Characterizing Semantic Mappings Adaptation via Biomedical KOS Evolution: A Case Study Investigating SNOMED CT and ICD

Julio Cesar Dos Reis, MA[1,2], Cédric Pruski, PhD[1], Marcos Da Silveira, PhD[1], Chantal Reynaud-Delaître, PhD[2]
[1]CR SANTEC – Public Research Centre Henri Tudor, Esch-sur-Alzette, Luxembourg
[2]LRI – University of Paris-Sud XI, Orsay, France

Abstract

Mappings established between Knowledge Organization Systems (KOS) increase semantic interoperability between biomedical information systems. However, biomedical knowledge is highly dynamic and changes affecting KOS entities can potentially invalidate part or the totality of existing mappings. Understanding how mappings evolve and what the impacts of KOS evolution on mappings are is therefore crucial for the definition of an automatic approach to maintain mappings valid and up-to-date over time. In this article, we study variations of a specific KOS complex change (split) for two biomedical KOS (SNOMED CT and ICD-9-CM) through a rigorous method of investigation for identifying and refining complex changes, and for selecting representative cases. We empirically analyze and explain their influence on the evolution of associated mappings. Results point out the importance of considering various dimensions of the information described in KOS, like the semantic structure of concepts, the set of relevant information used to define the mappings and the change operations interfering with this set of information.

Introduction

The misunderstanding of shared biomedical information caused by "terminological problems" is a well-known problem that has a great impact on the work of healthcare professionals[1]. Dictionaries have been proposed since years to cope with this problem. However, due to the advances of Informatics, terminological problems are also observed between biomedical information systems, which reduces semantic interoperability between them. Different schemas have been proposed to structure biomedical knowledge and to make it interpretable by computer systems. Classifications, taxonomies, thesauri and ontologies, as part of Knowledge Organization Systems (KOS), are examples of widely employed models that describe biomedical knowledge. Several KOS, such as Gene Ontology (GO), NCI Thesaurus (NCIt), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT®) and International Classification of Diseases (ICD), play an important role in the improvement of the semantic interoperability between biomedical systems.

Due to the overwhelming available data generated in the biomedical domain, KOS have become cornerstones for enabling adequate understanding, management, exploration and integration of these data[2]. Nevertheless, new challenges have been derived from the utilization of KOS, for instance, versioning and mapping maintenance. The former expresses the consequence of the evolution of KOS's content reflecting the dynamics of the biomedical domain. The later challenge, closely linked to versioning issues, is related to the consequence of the evolution of KOS on existing semantic correspondences between them. Since KOS play a critical part in biomedical software applications, it is important to evaluate and characterize the impact of KOS evolution on mappings in order to react, with regard to the occurring KOS changes, in the case of a (partial) invalidation of mappings.

Issues raised by KOS evolution have already been the source of many research efforts in the biomedical domain but also in the Semantic Web community, where ontology evolution is still under investigation. In that sense, the work conducted by Noy & Klein[3] proposes a first categorization of changes that can affect ontologies. This first attempt, aiming at supporting ontology versioning, defines two main categories of changes: atomic and complex changes. The first refers to the changes of only a single specific feature of the ontology model (*e.g.*, concepts, attributes or relationships) while the second denotes changes that are composed of multiple atomic ones. Although promising, and while this work allows the characterization of the ontology evolution, its impact has not been further studied and applied in the context of mappings.

Existing work in mapping maintenance attempts to fully or partially (based on concepts affected by changes) re-calculate the set of mappings[4] or to adapt them after KOS evolution[5,6]. Using KOS changes for maintaining mappings seems to be favorable, and it has been explored mainly for schema changes in the adaptation of database and XML schema mappings[7]. Nevertheless, the influence of KOS evolution on mappings is still not completely understood, and thus how to correlate the way mappings must be adapted according to different complex change behaviors is a real research challenge. Only few recent work in literature[8,9] has investigated this problem and we have proposed mapping adaptation actions to support automatic mapping maintenance[10]. However, existing studies have analyzed mappings evolution mainly considering changes affecting source concepts of mappings in an isolated way. In order to adequately propagate KOS changes on mappings, and thus to provide the appropriate use of changes for maintaining them, it is essential to better precise KOS evolution impact on how mappings change. Therefore, KOS complex changes need to be further investigated, observing in detail which information changes and how to correlate that with the adaptation of the associated mappings.

This article addresses how KOS changes influence the adaptation of mappings by means of different KOS evolution cases analysis. More precisely, we analyze and explain the impact of specific complex changes occurring in a KOS on associated mappings. This investigation is based on the analysis of the evolution of SNOMED CT (SCT) and ICD–9–CM (ICD) over a period of four years. We use various versions of official releases of SCT and ICD with their associated validated mappings in which the cases studied were observed. Previous work conducted on the quantitative analysis of KOS and mappings evolution[11] shows that split is a recurring complex change appearing in biomedical KOS. Consequently, the impact of this particular type of change on mappings deserves a closer attention due to the difficulties involved. For instance, when a concept is split into several ones, it is complicated to correctly adapt the early associated mappings, since there are many ways to change these mappings considering the concepts resulting from the split. This is why we focus on these particular complex changes in this article. Results highlight mainly that considering complex scenarios of change is important to drive how mappings need to be adapted in that context. This study enables a better understanding of mappings evolution in the context of KOS complex changes. This will be the basis of an approach to a (semi-) automatic mapping adaptation mechanism.

The remainder of this article is structured as follows: First, we describe the materials used and the procedure conducted. Afterwards, we present the results analyzing them. The discussion section provides the lessons learned we achieved from the results. Final remarks and future work conclude the article.

Method

In this investigation we aim at identifying and analysing cases of split change operations impacting mappings, in order to study how these impacted mappings are adapted. We considered SCT and ICD, since several versions of these KOS as well as the official mappings between them are available. We used six different KOS versions published between 2009 and 2011: four different versions of SCT and two different versions of ICD. We also selected four official versions of mappings (provided by IHTSDO): between the SCT releases of Jan/2010, Jul/2010 and the ICD release of 2009; and between the SCT releases of Jan/2011 and Jul/2011 and the ICD release of 2010.

We consider a mapping as a triple (*s, t, r*) where *s* is the "source concept" and denotes a concept belonging to the source KOS, *t* is the "target concept" and denotes a concept from the target KOS different from the source, and *r* is a *relation symbol* which represents the type of semantic relation between *s* and *t*. The types of semantic relations considered in this study are based on the ones proposed by IHTSDO: *Unmappable* (⊥) which means that a source concept cannot be linked to a specific target concept; *Equivalent* (=) such that the two concepts are identical or the source concept is listed as an inclusion within the target concept; *Narrow-to-broad* (≤) when the source concept is semantically more specific than the target concept; *Broad-to-narrow* (≥) which is the opposite proposition of narrow-to-broad; and *Partial Overlap* (≈) when there is a relation between the concepts but it is not one of the previous defined relations.

ICD and SCT rely on different knowledge representation models. SCT considers three main entities: *Concept, Description* and *Relationship*. *Concepts* are identified by a unique identifier and have attributes such as name and status. *Descriptions* are independent entities related to concepts and are composed of sets of terms that textually describe the concepts. Their type denotes either a preferred term or a synonym. *Relationships* connect two different concepts. The ICD model is composed of a pre-defined hierarchical structure including *Chapters*, *Blocks* and *Codes*. *Chapters* are the most general level of organization. *Blocks* always belong to a *Chapter*, and *Codes* are identified by a unique numerical identifier and belong to a unique *Block* and *Chapter*. *Codes* also contain attributes such as *title*, *notes*, *includes* and *excludes*. There are no explicit relationships between concepts in ICD, and mappings are always interrelated to the *Codes* level. We use SCT as source and ICD as target one, and thus in the adopted definition of mapping, the source concept is a concept in SCT while the target concept belongs to ICD.

Based on this material we conducted the following four-step procedure for both SCT and ICD in order to investigate cases of split changes impacting on mappings: (1) Automatic identification of complex changes; (2) Refinement of the complex changes previously identified; (3) Selection of representative cases impacting associated mappings and (4) Detailed analysis case-by-case of complex changes correlated to mappings adaptation.

1. Automatic identification of complex changes

This is usually referred as the *diff* calculation problem[12,13]. In this study, we have not used developed tools since they usually require input files in OWL or OBO formats. Since SCT and ICD are not available in these formats we had to identify the split complex changes by implementing a particular process.

We first need to define what is assumed to be a split in this work. Considering that the KOS_{v1} is a new version released of a KOS_{v0} and that the concept $c_1' \in KOS_{v1}$ has the same identifier but at least one attribute changed (*i.e.*, was added, deleted or the value was modified), compared with $c_1 \in KOS_{v0}$ (first situation), or $c_1' \notin KOS_{v1}$ (second situation). Thus, the split of c_1 into a non-empty subset of concepts in KOS_{v1} ($h \subseteq KOS_{v1}$, $h \neq \emptyset$) is identified when for each element $p \in h$, the following conditions are satisfied according to the situation.
For the first situation:

1. There is at least one *common super-concept* between p and c_1', or c_1' is the *super-concept* of p;
2. There is a semantic similarity between p and c_1, greater than a threshold σ.

For the second situation:

1. In the case where p is a modified concept there must exist at least one *common super-concept* between p and c_1, or c_1 is the *super-concept* of *p,* otherwise this condition is not considered;
2. There is a semantic similarity between p and c_1, greater than a threshold σ.

The result of the split is given by $h \cup c_1'$.

Based on this definition we describe the split identification procedure to recognize split change operations between two different versions of a KOS:

- First, we identify all concepts that were affected by the KOS evolution and we group them in the set *K*. We calculate a simple *diff* between all concepts of KOS_{v0} and KOS_{v1}. To this end, we compare the concepts' identifier to determine whether a concept is added or removed. For instance, if the identifier exists in KOS_{v0} and not in KOS_{v1} then we consider that the concept was removed, and the opposite for added concepts. For modified concepts, we compare the content of concepts that exist in both versions. In the set *K* each concept is associated to one type of change (add, remove, modify).
- Second, we filter the set *K* in order to keep only concepts that are related to the existing mappings. Since we are looking for correlations between KOS changes and mappings evolution, we limit the investigation to concepts that are associated to mappings. In other words, we verify whether each concept from the set *K* is the source concept of the mapping (if analyzing SCT) or the target one (for ICD). Concepts that are not related to mappings are excluded from *K*.
- Afterwards we use the given definition of split to identify concepts belonging to *K* that can be involved in a split change operation. We consider each concept from the set *K* and we start by verifying the first condition (super-concepts or siblings). When two concepts from *K* fulfill the first condition (*i.e.*, we find one added or modified concept from *K* that is sub-concept or sibling of another modified or removed concept of *K*), then we verify the second condition (similarity). If the second condition is fulfilled, we group both concepts into a "*pair*". For example, in ICD the concept codes 560.39 and 560.32 (Figure 1) belong to *K* since the former has been modified and the latter is a new added concept. In this case both conditions are fulfilled as these are sibling concepts and there is a similarity between them. We follow this order of verification of the conditions because it is more likely that neighbor concepts are involved in a split than the others, and because this is important for optimization, decreasing the quantity of similarity to be calculated between concepts. The details about the semantic similarity measures are explained later in this section.
- Finally, we analyze the set of pairs found in order to identify those concepts from KOS_{v0} that were split into more than one concept in KOS_{v1}. For this purpose, we group all pairs of similar concepts that share a common concept. In this case, the concept in KOS_{v0} is similar to one or more concept(s) in KOS_{v1} and we certify that the content of this concept has modified or the concept was removed from one version to another. These pairs are associated to one split operation.

The semantic similarity between concepts can be calculated using different techniques[14,15]. In general, the result expresses a weight of how much both concepts are semantically similar. We utilize a hybrid method considering syntactic and semantic information. Let c_{1_v0}, c_{2_v1} be two different concepts in two distinct versions of a KOS. The syntactic part of the method compares values of attributes of both c_{1_v0} and c_{2_v1} as strings using the well-known

Levenshtein distance measure. Concerning semantics, we used *MetaMAP*. If concepts share the same semantic type in UMLS, then c_{1_v0} and c_{2_v1} are considered as similar ones.

Elements used for calculating the similarity are different according to the considered KOS model. ICD and SCT are based on different knowledge models and they do not provide the same type of KOS elements. Concepts in the ICD, for instance, provide textual information such as values of *titles* and of attributes such as *notes, includes* and *excludes*. By contrast, we can explore further descriptions and structural information in SCT. Therefore, we define slightly different strategies to calculate the similarity in ICD and SCT based on the proposed hybrid method.

We calculate the similarity between c_{1_v0} and c_{2_v1} in ICD as follows:

- c_{1_v0} and c_{2_v1} in ICD are considered similar if they have their title attribute considered syntactically or semantically similar, or if they have at least one similar phrase in notes, includes and excludes.
  - We compare the values of the title of c_{1_v0} and c_{2_v1} using both syntactic and semantic methods. If a negative result is found then we try to compare information contained in notes, includes and excludes attributes in both c_{1_v0} and $c_{2\_v1.}$ For instance, a negative result is found comparing the value of the title of the concepts 560.39 ("other") and 560.32 ("fecal impaction"), but when comparing one of the notes of the former with the value of the title of the latter, an exact match is found.
  - We compute the *cartesian* product between these attributes. In this sense, we compare all notes of c_{1_v0} with all notes of $c_{2\_v1.}$ A similar approach is applied for includes and excludes. The value of these attributes is composed of a set of distinct phrases, and each phrase is composed of a set of words. Observing if at least one phrase of c_{1_v0} is similar to a phrase in c_{2_v1} is made using the syntactic method. We compare all sets of phrases from c_{1_v0} to all set of phrases of c_{2_v1} for each type of attributes, searching for a "true" similarity.

We calculate the similarity between c_{1_v0} and c_{2_v1} in SCT as follows:

- In order to consider that c_{1_v0} and c_{2_v1} are two similar concepts in SCT, one of the conditions must be fulfilled in the following order: (1) Syntactic comparison of the name; (2) Semantic comparison of the name; (3) Syntactic comparison of the descriptions; (4) Sematic comparison of the descriptions; and (5) Sharing of same relationships.
- Given two sets of descriptions, one belonging to c_{1_v0} and the other to c_{2_v1} we use the *cartesian* product between both sets in order to compare them based on the syntactic and semantic parts of the method.
- We also consider a similarity between c_{1_v0} and c_{2_v1} based on the relationships associated to these two concepts. For this purpose, the quantity of equal relationships shared between c_{1_v0} and c_{2_v1} is taken into account. Therefore, if the quantity of equal relationships shared between c_{1_v0} and c_{2_v1} is bigger than the half of the total of relationships associated to c_{2_v1} then they are considered similar.

2. <u>Refinement of the previously identified complex changes</u>

We manually refine the identified groups of concepts involved in the split operations. This step is important due to the possible inaccuracy of similarities, and to improve results in a re-organization of splits. In this analysis we might merge groups of concepts that appeared to belong to the same split operation. We might identify false positives groups and remove them. For instance, the case of ICD presented in Figure 3 had been firstly automatically identified as different split cases, and by the manual refinement it was realized they concerned the same split operation. We enrich the information about possible concepts involved in a split in adding, for instance, a new sibling concept that should be involved in a split operation and which was not assigned in the automatic step. For example, the concepts 752.45, 752.46 and 752.47 of ICD in Figure 2 were manually added since it was observed they shared a similarity with the concept 752.49. This step provides several cases of split to be further analysed.

3. <u>Selection of representative cases impacting associated mappings</u>

We associate all mappings with the concepts belonging to cases of split of the latter step. Note that the splits that do not contain associated mappings are not further considered. This reduces the quantity of cases having to be taken into account in this study. In this third step, we analyse the adaptation of mappings in the context of split operations. That is, we observe the type of changes occurred in mappings. For instance, we analyse mappings that are adapted to the resulting concepts of the split, those that are removed or with a modification in the type of their semantic relation. Based on this initial analysis we select the most representative cases for a detailed analysis (next step). For instance, we consider only one case among those containing repeated behaviours. We thus depict the behaviour illustrating a scenario before and after evolution of the selected cases which shows concepts of the splits, and the changes affecting the associated mappings.

4. Detailed analysis case-by-case of complex changes correlated to mappings adaptation

We analyse the final selected cases of ICD and SCT. This consists of observing the types of atomic changes affecting the split concepts. For instance, we observe the value of the attributes shared between the concepts of the splits. We explain the behaviour of the mappings correlating them to the (types of) change(s) affecting the concepts of the split. For instance, we try to understand the modifications of the semantic relations occurred in the mappings. Also, we search for the reasons that led mappings to be adapted toward one or the other concept resulting from the split. We compare the different cases, searching for contrasts between them and how we characterize all that. We also relate differences between the cases of SCT and ICD. The utmost goal in this step is to learn lessons from the selected and analysed cases to have adequate knowledge for designing an automatic mapping adaptation mechanism later.

Results

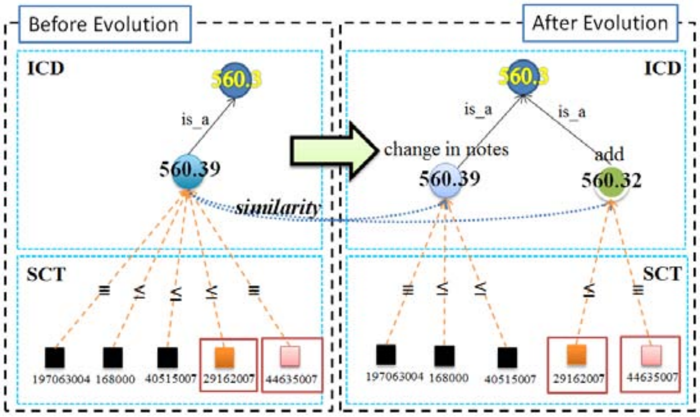

Figure 1: First case of split complex change in ICD

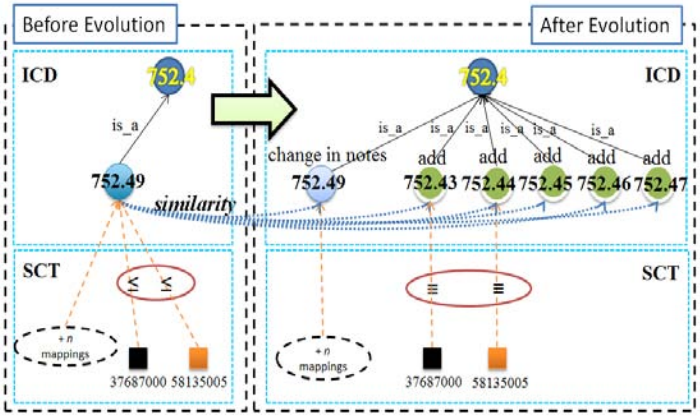

Figure 2: Second case of split complex change in ICD

This protocol allowed the identification of interesting variations of the split change involved in the evolution of mappings. As the split of concepts represents one of the most frequent and also the most complex case of KOS changes, we focus on this particular complex change operation. The significant number of split cases occurring in biomedical KOS is due to the dynamics of this domain. As new knowledge is permanently defined, the domain becomes more precise, forcing general concepts to be refined and consequently split into more fine-grained ones. Existing work on the biomedical KOS evolution provides tools for the identification of complex changes like split or merge of concepts[12], but the present investigation shows the necessity to further refine the definition of these operations in order to better exploit them for the maintenance of mappings. The conducted experiment mainly underlined eight assorted cases of concept splits , of which four affecting ICD and four occurring in SCT, having a different impact on the way mappings evolved over time. As ICD's model is structured according to a single hierarchy and concepts are described using pre-defined attributes, identifying a split of concepts was easier than for SCT. We selected the most representative cases of split from each KOS in order to thoroughly analyze them.

We depict the cases in Figures throughout this section, showing a scenario with concepts and associated mappings before evolution (left part of the figure) and the scenario after evolution with the updated status of concepts and mappings (right part of the figure). Concepts in ICD are represented as circles while those in SCT are represented as squares. Light blue concepts are modified concepts and green concepts (with larger borders) denote new concepts. Mappings are represented as orange arrow lines connecting the concepts between SCT and ICD. Note that the direction of the mappings is always from SCT to ICD. Therefore, a mapping of narrow-to-broad (≤) type means that the SCT concept is more semantically specific than the ICD concept. Blue arrow lines represent a similarity that was found between concepts before and after evolution. Analyzing the four different split cases affecting ICD (see

Figure 1, Figure 2, Figure 3 and Figure 4) and the four ones affecting SCT (see Figure 5, Figure 6, Figure 7 and Figure 8), we observed that no concepts are removed. More generally, the concrete removal of concepts rarely occurs in ICD and never occurs in SCT. Addition of concepts is the most frequent operation since it is more usual and natural that new knowledge is aggregated into the biomedical KOS over the time.

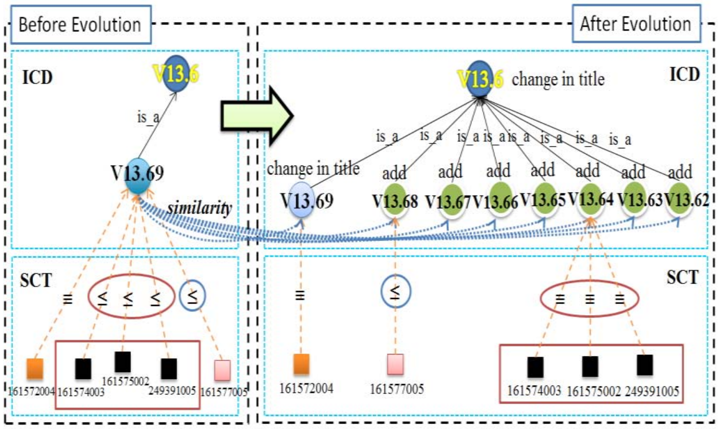

Figure 3: Third case of split complex change in ICD

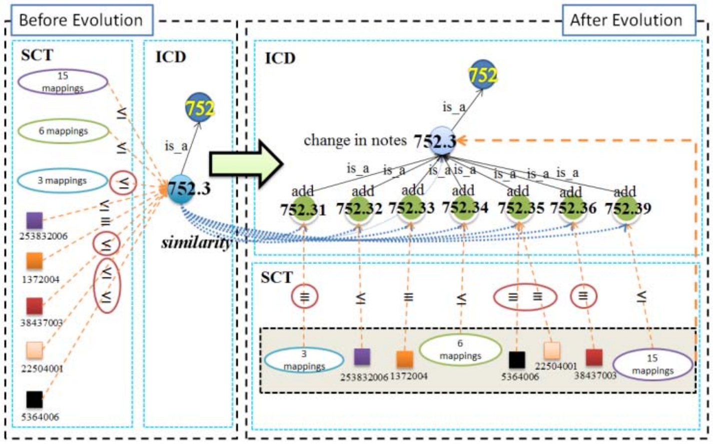

Figure 4: Fourth case of split complex change in ICD

The first case (Figure 1) highlights the split of the concept 560.39. In this case, the most interesting part is the attribute *notes.* Actually, before evolution, it contained three different *values* "Concretion of intestine", "Enterolith" and "Fecal impaction". After evolution, the latter mentioned value was deleted from the notes of the concept 560.39 and became the title of the new concept 560.32. A closer look at the five mappings linking the SCT concepts to the ICD concept 560.39 before evolution reveals that two of them have as SCT concept names "Fecal impaction (disorder)" and "Fecal impaction of colon (disorder)". After KOS evolution, these two mappings are directly moved (without modification of the type of the semantic relation) to the newly created ICD concept 560.32 that has "fecal impaction" as title. This operation means that the mapping has its source or target element changed. This particular case underlines the importance of considering the value of attributes for maintaining mappings valid over time, since mappings follow the flow of information they were attached to. Besides, the three mappings that remain unchanged involved "Enterolith (disorder)", "Typhlolithiasis (disorder)" and "Concretion of intestine (disorder)" of SCT are three names of concept that correspond to unmodified values of concepts attributes in ICD.

The second case (Figure 2) represents a generalization of the split change described in the first case. Note that instead of one new concept in the split there are many of them. We observed that part of the information contained in the *notes* attribute of the initial concept in ICD (*i.e.*, 752.49 before evolution) is distributed over five newly created concepts. After the evolution, the initial concept becomes semantically more general and the created concepts are semantically more specific. More precisely, information about "Absence of cervix" describing the initial concept has been split into two new concepts: 752.43 "Cervical agenesis" and 752.44 "Cervical duplication". These modifications caused the move of two of the existing mappings combined with an adaptation of the type of their semantic relation from (≤) to (=) since the two new concepts are more specific than the initial one. Observe that three new concepts remain without associated mappings after evolution, and *n* mappings associated with 752.49 before evolution remain unchanged. We noticed that these *n* mappings are associated with elements of the concept 752.49 that did not change. Consequently, in a situation where one of such content was deleted the adaptation of mappings could consider the removal of directly affected mappings. Therefore, in the context of a split change, mappings can either remain unchanged, or are moved towards a resulted split concept or are removed. Based on these observations, these strategies of adaptation could be automatically decided according to the content that mappings are associated with, and the flow of information over the concepts belonging to the complex change operation.

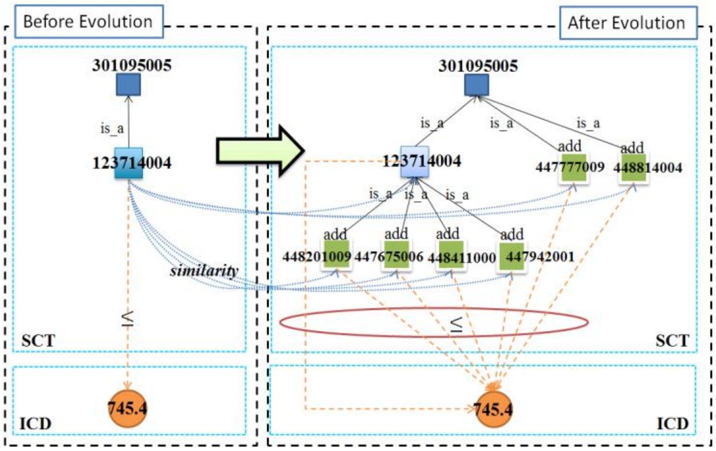

Figure 5: First case of split complex change in SCT

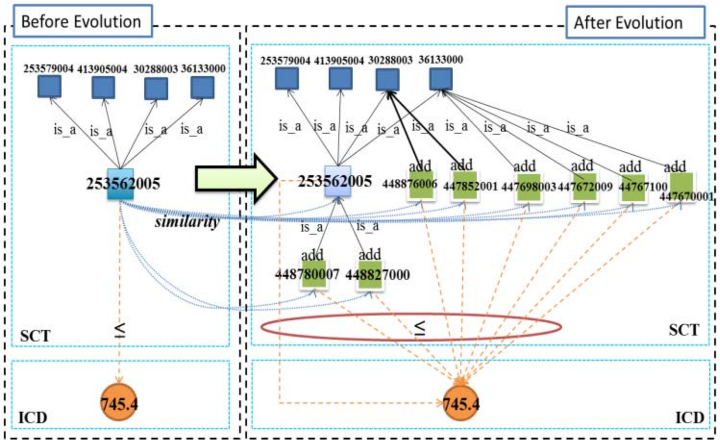

Figure 6: Second case of split complex change in SCT

The third case (Figure 3) is different from the two previous ones since move of mappings may be combined with a change of semantic relation. In the first case we have studied, any moved mappings changed the type of their semantic relation while in the second one, all moved mappings changed their semantic relation. In the third case, there is a mix of them. A potential explanation is that the split change generated new sibling concepts with more semantically specific titles more similar to some (but not all) concepts from SCT, causing the change in the relations. For instance, ICD concept V13.69 had the title changed from "Other congenital malformations" to "Personal history of other (corrected) congenital malformations". After evolution, the concept V13.69 was split into eight concepts. Associated with this split change, we found the new concept V13.68 whose title "Personal history of (corrected) congenital malformations of integument, limbs, and musculoskeletal systems" is mapped, after evolution, to the SCT concept "History of - congenital dislocation – hip" (a similarity between "hip" and the words "limbs" and "musculoskeletal" is found). In this case, the SCT concept is still more semantically specific than the ICD, and the relation between them did not change. However, the new concept V13.64 (also associated with the same split change operation) with the title "Personal history of (corrected) congenital malformations of eye, ear, face and neck" is mapped after evolution to the SCT concept "History of - cleft lip (situation)" (a better similarity between "lip" and the word "face" is found). Consequently, the type of the semantic relation in the adapted mapping needs to reflect this improvement of the similarity, thus the change from narrow-to-broad (≤) to equivalent (=) in the semantic relation. According to the given definition this third case corresponds to a split, but it could also be considered as a specialization of the super-concept V13.6. A closer look at this concrete example raises doubts about the border between a split and a specialization. Analyzing this case, we do not find an explicit transfer of information from one concept to the others belonging to the split, which could better characterize a split. A transfer of information explicitly denotes a content that was deleted from one concept and added into another. Moreover, the super-concept V13.6 also had some lexical modifications in its title (from "Congenital malformations" to "Congenital (corrected) malformations"), but this change does not really affect the semantics of this concept. In other words, the new sub-concepts could be considered as a specialization of this super-concept instead of a split of the concept V13.69. This highlights the intrinsic difficulties in the identification of complex changes and the crucial role played by semantic similarity measures.

The last case in ICD (Figure 4) describes a structural variant of the split change. Unlike in the previously presented cases, the KOS evolution leads to the creation of new sub-concepts that describe a refinement of the initial super-concept. Actually, the description of the super-concept 752.3 has been made more general from the semantic point of view, and new sub-concepts were created based on the information that has been removed from this initial concept. This reinforces the idea of a split since there is an explicit transfer of information from one concept to another. In fact, the title of the new sub-concepts is defined based on the content removed from the *notes* attribute of the concept 752.3. This has impacted the associated mappings by adapting the type of their semantic relations accordingly. In this case, mappings, associated with the super-concept before evolution, are duplicated to the new sub-concepts after evolution (*i.e.*, each new sub-concept has a copy of a sub-set of mappings from the super-concept). Observe that transfer and duplication of mappings are different since in the latter the original mappings are not deleted. Moreover, some of the duplicated mappings are also affected by a change in the type of their semantic relation from a narrow-to-broad (≤) to equivalent (=). However, it is logically inconsistent to have two different concepts from ICD (connected through an "is_a" relationship) associated with a mapping of (=) equivalent type with the same concept (in SCT).

The analysis of split cases occurring in ICD allows the identification of very interesting aspects for maintaining mappings valid over time. This study is enriched with the analysis of four cases of concept splitting that often occur in SCT according to our experiments. As the SCT model is richer than ICD, more possibilities regarding the behavior of the KOS evolution and mappings from the semantic and structural point of views are offered.

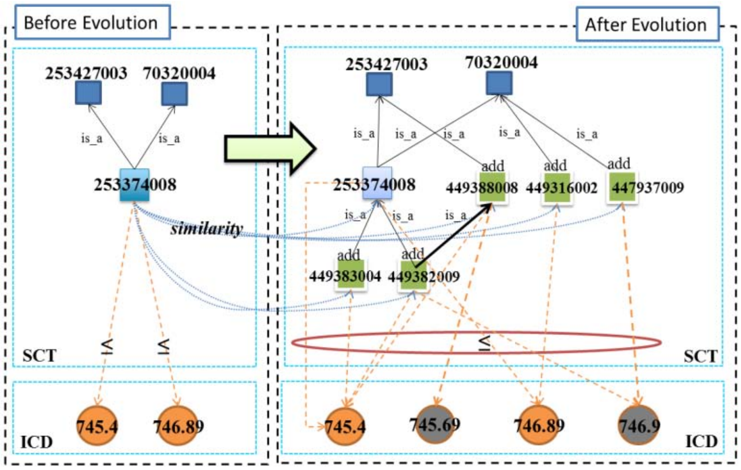

Figure 7: Third case of split complex change in SCT

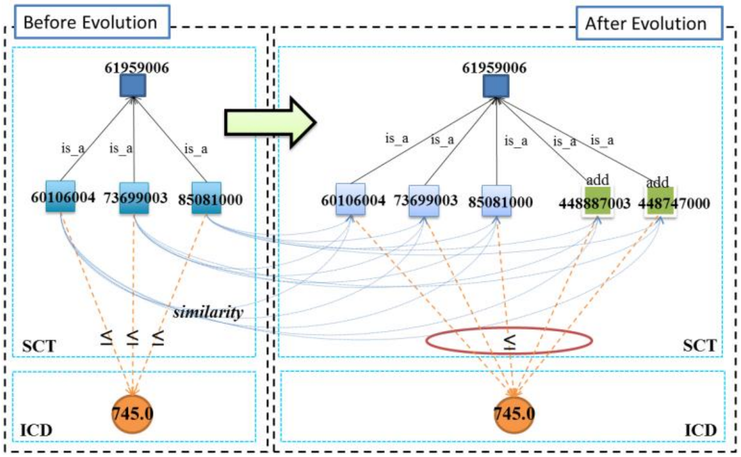

Figure 8: Fourth case of split complex change in SCT

The first and second cases of split we observed are depicted at the Figure 5 and Figure 6. The first case consists in a transfer of information contained in the name and description of the concept "Ventricular septal defect, spontaneous closure (disorder)" (123714004) to newly created concepts, in which some of them are sub-concepts and others are siblings. We observe a duplication of the existing mappings to these new concepts without any modification of their semantic relation (*i.e.*, the relation narrow-to-broad is maintained). From the semantic point of view, it is consistent that new more specific concepts have a duplication of a mapping of (≤) narrow-to-broad type which is not true for other types of semantic mappings. Since the super-concept is more specific than the target concept of the mappings (in ICD), sub-concepts are naturally even more specific than the target concept in ICD. The second case (Figure 6) represents a variant of the first case in the sense that the newly created concepts, siblings of the initial concept, do not all have the same super-concept. However, the adaptation of the affected mappings has the same behaviour as in the first case. A similar observation can be made for the third case (Figure 7), except that new sub-concepts, resulting from the split, can be linked through an "is_a" relationship with other concepts resulting from the split (*i.e.,* one concept can be the super-concept of the other one). Also in this case, new mappings involving concepts in ICD that were not mapped before evolution appear. Since in all three cases mapping adaptation behaves similarly, despite some minor differences in the split of each case, it raises an interesting fact that some aspects of the split cannot deeply affect the evolution of associated mappings, while others are determinant. For instance, new concepts belonging to the split having relationships to distinct super-concepts do not seem to be a determinant factor in the adaptation of mappings.

The last case of split in SCT (Figure 8) shows that information contained in several initial concepts of the split can be assembled into new sibling concepts (*i.e.,* new and initial concepts have the same level of generalization). This can occur without altering existing mappings, but by linking the newly created concepts with the ones already linked with the initial concepts of the split. In this case, the types of the semantic relation are also not modified. We observed that the studied cases in the SCT do not normally impact the modification of the semantic relation of the mappings because SCT is a much more specialized KOS than ICD, *i.e.,* there is a highly difference of granularity between both KOS.

Discussion

The results obtained through the analysis of the identified complex change cases in ICD and SCT, according to the proposed method of investigation, put the stress on various very important aspects for tackling the mapping maintenance problem:

1. Although mappings are established between KOS concepts to put them in correspondence in their entirety, all investigated cases reveal that mappings are defined based on information described partially within the concepts (*e.g.,* concept attributes). Moreover, the way mappings are adapted after the evolution of a KOS is strongly dependent on the modifications affecting this piece of information. For example, if a mapping is established based on a specific attribute of a concept that remains unchanged while other attributes evolve, there is no need to adapt such mapping.
2. It is really important to know which information serves to define semantic correspondences between KOS concepts, and to consider it as an additional (meta-) data of the mappings. To this end, the adopted definition of mappings in this article, but also the well accepted definition of ontology mappings provided by Euzenat & Shvaiko[16], could be enriched. This definition says that a mapping is defined as 5-tuple (*id,* e_1, e_2, *n, r*), where: *id* is a unique identifier of the given correspondence; e_1 and e_2 are elements of two different ontologies; *n* is a confidence measure holding for the correspondence between e_1 and e_2; *r* is the type of semantic relation holding between e_1 and e_2. Considering the obtained results to cope with the mapping maintenance problem, it should be interesting extending this definition adding information on elements that was useful to establish mappings between dynamic KOS.
3. The expressivity of the knowledge representation model of biomedical KOS such as the structural properties of KOS are also important regarding the evolution of KOS and its impact on mappings. Actually, our investigation shows that the re-construction of the KOS at evolution time (*e.g.,* creation of new concepts that can be either siblings, sub-concepts or both) causes different mappings adaptations, such as move or duplication of mappings depending on the type of structural modification affecting the KOS. The modification of the type of the semantic relation is also influenced by this kind of information. This highlights that when adapting mappings according to complex change operations, it is necessary to take into account the structural organization of the involved concepts in the change.
4. Changes interfering in a KOS can modify the semantics of its concepts leading either to a generalization or a specialization of the domain. It forces the re-definition of the semantic relations of mappings, which semantically interrelates concepts of different KOS. Most of the time, the values of concepts' names or attributes are suffering lexical changes. For instance, in Figure 1 the value of the notes attribute "Agenesis of cervix" of the concept 752.49 is transformed into a synonym "Cervical agenesis" which does not really impact the associated mapping. On the contrary, part of the information contained in a concept can be transferred to another concept involved in the complex change. It can make the initial concept semantically more general and, in consequence, mappings that have an equivalent type of semantic relation must change to broad-to-narrow, for instance. These remarks put the stress on the importance of the similarity shared between concepts involved in complex change operations that need to be further studied.

The conducted investigation with the obtained results shows the importance of considering various dimensions like the KOS structure, organizing the underlying concepts involved in complex changes, the semantic of the elements after KOS evolution, and the similarity between resulting concepts of complex changes. As a result, it could be important to consider the notion of change patterns, defined explicitly based on these dimensions, which makes it possible to characterize and formalize complex changes (*e.g.*, all variations of split or merge of concepts). Considering changes as such patterns will be the cornerstone of an approach to (semi-) automatically maintain mappings valid at KOS evolution time, especially for those complex scenarios of KOS changes for which a further understanding of the underlined changes is necessary in order to adequately adapt existing mappings.

Conclusion

Mapping maintenance is a very important research challenge, since KOS have been extensively implemented in a combined way in biomedical software applications. It is thus essential to keep the semantic validity of mappings, as these applications rely on them. Although ontology evolution has been under investigation for a long time, there are no approaches explicitly exploiting information learned from KOS evolution combined with information from existing mappings to tackle the mapping maintenance problem. In this article, we have shown that understanding and characterizing KOS evolution and especially the complex changes affecting KOS elements, taking different aspects of the underlined changes and existing mappings into account, is of utmost importance and shall be explored to adapt the associated mappings. The case of split of concepts, addressed in this investigation, has particularly highlighted that a fine-grained definition of complex changes as possible change patterns can be valuable to support the update of mappings according to KOS evolution. The further definition of change patterns, combined with the definition of heuristics that might drive the adaptation of mappings, will be the source for the development of a (semi-) automatic mechanism for adapting semantic mappings impacted by KOS evolution, which is subject of future research.

Acknowledgements

This research is entirely funded by the National Research Fund (FNR) of Luxembourg under the DynaMO project (grant C10/IS/786147).

References

1. van Bemmel JH, Musen MA. Handbook of Medical Informatics. Springer; 1997.
2. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Briefings in Bioinformatics. 2006;7(3):256-274.
3. Klein M, Noy NF. A Component-Based Framework for Ontology Evolution. Workshop on Ontologies and Distributed Systems at IJCAI-03 Acapulco, Mexico. 2003
4. Khattak A, Pervez Z, Latif K, Sarkar AMJ, Lee S, Lee Y-K. Reconciliation of Ontology Mappings to Support Robust Service Interoperability. IEEE International Conference on Services Computing; 2011
5. Martins N, Silva N. A User-driven and a Semantic-based Ontology Mapping Evolution Approach. 11th International Conference on Enterprise Information System Milano, Italy. 2009 p. 214-221
6. Gross A, Dos Reis JC, Hartung M, Pruski C, Rahm E. Semi-Automatic Adaptation of Mappings between Life Science Ontologies. Data Integration in the Life Sciences (DILS 2013) Montreal, Canada. Springer-Verlag; 2013
7. Velegrakis Y, Miller RJ, Popa L. Preserving mapping consistency under schema changes. The VLDB Journal. 2004;13(3):274-293.
8. Dos Reis JC, Pruski C, Da Silveira M, Reynaud-Delaître C. Analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems. Semantic Interoperability in Medical Informatics (SIMI) Heraklion, Greece. 2012
9. Groß A, Hartung M, Thor A, Rahm E. How do computed ontology mappings evolve? - A case study for life science ontologies. Joint Workshop on Knowledge Evolution and Ontology Dynamics @ ISWC; 2012
10. Dos Reis JC, Dinh D, Pruski C, Da Silveira M, Reynaud-Delaître C. Mapping Adaptation Actions for the Automatic Reconciliation of Dynamic Ontologies. ACM International Conference on Information and Knowledge Management (CIKM 2013) San Francisco. ACM; 2013
11. Dos Reis JC, Pruski C, Da Silveira M, Reynaud-Delaître C. Analyzing the Evolution of Semantic Correspondences between SNOMED CT and ICD-9-CM. Med-e-Tel Luxembourg. 2013
12. Hartung M, Gross A, Rahm E. COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. Journal of Biomedical Informatics. 2012.
13. Noy NF, Musen MA. Promptdiff: a fixed-point algorithm for comparing ontology versions. Eighteenth national conference on Artificial intelligence Edmonton, Alberta, Canada. American Association for Artificial Intelligence; 2002 p. 744--750
14. Agirre E, Alfonseca E, Hall K, Kravalova J, Pasca M, Soroa A. A study on similarity and relatedness using distributional and WordNet-based approaches. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Boulder, Colorado. Association for Computational Linguistics; 2009 p. 19-27
15. Stahl A, Gabel T. Using evolution programs to learn local similarity measures. Proceedings of the 5th international conference on Case-based reasoning: Research and Development Trondheim, Norway. Springer-Verlag; 2003 p. 537-551
16. Euzenat J, Shvaiko P. Ontology Matching. Heidelberg (DE): Springer-Verlag; 2007.