Intégration des ontologies de domaines aux bases de données relationnelles et modélisation conceptuelle : Survol critique de l'état de l'art

Oumar Sy

Section d'informatique

U.F.R de Sciences Appliquées et de Technologie
Université Gaston Berger de Saint-Louis
B.P. 234 Saint-Louis
SENEGAL

oumar.sy@ugb.edu.sn

RÉSUMÉ. De nombreux travaux ont consacré leurs efforts à l'intégration des ontologies de domaines aux bases de données. Mais à notre connaissance, aucune des approches proposées n'intègre entièrement les deux cycles de vie. Nous présentons ici un survol critique de l'état de l'art suivi d'une approche de conception parallèle pour l'intégration définitive de l'ontologie au schéma de la base de données. Notre approche, conduite par les vues des experts et des utilisateurs, est basée sur l'architecture ANSI/X3/SPARC. La sémantique des concepts est capturée dans les schémas externes et propagée aux niveaux conceptuel et logique. Les deux cycles de vie sont ainsi totalement intégrés dans un seul catalogue de base de données.

MOTS-CLÉS : ontologie de domaine, schéma de base de données, modélisation conceptuelle, cycle de vie, intégration de schémas, architecture ANSI.

ABSTRACT. Several works addressed the problem of domain ontology and databases integration. However, in our knowledge, no approach had fully integrated both life cycles. We present a survey of the state of the art and then we propose a conceptual approach for definitely integrating the domain ontology to the database schema. Our approach, leaded by the users and experts views, is based upon the well-known ANSI/X3/SPARC modeling framework. Concepts' semantics are captured into the external schemata and propagated to lower levels, namely conceptual and logical schemata. Both life cycles are then entirely integrated in a single database catalog.

KEYWORDS: domain ontology, database schema, conceptual modeling, life cycle, scheme integration, ANSI architecture.

1. Introduction

De nombreux travaux ont consacré leurs efforts à l'intégration des ontologies de domaines aux bases de données [1] [3] [4] [5] [10] dans le cadre de l'acquisition de la connaissance. Or, la connaissance, aujourd'hui une exigence des utilisateurs, dépend fortement du contexte d'utilisation de l'information et de son mode de formalisation. L'activité de modélisation de l'information joue ainsi un rôle de premier plan pour le développement des applications de bases de données et des ontologies de domaines.

Par ailleurs, il est nécessaire de permettre aux utilisateurs d'avoir un accès transparent aux masses de données hétérogènes disponibles sur le web. Ces données étant généralement stockées dans des bases de données, leur intégration requiert un consensus sur la terminologie, les catégories et les relations sémantiques entre eux [7].

Le problème posé est de concevoir un schéma de base de données (une ontologie de domaine) permettant l'accès aux connaissances ontologiques (aux données de la base).

Nous présentons, ici, un aperçu critique de l'état de l'art (Section 2) puis nous présentons une approche (Section 3) basée sur l'architecture ANSI/X3/SPARC pour l'intégration du cycle de vie de l'ontologie au cycle de vie de la base de données du domaine d'application. La section 4 conclut le papier.

2. Survol critique de l'état de l'art

La connexion entre les bases de données et les ontologies de domaines est abordée selon quatre axes de recherche : *mapping*, *transformation de schéma, extraction* de schéma et *modélisation conceptuelle*.

2.1. Le mapping

Une technique de mapping consiste, à partir d'une *ontologie de domaine existante* (classes, propriétés, types de données, contraintes) et d'une *base de données existante* (tables, colonnes, types de données, contraintes), à produire un modèle de correspondances (Figure 1 (a)) sur la base d'un ensemble de *règles*.

Il convient de noter qu'un groupe de travail du W3C, le «RDB2RDF Working Group», travaille sur un standard dans ce sens. Trois approches sont ainsi mises en œuvre pour le mapping du relationnel vers le format RDF [13] dont l'une est le mapping d'une base de données vers une ontologie.

2.2. La transformation de schéma

La transformation consiste à construire un schéma de base de données relationnelle à partir d'une *ontologie existante* [8].

2.3. L'extraction de schéma

L'extraction ou l'*acquisition* est l'opération qui consiste à construire («extraire») une ontologie [9][12] du domaine à partir d'une base de données relationnelle.

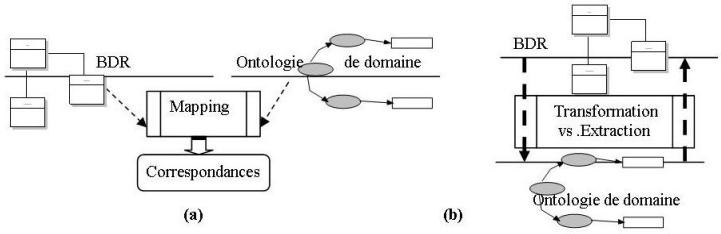

Figure 1 *Mapping (a) et Transformation vs. Extraction (b)*

Le RDB2RDF [13] est l'une des techniques de mapping du «RDB2RDF Working Group» qui prend en entrée une base de données relationnelles (le schéma et les données) pour produire un ou plusieurs graphes RDF.

Il ressort ainsi que le mapping est une technique transversale utilisée dans les techniques de transformation et d'extraction.

De notre point de vue, la transformation et l'extraction (Figure 1 (b)) sont assimilables à des techniques de *réingénierie* ayant l'inconvénient de produire des schémas incomplets. En effet, les règles-métiers, propres aux bases de données, sont inapplicables à l'ontologie de domaine, et inversement certaines contraintes sémantiques telles que la synonymie, permises sur les concepts d'ontologie, ne sont pas applicables à une base de données. Par exemple, *«L'agriculteur multiplicateur devra avoir un contrat de multiplication en bonne et due forme avec un établissement semencier agrée»* est une règle de gestion (règle-métier) qui concerne la dimension des données tandis que *«la semence est une «graine sélectionnée pour être semée»* est une définition propre à la terminologie ontologique.

Il existe, certes, des langages de description expressifs dédiés à la représentation des connaissances, au raisonnement et au web sémantique (tels que CycL, KIF, RDF-S, SPARQL, DAML+OIL, OWL).

Cependant, à notre connaissance, mis à par le système *QuOnto* de [3], il n'existe à ce jour ni un modèle ni «un système de gestion d'ontologies» reconnu à l'instar des SGBDR qui reposent sur un modèle universel robuste, en l'occurrence le modèle relationnel, pourvu d'une algèbre et d'un langage d'interrogation normalisé et *computable* : le langage SQL. De plus, *QuOnto* qui n'est pas un produit open source ne peut être évaluable étant donné qu'il n'a pas encore été commercialisé.

En résumé, l'intégration entre les ontologies de domaines et les bases de données relationnelles est prise en charge par des techniques [15] de description et de définition de différentes formes de vocabulaires (terminologies) dans un format standard tels que RDF et RDF-S, Simple Knowledge Organization System (SKOS), OWL et RIF (Rule Interchange Format). Le choix entre ces différentes technologies dépend de la complexité des applications.

Au total, ces trois approches (mapping, transformation et extraction) ont l'inconvénient commun de présenter des techniques différentes en fonction de la spécificité du domaine et de sa complexité. De plus, la transformation et l'extraction présentent deux autres limites : ***(i)*** non implication des utilisateurs dans le processus, ce qui introduit des biais dans les modèles; ***(ii)*** lourdeurs de maintenance si le contexte change.

Face à cette problématique, des approches conceptuelles ont été proposées [4] [5] [3][14]. Le système proposé par [3] a été brièvement évoqué dans la section précédente.

3. La modélisation conceptuelle

Traditionnellement les schémas des bases de données relationnelles sont conçues à l'aide des modèles sémantiques de données ER/EER ou avec le standard UML. Cependant, avec le besoin d'intégration des bases de données aux ontologies, de nombreuses critiques [4] [5] [11] ont été émises contre les modèles conceptuels. Paradoxalement, la plupart des approches [1] [3] [4] [5] [10] proposées accordent la priorité aux ontologies de domaine. La conséquence immédiate est la non prise en compte des vues des utilisateurs qui eux-mêmes ne sont pas impliqués dans tout le processus de conception. Comme inconvénient, les modèles élaborés sont généralement biaisés.

Pour pallier cet inconvénient, une approche conceptuelle a été proposée dans [14] pour l'intégration du cycle de vie de l'ontologie de domaine à celui de la base de données de son domaine d'application. Une synthèse de cette approche est faite dans la section 3.2. Auparavant, pour fixer les idées, nous faisons un récapitulatif des travaux connexes dans la section 3.1.

3.1. Travaux connexes

Comme l'approche [14], les travaux exposés dans [4] [5] abordent également l'intégration des bases de données aux ontologies de domaines par une approche conceptuelle. Pour montrer les limites notées dans [4] [5], nous rappelons, ci-après, le principe de chacune d'entre elles.

Dans [4], une extension de l'architecture ANSI/X3/SPARC avec une *«couche ontologique»* au-dessus du niveau conceptuel a été proposée. Or, l'architecture ANSI dédiée à la conception des bases de données relationnelles est une architecture tri-schématique comprenant : le niveau *externe*, le niveau *conceptuel* et le niveau *interne*.

En effet, contrairement à une pratique courante, le *niveau logique* est un niveau intermédiaire ajouté par les concepteurs de méthodes de développement [6] pour la description de la structure logique des données en fonction du modèle d'implémentation de la base de données.

Par conséquent, la *«couche ontologique»* devrait être un niveau intermédiaire entre le niveau externe et le niveau conceptuel et, l'architecture ANSI serait étendue à 4 niveaux. Dans ce cas, d'une part, le cycle vie de la conception deviendrait plus long et, d'autre part, le passage du niveau externe au niveau «ontologique» n'a pas été clarifié. L'omission du niveau externe qui est le niveau des vues présente une certaine ambigüité qui laisse suggérer que la «couche ontologique» se substitue au niveau externe. Or, le niveau externe ou ***niveau besoin***, est intrinsèque à l'activité de modélisation conceptuelle. Par conséquent, son omission introduit un biais dans les modèles d'analyse. De plus, contrairement à [4], nous convenons avec [2] qu'il n'existe pas d'ontologie canonique. Enfin, la structuration de la base de données intégrée à l'ontologie n'a pas été clairement définie dans [4].

Concernant [5], les auteurs se sont évertués à montrer les limites des modèles sémantiques de données par rapport aux langages de description logique tels que DL-Lite$_A$. Cette approche privilégie ainsi le raisonnement sur le schéma conceptuel plutôt que sur le niveau physique. Or, dans le monde des bases de données des SI, toute information qui n'est pas enregistrée dans une table de la base est considérée comme fausse. Autrement dit, c'est l'hypothèse du monde clos (Closed World Assumption) qui est prise en considération. Cette hypothèse se fonde sur la règle de l'information : (une des douze règles énoncées par Edgar Frank Codd, le fondateur du modèle relationnel) «*Toutes les informations dans une base de données relationnelle sont représentées de façon explicite au niveau logique et d'une seule manière : par des valeurs dans des tables»)*.

Dans la section suivante, nous rappelons les principes de base de l'approche conceptuelle que nous avons présentée dans [14].

3.2. Contribution: *OntoViews CM* (Ontology-based views conceptual modeling)

OntoViews CM est basée sur l'architecture ANSI/X3/SPARC dédiée à la conception des bases de données relationnelles. Le modèle relationnel repose sur la logique des prédicats du 1er ordre qui permet la description de l'univers du discours (objets-métiers et la terminologie de l'ontologie du domaine) en spécifiant formellement la structure et l'organisation des concepts du domaine. L'architecture ANSI permet aux administrateurs de bases de données, aux gestionnaires et aux utilisateurs de se référer, chacun de son point de vue, aux mêmes objets-métiers.

Étant donné un domaine du réel (partie du réel) et un contexte donné, le problème posé est de concevoir (développer) un schéma de base de données (une ontologie du domaine) intégrée à l'ontologie (à la base de données). Nous considérons un domaine comme un ensemble d'acteurs, d'objets-métiers, de processus et de tâches nécessitant une représentation adéquate des connaissances prenant en compte les rôles des acteurs dans le cadre de la réalisation de tâches de gestion spécifiques et/ou de processus.

La figure 2 illustre ce processus de conception axé sur l'architecture ANSI qui fonde le premier principe d'*OntoViews CM*.

L'intégration de l'ontologie de domaine à la base de données est prise en compte dès le niveau externe où les deux dimensions sont capturées dans les schémas externes. Au niveau interme, outre le schéma de la base de données (KRDB ou Knowledge-based Relational DataBase) et un référentiel de connaissances (KR ou Knowledge Repository) sont stockés dans le même catalogue [14].

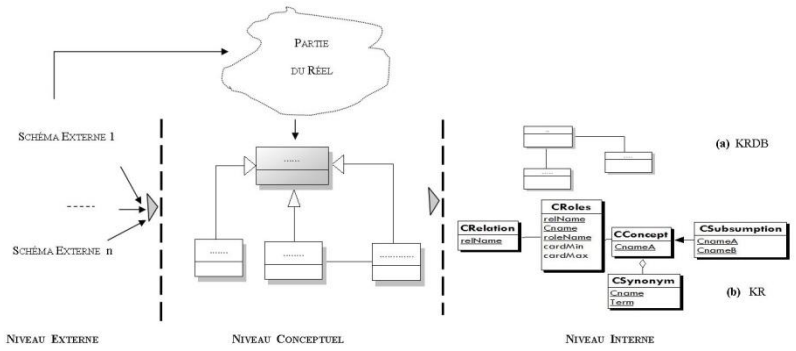

Figure 2 *OntoViews CM basée sur l'architecture ANSI/X3/SPARC*

Le schéma KR comporte les concepts de l'ontologie et, additionnellement, la sémantique du modèle ou du langage de représentation (ER/EER, UML) tels que les cardinalités et les rôles. De plus, le référentiel peut être étendu et enrichi pour capturer

toute la sémantique ontologique (propriétés et contraintes) au travers de la classe *CConcept.* Le processus de conception par OntoViews CM se résume comme suit:

1ère étape: Identification des vues référant les concepts de classes de l'ontologie de domaine et les objets-métiers (dimension données) de l'univers du discours considéré. A ce stade, l'ensemble Σ des concepts est défini.

2ème étape: Construction de la dimension des données. A chaque concept dans Σ, est association une relation dans la structure de la base de données. Par exemple, si *PERS* $\in\Sigma$ *et* $\leq_s$ *(PERS, GRST),* $\leq_s$ *(PERS, PHDST),* $\leq_s$*(PERS, TR)* alors, la description de la classe *PERS* entraîne la définition des classes *PHDST, TR* and *GRST.*

3ème étape: Construction des relations entres classes et sous-classes

***4ème étape*:** Extension de la *T-Box* avec deux classes: *CSynonym* de schéma *{CName, Term}* et *CRoles* ayant pour schéma *{CName, roleName, relName, cardMin, cardMax*). Dans *CSynonym*, pour chaque concept *(CName),* tous ses synonymes (*Term*) sont stockés. Dans *CRoles*, chaque concept *(CName)* est associé aux cardinalités minimale (*cardMin*) et maximale (*cardMax*) définissant ses rôles de participation (*roleName*) dans une relation *(relName*). De cette manière, tous les noms de relations sont stockés ensemble avec les noms des associations (liens d'association nommés) et les cardinalités correspondantes.

Cette approche d'intégration est similaire à celle adoptée par [3] dans sa finalité. Cependant, notre approche présente deux avantages majeurs : ***(i)*** OntoViews CM est basé sur des modèles facilement interprétables par des non-experts ; ***(ii)*** le standard UML prend en charge intégralement tout le processus de développement des applications des bases de données des plus simples aux plus complexes. En effet, UML peut capturer les sémantiques fonctionnelle et structurelle d'un SI quelle que soit sa complexité.

4. Conclusion

L'objectif, dans ce papier, était de donner un aperçu synthétique sur l'état de l'art concernant l'intégration des bases de données aux ontologies. Nous avons ainsi rappelé les techniques utilisées dans ce domaine (mapping, transformation et extraction).

Il ressort clairement que la pratique générale consiste au développement d'ontologies à partir de schémas conceptuels et/ou de bases de données existantes.

Nous avons ainsi rappelé, en résumé, notre approche conceptuelle pour l'intégration des cycles de vie des ontologies de domaines aux bases de données.

Par ailleurs, dans le contexte du web où la plupart des données requises sont stockées dans des bases de données relationnelles, il est nécessaire d'établir une connexion entre ces bases de données et un format compatible pour le data-web.

Il est également établi qu'une spécification RDF-S peut être entièrement définie à partir d'un modèle UML [15]. Par conséquent, une base de données conçue selon l'approche OntoViews CM se prête davantage à une extraction d'ontologies en vue de son exposition sur le data-web dans l'esprit de RDB2RDF. Il existe, en effet, une étroite similarité sur la manière dont les données sont structurées par SQL, d'une part, et RDF-S, d'autre part (voir Figure 3).

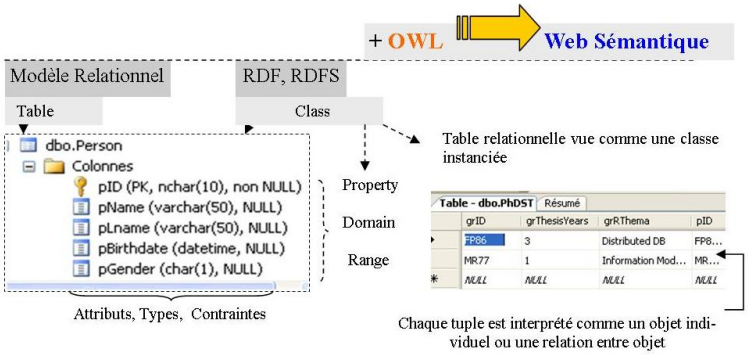

Figure 3 *De SQL à RDF-Schema*

5. Bibliographie

[1] CALVANESE, D., DE GIACOMO, G., LEMBO, D., LENZERINI, M., POGGI, A., ROSATI, R.: Linking Data to Ontologie, The Description Logic DL-LiteA

[2] COSTELLO, J. C, DAN, S., JEFF, G., MEHMET, D., «DATA-DRIVEN ONTOLOGIES», *Pacific Symposium on Biocomputing* 14:15-26, 2009.

[3] DE GIACOMO, G.: QuOnto: ontology-based data access and integration using relational technology, *Semantic Days*, SAPIENZA UNIVERSITA DI ROMA, 2009.

[4] FANKAM, C., JEAN, S., PIERRA, G., BELLATRECHE, L., AÏT-AMEUR, Y.: Towards Connecting Database Applications to Ontologies, *First Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA*, 2009.

[5] FRANCONI, E.: Ontologies and Databases: myths and challenges PVLDB '08, August 23-28, 2008, Auckland, New Zealand, Copyright 2008 VLDB Endowment, ACM 978-1-60558-306-8/08/08, 2008.

[6] FONSECA, F., MARTIN, J.: Learning the Differences between Ontologies and Conceptual Schemas Through Ontology-Driven Information Systems, *JAIS - Journal of the Association for Information Systems - Special Issue on Ontologies in the Context of IS,* Volume 8, Issue 2, Article 3, pp. 129–142, preprint version 1, 2007.

[7] Herman, I. Introduction to Semantic Web Technologies Ivan Herman, *W3C Semantic Technology Conference, San Franciso*, June 21-25, 2010.

[8] IRINA, A., KORDA, N., AND KALJA, A.: Storing OWL Ontologies in SQL Relational Databases, World Academy of Science, Engineering and Technology, pp.167-172, 2007.

[9] LUBYTE, L., TESSARIS, S. KRDB Research Centre Technical Report: Extracting Ontologies from Relational Databases, KRDB, 2007

[10] NETWORK INFERENCE. INC. Ontologies and Data Warehousing | www.networkinference.com, 2004

[11] PIROTTE, A., MASSART, D.: Integrating Two Descriptions of Taxonomies with Materialization, in Journal of Object Technology, vol. 3, no. 5, pages 143-149, http://www.jot.fm/issues_2004_05/article4, 2004.

[12] SHUFENG, Z. GUANGWU, M., HAIYUN, L. Ontologies Acquisition from Relational Databases, Computer and Information Sciences, Vol. 3, N°1, February 2010.

[13] SÖREN, A., LEE, F., DANIEL, M., ANGELA, F., JUAN, S., *Use Cases and Requirements for Mapping Relational Databases to RDF* W3C Working Draft 8 June 2010, http://www.w3.org/TR/2010/WD-rdb2rdf-ucr-20100608/

[14] SY, O., DUARTE, D., LO, M. "Integrating Ontologies in Database Scheme: Ontology-Based Views Conceptual Modeling, *the 6th International Conference on Signal-Image Technology & Internet-Based Systems SITIS*, 15-18 December, Kuala Lumpur (Malaysia), 2010.

[15] WALTER, W.C., A Discussion of the Relationship between RDF-Schema and UML, W3C Note 04-Aug-1998, http://www.w3.org/TR/1998/NOTE-rdf-uml-19980804