UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE D'INFORMATIQUE DE PARIS-SUD

LABORATOIRE DE RECHERCHE EN INFORMATIQUE

Discipline: Informatique

Synthèse en français

THÈSE DE DOCTORAT

Soutenue le 24 Octobre 2014

par

Julio Cesar DOS REIS

Adaptation des Mappings entre Systèmes d'Organisation de la Connaissance du domaine Biomédical

Directrice de thèse : Chantal Reynaud-Delaître - Professeur, Université de Paris-Sud, France
Co-encadrant de thèse : Cédric Pruski - Chargé de recherche, CRP Henri Tudor, Luxembourg

Composition du jury :

Rapporteurs : Nathalie Aussenac-Gilles - Directrice de recherche, IRIT-CNRS, France
Frank van Harmelen - Professeur, Université libre d'Amsterdam, Pays-Bas

Examinateurs : Christine Froidevaux - Professeur, Université de Paris-Sud, France
Erhard Rahm - Professeur, Université de Leipzig, Allemagne
Stéfan J. Darmoni - Professeur, Université de Rouen, France

Résumé

Les systèmes d'information biomédicaux actuels reposent sur l'exploitation de données provenant de sources multiples. Les Systèmes d'Organisation de la Connaissance (SOC) permettent d'expliciter la sémantique de ces données, ce qui facilite leur gestion et leur exploitation. Bénéficiant de l'évolution des technologies du Web sémantique, un nombre toujours croissant de SOCs a été élaboré et publié dans des domaines spécifiques tels que la génomique, la biologie, l'anatomie, les pathologies, *etc.* Leur utilisation combinée, nécessaire pour couvrir tout le domaine biomédical, repose sur la définition de mises en correspondance entre leurs éléments ou mappings. Les mappings connectent les entités des SOCs liées au même domaine via des relations sémantiques. Ils jouent un rôle majeur pour l'interopérabilité entre systèmes, en permettant aux applications d'interpréter les données annotées avec différents SOCs. Cependant, les SOCs évoluent et de nouvelles versions sont régulièrement publiées de façon à correspondre à des vues du domaine les plus à jour possible. La validité des mappings ayant été préalablement établis peut alors être remis en cause. Des méthodes sont nécessaires pour assurer leur cohérence sémantique au fil du temps. La maintenance manuelle des mappings est une possibilité lorsque le nombre de mappings est restreint. En présence de SOCs volumineux et évoluant très rapidement, des méthodes les plus automatiques possibles sont indispensables.

Cette thèse de doctorat propose une approche originale pour adapter les mappings basés sur les changements détectés dans l'évolution de SOCs du domaine biomédical. Notre proposition consiste à comprendre précisément les mappings entre SOCs, à exploiter les types de changements intervenant lorsque les SOCs évoluent, puis à proposer des actions de modification des mappings appropriées. Nos contributions sont multiples : (i) nous avons réalisé un travail expérimental approfondi pour comprendre l'évolution des mappings entre SOCs; nous proposons des méthodes automatiques (ii) pour analyser les mappings affectés par l'évolution de SOCs, et (iii) pour reconnaître l'évolution des concepts impliqués dans les mappings via des patrons de changement; enfin (iv) nous proposons des techniques d'adaptation des mappings à base d'heuristiques. Nous proposons un cadre complet pour l'adaptation des mappings, appelé *DyKOSMap*, et un prototype logiciel. Nous avons évalué les méthodes proposées et le cadre formel avec des jeux de données réelles contenant plusieurs versions de mappings entre SOCs du domaine biomédical. Les résultats des expérimentations ont démontré l'efficacité des principes sous-jacents à l'approche proposée. La maintenance des mappings, en grande partie automatique, est de bonne qualité.

Mots-clés : Système d'Organisation de la Connaissance; Domaine biomédical; Ontologies biomédicales; Evolution d'ontologies; Evolution de SOCs dans le domaine biomédical; Changements dans les ontologies; Mappings entre ontologies; Maintenance des mappings; Adaptation des mappings; Evolution des mappings.

1. Introduction

Les avancées dans le domaine biomédical provoquent régulièrement l'apparition de nouveaux concepts médicaux et de nouvelles méthodes de traitement des maladies sont découvertes grâce aux recherches effectuées dans le domaine. Ces traitements peuvent cependant devenir moins efficaces au fil du temps car les causes, voire les symptômes, de certaines maladies peuvent changer sous l'effet de bouleversements d'ordre social ou environnemental liés, par exemple, à l'urbanisation, la pollution, le déboisement. Ainsi le virus H7N9, qui est une sorte de *virus grippal de type A* de la maladie *grippe aviaire*, a été détecté pour la première fois chez l'Homme à Shanghaï, en mars 2013. Selon l'OMS au 29 mai 2013, il est déjà à l'origine de 37 décès[1].

Des standards terminologiques, gérés par plusieurs établissements dans le monde, ont été créés pour aider à gérer les concepts biomédicaux (maladies, causes, symptômes, *etc.*) utilisés par les professionnels de la santé. La Classification Internationale des Maladies[2] (CIM) en est un exemple. Publiée par l'Organisation Mondiale de la Santé (OMS), elle traite des causes de morbidité et de mortalité. Le terme « somnambulisme », qui désigne un concept lié à la maladie ou au trouble du sommeil appartenant à la famille des parasomnies, y figure associé au code « 307.4 » dans la CIM-9-CM et « F51.3 » dans la CIM-10. La nomenclature SNOMED-CT, actuellement gérée et distribuée par l'organisme IHSTDO[3], en est un autre exemple. Elle fournit des codes, des termes, des synonymes et des définitions de centaines de milliers de concepts (environ 400 000 concepts en 2013) liés aux maladies[4], par exemple « paludisme » [61462000], « rougeole » [14189004], « reflux gastro-œsophagien » [54856001], des concepts liés aux substances, par exemple « lactoferrine » [10267005], « beta-N acetylhexosaminidase A » [102784005], « ribose-5-phosphate isomérase » [412004], etc. En pratique, pour réduire l'ambiguïté des termes de cette nomenclature, chaque terme est associé à un suffixe (*e.g.*, maladie, procédure, substance, *etc.*). Ainsi le concept ayant l'identifiant [6146200] est dénoté par le terme « paludisme (*maladie*) » tandis que le concept identifié par le code [10267005] est dénoté par le terme « lactoferrine (*substance*) ».

La taille importante ainsi que la complexité du domaine biomédical conduisent généralement à concevoir des systèmes d'information combinant plusieurs Systèmes d'Organisation de la Connaissance (SOCs). Dans le guide de bonnes pratiques des vocabulaires publié par l'organisation HL7[5], un document peut contenir des champs codés associés à des termes issus

[1] http://www.who.int/csr/don/2013_05_29/fr/

2 http://www.who.int/classifications/icd/en/index.html

[3] http://www.ihtsdo.org

[4] Notons que la traduction en langue française de la SNOMED-CT est en cours jusqu'en mai 2013

[5] http://www.hl7.org

d'une ou de plusieurs terminologies dont certaines peuvent être créées par l'utilisateur lui-même (par exemple pour représenter la localisation ou la structure d'un établissement de la santé). Les informations codées par la CIM-9-CM peuvent être utilisées pour l'analyse statistique de la morbidité et de la mortalité des maladies, le remboursement des frais médicaux, ou l'aide à la prise de décision médicale. La nomenclature standard SNOMED-CT couvre différents domaines cliniques comme les maladies, les symptômes, les traitements, les matériels, les substances, *etc.* Elle aide à organiser le contenu des dossiers médicaux des patients et permet d'échanger l'information médicale en facilitant l'interopérabilité entre les systèmes d'information.

Afin d'exploiter de manière efficace plusieurs SOCs, il est nécessaire d'établir des connexions entre elles. Ces correspondances sémantiques, plus communément appelées mappings, définissent des relations sémantiques (équivalence, plus générique, plus spécifique, *etc.*) entre leurs éléments, généralement entre des concepts. A titre d'exemple, l'organisme IHSTDO[6] a créé 86 638 mappings entre la version de janvier 2012 de la nomenclature SNOMED-CT[7] et la version 2011 de la classification des maladies CIM-9-CM[8]. Cependant, les connaissances évoluent. Cela entraîne des modifications dans les SOCs qui pouvent invalider les mappings établis, et par conséquent perturber le fonctionnement des applications logicielles qui les exploitent.

Ainsi, la maintenance des mappings est une tâche primordiale. Plusieurs aspects doivent être pris en compte, en particulier les informations responsables de l'évolution des SOCs. En effet, l'alignement de concepts s'explique bien souvent par l'existence de relations sémantiques entre des parties d'information (des attributs par exemple) les décrivant. Ainsi, lorsqu'un concept évolue, l'identification de l'information conceptuelle modifiée est importante car, combinée à la connaissance expliquant les mappings préexistants, elle permet de prévoir l'évolution de ces derniers.

Dans cette thèse, nous abordons le problème de l'adaptation des alignements sémantiques rendus incorrects par l'évolution des SOCs auxquels ils sont rattachés. L'approche *DyKOSMap* que nous proposons s'appuie sur l'utilisation d'heuristiques permettant de combiner des informations conceptuelles définissant les mappings, notamment la relation sémantique qui lie les éléments mis en correspondance, et principalement celles provenant de l'évolution de la valeur des attributs de concepts. Nous avons défini des *patrons de changement* à partir des informations provenant des mappings eux-mêmes pour décider de l'adaptation des mappings. Nous

[6] http://www.ihtsdo.org
[7] http ://www.ihtsdo.org/snomed-ct
[8] http ://www.cdc.gov/nchs/icd/icd9cm.htm

proposons également une validation expérimentale de notre approche appliquée à des ensembles de mappings reliant plusieurs SOCs du domaine biomédical.

Ce document présente un résumé du manuscrit original. La suite de ce document est structurée comme suit : la section 2 présente une définition du problème de l'adaptation des mappings. La section 3 traite de l'analyse expérimentale du problème que nous avons menée. La section 4 présente la méthode pour l'identification des informations conceptuelles expliquant les mappings. La section 5 présente notre approche pour la caractérisation de l'évolution des éléments des SOCs. La section 6 définit les actions d'adaptation de mappings et la section 7 présente le cadre formel *DyKOSMap*. La section 8 conclut et énonce quelques perspectives.

2. La problématique de l'adaptation des mappings

Dans notre contexte de travail, nous définissons un mapping m_{st} comme un triplet (c_s, c_t, r) où c_s représente un concept d'un SOC source (SOC_A), c_t un concept d'un SOC cible (SOC_B) et r la relation sémantique entre c_s et c_t. Le problème traité dans ces travaux (*cf.* **Error! Reference source not found.**) consiste à faire évoluer de façon semi-automatique la définition d'un mapping pour que celui-ci reste valide après évolution des SOCs alignés. En d'autres termes, nous nous proposons de modifier c_s et/ou r pour que la nouvelle version du mapping soit valide sémantiquement. La Figure *2* présente un exemple de la problématique appliquée à des concepts du domaine biomédical.

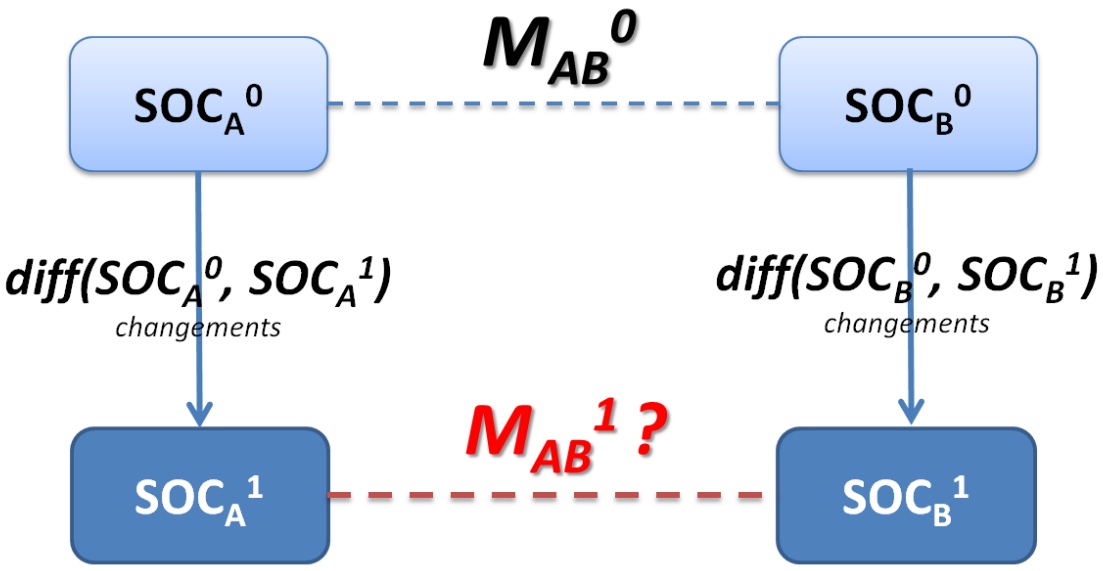

Figure 1. La problématique de l'adaptation des mappings

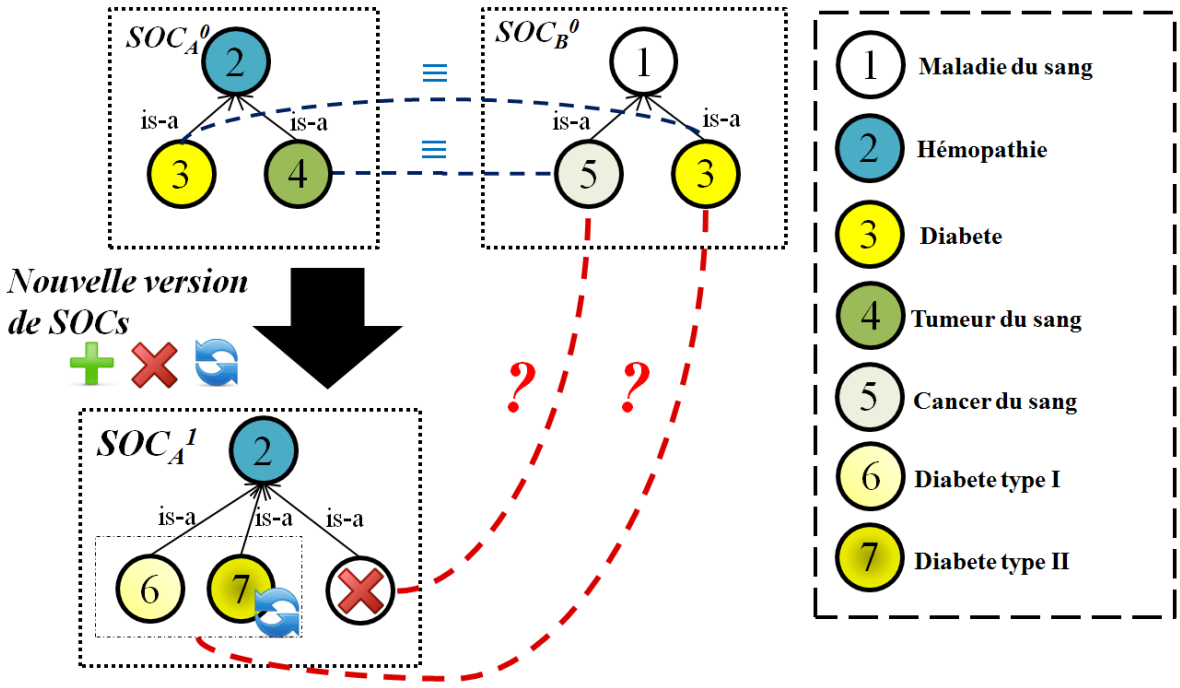

Figure 2. Un exemple concret de la problématique

3. Analyse expérimentale de l'évolution des mappings

Nous avons débuté notre travail par une étude expérimentale ayant pour objectif de comprendre comment se comportent les mappings lorsque des SOCs du domaine biomédical évoluent [5, 6]. Dans ce but, nous avons mis en place une série d'expérimentations reposant sur des versions successives de la SNOMED CT, de la CIM-9-CM et des mappings connectant leurs concepts respectifs. Les analyses quantitative et qualitative de l'influence de l'évolution des SOCs sur les mappings ont montré deux choses essentielles :

1. La corrélation entre la façon dont les concepts des SOCs évoluent et le comportement des mappings qui leur sont associés;
2. La très forte influence de l'évolution de la valeur des attributs de concepts sur le comportement des mappings.

Cette dernière observation a conditionné la suite de nos travaux et nous distingue des approches existantes [8-10] du fait que nous allons aborder la problématique de l'adaptation des mappings et de l'évolution des SOCs en travaillant au niveau des attributs de concepts et non pas au niveau des concepts (niveau plus abstrait).

4. Identification des informations conceptuelles définissant les mappings

L'étude menée à l'étape précédente a montré que lorsque des concepts sont alignés, très souvent seule une partie des informations les décrivant explique le mapping. Il est alors primordial d'identifier cette partie d'information pour pouvoir, dans un deuxième temps, caractériser son évolution et en déduire l'évolution du mapping associé.

Pour identifier les informations conceptuelles définissant les mappings, nous avons proposé une méthode utilisant des mesures de similarité lexicales (distance de Levenshtein), syntaxiques (travaillant au niveau des mots) et sémantiques (exploitant un corpus de documents annotés) [1]. Ce choix a été motivé par le fait que les méthodes classiques d'alignement de SOC s'appuient sur ce type de métriques pour déterminer les mappings entre concepts. Notre algorithme calcule la similarité entre les valeurs des attributs définissant les concepts alignés et classe, suivant la valeur de la mesure obtenue, les attributs des deux concepts du mapping. Ainsi, les attributs dont les valeurs sont les plus similaires sont considérés comme les plus pertinents pour la définition des mappings. L'algorithme *TopA* est paramétrable dans la mesure où nous pouvons fixer le nombre d'attributs pertinents et nous avons montré expérimentalement [1] que 3 est cette valeur optimale.

5. Caractérisation de l'évolution des SOCs

Les expérimentations effectuées dans nos tâches d'analyses préliminaires ont clairement montré une forte corrélation entre l'évolution des SOCs et le comportement des mappings à travers le temps [6]. De ce fait, nous nous sommes concentrés sur la caractérisation de l'évolution des informations conceptuelles, retournées par notre algorithme *TopA*, et nous nous sommes focalisés sur l'évolution des valeurs des attributs. Nos propositions reposent sur l'analyse d'un corpus construit à partir de versions successives de SOCs, comme la SNOMED CT, la CIM-9-CM et le MeSH[9]. Nous avons défini 8 *patrons de changement* (4 au niveau lexical et 4 au niveau sémantique) [3] caractérisant formellement l'évolution des valeurs des attributs de concepts afin de les exploiter ensuite pour la phase d'adaptation des mappings invalides.

[9] http://www.ncbi.nlm.nih.gov/mesh

A travers de nos expérimentations, nous avons observé que, d'un point de vue lexical, 4 types de changement affectaient les valeurs des attributs de concept. Celles-ci pouvaient être *copiées* d'un attribut de concept à un autre soit *totalement* soit *partiellement*. Par exemple, au temps j un attribut a_1 peut avoir comme valeur « portal systemic encephalopathy » et au temps j+1, a_1 conserve cette valeur, mais un attribut a_2 du même SOC aura également cette même valeur (cas d'une copie totale). Un *transfert partiel* ou *total* de la valeur d'un attribut de concept peut également être observé lors de l'évolution des SOCs. Dans le cas d'un transfert, contrairement au patron de changement précédent *copie*, la valeur de l'attribut a_1 est effacée.

D'un point de vue sémantique, les 4 patrons de changement que nous proposons permettent d'évaluer si les changements au niveau des valeurs des attributs considérés modifient le niveau d'abstraction des concepts qui leurs sont associés. Ceci nous permet de comprendre si les concepts deviennent *plus spécifiques*, *plus généraux*, restent *équivalents* ou sont en relation (sans que celle-ci ne soit précisément définie) lors de leur évolution. Par exemple, un attribut a_1 peut avoir « kappa light chain disease » comme valeur au temps j, et « kappa chain disease » au temps j+1. La perte du qualificatif « light » rend la valeur au temps j+1 *plus générale* que celle au temps j, et cette modification se répercute sur les concepts concernés.

Nous avons proposé une méthode de reconnaissance automatique de ces patrons de changement et validé empiriquement l'ensemble de nos propositions sur un corpus d'environ 1000 attributs, construit manuellement par des experts du domaine. Ces derniers ont également permis de confirmer les résultats retournés par nos algorithmes.

6. Les actions d'adaptation des mappings

Nos patrons de changement nous ont permis de caractériser l'évolution des informations conceptuelles des concepts définissant un mapping. De la même manière, nous avons expérimentalement mis en évidence le comportement des mappings affectés par l'évolution des SOCs. Ce travail a permis de définir, dans le même formalisme logique que nos patrons de changement, un ensemble de 6 actions s'appliquant aux mappings pour garantir leur validité [2, 7]. Des actions simples, comme *AdditionM* ou *RemoveM,* ajoutent et suppriment respectivement des mappings si l'évolution des SOCs le justifie. Des actions plus complexes ont également été définies. *MoveM* spécifie le remplacement du concept source c_s par un concept c_s' dans la nouvelle version du mapping. *DeriveM* entraîne une duplication d'un mapping existant et le changement du concept source c_s dans la version dupliquée. *ModSemTypeM* change le type de

relation sémantique reliant c_s et c_t dans la nouvelle version du mapping. Enfin, des changements au niveau des concepts sous-jacents peuvent ne pas entraîner de modification au niveau du mapping, auquel cas *NoAction* est appliquée.

Les actions d'adaptation de mapping ainsi définies ont été largement observées sur un corpus composé de plusieurs versions successives de mappings reliant la SNOMED CT, la CIM-9-CM, le MeSH mais aussi le thesaurus NCI[10] et la CIM-10-CM.

7. Relations entre patrons de changement et actions d'adaptation de mappings

Le fait d'être capable de caractériser, d'une part l'évolution des informations conceptuelles définissant les mappings sous la forme de patrons de changement et d'autre part, le comportement des mappings à travers le temps, permet de maintenir les mappings uniquement à partir du moment où le lien entre patrons de changement et action d'adaptation des mappings est établi. Dans ce but, nous avons proposé un ensemble d'heuristiques formalisant ce lien. Ces heuristiques établissent une relation entre la façon dont les informations des SOCs évoluent et leur impact sur la nouvelle définition des mappings. Ces heuristiques ont été formalisées en logique du premier ordre. Elles ont été validées expérimentalement sur plusieurs versions successives de mappings officiels établis entre la SNOMED CT et la CIM-9-CM. Nous avons appliqué ces heuristiques à un ensemble de mappings au temps j et comparé les résultats obtenus avec ceux de la version officielle des mappings au temps j+1, en utilisant les mesures classiques de précision, rappel et F-Score.

8. Le cadre formel *DyKOSMap* pour l'adaptation des mappings

L'approche originale que nous proposons pour l'adaptation des mappings entre SOCs du domaine biomédical a conduit à définir le cadre *DyKOSMap* (*cf.* Figure 3) [4][11]. Le but est d'intégrer tous les éléments mis en jeu au sein de ce travail de doctorat. Cette intégration a été rendue possible grâce à l'utilisation d'un formalisme logique homogène pour représenter les patrons de changement, les actions d'adaptation des mappings et les heuristiques liant les patrons de changements aux actions. Notre approche a été mise en œuvre au sein d'une application logicielle développée en langage Java.

[10] http://ncit.nci.nih.gov

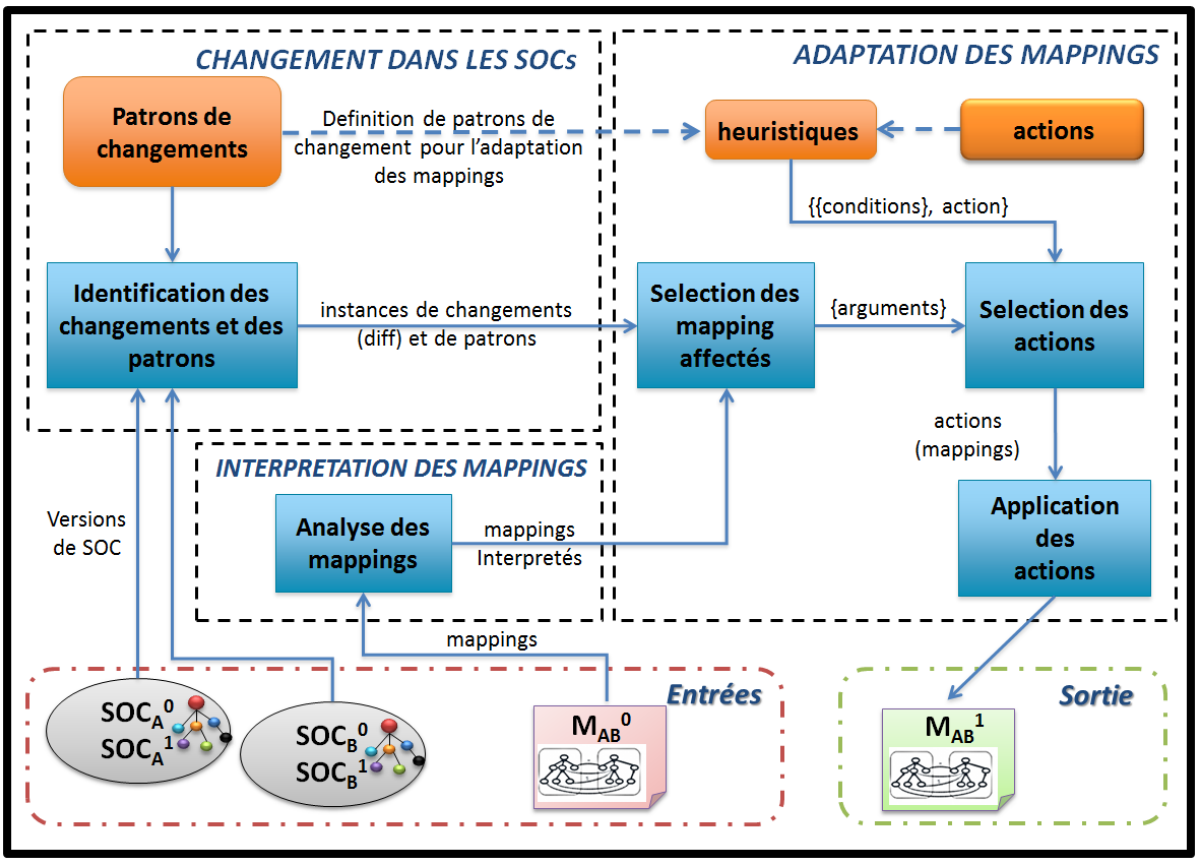

Figure 3. Le cadre formel *DyKOSMap*

En plus des différentes expérimentations menées tout au long de cette thèse, le cadre *DyKOSMap* a également été validé expérimentalement sur les mappings existants entre la SNOMED CT et la CIM-9-CM ainsi qu'entre la SNOMED CT et le thésaurus NCI. Les résultats obtenus montrent les bénéfices de notre approche pour la maintenance des mappings par rapport aux approches existantes, ce qui témoigne de la qualité de nos travaux de recherche.

9. Conclusion

Dans cette thèse nous avons proposé une approche originale pour l'adaptation des alignements sémantiques établis entre les concepts des SOCs du domaine biomédical. L'approche *DyKOSMap* que nous avons définie exploite à la fois les informations provenant de la caractérisation de l'évolution des informations conceptuelles définissant les mappings, et les différents comportements d'adaptation de mappings afin de garantir la validité, d'un point de vue sémantique, de ces derniers au cours du temps. Un prototype supportant *DyKOSMap* a été développé et a servi à une validation expérimentale approfondi sur les principaux SOCs du

domaine biomédical. Ces travaux d'investigation ont permis d'ouvrir un grand nombre de perspectives pour les travaux futurs, parmi lesquelles on retrouve principalement l'enrichissement des mappings grâce aux informations obtenues sur l'évolution des concepts.

Références

[1] D. Dinh, J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Identifying relevant concept attributes to support mapping maintenance under ontology evolution," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2014.

[2] J. C. Dos Reis, D. Dinh, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Mapping Adaptation Actions for the Automatic Reconciliation of Dynamic Ontologies," in ACM International Conference on Information and Knowledge Management (CIKM 2013), San Francisco, 2013, pp. 599-608.

[3] J. C. Dos Reis, D. Dinh, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Identifying change patterns of concept attributes in ontology evolution," in ESWC, Anissaras, Crete, (Greece), 2014, pp. 768-783.

[4] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Analyzing and Supporting the Mapping Maintenance Problem in Biomedical Knowledge Organization Systems," *Workshop on Semantic Interoperability in Medical Informatics (SIMI 2012) - 9th Extended Semantic Web Conference*, 2012, pp. 25-36.

[5] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Characterizing Semantic Mappings Adaptation via Biomedical KOS Evolution: A Case Study Investigating SNOMED CT and ICD," in AMIA 2013 Annual Symposium, Washington DC (USA), 2013, pp. 333-342.

[6] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Understanding Semantic Mappings Evolution by Observing Changes in Biomedical Ontologies," *Journal of Biomedical Informatics,* vol. 47, pp. 71-82, 2014.

[7] A. Gross, J. C. Dos Reis, M. Hartung, C. Pruski, and E. Rahm, "Semi-Automatic Adaptation of Mappings between Life Science Ontologies," in Data Integration in the Life Sciences (DILS 2013), Montreal, Canada, 2013, pp. 90-104.

[8] A. Khattak, Z. Pervez, K. Latif, and S. Lee, "Time efficient reconciliation of mappings in dynamic web ontologies," *Knowl.-Based Syst.,* vol. 35, pp. 369-374, 2012.

[9] H. Martins, and N. Silva, "A User-Driven and a Semantic-Based Ontology Mapping Evolution Approach," *11th International Conference on Enterprise Information Systems*, 2009, pp. 6-10.

[10] C. Meilicke, H. Stuckenschmidt, and A. Tamilin, "Reasoning Support for Mapping Revision," *Journal of Logic and Computation,* vol. 19, no. 5, pp. 807-829, 2008.

[11] J. C. Dos Reis, "Maintaining Mappings valid between dynamic KOS" in ESWC, Montpellier (France), 2013, pp. 650-655.