doit être créée et alimentée avant pour l'alimentation de la classe-fait « Ventes ». Sa représentation textuelle incluant la relation de corrélation est :*

$$
\begin{array}{l}
Ventes[< \ no\_contrat, taux\_change \ > \\
, < \ disponible(20, heure, jour), \ historiser(annee, 3) \ > \\
, < \ CA, Benefices, Quantite \ >, < \ Achats \ > \\
, < \ Geographie.Ville, Produits.Produit, Commerciaux.Commercial, Temps.Date \ > \\
]
\end{array}
$$

Définition 3 *Une relation dite de **dégénération** existe entre un fait F_3 et un fait F_4 par rapport à une dimension D si les attributs et les mesures de F_3 dépendent du fait F_4 et d'une dimension D de F_4.*

Cette relation est graphiquement représentée par une classe d'association UML entre la classe-fait F_4 et la classe-dimension D et la relation porte le stéréotype « dégénération ». Mais, contrairement à la classe d'association UML une classe-fait dégénérée ne peut pas être liée à d'autres classes car elle ne peut être analysée qu'en fonction des mêmes dimensions que la classe-fait associée. Pour une meilleure compréhension du schéma, le stéréotype << FaitDégénéré >> est associé à F_3. Elle est textuellement définie dans la modélisation de F_4, par l'ajout du nom de la dimension D en indice du nom du fait F_3.

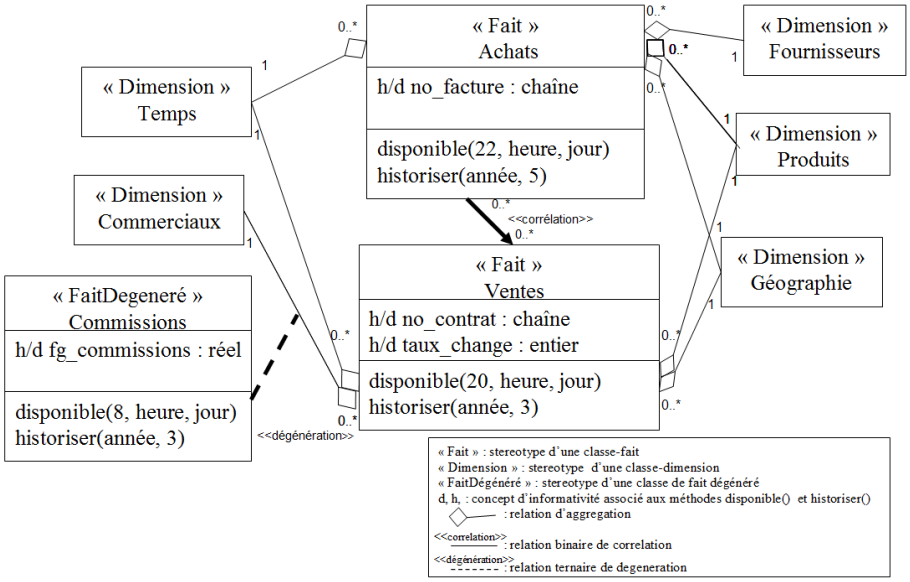

Figure 3. *Classes-faits « Ventes » et « Achats » connectées par une relation de corrélation et la classe-fait dégénérée « Commissions » intervenant dans une relation de dégénération.*

Exemple 3 *La classe-fait dégénérée « Commissions » est une classe d'association entre la classe-fait « Ventes » et la classe-dimension « Commerciaux » reliée par la relation de dégénération car les commissions d'une vente sont associées aux commerciaux qui ont réalisé la vente (cf. figure 3) et les détails de cette commissions sont nécessaires. La représentation textuelle de Ventes incluant la relation de dégénération est :*

$$
\begin{array}{l}
Ventes[< \ no\_contrat, taux\_change \ > \\
, < \ disponible(20, heure, jour), \ historiser(annee, 3) \ > \\
, < \ CA, Benefices, Quantite \ >, < \ Achats, Commissions_{Commerciaux} \ > \\
, < \ Geographie.Ville, Produits.Produit, Commerciaux.Commercial, Temps.Date \ > \\
]
\end{array}
$$

5. Modélisation des mesures et des liens associés

Dans cette section, nous proposons une modélisation originale des mesures qui permet aussi bien aux concepteurs décisionnels qu'aux décideurs de suivre les corrélations et les impacts des modifications de données des mesures *via* les liens entre mesures et ainsi de gérer la fiabilité de leurs données.

5.1. *Modélisation d'une mesure*

Une mesure est une caractéristique du fait. A toute mesure est associée une classe multidimensionnelle appelée « classe-mesure » dont la modélisation textuelle est présentée dans la définition 4. Une classe-mesure ne peut pas être connectée à une classe-dimension ou à des classes-niveaux afin d'assurer la lisibilité et la maintenance du schéma. De plus, toute mesure est analysée par rapport aux dimensions connectées au fait qui lui est associé en fonction des processus de préparation des données.

Définition 4 *Soit la mesure « M » définie par la classe-mesure* $M[< am_1, ..., am_v >, < md_1, ..., md_w, mp_1, ..., mp_x >, < M_1, ..., M_y >]$ *telle que :*

– M : le nom de la mesure,

– $< am_1, ..., am_v >$ *: les attributs spécifiques à la mesure,*

– $< md_1, ..., md_w, mp_1, ..., mp_x >$ *: les méthodes associées aux processus de dérivation et de préparation des données,*

– $< M_1, ..., M_y >$ *: les autres mesures qui participent à des processus de dérivation ou de préparation de la mesure.*

La modélisation graphique d'une classe-mesure « M » est une classe multidimensionnelle avec le stéréotype « Mesure ». Le compartiment attribut comprend ses attributs spécifiques $< am_1, ..., am_v >$ précédés des concepts d'informativité associés aux processus dont ils font l'objet. Le compartiment méthode comprend les méthodes de dérivation $< md_1, ..., md_w >$ et les méthodes de préparation des données $< mp_1, ..., mp_x >$. La mesure « M » est liée au fait « F » par l'une des relations définies dans la sous-section 5.2. Elle est liée aux autres mesures $< M_1, ..., M_y >$ par des relations définies dans la sous-section 5.3.

Exemple 4 *Considérons la mesure « CA » du fait « Ventes » présentée dans la figure 4. Elle a un attribut spécifique « m_CA ». Elle a une méthode de dérivation « harmoniser(1,1) » qui signifie que le « CA » est dérivé à partir d'une règle d'extraction simple. Elle a aussi une méthode de préparation « calculer(Quantité, Prix) » qui signifie que la mesure « CA » est calculée à partir de la « Quantité » (quantité de produits vendus) qui est une mesure de « Ventes » et de la donnée issue directement des systèmes sources « Prix » (prix des produits vendus). Les méthodes de dérivation et de préparation de la mesure « CA » requièrent la mesure « Quantité », il y a donc une relation entre « CA » et « Quantité » que nous définissons dans la section 5.3. La représentation textuelle de CA n'incluant pas le lien avec « Quantité » est :*

$CA[< m\_CA >, < harmoniser(1,1), calculer(Quantite, Prix) >$
$, < Quantite >]$

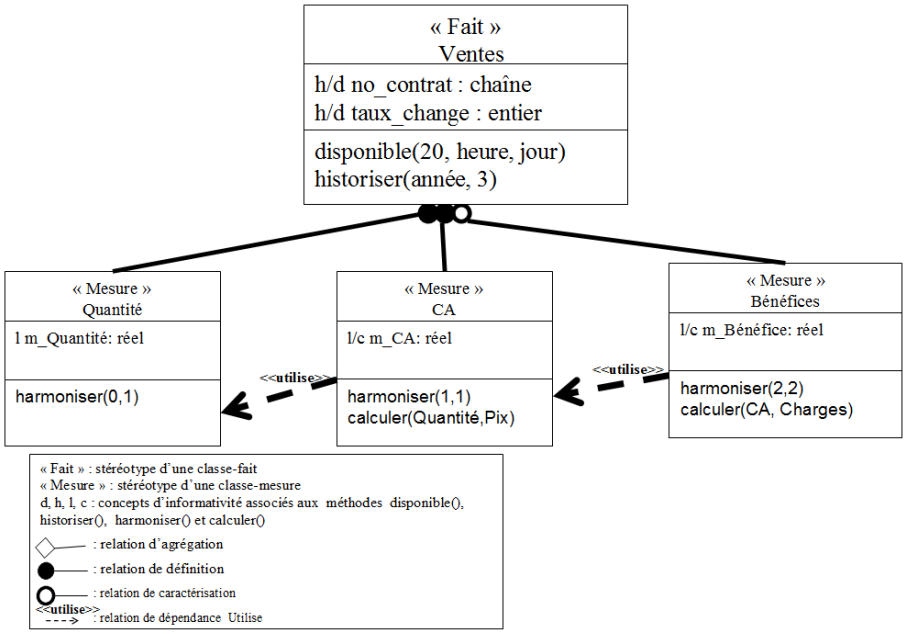

Figure 4. *Les liens entre les classes-mesures « CA », « Bénéfices » et « Quantite » et ceux avec la classe-fait « Ventes ».*

5.2. *Modélisation des liens entre faits et mesures*

Une mesure décrit un fait et elle n'est pas un élément constitutif du fait au même titre qu'un niveau est un élément constitutif d'une hiérarchie d'une dimension. De plus, certaines mesures sont nécessairement définies pour un fait car elles ne peuvent pas être nulles dans le sens où elles portent la sémantique du fait. Une mesure peut décrire un fait de deux façons car le fait est un concept associé à l'activité analysée et la mesure est un concept associé aux indicateurs de performance de cette activité. D'une part, une mesure M_A peut nécessairement être définie pour un fait F_1 dans le cas où la mesure est représentative du fait. D'autre part, certaines mesures peuvent ne pas être définies à la création ou à l'alimentation d'un fait, ce qui signifie que les valeurs des mesures ne sont pas connues lors de la création et de la mise à jour du fait.

Ainsi, pour indiquer qu'une classe-mesure caractérise ou définit une classe-fait, nous proposons deux nouvelles relations appelées **caractérisation** et **définition**. Une classe-fait et une classe-mesure ne peuvent être reliées que par une seule relation. Nous définissons ces relations en s'inspirant respectivement des relations d'agrégation et de composition UML car il existe un lien qui peut être faible ou fort entre un fait et une mesure. Mais, comme la propagation des valeurs d'attributs n'est pas pertinente entre des classes-faits et des classes-mesures, nous avons défini de nouvelles relations.

Définition 5 *Une relation dite de **caractérisation** existe entre une classe-fait F_1 et une classe-mesure M_A si la classe-mesure M_A peut ne pas être définie pour cette classe-fait F_1. Cette relation entre F_1 et M_A signifie que l'indicateur n'est pas un indicateur de référence pour le métier analysé.*

Cette relation est graphiquement représentée par un trait entre M_A et F_1 terminé par un rond vide placé du côté de la classe-fait F_1 car la relation de caractérisation est un lien faible entre un fait et une mesure. Elle est textuellement définie dans la modélisation de F_1, par l'ajout de la lettre « c » pour « caractérisation » en indice du nom de la mesure M_A.

Exemple 5 *Dans la figure 4, la classe-fait « Ventes » est reliée à la classe-mesure « Bénéfices » par une relation de caractérisation alors qu'elle est reliée aux classes-mesures « CA » et « Quantité » par des relations de définition. Le « CA » et la « Quantité » sont des indicateurs de référence, mais le « Bénéfices » est un indicateur qui est analysé dans un deuxième temps car il requiert que le calcul des charges provenant des sources soit réalisé lors de l'alimentation du fait « Ventes ». Cependant, le calcul des charges nécessite toujours un laps de temps voire des itérations de modification pour une évaluation complète et exacte.*

Définition 6 *Une relation dite de **définition** existe entre une classe-fait F_1 et une classe-mesure M_B si la classe-mesure M_B est nécessairement définie pour cette classe-fait F_1.*

Cette relation est graphiquement représentée par un trait entre M_B et F_1 terminé par un rond noirci placé du côté de la classe-fait F_1 car la relation de définition est lien fort entre un fait et une mesure. Elle est textuellement définie dans la modélisation de F_1, par l'ajout de la lettre « d » pour « définition » en indice du nom de la mesure M_B.

Exemple 6 *La modélisation textuelle complète du fait « Ventes » est la suivante :*

$$
\begin{array}{l}
Ventes[< no\_contrat, taux\_change > \\
, < disponible(20, heure, jour), historiser(annee, 3) > \\
, < CA_d, Benefices_c, Quantite_d >, < Achats, Commissions_{Commerciaux} > \\
, < Geographie.Ville, Produits.Produit, Commerciaux.Commercial, Temps.Date > \\
]
\end{array}
$$

5.3. *Modélisation des liens entre mesures*

Nous avons constaté que les liens entre les mesures d'un fait, qui sont des mesures dérivées ou qui participent aux calculs de ces mesures dérivées, correspondent à deux types de dépendances. Nous définissons donc deux relations de dépendance entre des classes-mesures, soient la **dépendance utilise** et la **dépendance hiérarchique**.

Définition 7 *Une relation dite de **dépendance utilise** existe entre deux classes-mesures M_A et M_B si la mesure associée à M_A est calculée à partir de la mesure associée à M_B (ou d'une ou de plusieurs classes-mesures $\{M_B\}^+$) et d'autres données provenant uniquement des systèmes sources non présentes dans le SID.*

Cette relation est graphiquement représentée par la relation de dépendance UML avec le stéréotype « utilise » entre M_A et M_B afin de réutiliser la relation existante et déjà maîtrisée qui porte le même sens et aussi éviter la conception d'un modèle complexe. Elle est textuellement définie dans la modélisation de la classe-mesure M_A, par l'ajout de la lettre « u » pour « utilise » en indice du nom de la mesure M_B.

Exemple 7 *Dans la figure 4, les classes-mesures « CA » et « Quantité » sont reliées par des relations de dépendance utilise car la mesure « CA » est calculée à partir de la mesure quantité de produit vendus et du prix de ces produits qui est une donnée existant que dans les sources. De même, la classe-mesure « Bénéfices » est reliée à la classe-mesure « CA » par une relation de dépendance utilise. La représentation textuelle des mesures est :*

$Quantite[< m\_Quantite >, < harmoniser(0,1) >], CA[< m\_CA >$
$, < harmoniser(1,1), calculer(Quantite, Prix) >, < Quantite_u >]$ *et*
$Benefices[< m\_Benefices >$
$, < harmoniser(2,2), calculer(CA, Charges) >, < CA_u >]$

Définition 8 *Une relation dite de **dépendance hiérarchique** existe entre deux classes-mesures M_C et M_D si la mesure associée à M_C est calculée uniquement à partir d'autres classes-mesures $\{M_D\}^+$ (soit une ou plusieurs classes-mesures M_D).*

Cette relation est graphiquement représentée par une extension de la relation de composition UML. L'extension est définie par un losange rempli et précédé par une flèche pointant vers la classe-mesure composée, soit la classe mère composante. Elle est textuellement définie dans la modélisation de la classe-mesure M_C, par l'ajout de la lettre « h » pour hiérarchique en indice du nom de la mesure M_D.

Exemple 8 *La figure 5 concerne le domaine de la comptabilité, la classe-mesure « Biens » est reliée aux classes-mesures « Exports » et « Imports » par des relations de dépendance hiérarchique car la mesure « Biens » est calculée à partir des exportations et des importations de marchandises suivant le calcul de la balance commerciale d'un pays. Ces deux données sont contenues dans le SID et sont des mesures du fait « Compta ». La représentation textuelle des mesures est :*

$Services[< m\_Services >, < harmoniser(3,3) >],$
$Biens[< m\_Biens >, < harmoniser(2,2), calculer(Exports, Imports) >$
$, < Exports_h, Imports_h >],$
$Exports[< m\_Exports >, < harmoniser(3,2) >]$ *et*
$Imports[< m\_Imports >, < harmoniser(3,4) >]$

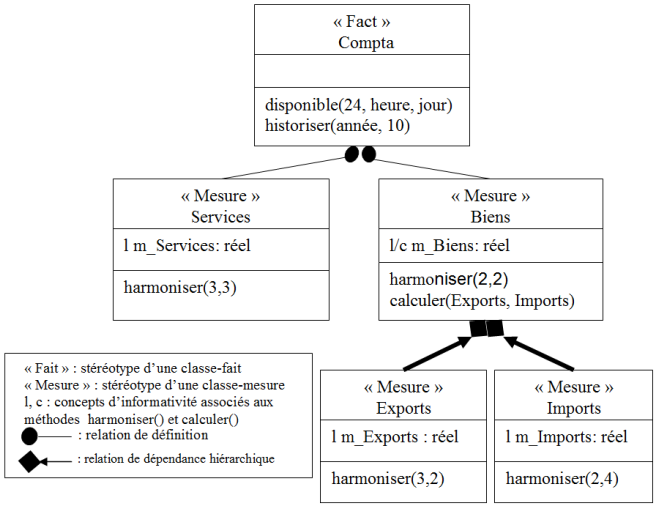

Figure 5. *Dépendances hiérarchiques entre les classes-mesures « Biens » et « Exports » et entre « Biens » et « Imports ».*

6. Conclusion

Dans ce papier, nous avons proposé une modélisation textuelle et graphique des faits et des mesures dans des structures distinctes telles que les mesures soient à l'extérieur du fait. Les liens entre faits, entre mesures, entre mesures et faits ainsi que les processus à l'origine de ces concepts peuvent donc être modélisés sur le schéma du système d'information décisionnel (SID). L'originalité de notre modélisation textuelle est qu'elle décrit la structure complexe des faits et des mesures par l'ajout de la première lettre du nom des relations en indice du nom des concepts. La force de notre représentation graphique est qu'elle repose sur un seul type de structure, soit la classe multidimensionnelle. Elle utilise aussi des relations spécifiques distinctes de celles utilisées pour la modélisation de la structure complexe des dimensions. Ainsi, d'un seul coup d'oeil sur un schéma multidimensionnel les concepts relatifs aux faits et aux dimensions sont distingués. Ces relations ont été définies en s'inspirant ou en réutilisant des relations UML existantes afin de faciliter l'apprentissage à notre modèle. De plus, au regard de la fonction première des SID, note modèle facilite le suivi des modifications de données et contribue à la fiabilité des données pour la prise de décision.

Cette proposition est énoncée dans le cadre de la définition d'un modèle multidimensionnel complet permettant de représenter l'ensemble des spécificités des SID définies dans les listes de propriétés en vue de répondre au manque d'un modèle multidimensionnel standard. Par ailleurs, bien que les liens entre les concepts et les processus de dérivation et de préparation des données facilitent le suivi de la fiabilité des

données, ils n'informent pas sur la qualité des données. En ce sens, nous projetons de travailler sur la définition de techniques et d'outils pour assurer la qualité des données du SID.

7. Bibliographie

Abelló A., Samos J., Saltor F., « YAM^2 : a multidimensional conceptual model extending UML. », *Inf. Syst. journal*, vol. 31, n° 6, p. 541-567, 2006.

Annoni E., Ravat F., Teste O., Zurfluh G., « Towards multimensional Requirement Design. », *DaWak 2006, Poland. LNCS 4081 Springer 2006*, vol. 4081/2006, p. 75-84, 2006.

Franconi E., Kamble A., « A Data Warehouse Conceptual Data Model. », *SSDBM*, IEEE Computer Society, p. 435-436, 2004. 0769521460.

Ghozzi F., Ravat F., Teste O., Zurfluh G., « Modèle multidimensionnel à contraintes », *EGC RSTI - série RIA ECA*, vol. 17, Hermes Science Publications, p. 43-55, 2003.

Golfarelli M., Rizzi S., « Methodological Framework for Data Warehouse Design. », *DOLAP*, p. 3-9, 1998.

Inmon B., « The Data Warehouse Budget », 1997. http ://www.dmreview.com/article_sub.cfm ?articleId=1315, DM Review Magazine Publisher.

Kimball R., *The data warehouse toolkit : practical techniques for building dimensional data warehouses*, John Wiley & Sons, Inc., New York, NY, USA, 1996.

Lehner W., « Modelling Large Scale OLAP Scenarios. », *EDBT, LNCS 1377 Springer*, p. 153-167, 1998.

Luján-Mora S., Trujillo J., Song I.-Y., « A UML Profile for Multidimensional Modelling in Data Warehouses. », *Data & Knowledge Engineering*, vol. 59, n° 3, p. 725-769, 2006.

Luján-Mora S., Vassiliadis P., Trujillo J., « Data Mapping Diagrams for Data Warehouse Design with UML. », *ER*, p. 191-204, 2004.

Pedersen T. B., Jensen C. S., « Multidimensional Databases. », *in* , R. Zurawski (ed.), *The Industrial Information Technology Handbook*, CRC Press, p. 1-13, 2005. 0849319854.

Rafanelli M., *Multidimensional Databases : Problems and Solutions*, Idea Group, 2003.

Ravat F., Teste O., Tournier R., Zurfluh G., « Algebraic and graphic languages for OLAP manipulations », *JDWM*, vol. 4, Idea Group, p. 17-46, janvier, 2008.

Schneider M., « A general model for design of data warehouses. », *International journal of Production Economics*, 2007.

Tryfona N., Busborg F., Christiansen J. G. B., « StarER : A Conceptual Model for Data Warehouse Design. », *DOLAP*, ACM, p. 3-8, 1999.

Vassiliadis P., Sellis T. K., « A Survey of Logical Models for OLAP Databases. », *SIGMOD Record*, vol. 28, n° 4, p. 64-69, 1999.

Vassiliadis P., Simitsis A., Skiadopoulos S., « Conceptual modeling for ETL processes. », *in* , D. Theodoratos (ed.), *DOLAP*, ACM, p. 14-21, 2002. 1581135904.