HAL
open science

Identification des informations conceptuelles définissant un alignement entre ontologies médicales

Dinh Duy, Julio Cesar dos Reis, Marcos da Silveira, Cédric Pruski, Chantal Reynaud-Delaître

To cite this version:

Dinh Duy, Julio Cesar dos Reis, Marcos da Silveira, Cédric Pruski, Chantal Reynaud-Delaître. Identification des informations conceptuelles définissant un alignement entre ontologies médicales. Atelier SIIM organisé conjointement à IC 2013, Jul 2013, Lille, France. hal-00945874

HAL Id: hal-00945874

https://inria.hal.science/hal-00945874

Submitted on 13 Feb 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identification des informations conceptuelles définissant un alignement entre ontologies médicales

Duy Dinh[1],Julio Cesar Dos Reis[1,2], Marcos Da Silveira[1], Cédric Pruski[1] et Chantal Reynaud-Delaître[2]

[1] CR SANTEC, Centre de Recherche Public Henri Tudor,
6 avenue des hauts-fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
{duy.dinh,julio.dosreis,marcos.dasilveira,cedric.pruski}@tudor.lu
[2] LRI, Université Paris-Sud XI,
Bâtiment 650 (PCRI), F-91405 Orsay, France
chantal.reynaud@lri.fr

Résumé : La quantité d'information médicale disponible de nos jours et la multitude de sources diverses poussent les utilisateurs à utiliser les ontologies pour des besoins d'interopérabilité sémantique. L'utilisation combinée de plusieurs ontologies est nécessaire de par la taille du domaine médical mais n'est possible qu'à travers la définition de relations sémantiques (ou mappings) entre leurs éléments. Cependant, les connaissances médicales évoluent, ce qui entraîne une révision constante des ontologies et, par conséquent, fragilise la cohérence des mappings ayant pu être établis. Cet article propose une méthode pour identifier les informations décrivant les concepts dans les ontologies sur lesquelles repose la définition des mappings. Cette approche s'appuie sur l'utilisation de la similarité lexicale et est évaluée sur la base des terminologies SNOMED CT, CIM-9 et les mappings officiels existants entre ces deux ressources.

Mots-clés : Evolution d'ontologies médicales, mesures de similarité, maintenance des mappings, adaptation et évolution des mappings.

1 Introduction

Avec les avancés dans le domaine biomédical, de nouveaux concepts liés aux maladies ainsi que les méthodes pour traiter ces maladies ont été trouvés grâce aux résultats de la recherche dans ce domaine. Cependant, ces traitements peuvent devenir moins efficaces au fil du temps car les causes et symptômes de certaines maladies peuvent changer en raison des changements sociaux et environnementaux liés à l'urbanisation, la

pollution et le déboisement. Par exemple, le virus H7N9, qui est un sous-type du *virus grippal de type A* de la maladie *grippe aviaire*, a été détecté pour la première fois chez l'Homme à Shanghaï, au mois de mars 2013 et déjà responsable de 37 décès confirmés[1] (selon l'OMS au 29 mai 2013). Afin de gérer les concepts biomédicaux (maladies, causes, symptômes, etc…) utilisés par les professionnels de la santé, les terminologies standards ont été créées et gérées par plusieurs établissements dans le monde entier. A titre d'exemple, la Classification Internationale des Maladies[2] (CIM), qui est publiée par l'Organisation mondiale de la santé (OMS), traite des causes de morbidité et de mortalité. Par exemple, le terme « somnambulisme », qui désigne un concept lié à la maladie ou au trouble du sommeil appartenant à la famille des parasomnies, est associé au code « 307.4 » dans la CIM-9 ou « F51.3 » dans la CIM-10.

La nomenclature SNOMED-CT, actuellement gérée et distribuée par l'organisme IHSTDO[3], fournit des codes, termes, synonymes et définitions des centaines de milliers de concepts (environ 400,000 concepts en 2013) liés aux maladies[4] (e.g., « paludisme » [61462000], « rougeole » [14189004], « reflux gastro-œsophagien » [54856001]…), aux substances (e.g., « lactoferrine » [10267005], « beta-N acetylhexosaminidase A » [102784005], « ribose-5-phosphate isomérase » [412004]…), etc. En pratique, pour réduire l'ambiguïté des termes désignant les concepts biomédicaux, chaque terme est associé à un suffixe (e.g., maladie, procédure, substance, etc.). Par exemple, 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 nécessite l'utilisation combinée de plusieurs ressources termino-ontologiques. En effet, selon le guide de bonnes pratiques de vocabulaires publié par l'organisation HL7[5], un document peut contenir certains champs codés qui contiennent des termes issus d'une ou de plusieurs terminologies et parfois les terminologies peuvent être créées par l'utilisateur lui-même (par exemple pour représenter les locations ou la structure de l'établissement de la santé). Les informations codées par la CIM-9-CM peuvent être utilisées pour de divers objectifs comme par exemple 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, matériels, substances … Elle assiste à l'organisation du

---

[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/

contenu des dossiers médicaux de patients et fournit un mécanisme consistant à échanger de l'information médicale en facilitant l'interopérabilité entre les systèmes d'informations.

Afin d'exploiter de manière efficace plusieurs ressources termino-ontologiques, il est nécessaire d'établir des connexions entre elles. Ces correspondances sémantiques, plus communément appelées mappings, définissent des relations sémantiques (*e.g.* équivalence, plus générique, plus spécifique…) entre leurs éléments, ou 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](abrégée SCT) et la version 2011 de la classification des maladies CIM-9-CM[8] (abrégée CIM). Cependant, les connaissances évoluent, ce qui entraîne des modifications dans les ontologies pouvant invalider les mappings existants, et par conséquent, perturber le fonctionnement des applications logicielles qui les exploitent.

De ce fait, la maintenance des mappings devient une tâche primordiale. Plusieurs aspects doivent être pris en compte, en particulier les informations responsables de l'évolution des ressources. 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 cet article, nous abordons le problème de l'identification des informations conceptuelles expliquant des mappings. L'approche que nous proposons s'appuie sur l'utilisation de mesures de similarité. Nous étudions trois types différents de mesures de similarité, la *similarité lexicale*, la *similarité syntaxique* et la *similarité sémantique*. Nous évaluons cette approche sur plusieurs versions de la SCT et de la CIM et des mappings officiels qui leur sont associés. Ces expérimentations montrent une corrélation entre les changements affectant les informations conceptuelles que la technique proposée est capable d'identifier et le comportement des mappings lorsque ceux-ci évoluent. Les résultats de ces expériences serviront, par la suite, à définir une approche (semi-)automatique pour l'adaptation des mappings (Dos Reis, Pruski, Da Silveira, & Reynaud-Delaître, 2012).

La suite de cet article est structurée comme suit : la section 2 introduit les travaux relatifs à la maintenance des mappings. La section 3 présente une définition du problème de maintenance des mappings ainsi que notre approche générale. La section 4 traite de l'identification des informations

---

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

[7] http ://www.ihtsdo.org/snomed-ct/

[8] http ://www.cdc.gov/nchs/icd/icd9cm.htm

conceptuelles expliquant les mappings dans le but de les maintenir. La section 5 présente une évaluation expérimentale de l'approche proposée soulignant la corrélation entre la façon dont les mappings évoluent et les changements affectant les informations conceptuelles identifiées. La section 6 conclut l'article et énonce quelques perspectives.

2 Etat de l'art

Les travaux qui s'intéressent à la maintenance des mappings se distinguent selon la méthode mise en œuvre : *révision des mappings* (Castano, Ferrara, Lorusso, Näth, & Möller, 2008 ; Meilicke, Stuckenschmidt, & Tamilin, 2008), *re-calcul des mappings* (Khattak, Pervez, Latif, & Lee, 2012) ou *adaptation des mappings* (Tang & Tang, 2010 ; Velegrakis, Miller, & Popa, 2004 ; Yu & Popa, 2005).

La révision semi-automatique des mappings vise à identifier et réparer les mappings non valides. Meilicke *et al.* (Meilicke *et al*., 2008) proposent d'utiliser des méthodes formelles s'appuyant sur les logiques de description pour assister l'utilisateur dans la révision des mappings invalides, générés automatiquement par des algorithmes d'alignement. De même, Castano *et al.* (Castano et al., 2008) suggèrent une approche probabiliste pour effectuer la validation des mappings selon la sémantique des ontologies impliquées. Ces techniques peuvent être appliquées pour détecter les mappings non valides après l'évolution de l'ontologie. Cependant, elles requièrent des ontologies très formelles exprimées à l'aide de langages logiques. De plus, les actions de maintenance se limitent à la suppression des mappings non valides.

Le re-calcul complet des mappings ne tient pas compte des informations provenant de l'évolution de l'ontologie ou des mappings existants, et c'est un processus coûteux. Or, lorsque de nouvelles versions de l'ontologie sont fréquemment publiées, les changements sont en général peu nombreux et ne justifient nullement ce re-calcul complet des mappings (Dos Reis *et al.*, 2012). Une approche de re-calcul partiel est proposée par Khattak *et al.* (Khattak *et al.*, 2012). Elle est appliquée aux seuls éléments modifiés dans la nouvelle version de l'ontologie. Des algorithmes d'alignement sont utilisés pour re-créer les mappings entre le sous-ensemble d'éléments modifiés de l'ontologie initiale et les éléments de l'ontologie cible.

L'adaptation des mappings repose sur des approches visant à changer les mappings affectés par l'évolution des ontologies sans, pour autant, effectuer un re-calcul des mappings. Pour cela, les modifications dans les ontologies sont généralement examinées avant de choisir la méthode d'adaptation des mappings la plus appropriée. Les premières propositions dans ce sens sont apparues dans le contexte des mappings entre schémas de base de données (Velegrakis, Miller, & Popa, 2003), avec des actions

d'adaptation pour chaque modification élémentaire du schéma. Une autre approche consiste à représenter les modifications de manière incrémentale (Yu & Popa, 2005) en utilisant les relations qui permettent de passer d'une version d'un schéma à une autre. Tang & Tang (Tang & Tang, 2010) ont proposé une méthode pour faire évoluer une ontologie en minimisant l'impact des changements sur les éléments ontologiques, dans le but de préserver la validité des mappings existants. Ils supposent que seule la suppression des axiomes de l'ontologie peut avoir un impact sur les mappings. De même, Martins & Silva (Martins & Silva, 2009) considèrent que l'évolution des mappings suit la même stratégie d'évolution que celle appliquée à l'ontologie. Plus précisément, lors de la suppression d'un concept, les concepts liés sont rattachés au super-concept du concept supprimé ou à un sous-concept, et le concept supprimé est remplacé par ce super-concept ou ce sous-concept dans les mappings dans lesquels le concept supprimé intervenait. Plus récemment, Groß *et al.* (Groß, Hartung, Thor, & Rahm, 2012) ont effectué une étude sur l'évolution des mappings entre des ontologies des Sciences de la Vie, visant une meilleure compréhension de l'évolution des mappings.

Dans notre approche, nous étendons les travaux cités précédemment. Nous considérons toutes les modifications de l'ontologie, pas seulement les suppressions de concepts. Nous avons défini l'approche DyKOSMap (Dos Reis *et al.*, 2012) dans l'objectif d'adapter les mappings suite à une évolution des ontologies. Elle s'appuie sur des informations conceptuelles décrivant les concepts de l'ontologie jugées suffisantes pour justifier les mappings établis entre les concepts. Ces informations seront notées ICS. Dans cet article, nous proposons une technique originale pour l'identification de ces ICS dans le but d'adapter les mappings de manière (semi-)automatique.

3 L'adaptation des mappings

Nous commençons par introduire les notions fondamentales sur lesquelles nous nous appuyons. Dans nos travaux, une ontologie *O* spécifie une conceptualisation d'un domaine. Nous représentons une ontologie sous la forme d'un triplet *(C, R, A),* où *C* représente un ensemble de concepts c_i ; *R* dénote l'ensemble des relations sémantiques reliant les éléments de *C* entre eux ; *A* représente l'ensemble des attributs caractérisant les concepts de *C*. Chaque concept $c_i \in C$ possède un identifiant unique et est caractérisé par un ensemble d'attributs $A(c_i) = \{a_{i1}, a_{i2}, \ldots, a_{ip}\}$ (*e.g.*, label, synonyme, définition...), où *p* est le nombre d'attributs du concept c_i. Chaque attribut est porteur d'information, *e.g.*, « label » désigne le nom du concept, « définition » précise le sens du concept et le contexte dans lequel il est utilisé. Chaque relation $r_i \in R$ est

un triplet (c_1, c_2, t) avec $c_1, c_2 \in C$ et t la relation sémantique qui lie c_1 à c_2, par exemple, « is_a » ou « part_of ».

On définit le contexte $CT(c_i)$ d'un concept $c_i \in C$ d'une ontologie O comme l'ensemble de ses supers-concepts, de ses sous-concepts et de ses concepts frères. Une des originalités de notre approche consiste en effet à tenir compte des changements éventuels du contexte d'un concept donné lors de l'évolution d'une ontologie. Formellement :

$$CT(c_i) = sup(c_i) \cup sub(c_i) \cup sib(c_i) \text{ avec,}$$

$sup(c_i) = \{c_j | \ c_j \in C, c_i \neq c_j \wedge is_a(c_i, c_j)\}$
$sub(c_i) = \{c_j | \ c_j \in C, c_i \neq c_j \wedge is_a(c_j, c_i)\}$
$sib(ci) = \{c_j \ | \ c_j \in C, c_i \neq c_j \wedge sup(c_j) = sup(c_i) \wedge sup(c_j) \neq \emptyset\}$

Étant donné deux ontologies $O_s=(C_s, R_s, A_s)$et $O_c=(C_c, R_c, A_c)$, nous définissons un mapping m_{st} entre deux concepts $c_{ss} \in C_s$ et $c_{cc} \in C_c$ comme un quadruplet $m_{st} = (c_{ss}, c_{cc}, rel, conf)$, où *rel* symbolise la relation sémantique qui relie c_{ss} et c_{cc} et *conf* représente la valeur de similarité sémantique entre c_{ss} et c_{cc}. Dans nos travaux, nous considérons les types de relation sémantique suivants : *unmappable* [⊥], *equivalent* [≡], *narrow-to-broad* [≤], *broad-to-narrow* [≥] et *overlapped* [≈]. En effet, il s'agit des relations sémantiques des mappings officiels existants entre la nomenclature SCT et la classification CIM fournis par l'IHTSDO[9].

Nous supposons, sans perte de généralité, que les ontologies évoluent à un rythme différent mais jamais de manière simultanée. Ainsi, nous ne considérons que les changements apportés à une version d'une seule ontologie. Étant donné deux versions de la même ontologie source, à savoir O_s^0à l'instant t_0 et O_s^1 à l'instant t_1, et un ensemble initial de mappings M_0 entre O_s^0 et O_c^0 à l'instant t_0, notre objectif est d'adapter les mappings de M_0 pour les faire évoluer vers une nouvelle version M_1 contenant uniquement des mappings valides. La validité correspond à la cohérence logique des mappings. Un mapping $m_{sc} = (c_{ss}, c_{cc}, rel, conf)$ n'est pas valide, par exemple, si le concept c_{ss} (concept source) a été supprimé, ou si la relation *rel* entre les concepts c_{ss} et c_{cc} est devenue inappropriée du fait des changements apportés au contenu dec_{ss}. La Figure 1 illustre le problème d'adaptation auquel nous nous intéressons.

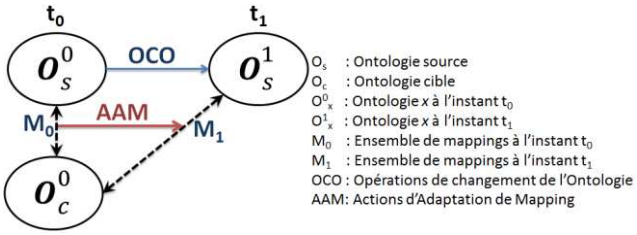

Figure 1. L'adaptation des mappings

[9] http://www.ihtsdo.org/about-ihtsdo/

Le processus d'adaptation des mappings que nous considérons ici s'inscrit dans l'approche générale DyKOSMap décrite dans (Dos Reis *et al.*, 2012). Elle se déroule suivant plusieurs étapes en tenant compte de plusieurs aspects. L'un d'entre eux correspond aux types de changements pouvant affecter les éléments d'une ontologie en considérant principalement les changements complexes (*e.g.* éclatement/fusion de concepts). Un premier travail a montré comment les changements complexes affectant les concepts d'une ontologie peuvent être exploités pour mieux adapter les mappings existants entre des ontologies médicales (Gross, Dos Reis, Hartung, Pruski, & Rahm, 2013).

Même si les mappings sont établis entre les concepts considérés dans leur globalité, une observation empirique et fine des changements complexes au sein d'une ontologie permet cependant de s'apercevoir que les mappings sont établis sur la base d'une similarité sémantique existant entre certaines informations caractérisant le concept cible, le concept source et leur contexte (i.e. les concepts pères, fils et frères). A titre d'exemple, le concept « 560.39 » de la CIM-9 version 2009 est caractérisé par trois attributs dont une des valeurs est « *Fecal impaction* ». 5 mappings lient ce concept à d'autres, dont un établi avec « *Fecal impaction (disorder)* » de la SCT. Après évolution, (*i.e.*, CIM version 2010) ce concept n'est plus décrit par l'attribut « *Fecal impaction* » (*i.e.*, l'attribut a été supprimé) et le mapping mentionné précédemment n'existe plus. Par contre, nous observons la création d'un nouveau concept dont le label est « *Fecal impaction* » dans la version 2010 de la CIM-9, relié par une relation d'équivalence à « *Fecal impaction (disorder)* » de la SCT. Cet exemple illustre l'importance des attributs des concepts pour définir les alignements entre ressources termino-ontologiques du domaine médical.

Suite à ces observations, nous proposons dans la suite de cet article une approche originale, basée sur l'utilisation de mesures de similarité. Elle permet d'identifier les informations conceptuelles sur lesquelles reposent les mappings existants. Cette approche permettra de raffiner les informations à prendre en compte pour la définition des stratégies d'adaptation de mappings intégrée à la plate-forme DyKOSMap.

4 Identification des informations conceptuelles définissant les mappings

Dans ce contexte, nous supposons que les modifications des concepts d'une ontologie peuvent invalider les mappings les concernant. Nous nous proposons de définir un algorithme (voir ci-dessous) permettant d'identifier ce que nous appelons les informations conceptuelles

suffisantes (ICS). Elles correspondent à l'ensemble des attributs sur lesquels est basée la relation sémantique d'un mapping. Si ces informations sont modifiées, elles invalident potentiellement le mapping. Les ICS sont donc importantes pour l'adaptation des mappings.

L'algorithme calcule l'ensemble des *n* attributs les plus pertinents composant les ICS d'un concept c_s donné à travers la fonction $top_A(c_s, c_c, n)$. Dans cette fonction, c_s et c_c sont respectivement les concepts de l'ontologie source et cible. La valeur de *n* est définie empiriquement, mais nous envisageons par la suite de définir une méthode automatique permettant de trouver la meilleure valeur de *n* pour un mapping donné.

L'algorithme proposé s'appuie sur l'utilisation des mesures de similarité couramment implémentées dans les techniques d'alignement d'ontologies, et considère non seulement c_s mais également son contexte $CT(c_s)$. Chaque attribut composant l'ensemble ICS est représenté par le triplet (a_q, s_q, ct_q) qui indique que l'attribut a_q caractérisant un des concepts du contexte ct_q a une similarité s_q avec un attribut du concept c_c. Le niveau de similarité entre les attributs varie entre 0 et 1 (1 représentant une équivalence exacte entre les attributs/concepts), et il est calculé par la fonction $sim(a_p,a_q)$. Trois mesures de similarité différentes ont été utilisées : (1) la mesure de *Levenshtein* évaluant la distance entre deux chaînes de caractères (Yujian & Bo, 2007) ; (2) la distance au niveau des mots (Maedche & Staab, 2002) ; (3) la similarité conceptuelle entre les phrases associée à *WordNet* (Jiang & Conrath, 1997).

```
topA (cs, cc, n)
    Att←∅ //initialisation des attributs de ICS
    Pour chaque ap∈A(cs) faire
          maxSim←0
          Pour chaque aq∈A(cc) faire
                s←sim(ap,aq)
                Att←add(ap,s,cti)
                Si maxSim< s alors
                      maxSim←s
                fsi
          fpour
    fpour
    Si maxSim< 1 alors
          Pour chaque aw∈A(ctj)ctj∈ CT(cs) ;faire
                Pour chaque aq∈A(cc) faire
                      s←sim(aw,aq)
                      Att←add(aw,s,ctj)
                fpour
          fpour
    fsi
    Retourner Trier(Att,n)
```

Algorithme 1–Sélection de *top* n attributs affectant un mapping *particulier*

5 Evaluation expérimentale

Les expérimentations que nous avons menées ont pour objectif de montrer la corrélation qui existe entre les ICS et l'adaptation des mappings entre deux ontologies médicales. Nous cherchons donc à valider notre algorithme d'identification des ICS par rapport à notre besoin d'adaptation des mappings. Dans notre contexte, la corrélation que nous voulons observer s'exprime à travers le nombre de mappings modifiés du fait d'un changement affectant les attributs des concepts sources. Les mappings considérés comme modifiés sont : (i) ceux qui ont été supprimés, ou (ii) ceux dont le concept source a changé ou (iii) ceux dont la relation a été modifiée. Formellement :

$$\#\text{correlation}=||\{m_i^0 \in M_0 | \exists a_k \in A_r(c_S^0), a_k \notin A_r(c_S^1) \wedge m_i^0 \neq m_i^1\}||$$

De plus, ces expérimentations doivent nous permettre de vérifier l'hypothèse de départ, à savoir l'influence des informations contextuelles (*CT*) venant des sous/super-concepts et des concepts frères du concept source sur le comportement des mappings. Afin de mesurer cette corrélation et valider cette hypothèse, nous avons suivi la démarche expérimentale suivante : nous utilisons deux versions de la SCT, celle de janvier 2010 (dénotée SCT10) et celle de janvier 2012 (dénotée SCT12), deux versions de la CIM, celle de 2009 (dénotée CIM09) et celle de 2011 (dénotée CIM11) et les deux versions de mappings officiels établis d'une part entre SCT10 et CIM09 et d'autre part entre SCT12 et CIM11. La première version de mappings contient 84519 mappings tandis que la deuxième version en contient 86638. Dans le but de réduire l'ensemble des mappings de départ à un ensemble de mappings qui ont probablement changé à cause de changements d'attributs, seuls ceux ayant été modifiés ont été considérés, soit un total de 3466 mappings.

Dans cet article, nous nous limitons à l'utilisation exclusive de la mesure de *Levenshtein* normalisée au niveau des caractères (dénotée *Character-based ED*), et à l étude de son impact sur l'identification des ICS. Pour chaque mapping de notre ensemble de départ, nous analysons d'une part les changements, du point de vue lexical, des attributs du concept source (sans tenir compte de son contexte, dénoté *NOCT*), des concepts formant son contexte (concepts parents directs, concepts fils directs, et concepts frères, dénotés respectivement *SUP*, *SUB*, *SIB*) et de la combinaison des deux (dénotée *ALL*), et d'autre part, nous étudions le lien entre ces changements au niveau des concepts et la façon dont les mappings associés évoluent. La Figure 2 présente les résultats obtenus. *#TopA* représente le nombre d'attributs formant l'ensemble ICS et *#Correlation* représente le nombre de mappings affectés par les changements d'attributs d'un concept et/ou de son contexte.

Les résultats montrent que sur les données manipulées, il faut considérer au moins 27 attributs associés à un concept, sans tenir compte

de son contexte, pour expliquer tous les changements de mappings. En général, la plupart des informations nécessaires à l'adaptation des mappings proviennent du concept source lui-même mais les informations relatives au contexte du concept source ont quand même une certaine influence. Les attributs des sous-concepts et des concepts frères ont une influence mineure dans les expérimentations réalisées. Ils n'expliquent que 4 changements de mappings. En revanche, les attributs du super-concept du concept source ont une influence beaucoup plus importante. Ils expliquent 25 mappings modifiés. Les concepts frères du concept source sont moins importants pour l'adaptation des mappings. Ils sont uniquement à l'origine de la modification de trois mappings. Enfin, bien que les attributs du contexte du concept source soient d'une moindre importance que les informations sur l'évolution du concept source en lui-même, la prise en compte de l'ensemble de tous les attributs (*ALL*) permet d'expliquer davantage de mappings.

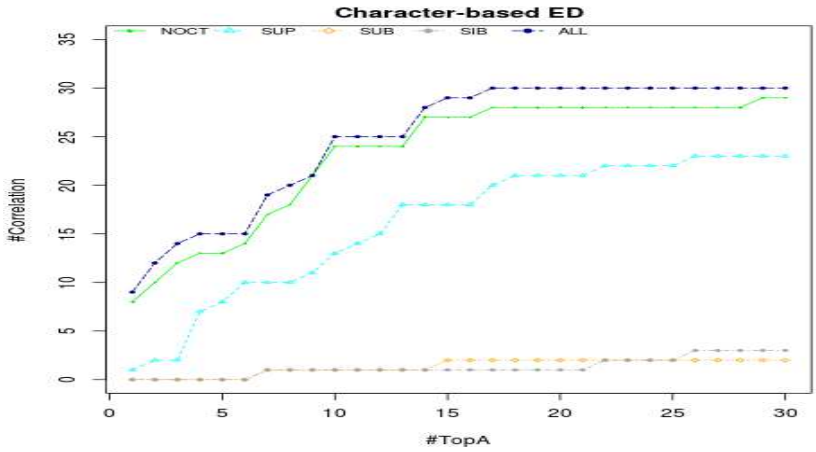

Figure 2- Résultats expérimentaux.

Ces expériences, qui ne considérent que la distance de *Levenshtein*, montrent qu'il est important de prendre en compte la similarité entre attributs de concepts pour définir le comportement des mappings à travers le temps. Dans la suite de nos travaux, nous allons analyser les résultats des expériences menées en calculant deux autres mesures de similarité (la mesure de similarité sur les mots et la mesure de similarité sémantique). Ces analyses nous permettront de déterminer plus précisément l'influence du type de similarité liant les concepts sur les mappings, et donc la façon dont les mappings doivent évoluer.

6 Conclusion

Dans cet article, nous avons proposé une méthode originale pour identifier les informations conceptuelles suffisantes sur lesquelles repose la définition des mappings entre ressources termino-ontologiques du domaine médical. Cette approche s'appuie sur l'utilisation de mesures de similarité entre les informations décrivant le concept source et son contexte, et le concept cible des mappings. Les résultats obtenus ont montré que les changements dans les valeurs des attributs expliquent les changements dans les mappings. Nous avons aussi observé que les modifications affectant le contexte d'un concept source influence les mappings. Les expérimentations supplémentaires, que nous sommes entrain de mener, basées sur le calcul d'autres mesures de similarité, serviront à confirmer ces conclusions et à montrer le rôle du type de similarité entre les attributs sur la définition des mappings existants. Notre objectif est ensuite, d'adapter, en conséquence, le processus d'adaptation (semi-)automatique des mappings entre ressources termino-ontologiques du domaine médical mis en œuvre dans DyKOSMap.

Remerciements

Les travaux présentés dans cet article ont été réalisés dans le cadre du projet DynaMO entièrement financé par le Fonds National de la Recherche du Luxembourg.

Références

CASTANO, S., FERRARA, A., LORUSSO, D., NÄTH, T. H., & MÖLLER, R. (2008). *Mapping validation by probabilistic reasoning*. Paper presented at the Proceedings of the 5th European semantic web conference on The semantic web: research and applications.

DOS REIS, J. C., PRUSKI, C., DA SILVEIRA, M., & REYNAUD-DELAÎTRE, C. (2012). *Analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems*. Paper presented at the Semantic Interoperability in Medical Informatics (SIMI).

GROSS, A., DOS REIS, J. C., HARTUNG, M., PRUSKI, C., & RAHM, E. (2013). *Semi-Automatic Adaptation of Mappings between Life Science Ontologies*. Paper presented at the Data Integration in the Life Sciences (DILS 2013).

GROß, A., HARTUNG, M., THOR, A., & RAHM, E. (2012).*How do computed ontology mappings evolve? - A case study for life science ontologies*. Paper presented at the Joint Workshop on Knowledge Evolution and Ontology Dynamics @ ISWC.

HARTUNG, M., GROSS, A., & RAHM, E. (2012). COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Journal of Biomedical Informatics*.

JIANG, J., & CONRATH, D. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*, Tapei, Taiwan: Academia Sinica.

KHATTAK, A., PERVEZ, Z., LATIF, K., & LEE, S. (2012). Time efficient reconciliation of mappings in dynamic web ontologies. *Knowl.-Based Syst., 35*.

MAEDCHE, A., & STAAB, S. (2002). *Measuring similarity between ontologies*. Paper presented at the Knowledge engineering and knowledge management: Ontologies and the semantic web (EKAW).

MARTINS, N., & SILVA, N. (2009). *A User-driven and a Semantic-based Ontology Mapping Evolution Approach*. Paper presented at the 11th International Conference on Enterprise Information System.

MEILICKE, C., STUCKENSCHMIDT, H., & TAMILIN, A. (2008). Reasoning Support for Mapping Revision. *Journal of Logic and Computation, 19*(5).

TANG, F., & TANG, R. (2010). *Minimizing Influence of Ontology Evolution In Ontology-based Data Access System*. Paper presented at the IEEE International Conference on Progress in Informatics and Computing (PIC).

VELEGRAKIS, Y., MILLER, R. J., & POPA, L. (2003). *Mapping adaptation under evolving schemas*. Paper presented at the Proceedings of the 29th international conference on Very large data bases - Volume 29.

VELEGRAKIS, Y., MILLER, R. J., & POPA, L. (2004). Preserving mapping consistency under schema changes. *The VLDB Journal, 13*(3), 274-293.

YU, C., & POPA, L. (2005). *Semantic Adaptation of Schema Mappings when Schemas Evolve*. Paper presented at the Proceedings of the 31st international conference on Very large data bases.

YUJIAN, L., & BO, L. (2007). A Normalized Levenshtein Distance Metric *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1091-1095.