Identifying relevant concept attributes to support mapping maintenance under ontology evolution

Duy Dinh*, Julio Cesar Dos Reis*, Cédric Pruski*, Marcos Da Silveira*, Chantal Reynaud-Delaître

*Public Research Centre Henri Tudor, 6 avenue des Hauts-fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
bLRI, University of Paris-Sud XI, Bât 650, 91405 Orsay Cedex, France

Abstract

The success of distributed and semantic-enabled systems relies on the use of up-to-date ontologies and mappings between them. However, the size, quantity and dynamics of existing ontologies demand a huge maintenance effort pushing towards the development of automatic tools supporting this laborious task. This article proposes a novel method, investigating different types of similarity measures, to identify concepts’ attributes that served to define existing mappings. The obtained experimental results reveal that our proposed method allows to identify the relevant attributes for supporting mapping maintenance, since we found correlations between ontology changes affecting the identified attributes and mapping changes.

Keywords: Ontology mappings, Ontology evolution, Mapping adaptation/maintenance

1. Introduction

The evolution of semantic technologies has led to the development and publication of a huge amount of ontologies, allowing information systems to better describe data and search for relevant information on the Web. Ontologies offer means to make the semantics of data explicit which, in turn, facilitates its exploitation and management. However, mainly for semantic interoperability issues, we need to establish semantic correspondences between ontologies, named mappings, to allow software applications to explore data annotated using various ontologies. The ever increasing number of large ontologies underlines the major role played by mappings [40].

The dynamic nature of domain knowledge induces continuous changes in existing ontologies like adding, removing or modifying ontology elements (e.g., classes, properties, etc.) [22]. These changes impact dependent artefacts such as mappings, making them invalid. In consequence, domain experts must repair affected mappings taking ontology changes into account. This laborious task consists in identifying changes affecting elements of ontologies, and to adapt mappings impacted by these changes accordingly. This process can be performed manually on small ontologies with a restricted number of mappings, but large and highly dynamic ontologies, like those of the life sciences, require appropriate methods and automatic tools.

Existing tools compute mappings between concepts in an (semi-)automatic way to create semantic correspondences between them [1]. Although concepts are considered in their entirety, a closer empirical analysis of mappings reveals that only partial textual statements characterizing concepts are used to define the semantic correspondences [14, 13].

When analysing two consecutive versions of the same ontology, we found cases, for instance, where concepts’ attribute values are completely transferred from one concept to its siblings. This had affected the associated mappings since their definition relies on such textual statement. For example, we observed this case with the concept “560.39” of the ICD-9-CM1 (ICD) biomedical ontology. Such concept contains three attributes and one of them has as value “Fecal impaction” (release 2009). Five

---

1http://www.cdc.gov/nchs/icd/icd9cm.htm
mappings are defined with this concept as domain, and one of these mappings has a range called “Fecal impaction (disorder)”, from SNOMED CT\(^2\) (SCT).

After evolution (i.e., ICD release 2010), the attribute value “Fecal impaction” is no longer associated with the ICD concept and the previously mentioned mapping has been removed. Moreover, the concept “Fecal impaction” has been newly created in ICD (release 2010) and is remapped to “Fecal impaction (disorder)” of SCT. This illustrates the major role played by concepts’ attributes in the definition of mappings [13]. However, when matching systems create a mapping, they fail to keep the ontological entities used to justify such mapping in its definition, preventing thus any future use for maintenance purpose.

In this article, we address this issue by investigating techniques suited to identify textual statements in concepts that might represent the most meaningful attributes for a given correspondence, but lack in its original description. We hypothesize that adequately supporting the mapping adaptation task requires the correct identification of these statements [13, 12]. Our identification method relies on the adaptation of various semantic similarity measures targeting: the lexical level [27], the syntactic level [29] and the semantic [23] level. These measures might support our method to identify a sufficient Subset of Concept Attributes (SCA) relevant for interpreting mappings. We conduct a set of experiments to assess the quality of the results yielded by the identification method using two life science ontologies (SCT and ICD) and their associated mappings. In particular, we measure correlations between ontology changes affecting the identified attributes and adaptation of associated mappings. We further study the stability of the underlying similarity measures analysed.

We structure the remainder of this article as follows: Section 2 presents the related work. Section 3 introduces the preliminaries. Section 4 describes our approach to identifying the set of relevant attributes denoting concepts that can be served to define a mapping between a source and a target concept. Section 5 presents the set of experiments conducted to empirically evaluate the proposed approach. Section 6 discusses the obtained results. Section 7 wraps up with concluding remarks and outlines some directions for future work.

2. Related work

Semantic interoperability among heterogeneous systems has increasingly pushed the efforts on semi-automatic matching approaches to aligning ontologies, by finding correspondences between concepts [16]. Ontology mappings play a key role in the biomedical domain [5] where various efforts such as Biopotal [35] and UMLS [4] aim at supporting interoperability.

The Unified Medical Language System (UMLS) [4], developed by the U.S. National Library of Medicine\(^3\), aims at facilitating the exchange of clinical data and improve retrieval of health information. UMLS includes over than one hundred of ontologies and mappings between them. For example, UMLS also integrates the mappings between SNOMED-CT and ICD9-CM which are created and maintained in collaboration with the IHTSDO organisation\(^4\). These mappings support a transition from the use of legacy ICD9-CM procedure codes to SNOMED-CT\(^5\).

A large number of work have investigated approaches to ontology matching and alignment [36, 34, 41, 43]. Various approaches focus on string-based similarity metrics to establish mappings between ontologies. Some surveys have reported on the performance of these metrics in the course of developing ontology alignment systems [3, 26, 7].

Indeed, despite advancements over the last years on ontology alignment, many issues remain open and represent a real challenge for the Semantic Web community [40]. Unaddressed aspects such as user’s interaction [17] and crowdsourcing [39] have recently gained interest and the mapping maintenance under evolving ontologies remains an open research problem [11]. To the best of our knowledge, only few studies have investigated fully automatic methods to keep ontology mappings semantically valid over time.

We distinguish three main categories of approaches for mapping maintenance. The first one relies on the revision of mappings by identifying and repairing invalid mappings. For example, Meilicke et al. [32] propose an automatic debugging of mappings between expressive ontologies eliminating inconsistencies, caused by erroneous mappings,

---

\(^2\)http://www.ihtsdo.org/snomed-ct

\(^3\)http://www.nlm.nih.gov

\(^4\)http://www.ihtsdo.org/snomed-ct

by means of logical diagnostic reasoning. Similarly, Castano et al. [6] suggest a probabilistic reasoning approach to performing the validation of mappings. These techniques can be applied after ontology evolution to detect invalid mappings. However, they require logically expressive ontologies at a high level of formalization, making this approach unavailable for information systems that rely on semantic resources of low level of formalization such as nomenclatures, thesauri, etc.

The second category performs a full or a partial re-calculation of mappings. While the former fails to consider any information from ontology evolution nor existing mappings, the latter aims at exploiting those information for recreating only mappings that are associated with changed concepts in ontologies. In addition, if large ontologies are frequently released, fully re-calculating mappings becomes less flexible than a partial re-calculation approach because the cost in terms of processing time for re-aligning ontologies still remains too expensive. Khattak et al. [24] propose a partial re-calculation approach, which re-creates only those mappings associated with concepts whose elements have changed. They use matching algorithms to perform a new alignment between changed concepts issued from source ontology and the whole target ontology. However, large ontologies, as in the biomedical domain, so far represent a big challenge for methods relying on mapping calculation [40].

The third category concerns approaches that attempt to adapt semantic mappings in response to ontology evolution. These approaches usually use ontology changes to support mapping adaptation, avoiding to perform calculations for re-aligning ontologies. The first propositions appeared in the context of database schema mappings [44] based on primitive schema changes. Composition of mappings [45] explores mappings between different schema versions for adapting mappings. Concerning ontologies, Tang & Tang [42] propose a method for ontology evolution to find the minimal impact of ontology change propagation. Nevertheless, they assume that only removal of axioms can impact mappings. Martins & Silva [30] propose that evolution of mappings should behave similarly with the strategies applied for ontology evolution, but their method only adapts mappings when concepts are removed from the ontology. More recently, researches have empirically investigated the evolution of life science ontology mappings to understand the mapping evolution phenomenon [20, 13]. On this basis, preliminary investigations have proposed ontology changes-based techniques to adapt mappings [19].

Despite these recent investigations, it still lacks adequate methods to fully perform mapping adaptation in an automatic way according to ontology evolution. Mapping adaptation under evolving ontologies should consider the whole set of possible types of ontology changes rather than only concept removal. We have defined the DyKOSMap framework [15, 12] for handling the adaptation of mappings based on ontology changes and mapping interpretation. We believe that the similarity between concept’s attributes that have been changed constitutes a key factor that we might take into account in automatic mapping adaptation. Beyond the literature, this article proposes a novel technique for identifying relevant attributes that we could exploit for adapting mappings under evolving ontologies.

3. Preliminaries

We first introduce the notions of ontology and mapping and then we define the problem of mapping adaptation. Afterwards, we define the specific problem of identifying relevant concept’s attributes defining mappings, which we deem necessary to interpret established mapping for maintenance purpose.

Definition 1. An ontology $O$ specifies a conceptualization of a domain in terms of concepts, attributes and relationships [21]. In other words, an ontology $O$ consists of a set of concepts interrelated by directed relationships. We define a set of all concepts of an ontology $O$ at time $h$ (denoted as $O^h$) as $\text{Concepts}(O^h) = \{c^h_i | i \in \mathbb{N}\}$. Each concept $c^h_i$ (i.e., concept $c_i$ at time $h$) has a unique identifier and is associated with a set of attributes $A(c^h_i) = \{a^h_i | i \in \{1..p\}\}$ (e.g., label, synonym, definition, etc.), where $a^h_i$ is attribute $a_i$ at time $h$ and $p$ is the number of attributes denoting concept $c^h_i$. Furthermore, each attribute is defined for a particular objective, e.g., “label” for denoting concept names or “definition” for giving the meaning in the context where the concept is used. A relationship between two concepts $c_1$ and $c_2$, denoted as $r(c_1, c_2)$, is defined for interrelating a particular concept and another one in the same ontology, e.g., $r(“cancer”, “brain cancer”) = “is-a”$. Our adopted definition of ontology is motivated by
Medical Subject Headings (MeSH). Each concept in MeSH is denoted by an attribute (preferred term), which can have one or more synonyms (non-preferred terms). Concepts are linked together by “is-a” relationship.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Preferred Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular Diseases [C14]</td>
<td></td>
</tr>
<tr>
<td>Vascular Diseases [C14.907]</td>
<td></td>
</tr>
<tr>
<td>Aneurysm [C14.907.055] +</td>
<td></td>
</tr>
<tr>
<td>Angiodysplasia [C14.907.075] +</td>
<td></td>
</tr>
<tr>
<td>Angiomatosis [C14.907.077] +</td>
<td></td>
</tr>
<tr>
<td>Angioedema [C14.907.079] +</td>
<td></td>
</tr>
<tr>
<td>Aortic Diseases [C14.907.109] +</td>
<td></td>
</tr>
<tr>
<td>Arterial Occlusive Diseases [C14.907.137] +</td>
<td></td>
</tr>
<tr>
<td>Arteriovenous Malformations [C14.907.150] +</td>
<td></td>
</tr>
<tr>
<td>Arteritis [C14.907.184]</td>
<td></td>
</tr>
<tr>
<td>AIDS Arteritis, Central Nervous System [C14.907.184.140]</td>
<td></td>
</tr>
<tr>
<td>Endarteritis [C14.907.184.281]</td>
<td></td>
</tr>
<tr>
<td>Giant Cell Arteritis [C14.907.184.438]</td>
<td></td>
</tr>
<tr>
<td>Polyarteritis Nodosa [C14.907.184.595]</td>
<td></td>
</tr>
<tr>
<td>Takayasu Arteritis [C14.907.184.800]</td>
<td></td>
</tr>
<tr>
<td>Capillary Leak Syndrome [C14.907.218]</td>
<td></td>
</tr>
<tr>
<td>Cerebrovascular Disorders [C14.907.253] +</td>
<td></td>
</tr>
<tr>
<td>Colitis, Ischemic [C14.907.286]</td>
<td></td>
</tr>
</tbody>
</table>

International Classification of Diseases - Clinical Modification (ICD9-CM). One concept has only one or more direct children concepts.

Figure 1: Examples of life science ontologies

Current life science ontologies which are less logically rigorous ones and instances are usually not available. Figure 1 depicts the structure of two large life science ontologies namely Medical Subject Headings (MeSH) and International Classification of Diseases - Clinical Modification (ICD9CM).

Note that in the biomedical domain, ontologies are defined and maintained by different organizations. For this reason, life sciences ontologies vary in terms of formats and structures.

Definition 2. We define the context of a particular concept $c_i \in \text{Concepts}(O)$, denoted as $CT(c_i)$,
as the union of the sets of \( \text{sup}(c_i) \) (direct super concepts), \( \text{sub}(c_i) \) (direct children concepts) and \( \text{sib}(c_i) \) (sibling concepts) of \( c_i \) as following:

\[
CT(c_i) = \text{sup}(c_i) \cup \text{sub}(c_i) \cup \text{sib}(c_i)
\]  

where

\[
\begin{align*}
\text{sup}(c_i) &= \{ c_j | c_j \in C, c_i \neq c_j \land c_i \sqsubset c_j \} \\
\text{sub}(c_i) &= \{ c_j | c_j \in C, c_i \neq c_j \land c_i \sqsupset c_j \} \\
\text{sib}(c_i) &= \{ c_j | c_j \in C, c_i \neq c_j \land \text{sup}(c_j) \cap \text{sup}(c_i) \neq \emptyset \}
\end{align*}
\]  

(1)

In Formula 2, the notation \( c_i \sqsubset c_j \) stands for “\( c_i \) is narrower or more specific than concept of \( c_j \), e.g., “hypotension” is more specific than “vascular disease”. These definitions rely on the “is-a” relationship that forms the hierarchical structure of the ontology. Moreover, this structure consists in a directed acyclic graph, which prevents circular definition of concepts.

The context \( CT(c_i) \) includes concepts in the neighborhood of \( c_i \), i.e., direct parents, direct children and sibling concepts. This excludes concepts linked to \( c_i \) by other relationships than “is-a” relationship. Indeed, in our previous investigation on mapping evolution [14], we pointed out that concepts outside its context have a very low impact on mapping evolution and thus are less relevant for this study.

**Definition 3.** We define the set of attributes in the context of concept \( c_i \) as \( A(CT(c_i)) \). Formally:

\[
A(CT(c_i)) = \bigcup_{j=1..n} A(c_j), c_j \in CT(c_i)
\]  

(3)

We denote the set of all attributes of concept \( c_i \) and the ones in its context, denoted as \( A_{all}(c_i) \), as follows:

\[
A_{all}(c_i) = A(c_i) \cup A(CT(c_i))
\]  

(4)

where \( A(c_i) \) is the set of attributes of concept \( c_i \) and \( A(CT(c_i)) \) refers to the set of attributes from each concept in the context of the concept \( c_i \).

**Definition 4.** We denote \( O_X^h \) as an ontology \( O_X \) (if we do not consider its version) at time \( h \), where \( X \) is the name of the ontology. We consider only the two successive versions of the same ontology, i.e., \( h = 0 \) or \( h = 1 \). Given two ontologies namely \( O_S^h \) and \( O_T^h \), a mapping \( m_{ST}^h \), established at time \( h \), between two concepts \( c_S \in \text{Concepts}(O_S^h) \) and \( c_T \in \text{Concepts}(O_T^h) \) is given by:

\[
m_{ST}^h = (c_S^h, c_T^h, \text{semType}_{ST}^h, \text{confidence}^h)
\]  

(5)

where \( \text{semType}_{ST} \) refers to the semantic relation between \( c_S \) and \( c_T \). The confidence value represents the semantic similarity between \( c_S \) and \( c_T \) (indicating the confidence of their relation [16]). The higher the value is, the more the concepts are related.

We consider the following semantic relations: un-mappable \([\perp]\), equivalent \([\equiv]\), narrow-to-broad \([\leq] \), broad-to-narrow \([\geq] \) and overlapped \([\approx]\). We define \( M_{ST}^h = \{ m_{ST}^h | i \in N \} \) as the set of mappings at time \( h \) between two ontologies \( O_S^h \) and \( O_T^h \).

**Definition 5.** An ontology change operation (OCO) characterizes the evolution of an ontology in terms of operations applied to concepts, relationships and attributes [19, 22]. They are classified into two main categories: atomic and complex changes. Each operation in the former category cannot be divided into smaller operations while each one of the latter category is composed of more than one atomic operations. For instance, the complex operation “change in attribute value” is composed of two atomic operations “delete attribute value” and “add attribute value” [22]. Table 1 and 2 present the possible atomic and complex ontology change operations, respectively.

**Definition 6.** Mapping adaptation problem: Figure 2 depicts the general scenario investigated in this article. We denote \( O_S \) as the source ontology and \( O_T \) is the target ontology in mappings. Since we focus in the evolution of ontologies and mappings, we examine different versions of each ontology. To better study correlations between ontology evolution and mapping evolution, we hypothesize that the target ontology remains unchanged (i.e., \( O_T^0 \equiv O_T^1 \equiv O_T \)) while the source evolves (or vice-versa). If the source ontology evolves, we might adapt the mappings in \( M_{ST}^0 \) accordingly using mapping adaptation actions (MAA) [12], for each source concept impacted by a change. The results of mapping adaptation consist of a set of up-to-date mappings in \( M_{ST}^1 \).

Given two versions of the same source ontology, namely \( O_S^0 \) at time \( h_0 \) and \( O_S^1 \) at time \( h_1 \), we always have at least one target ontology \( O_T \) and an initial set of valid mappings \( M_{ST}^0 \) between \( O_S^0 \) and \( O_T \) at time \( h_0 \). Since this evolution probably impacts the mappings \( M_{ST}^0 \), the necessary mapping adaptation actions are applied to \( M_{ST}^0 \) generating
Change operation | Description
--- | ---
addC(c) | Addition of a new concept $c \in \text{Concepts}(O_X^1)$
delC(c) | Deletion of an existing concept $c \in \text{Concepts}(O_X^1)$
addA(a, c) | Addition of a new attribute $a$ to a concept $c \in \text{Concepts}(O_X^1)$
delA(a, c) | Deletion of an attribute $a$ from a concept $c \in \text{Concepts}(O_X^1)$
addR(r, c1, c2) | Addition of a new relationship $r$ between two concepts $c_1, c_2 \in \text{Concepts}(O_X^1)$
delR(r, c1, c2) | Deletion of an existing relationship $r$ between two concepts $c_1, c_2 \in \text{Concepts}(O_X^1)$

\[ h_0 \quad O_X^0 \quad O_X^0 \quad M_{ST}^0 \quad O_X^1 \quad O_X^1 \quad h_1 \]

- $O_S$ : source ontology
- $O_T$ : target ontology
- $O_X^0$ : ontology $O_X$ at time $h_0$
- $O_X^1$ : ontology $O_X$ at time $h_1$
- $M_{ST}^0$ : set of mappings at time $h_0$
- $M_{ST}^1$ : set of mappings at time $h_1$
- OCO : Ontology Change Operations
- MAA : Mapping Adaptation Actions

Figure 2: Mapping adaptation based on ontology changes

Table 1: Atomic ontology change operations related to ontology $O_X$ [22].

<table>
<thead>
<tr>
<th>Change operation</th>
<th>Description</th>
</tr>
</thead>
</table>
| chgA(c_k, a_i, v_h) | Change of attribute $a_i$ in concept $c_k \in \text{Concepts}(O_X^1)$ with the new value $v_h$
| moveC(c_k, p_1, p_2) | Moving of concept $c_k$ (and its subtree) from concept parent $p_1 \in \text{Concepts}(O_X^1)$ to concept parent $p_2 \in \text{Concepts}(O_X^1)$
| substitute(c_k, c_j) | Replacement of concept $c_k \in \text{Concepts}(O_X^1)$ by concept $c_j \in \text{Concepts}(O_X^1)$
| merge(ζ_u, c_k) | Fusion of a set of multiple concepts $ζ_u \subset \text{Concepts}(O_X^0)$ into concept $c_k \in \text{Concepts}(O_X^1)$
| split(c_k, ζ_u) | Split of concept $c_k \in \text{Concepts}(O_X^0)$ into a set of resulting concepts $ζ_u \subset \text{Concepts}(O_X^1)$
| toObsolete(c_k) | Sets status of concept $c_k$ to obsolete ($c_k$ is no longer active)
| delInnerC(c_k, p_l) | Deletion of concept $c_k$ where $p_l \in \text{sup}(c_k)$ and $\text{sub}(c_k) \neq \emptyset$ from ontology $O_X^0$
| addInnerC(c_k, p_l) | Addition of a sub concept $c_k$ under the concept $p_l \in \text{sup}(c_k)$ to the ontology $O_X^1$
| addLeafC(c_k, p_l) | Addition of leaf concept $c_k$ where $p_l \in \text{sup}(c_k)$ and $\text{sub}(c_k) = \emptyset$ to the ontology $O_X^1$
| revokeObsolete(c_k) | Revokes obsolete status of concept $c_k$ (i.e., $c_k$ becomes active)

Table 2: Complex ontology change operations related to ontology $O_X$ [22].
In this study, we suppose that changes in ontology entities affect any of those identified attributes belonging to a source concept that could be used as relevant concept’s attributes for interpreting mappings associated with this concept. This objective requires to investigate the similarity between attributes in the context of the source concept and the ones belonging to the target concept. We consider the similarity as a criteria for ranking candidates of relevant attributes.

Given a mapping \( m_{ST} \) between two concepts \( c_S \in \text{Concepts}(O_S) \) and \( c_T \in \text{Concepts}(O_T) \), we investigate the task of determining a set of metadata containing a sufficient number of attributes of \( c_S \) or those from its context \( CT(c_S) \). In Section 4, we describe our algorithm for identifying this set of metadata relying on different similarity measures.

### 4. Identifying candidate attributes interpreting mappings

We assume that mappings are established according to the degree of similarity between a subset of attributes, belonging to the interrelated concepts or their context. Mappings are defined based on a set of most similar concepts’ attributes, more precisely, relying on the attributes’ values of concepts involved in mappings. Despite the importance of this subset of textual statements during the mapping adaptation process (cf. Section 1), mappings are released along with the source and target ontologies without these metadata defining mappings.

Given a mapping, this requires to (re-)identify the most relevant attributes of \( c_S \in \text{Concepts}(O_S) \) and/or \( c_T \in \text{Concepts}(O_T) \) that optimize its semantic confidence. For this purpose, we use similarity measures to quantify the semantic relatedness between concepts’ attributes. On this ground, we aim to empirically observe whether an ontology change, affecting any of those identified attributes related to a mapping, leads to the adaptation of such mappings. To examine the correlation between ontology changes and mapping adaptation in this study, we suppose that \( c_T \) remains unchanged while \( c_S \) evolves.

We define \( top_A(c_S, c_T, n) \) as the method to detect the set of top \( n \) attributes that may come from concept \( c_S \) or its context. Thus, its outcome corresponds to the subset of concept’s attributes (SCA) defining a mapping. Each attribute \( a_p \in A(c_S) \) can have a particular similarity with each attribute \( a_q \in A(c_T) \). If the target concept \( c_T \) (all its attributes) remains unchanged and the source concept \( c_S \) (at least one attribute) changes, we denote \((s_{a_p}, ct_{a_p})\) as the couple of parameters related to attribute \( a_p \in A(c_S) \) or \( a_p \in A(CT(c_S)) \). The similarity values \( s_{a_p} \) refers to best similarity of attribute \( a_p \) with the most similar one in concept \( c_T \). The argument \( ct_{a_p} \) accounts for the context where we find the attribute \( a_p \). We set \( ct_p \rightarrow NOCT \) (not in the context) if \( a_p \in A(c_S) \), i.e., attribute \( a_p \) belongs to concept \( c_S \), but not to its context. We compute the similarity value between each attribute \( a_p \in A(CT(c_S)) \) and \( a_q \in A(c_T) \). Figure 3 illustrates a scenario of identifying the concept’s attributes where we highlight a few relevant attributes.

Algorithm 1 presents the designed procedure for selecting the top attributes of the source concept \( c_S \) or its context \( CT(c_S) \). After calculating the similarity between each attribute of \( c_S \) and the ones in \( c_T \) (lines 3-12), the algorithm tries to find the best candidate attribute in \( c_S \). If there is an exact match between any couple of attributes \( a_p \) and \( a_q \), i.e., the similarity \( s_{a_p} \) between them is equal to 1 (line 14), then the algorithm does not identify more attributes in the context \( CT(c_S) \). Otherwise, the algorithm calculates the similarity between each attribute in the context of \( c_S \) and each attribute of \( c_T \) (lines 15-21) to retrieve more attributes from the context \( CT(c_S) \). Finally, the algorithm sorts and returns the top \( n \) attributes.

To examine the adequate similarity measures, we selected and evaluated different well-known measures. We discuss the motivations leading to the chosen measures. Note that the similarity measure is customizable in our algorithm. We consider three different similarity measures: character-based, word-based and semantic-based similarity.

First, we measure the edit-distance between the value of attribute \( a_p \) in concept \( c_S \) or its context (denoted by string \( X \)) and the value of attribute \( a_q \) issued from concept \( c_T \) (denoted by string \( Y \)). Formally:

\[
\begin{align*}
\text{\( X = \text{value}(a_p), a_p \in A(c_S) \lor a_p \in A(CT(c_S)) \)} \\
\text{\( Y = \text{value}(a_q), a_q \in A(c_T) \)}
\end{align*}
\] (6)
Figure 3: Identification of relevant concept attributes defining the mapping between concept $c_S$ and $c_T$. Candidate attributes may come from concept $c_S$ or its context. The most similar candidate attributes constitute SCA, e.g., $top_A(c_S, c_T, 2) = \{(a_{c_Sp}, a_{c_Tp})\}$ because $\text{sim}_{a_{c_S}}$ and $\text{sim}_{a_{c_T}}$ are the best similarity values reflecting the similarity between $c_S$ and $c_T$.

Algorithm 1: Select top $n$ attributes defining a mapping

Input: $m_{ST} = \langle c_S, c_T, \text{semType}_{ST} \rangle \in M^0_{ST}$; $c_S \in \text{Concepts}(O^0_S)$; $c_T \in \text{Concepts}(O^0_T)$, $n \in \mathbb{N}$

Output: $\text{SCA} = \{(s_{a_1}, c_{t_1}), (s_{a_2}, c_{t_2}), \ldots, (s_{a_n}, c_{t_n})\}$

1. $\text{SCA} \leftarrow \emptyset$; {Initialize the final result set}
2. {Compute similarity between attributes in $c_i$ and $c_j$}
3. for all $a_p \in A(c_S)$ do
4. $\text{maxSim} \leftarrow 0$;
5. for all $a_q \in A(c_T)$ do
6. $s_p \leftarrow \text{sim}(a_p, a_q)$;
7. $\text{SCA} \leftarrow \text{SCA} \cup \{(s_{a_p}, \text{NOCT})\}$;
8. if $\text{maxSim} < s_p$ then
9. $\text{maxSim} \leftarrow s_p$;
10. end if
11. end for
12. end for
13. {Select attributes in context if exact matches are not found}        
14. if $\text{maxSim} < 1.0$ then
15. for all $a_k \in A(c_T(c_S))$ do
16. for all $a_q \in A(c_T)$ do
17. $s_k \leftarrow \text{sim}(a_k, a_q)$;
18. $\text{SCA} \leftarrow \text{SCA} \cup \{(s_{a_k}, c_{t_k})\}$;
19. end for
20. end for
21. end if
22. $\text{SCA} \leftarrow \text{sort}(\text{SCA}, n)$; {Select top $n$ attributes}

where $\text{value}(a_p)$ (resp. $\text{value}(a_q)$) is the value of attribute $a_p$ (resp. $a_q$). For example, given two attribute’s values $X$—“tracheal stenosis following tracheostomy” and $Y$—“tracheal stenosis due to tracheostomy”, the similarity function must express to which extent these values are similar or different in terms of lexical similarity. The more similar they are, the more they are considered as related in the sense that we can account a semantic relation between the two underlying concepts. For this purpose, we use the well-known string edit-distance measure, but at the level of characters [27] (Section 4.1).

Second, the similarity function must also take into account the level of words in addition to the differences at the level of characters. For example, we must consider the following attribute’s values “skin cancer” and “cancer of the skin” as equivalent ones. Indeed, the edit distance at the level of characters fails to allow coping with the word order issue in those terms. Therefore, we should consider the distance at the level of words in order to provide a summarizing figure for the lexical level [29] (Section 4.2).

Third, we also evaluate the impact of using semantic similarity for identifying relevant attributes (i.e., SCA) related to each mapping, by considering the semantic information in common between the attributes’ value. This refers to the conceptual similarity between strings, to which Resnik [37] proposed to determine by using the information content approach. For example, the two following attribute’s values “bone of the extremity” and “limb bone” semantically refer to the same concept and shall be considered as equivalent because “extremity” and “limb” are semantically related. To compute this semantic distance, we further need an external lexical knowledge resource, e.g., WordNet [18] and an annotated corpus such as SemCor [33] (Section 4.3).

4.1. Character-based edit-distance similarity

Many related applications require to determine the similarity between two strings, such as pattern recognition, information retrieval, ontology alignment, etc. A widely-used notion of string similarity stands for the edit distance or the Levenshtein [27].
It corresponds to the minimum number of edit operations (insertion, deletion, and substitution of individual symbols) required to transform one string into the other. Mathematically, the string-based or character-based edit distance between two strings \(X\) and \(Y\), denoted as \(ED(X, Y)\), can be defined as the minimum weight of transformation through a sequence of weighted edit operations.

A normalised \(ED(X, Y)\) aims at ensuring that it provides values in the interval \([0, 1]\). We adopt the normalisation of the character-based edit distance similarity measure between two strings \(X\) and \(Y\), denoted as \(sim_{CED}(X,Y)\), defined as follows:

\[
sim_{CED}(X,Y) = \frac{2 \cdot ED(X, Y)}{\|X\| + \|Y\| + ED(X, Y)} \quad (7)
\]

The advantage of such normalisation relies on the fact that it allows to measure the similarity between strings using a genuine metric [31] with a degree of similarity between 0 and 1, where 1 stands for a perfect/exact match. We formally define the edit distance between two strings \(X\) and \(Y\) as follows:

\[
ED(X,Y) = \min \{ \gamma(T_{X,Y}) \} \quad (8)
\]

where

- \(T_{X,Y} = T_1T_2...T_n\) is an edit path, i.e. the sequence of atomic edit operations transforming \(X\) into \(Y\);
- \(\gamma(T_{X,Y}) = \sum_{i=1}^{n} \gamma(T_i)\) corresponds to the cost function of an edit transformation \(T_{X,Y}\) from \(X\) to \(Y\), which is the sum of the individual cost of each edit operation \(\gamma(T_i)\). More details about the estimation of this cost function can be found in [31].

An atomic edit operation \(T_i\) can be among the three following ones: insertion, deletion and substitution. Formally:

\[
\begin{align*}
\lambda \rightarrow a & \quad (\forall a \in \Omega) : & \text{insertion} \\
 a \rightarrow \lambda & \quad (\forall a \in \Omega) : & \text{deletion} \\
 a \rightarrow b & \quad (\forall a,b \in \Omega) : & \text{substitution}
\end{align*} 
\]

where

- \(\Omega\) is a finite set of characters or symbols, e.g. \(\Omega = \{a, b, c, ..., 1, 2, 3, ...\}\).
- \(a, b\) are symbols are characters belonging to \(\Omega\).

\(\lambda\) is the null symbol or character.

\(ED(X,Y)\) becomes a metric if the following conditions are satisfied:

\[
\begin{align*}
\forall a,b \in \Omega, \quad & \gamma(a \rightarrow a) = 0 \\
 & \gamma(a \rightarrow b) > 0 \quad \text{if} \quad ((a \neq b) \land (\gamma(a \rightarrow b) = \gamma(b \rightarrow a)))
\end{align*} 
\]

(10)

4.2. Word-based edit-distance similarity

The character-based edit distance (cf. Section 4.1) provides an effective way to compute the semantic relatedness between similar strings, especially for single-word strings. However, it fails to take into account the syntactic information in compound terms. For example, it results in a very low degree of similarity between two synonymous terms, but with a different word order, e.g., “cancer of the skin in the face” vs. “face skin cancer”.

To cope with this issue, we examine another similarity function that determines the extent to which the symbols in each substring \(W_i\) (separated by a set of delimiters, e.g., a white space) in the first string \(X\) are similar to the symbols in each substring \(W_j\) of the second string \(Y\). The underlying assumption states that: if the first string contains all substrings of the other, it is likely that they are semantically related [29]. Originally proposed in [29], we adapt the word-based edit-distance similarity measure, by extending the character-based similarity. This enables us to evaluate the performance of our algorithm 1, by comparing these similarity measures that underlying the algorithm. We define the word-based edit-distance similarity measure (namely syntactic measure) between two strings \(X\) and \(Y\), denoted \(sim_{WED}(X,Y)\), as follows:

\[
sim_{WED}(X,Y) = \frac{1}{\|L_X\|} \sum_{W_i \in L_X} \max_{W_j \in L_Y} sim_{CED}(W_i,W_j) 
\]

(11)

where

- \(sim_{CED}(W_i,W_j)\) refers to the normalised character-based edit-distance similarity between single words (substrings or tokens) in \(W_i\) and \(W_j\) in strings \(X\) and \(Y\) respectively (cf. Formula 10);
- \(L_X\) and \(L_Y\) are the sets of words in \(X\) and \(Y\), respectively;
• $\|L_X\|$ is the length of string $X$ in terms of tokens.

Note that in Formula 10, the normalisation by $\|L_x\|$ aims to ensure that $sim_{WED}(X, Y)$ provides values in the interval $[0, 1]$. The max function ensures that $W_i$ and $W_j$ constitute the couple of most similar strings in $X$ and $Y$.

### 4.3. Knowledge-based similarity

Using lexical knowledge resources to compute the similarity can help to determine the semantic relatedness in a semantic network, constructed by a taxonomy of concepts or classes. Combining corpus statistical information can enhance the resources, because the computational evidence derived from a distributional analysis of corpus data can better quantify the semantic distance between nodes in the semantic network. We propose to adapt the original Jiang-Conrath similarity [23] measure, by normalizing the similarity with the length of the source string, so that the similarity values stay between 0 and 1. We define the semantic similarity between two strings $X$ and $Y$, denoted as $sim_{SEM}(X, Y)$, as follows:

$$sim_{SEM}(X, Y) = \frac{1}{\|L_X\|} \sum_{W_i \in L_X} \max_{W_j \in L_Y} sim_{JCN}(W_i, W_j)$$  \hspace{1cm} (12)

where $sim_{JCN}(W_i, W_j)$ stands for the Jiang & Conrath’s knowledge based similarity between two words $W_i$ and $W_j$ referring to two concepts in the same ontology. Formally:

$$sim_{JCN}(W_i, W_j) = \max\{IC(W_i) + IC(W_j) - 2 \cdot IC(lcs(W_i, W_j))]^{-1}\}$$  \hspace{1cm} (13)

where

- $IC(W_k)$ refers to the information content of word $W_k$ denoting a particular concept in the semantic network, computed as follows:

  $$IC(W_k) = -\log P(W_k)$$  \hspace{1cm} (14)

where $P(W_k)$ calculates the probability of encountering an instance of a concept denoted by the single-word term $W_k$ in a corpus [23].

- $lcs(W_i, W_j)$ stands for the lowest common subsumer, i.e., the lowest node or concept in the hierarchy subsuming both concepts denoted by $W_i$ and $W_j$.

The calculation of the information content in Formula 14 requires a large corpus [38]. In the biomedical domain, there exists an annotated corpus namely GENIA corpus [25]. However, the semantic annotation is only limited to entities of interest in molecular biology such as proteins, genes and cells constituting the GENIA ontology. In our experiments, we accept SemCor as the underlying corpus because it has been widely used for computing the similarity measures based on external resources [2, 28]. SemCor consists of a text corpus which has been semantically annotated with information about Part-Of-Speech (i.e., noun, verb, adjective and adverb), lemma and WordNet synset [33]. This is composed of 352 texts and includes more than 200,000 sense-tagged words.

### 5. Experimental evaluation

This section presents the set of experiments conducted to empirically evaluate the proposed approach to identifying concept’s attributes defining mappings. We compare the performance of similarity measures by exploiting the contextual information of interrelated concepts, i.e., attributes denoting concepts in the neighborhood of a concept involved in mapping. We believe that studying the evolution at the level of attributes provides an evidence for better supporting the mapping adaptation. Therefore, we study the impact of the different similarity measures (cf. Sections 4.1, 4.2, 4.3) considered through the correlation between changes in mappings (MAAs [12]) and modifications in candidate attributes (OCOs [22]).

We performed a series of experiments to achieve these objectives. First, we describe the used dataset and the methodology (Section 5.1). Then, we evaluate to which extent changes affecting relevant attributes, susceptible for defining mappings, influence their adaptation (Section 5.2). Furthermore, we analyse the stability of the examined similarity measures (Section 5.3).
### 5.1. Experimental dataset and methodology

The conducted evaluation consider two ontologies of the biomedical domain namely SNOMED-CT (SCT)\(^7\) and ICD-9-CM (ICD)\(^8\) and the official mappings between them\(^9\). Indeed, mappings between SCT and ICD are provided for each release of SCT by the IHTSDO organisation. We use the following versions of those ontologies: SCT released in January 2010 and in January 2012, and ICD released in 2009 and in 2011. Therefore, our experiments are based on the two sets of mappings between SCT and ICD, which have been established by experts at IHTSDO: \(M^0_{ST} = (SCT \text{ Jan. 2010, ICD 2009})\) and \(M^1_{ST} = (SCT \text{ Jan. 2012, ICD 2011})\). Table 3 presents some statistics about the dataset.

In the conducted experiments, we focus on studying the evolution of both ontologies; in particular on source concepts involved in mappings. We conducted the following experimental procedure to select the adequate set of mappings:

1. Based on the two releases of the same ontology, we identify the set of concepts that have changed from one version to another. We name this subset of concepts as \(C_{diff}(O^0, O^1)\) that is obtained using the Conto-Diff tool [22] (cf. Table 4). We are particularly interested in exploiting the two OCOs namely \(delA\) and \(chgA\) because they directly impact on the values of attributes that can serve for defining mappings. Therefore, although Table 4 shows an overview about ontology changes, in our experiments, we only consider \(delA\) and \(chgA\) as relevant for detecting changes in concept’s attributes values.

2. From the \(C_{diff}(O^0, O^1)\), we remove all unassociated concepts with a mapping, since we consider that only concepts associated with mappings can impact on the validity of mappings.

3. We further remove from the \(C_{diff}(O^0, O^1)\) added and removed concepts. The correlations between these two change operations and mapping adaptation was already studied in our previous work [14]. We focus on analysing mapping adaptation particularly on those mappings associated only with concepts that had

---

\(^7\)http://www.ihtsdo.org/snomed-ct  
\(^8\)http://www.cdc.gov/nchs/icd/icd9cm.htm  
\(^9\)http://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>(M^0_{ST})</th>
<th>(M^1_{ST})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr. of concepts</td>
<td>390,022 (\uparrow) 12,734</td>
<td>395,346 (\uparrow) 13,059</td>
</tr>
<tr>
<td>Nr. of relationships</td>
<td>530,433 (\uparrow) 11,619</td>
<td>567,719 (\uparrow) 11,962</td>
</tr>
<tr>
<td>Nr. of attributes</td>
<td>1,547,855 (\uparrow) 34,046</td>
<td>1,570,504 (\uparrow) 34,944</td>
</tr>
<tr>
<td>Nr. of mappings</td>
<td>84,519</td>
<td>86,638</td>
</tr>
</tbody>
</table>

Table 3: Statistics about the dataset including ontologies and associated mappings

<table>
<thead>
<tr>
<th>OCO</th>
<th>(O^0)</th>
<th>(O^1)</th>
<th>(O^0)</th>
<th>(O^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>merge</td>
<td>0</td>
<td>0</td>
<td>3,649</td>
<td>140</td>
</tr>
<tr>
<td>split</td>
<td>134</td>
<td>45</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>substitute</td>
<td>0</td>
<td>0</td>
<td>4,327</td>
<td>110</td>
</tr>
<tr>
<td>toObsolete</td>
<td>794</td>
<td>0</td>
<td>delR</td>
<td>7,210</td>
</tr>
<tr>
<td>revokeObsolete</td>
<td>20</td>
<td>0</td>
<td>move</td>
<td>7,140</td>
</tr>
<tr>
<td>delInner</td>
<td>0</td>
<td>0</td>
<td>addA</td>
<td>7,720</td>
</tr>
<tr>
<td>delLeaf</td>
<td>2</td>
<td>0</td>
<td>delA</td>
<td>4,003</td>
</tr>
<tr>
<td>delSubGraph</td>
<td>0</td>
<td>0</td>
<td>chgA</td>
<td>950</td>
</tr>
<tr>
<td>addInner</td>
<td>1,348</td>
<td>26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Ontology change operations (OCOs) identified between the two versions of the studied ontologies. Table 1 and 2 provide more details about the description of each OCO. We consider \(delA\) and \(chgA\) as the most relevant OCOs for identifying changes in concept’s attributes values.
their contents somehow modified, i.e., affected by the change operations concerning addition and removal of attributes.

4. Finally, to avoid misunderstandings on the results, we consider only those mappings where the target concepts remained unchanged from one version to another. Therefore, we removed from the analysis mappings where source and target concepts simultaneously change. This result in a final subset of modified concepts from $C_{diff}(O^0, O^1)$ containing mappings associated with them. These mappings are represented by the set $M_{diff}$.

We conduct this experimental procedure with both ontologies in an isolated way. We present the achieved results in Sections 5.2 and 5.3. In the following of this article, we only refer to the set of mappings $M_{diff}$ associated with the concepts of the calculated and filtered $C_{diff}(O^0, O^1)$ for both ontologies. We analyze a total number of 6,672 mappings in the resulted set of mappings regarding SCT and 3,788 for ICD, respectively. Note that no intersection exists between these sets of mappings.

5.2. Evaluating the correlation between changes in SCA identified and changes in mappings

We aim to examine the way that we can adapt mappings under evolving ontologies, by considering changes in the sufficient Subset of Concept Attributes (SCA). For this purpose, we measure the correlation between changes detected in identified attributes (i.e., delA, chgA) and the adaptation of the associated mapping. We also assume as a correlation when both the identified attributes and the associated mapping remain unchanged. The objectives of our experiments are three-fold:

1. We evaluate the utility of using different similarity measures to identify relevant attributes (SCA) for supporting mapping adaptation.
2. We evaluate the impact of using the ontology context (cf. Formula 2) of source concepts on the yielded correlations.
3. We compare, with respect to the studied task, the quality and the stability of each similarity measure (Character-based ED, Word-based ED and Knowledge-based Similarity).

The experiments may allow us to empirically observe various aspects: (1) the quality of the identified SCA by the proposed technique (cf. Algorithm 1); (2) a comparison among the three types of similarity measures and the two ontologies of the experimental dataset; and most importantly (3) to which extent mapping adaptation should rely on the identification of relevant attributes (SCA).

**Evaluation method.** We calculate the different types of correlations between changes affecting the attributes in SCA and mapping adaptation. To this end, we count the number of mappings that changed when at least one attribute of SCA has changed (i.e., its value has been modified or deleted), in addition to the number of mappings that remained unchanged when particularly any attribute of SCA changed. Note that the source concept of all the analysed mappings belongs to the subset $C_{diff}(O^0, O^1)$, which means that at least one attribute of such concept has been changed. However, this attribute may belong to SCA or not, which justifies the different types of correlations we propose. We characterize a mapping change when a mapping is removed from one release to another, or when the source concept is replaced by another one, or either the semantic relation in the mapping is modified. Formally, we define the different types of correlations as follows:

\[
\#\text{change-correlation} = \|\{m^0_{ST}, m^0_{ST}, a_k \notin A_{diff}(c^1_S)\}\|/\|M^0_{ST}\| \quad (15)
\]

\[
\#\text{unchange-correlation} = \|\{m^1_{ST}, m^1_{ST}, a_k \in A_{diff}(c^1_S)\}\|/\|M^1_{ST}\| \quad (16)
\]

\[
\#\text{total-correlation} = \#\text{change-correlation} + \#\text{unchange-correlation} \quad (17)
\]

The change-correlation in Formula 15 corresponds to the percentage of adapted mappings that have at least one modified attribute in SCA. Similarly, the unchange-correlation in Formula 16 corresponds to the percentage of unchanged mappings because no changes occurred in attributes identified in SCA. Finally, the total-correlation in Formula 17 calculates the sum of the two previous types of correlation.
For each one of the considered ontologies, we first compute the total-correlation values for mappings of $m_{ST} \in M_{aff}$ without using the context of source concepts (denoted as NOCT). Afterwards, we measure the correlations obtained by using solely the different concepts from the context (denoted as SUP for super concepts, SUB for sub concepts, and SIB for sibling concepts). For the unchange-correlation, we measure the context influence considering the retrieved attribute with the highest similarity value. For instance, if our Algorithm 1 finds such attribute in a super concept, then we compute the correlation for this type of concept. We compare the contributions of the different types of context (SUP, SUB, SIB) separately, and with the obtained results without using the context (NOCT). Finally, we combine all conceptual information together to build the SCA as a whole (denoted as ALL). This allows to evaluate the influence of the context for supporting mapping adaptation.

To have an overview, we first present the results concerning the analysis of total-correlation (Figure 4 and Figure 5). Afterwards, we present the results with respect to a more detailed analysis on the change-correlation (Figure 6 and Figure 7). The size of the set of relevant attributes in SCA examined (denoted as $\#topA$) refers to a parameter taking integer values ranging from 1 to 10.

**Results.** Figure 4 depicts the results obtained by computing the total-correlation (cf. Formula 17) with respect to SCT and Figure 5 presents the results for ICD. More specifically, Figure 4.1 and Figure 5.1 present the results without considering the context (NOCT) by comparing the different similarity measures. It allows to point out three main aspects:

- First, the similarity measures perform differently for SCT and ICD. If we do not consider the information about the context of mapped concepts (without CT, cf. Figure 4.1), we observe that the Character-based ED outperforms the Word-based ED and Knowledge-based similarity for SCT while we observe the inverse findings for ICD (cf. Figure 5.1).

- Second, if we consider the parents of mapped concepts (SUP), the Knowledge-based similarity significantly outperforms the two other similarity measures with an improvement rate of $\sim9\%$ for SCT while no improvement is observed for ICD. In addition, children (SUB) and sibling (SIB) concepts are not useful for identifying SCA. This indicates the usefulness of considering the context, more precisely the super concepts of mapped concepts in the mapping adaptation process.

- Third, the percentage of found correlations with respect to the number of analysed mappings differs comparing SCT ($\sim25\%$ without context and $\sim90\%$ with context in addition) and ICD ($\sim3\%$ without context and $\sim5\%$ with context in addition). This observation is coherent with the statistics about SCT and ICD. Indeed, SCT stands for a much larger ontology than ICD in terms of number of attributes and concepts, number of associated mappings (cf. Table 3) as well as in numbers of ontology change operations (cf. Table 4).

The results also indicate several interesting facts by analysing the performance of the similarity measures, considering the different concepts retrieved from the context for the source concepts in mappings. Figure 4.2 and Figure 5.2 show that by considering ALL attributes in mapped concepts and their context, we obtain the best correlation counts both for SCT and ICD. These results also show that while the super concepts play a major role in SCT (cf. Figure 4.2), in ICD the correlations regarding the sibling concepts dominated, but they still remain less expressive than correlations found only considering the source concept (NOCT) for ICD (cf. Figure 5.2).

At this level, we aim to observe the results only for the change-correlation in SCT (cf. Figure 6) and ICD (cf. Figure 7), respectively. We skip the unchange-correlation analysis because this seems a more trivial scenario for us that, whether attributes denoting a mapped concept remain unchanged, the associated mappings might also remain unchanged. Results illustrate that the Word-based ED similarity performs better than the Character-based ED and the Knowledge-based similarity measure for SCT, while the Character-based ED similarity performs better than the other similarity measures for ICD. Both lexical and syntactic measures outperform the Knowledge-based similarity. The latter is mainly based on the similarity between senses of concepts in WordNet, and the exploitation of the term co-occurrence in the SemCor corpus. The
### 1) Similarity measures (without CT)

<table>
<thead>
<tr>
<th>Character based ED</th>
<th>Knowledge based Similarity</th>
<th>Word based ED</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>#topA</th>
<th>#total-correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15 20 25 30 35 40</td>
</tr>
</tbody>
</table>

### 2) Character-based ED (with CT)

<table>
<thead>
<tr>
<th>NOCT SUP SUB SIB ALL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>#topA</th>
<th>#total-correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15 20 25 30 35 40</td>
</tr>
</tbody>
</table>

### 3) Knowledge-based Similarity (with CT)

<table>
<thead>
<tr>
<th>NOCT SUP SUB SIB ALL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>#topA</th>
<th>#total-correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15 20 25 30 35 40</td>
</tr>
</tbody>
</table>

### 4) Word-based ED (with CT)

<table>
<thead>
<tr>
<th>NOCT SUP SUB SIB ALL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>#topA</th>
<th>#total-correlation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15 20 25 30 35 40</td>
</tr>
</tbody>
</table>

Figure 4: Performance of our Algorithm 1 using three different similarity measures and the impact of using context (CT) for understanding the correlation of mapping evolution with the evolution of SCT ontology. The #topA corresponds to the number of candidate attributes. The #total-correlation refers to the percentage of correlations between the number of changed and unchanged mappings and the number of changed and unchanged attributes in mapped concepts and/or their context. The plot on the top left (1) compares the performance of the three similarity measures, while the remainder plots (2, 3, 4) show the performance of each similarity measure individually by comparing the different type of context where the candidate attribute is found (SUP, SUB, SIB concept or from the source concept without using its context (NOCT) or from both (ALL)).

In low performance of the Knowledge-based similarity measure quality is probably because of the lack of domain information, i.e., the biomedical knowledge available in the training data as well as in the underlying knowledge source.

Analysing the influence of context for identifying relevant attributes, we observe that attributes in the source concepts yield the best correlations rather than the ones in their context. This observation is coherent for the three similarity measures used as depicted in Figures 6.2, 6.3 and 6.4 for SCT. Indeed, the NOCT curve, which corresponds to the scenario where no context is used, remains always above other curves corresponding to each type of context (i.e., SUP, SUB, SIB). We observe the same results on ICD with the Character-based ED similarity (cf. Figure 7.2) with the exception that NOCT and SUP tend to perform similarly both for Knowledge-based and Word-based ED similarity measures (cf. Figure 7.3 and Figure 7.4).

5.3. Analyzing the stability of similarity measures

We aim to study the stability of the similarity measures because an important issue of mapping adaptation concerns determining in which situation a mapping should be adapted. In some cases, the decision could easily be made, e.g., when a source concept is removed, the mapping is removed. However, other cases represent much more difficulties to determine the impact of change on the associated mappings, e.g., where concepts are only affected by a change in attributes. In such situations, the SCA relevant attributes identified could help, since we can exploit changes in
attributes’ values as well as their similarity with attributes in the corresponding target concept. We consider unaffected mappings those with attributes that remain unchanged from one ontology version to another. On the other hand, it is more likely to adapt a mapping when a change affects an identified attribute for such mapping. Even in this situation, the decision on adaptation still appears complex. Therefore, we aim to observe whether found correlations relate to the similarity values described in the identified SCA. For this purpose, we study the stability of the similarity measures used for identifying the relevant attributes based on different aggregation functions in statistics.

Evaluation method. We conduct a more fine-grained analysis of correlations based on different intervals of similarity values between attributes of each source concept (including its context) and the corresponding target concept from each mapping. More specifically, for each interval of similarity values ranging from 0 to 1, we compute the frequency of the correlations obtained using the ALL attributes configuration, i.e., either from the source concepts of mappings as well as the ones issued from their context (SUP, SUB and SIB). We use four aggregate functions (MAX, MEDIAN, ANZ and AVG) for selecting the appropriate candidate attribute in SCA to study its similarity value. We chose these functions because they have been widely used for selecting best candidates from a list of items in related research areas, e.g., information retrieval [8, 9, 10]. The MAX function corresponds to the maximum similarity value from topA. The MEDIAN and AVG functions compute the median and average of similarity values obtained in topA. The ANZ function is similar to the AVG function with the exception that attributes in topA with similarity values of zero are ignored. We divide the similarity values into 10 groups from 0.0 to 1.0, each of which contains a set of attributes whose similarity value ranges from $x$ to $y$ with a step of 0.1 ($y - x = 0.1$).
Results. We plot the percentage of the total-correlation observed using each similarity measure over the similarity threshold from 0.1 to 1.0 with the step of 0.1. We examine the similarity values from the list of candidate attributes in SCA by using each aggregate function ANZ, AVG, MAX and MEDIAN, respectively. Figure 8 compares the frequency of correlations obtained for each type of aggregate functions observing the similarity measures when studying SCT. Similarly, Figure 9 presents the results analysing the frequency of correlations obtained when studying ICD.

The results indicate that the Word-based ED is the most appropriate similarity measure for identifying the correlations between attribute changes and mapping changes. Indeed, for most of the aggregate functions, the Word-based ED performs better than the two other similarity measures, considering higher values of similarity ([0.5-1.0]) (cf. Figure 8). We notice that the higher number of correlations occur in groups of high similarity values. It reinforces the proposed suggestion that semantic mappings are established mainly based on the similarity between attributes. Therefore, the higher the similarity value of the changed attribute is identified, the greater is the probability of adapting the associated mapping. Hence, if a change occurs in an attribute identified in a source concept $c_S$ or in its context $CT(c_S)$, we should adapt such mapping in an appropriate way (e.g., replacement of source concept, modification of mapping semantic relation or removal). This suggests that correlations found in groups of high values of similarity are important evidences for triggering an adaptation of mapping.

6. Discussion

The continuous evolution of ontologies requires adapting mappings over time. Coping with this issue remains so far an open research problem demanding serious investigations to guarantee the interoperability in the Semantic Web. To investigate the impact of ontology changes affecting concept’s attributes on mapping adaptation, this article proposed a novel approach to exploring similarity between concepts’ attributes of source and target con-

Figure 6: Analysis of change-correlation (cf. Formula 15) w.r.t SCT. These plots are similar to the ones in Figure 4 with the exception that we only plot the percentage of change-correlation.
cepts in established mappings. Our investigation relies on previous results of our empirical study on the analysis of mapping evolution [14].

Through the experiments reported in this article, we hypothesized that identifying a set of candidate attributes (in the mapping’s source concept or its context), namely SCA, might play a central role for automatically adapting mappings. Indeed, according to the obtained results, the SCA identified for each mapping provides an evidence for better interpreting mappings. To this end, we investigated the influence of different similarity measures, at three different linguistic levels, on our proposed method for identifying SCA (cf. Algorithm 1).

The results underscored the quality of the attributes identified by our algorithm and the relevance of exploiting them for supporting mapping adaptation. We found correlations between changes affecting identified attributes and modifications in the corresponding mapping interconnecting source and target concepts. This allows to understand the influence of ontology changes specifically affecting concept’s attributes on mapping adaptation.

According to the results, we found relevant to combine the attributes from the concepts of the context with the source concept’s attributes for boosting the identification of SCA. Using the context revealed more relevant to find unchange-correlation than change-correlation. Our method performed well without using context relying either on Character-based ED or Word-based ED. Indeed, in some cases, the Word-based ED allows to efficiently cope with the problem of word order in attributes such as “skin cancer” and “cancer in the skin” in comparison to the Character-based ED.

When considering the context of interrelated concepts for studying the mapping adaptation, results demonstrated the convenience of exploiting the similarity based on the background knowledge such as WordNet, even that this consists of a general resource lacking domain information. To improve these results, we plan to investigate integrating domain-specific resources in our background knowledge similarity, but the reliable semantically annotated corpus underlying such resources remain so far scarce and needs more supports from the com-
Figure 8: Analysing similarity values found by the studied similarity metrics by using aggregate functions (ANZ, AVG, MAX and MEDIAN) in total-correlation w.r.t SCT

Figure 9: Analysing similarity values found by the studied similarity metrics by using aggregate functions (ANZ, AVG, MAX and MEDIAN) in total-correlation w.r.t ICD
munity. Moreover, we observed that results vary according to the studied ontologies. We should particularly chose the adequate similarity measure according to characteristics of the involved ontologies, but this requires further research. Therefore, we judge more appropriated to have the similarity metric as a parameter in the proposed method.

7. Conclusion

Mappings interconnecting ontologies play a key role for software applications dealing with tasks related to information integration and sharing on the Web. Nevertheless, ontology evolution potentially makes existing mappings semantically invalid or outdated. In this article, we proposed an original method for identifying the most relevant subset of concept’s attributes, namely SCA, which is useful for interpreting the evolution of mappings under evolving ontologies. Our research aims to facilitate the maintenance of mappings based on the detected attributes. We conducted a series of experiments investigating the influence of similarity measures underlying our method at three different linguistic levels: lexical, syntactic and semantic level. Our achieved results provided empirical evidences of the relevance of identifying the relevant attributes for supporting mapping adaptation. The conducted evaluation shows to which extent the similarity is valuable between source and target concepts involved in mappings, especially at the level of attributes in concepts, to support the decisions on mapping adaptation.

For future work, we aim to study additional factors that might influence the adaptation of mappings due to ontology evolution. For instance, when a concept is removed (or is set to obsolete), we plan to investigate the similarity between the source concept and the ones in its context in a new ontology version to decide whether mappings should be removed or moved to a most related concept in the context. In addition, we aim to examine the influence of other types of ontology change operations, such as those affecting relationships (e.g., delR, addR) to determine precise decisions on mapping adaptation actions based on ontology changes.

Acknowledgment

The National Research Fund (FNR) of Luxembourg entirely supports this research under the DynaMO project (grant #C10/18/786147).

References

[14] Dos Reis, J. C., Pruski, C., Da Silveira, M., Reynaud-


