

Updating Ontology Alignments in Life Sciences based on New Concepts and their Context

Victor Eiti Yamamoto, Julio Cesar dos Reis

Institute of Computing, University of Campinas, Campinas - SP, Brazil
eitiyamamoto@gmail.com, jreis@ic.unicamp.br

Abstract. Ontologies and their associated mappings in life sciences play a central role in several semantic-enabled tasks. However, the continuous evolution of these ontologies requires updating existing concept alignments. Whereas mapping maintenance techniques have mostly handled revision and removal type of ontology changes, the addition of concepts demands further studies. This article proposes a technique to refine a set of established mappings based on the evolution of biomedical ontologies. We investigate ways of suggesting correspondences with the new version of the ontology without applying a matching operation to the whole set of ontology entities. Obtained results explore the neighbourhood of concepts in the alignment process to update mapping sets. Our experimental evaluation with several versions of aligned biomedical ontologies shows the effectiveness in considering the context of new concepts.

Keywords: ontology alignment; ontology evolution; mapping refinement; concept addition; biomedical vocabulary

1 Introduction

Over the last decade the biomedical domain has exploited ontologies and their capabilities for various purposes ranging from information retrieval to data management and sharing. However, the size of this domain often requires the use of several ontologies whose elements are linked through mappings. Mappings are the materialization of semantic relations between elements of interrelated ontologies [14].

Creating mappings between ontologies is a complex task especially due to the increasing size of biomedical ontologies. Several automatic ontology alignment techniques have been proposed [10]. Nevertheless, significant manual efforts of validation are still demanded if a certain level of quality is required. This prevents software applications relying on mappings to fully take advantage on them.

Ontologies in life sciences evolve over time to keep them up-to-date according to the domain knowledge. Ontology changes may affect mappings already established or can be a source for treating mapping refinement. In this context, in order to avoid the costly ontology re-alignment process, it is crucial to have adequate mapping techniques to keep mappings semantically valid over time [1]. Manual mapping maintenance is possible only if modifications are applied to

a restricted number of mappings. Otherwise automatic methods are required for large and highly dynamic ontologies. Biomedical ontologies usually contain hundred of thousands of concepts interconnected via mappings.

Coping with the mapping reconciliation problem in a semi-automatic way entails many research challenges. First, it is difficult to evaluate the real impact of the ontology evolution on mappings. For instance, changing an attribute value may lead to invalidate a mapping in some cases. In these situations, the challenging issue is to identify and classify the different cases. Second, several types of ontology changes can be applied to an ontology, but it is unknown how these different types of operations should be duly taken into consideration for mapping reconciliation [4].

The design of techniques for mapping adaptation according to the different types of ontology changes has been coped within existing approaches. Previous work presented a mapping adaptation strategy for two out of three categories of ontology evolution: *removal* of knowledge and *revision* of knowledge [1]. For example, when concepts are removed, heuristics were designed to automatically apply adaptation actions over mappings. The *addition* of knowledge (third category) is the most frequent type of change occurred in ontology evolution. New concepts or attributes in concepts are added to comply with new domain knowledge. Such new knowledge needs to be aligned with the interrelated ontologies.

In this paper, we propose a mapping refinement methodology to update mapping sets taking ontology changes into account (based on new concepts added in ontology evolution). We study the use of conceptual information related to neighbour concepts for enhancing the mapping completeness. For this purpose, we investigate a technique to reuse already established mappings and to explore the role of neighbour concepts to derive new mappings. Our proposal allows suggesting new correspondences without applying a matching operation with the whole set of ontology entities.

Our experimental evaluation explored real-world biomedical ontologies and mappings established between them. We examine the quality of the automatically-suggested enriched set of mappings with respect to the set of new correspondences observed in the official updated release of mappings via standard evaluation metrics. The achieved results show innovative findings regarding the way mappings can be refined based on new concepts added. We demonstrate that the local matching considering neighbour concepts is competitive with a matching operation with the whole target ontology.

The remainder of this article is organized as follows: Section 2 presents the related work; Section 3 presents the formal definitions and problem statement. Section 4 reports on our approach to refine ontology mappings under ontology evolution. Section 5 shows the used materials and the results obtained. Section 6 discusses the findings whereas Section 7 draws conclusions and future work.

2 Related Work

Previous studies have investigated semi-automatic approaches to adapting ontology mappings when at least one of the mapped ontologies evolves [4]. Dos Reis *et al.* conceptualized the *DyKOSMap* framework [1] for supporting the adaptation of semantic mappings highlighting different aspects such as: the role of different types of ontology changes, the importance in considering the conceptual information which established mappings are related to, as well as the relevance of the different types of semantic relation of mappings.

Some techniques have used external resources aiming to improve and increase the number and precision of established mappings. Stoutenburg [15] argued that the use of upper ontologies (an ontology which consists of very general terms that are common across all domains), and linguistic resources can enhance the alignment process.

The *TaxoMap* matching tool [6] explored pattern-based refinement techniques. The mapping is generated, with initial proposed relations (correspondences found are equivalence relations, subsumption relations and their inverse, or proximity relations). A domain expert manually validates the generated mappings and correct problems, grouping the identified problems together when they correspond to a similar case. The tool generates patterns based on groups of similar cases, which can be applied to other mappings in the same domain.

Other approaches have combined lexical-based and semantic-based algorithms, mostly using resources available in the *Unified Medical Language System* (UMLS)¹ for generating mappings. The use of UMLS as an external resource can be interesting in various aspects: (1) favors an increase in the number of mapping, (2) provides different synonyms terms for a given concept, and (3) defines relations between concepts in a semantic network. Zhang and Bodenreider [16] explored UMLS to improve alignment between anatomical ontologies. They showed that domain knowledge is a key factor for the identification of additional mappings compared with the generic schema matching approach.

Sekhavat and Parsons [13] explored conceptual models (*e.g.*, Entity Relationship, Class diagrams or domain ontologies) as background knowledge to enrich database schema mappings and resolve ambiguous mappings. Their approach used conceptual models as external resources to capture semantics of schema elements, for instance, a pair of concepts a_1 and a_2 where a_1 is a subclass and a_2 is a superclass in a conceptual model. This information was used to enrich the schema before mapping, marking the foreign keys corresponding to a_1 and a_2 as generalizations. As a consequence, the relationship identified in the schema mapping is a generalization (*is-a*) instead of equivalence.

Pruski *et al.* [11] proposed exploiting domain-specific external source of knowledge to characterize the evolution of concepts in dynamic ontologies. The technique analyzed the evolution of values in concept attributes. The approach used

¹ UMLS is a collection of health and biomedical vocabularies and standards. URL: www.nlm.nih.gov/research/umls/

ontological properties and mappings between ontologies from online repositories to deduce the relationship between a concept and its successive version.

Noy *et al.* [9] and Seddiqui *et al.* [12] explored anchor concepts to obtain mappings. They use a set of concept pairs aligned to obtain other mappings based on these pairs. These approaches calculate new alignment for all concepts from the involved ontologies, but they are not used for ontology evolution.

In this investigation, we explore ontology change operations to leverage refinement, in particular, concept addition. We contribute with a methodology to consider newly added concepts and investigate the context of candidate target concepts of existing mappings for refinement over time. We further evaluate the proposed algorithms by measuring the effectiveness of our mapping refinement approach on real-world biomedical ontologies.

3 Preliminaries

Ontology. An ontology \mathcal{O} specifies a conceptualization of a domain in terms of concepts, attributes and relationships [5]. Formally, an ontology $\mathcal{O} = (\mathcal{C}_{\mathcal{O}}, \mathcal{R}_{\mathcal{O}}, \mathcal{A}_{\mathcal{O}})$ consists of a set of concepts $\mathcal{C}_{\mathcal{O}}$ interrelated by directed relationships $\mathcal{R}_{\mathcal{O}}$. Each concept $c \in \mathcal{C}_{\mathcal{O}}$ has a unique identifier and is associated with a set of attributes $\mathcal{A}_{\mathcal{O}}(c) = \{a_1, a_2, \dots, a_p\}$. Each relationship $r(c_1, c_2) \in \mathcal{R}_{\mathcal{O}}$ is typically a triple (c_1, c_2, t) where t is the relationship (*e.g.*, “is_a”, “part_of”, “advised_by”, *etc.*) interrelating c_1 and c_2 .

Context of a concept. We define the context of a particular concept $c_i \in \mathcal{C}_{\mathcal{O}}$ as a set of *super concepts*, *sub concepts* and *sibling concepts* of c_i , as following:

$$CT(c_i, \lambda) = sup(c_i, \lambda) \cup sub(c_i, \lambda) \cup sib(c_i, \lambda) \quad (1)$$

where

$$\begin{aligned} sup(c_i, \lambda) &= \{c_j | c_j \in \mathcal{C}_{\mathcal{O}}, r(c_i, c_j) = \text{“}\sqsubset\text{”} \wedge length(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \\ sub(c_i, \lambda) &= \{c_j | c_j \in \mathcal{C}_{\mathcal{O}}, r(c_j, c_i) = \text{“}\sqsubset\text{”} \wedge length(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \\ sib(c_i, \lambda) &= \{c_j | c_j \in \mathcal{C}_{\mathcal{O}}, ((sup(c_j) \cap sup(c_i)) \vee (sub(c_j) \cap sub(c_i))) \\ &\quad \wedge length(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \end{aligned} \quad (2)$$

where λ is the level of the context. It represents the maximum value for the length between two concepts (in terms of their shortest relationship distance in the hierarchy of concepts) and the “ \sqsubset ” symbol indicates that “ c_i is a sub concept of c_j ”. This definition of $CT(c_i, \lambda)$ is specially designed as the relevant concepts to be taken into account in the settings of this investigation on mapping refinement.

Similarity between concepts. Given two particular concepts c_i and c_j , the similarity between them can be defined as the maximum similarity between each couple of attributes from c_i and c_j . Formally:

$$sim(c_i, c_j) = \arg \max sim(a_{ix}, a_{jy}) \quad (3)$$

where $sim(a_{ix}, a_{jy})$ is the similarity between two attributes a_{ix} and a_{jy} denoting concepts c_i and c_j , respectively.

Mapping. Given two concepts c_s and c_t from two different ontologies, a mapping m_{st} can be defined as:

$$m_{st} = (c_s, c_t, semType, conf) \quad (4)$$

where $semType$ is the semantic relation connecting c_s and c_t . In this article, we differentiate *relation* from *relationship*, where the former belongs to a mapping and the later to an ontology. The following types of semantic relation are considered: *unmappable* [\perp], *equivalent* [\equiv], *narrow-to-broad* [\leq], *broad-to-narrow* [\geq] and *overlapped* [\approx]. For example, concepts can be equivalent (*e.g.*, “head” \equiv “head”), one concept can be less or more general than the other (*e.g.*, “thumb” \leq “finger”) or concepts can be somehow semantically related (\approx). The $conf$ is the similarity between c_s and c_t indicating the confidence of their relation [3]. We define \mathcal{M}_{ST}^j as a set of mappings m_{st} between ontologies \mathcal{O}_S and \mathcal{O}_T at a given time j . We assume $j \in N$ the version of the ontology release \mathcal{O}_S^j . Ontology \mathcal{O}_S^0 is the version 0 whereas \mathcal{O}_S^1 is the version 1 of the same ontology.

Ontology change operations (OCO). An ontology change operation (OCO) is defined to represent a change in an attribute, in a set of one or more concepts or in a relationship between concepts. OCOs are classified into two main categories: *atomic* and *complex* changes. Each OCO in the former cannot be divided into smaller operations while each one of the latter is composed of more than one atomic operation. In this paper, we pay further attention to the operations of concept addition which is an atomic operation.

Problem statement. Consider two versions of the same source ontology \mathcal{O}_S^j at time j and \mathcal{O}_S^{j+1} at time $j+1$, a target ontology \mathcal{O}_T^j , and an initial set of mappings \mathcal{M}_{ST}^j between \mathcal{O}_S^j and \mathcal{O}_T^j at time j . Suppose that the frequency of new releases of \mathcal{O}_S and \mathcal{O}_T is different and at time $j+1$ only \mathcal{O}_S evolves. Since this evolution is likely to impact the mappings \mathcal{M}_{ST}^j , it is necessary to refine \mathcal{M}_{ST}^j to guarantee the quality and completeness of \mathcal{M}_{ST}^{j+1} . The quality is related to the consistency of mappings and it can be measured using precision. For instance, mappings cannot be established between removed concepts. The completeness refers to the recall of aligned concepts in \mathcal{M}_{ST}^{j+1} . In this investigation, we study how \mathcal{M}_{ST}^j can be refined (*e.g.*, new mappings derived) based on ontology changes related to *addition of knowledge*. We address the following research questions:

- How to exploit existing mappings for mapping refinement based on new concepts added?
- Is it possible to reach mapping refinement for alignment of new concepts without applying a matching operation in the whole target ontology?
- What is the impact of using the context of concepts $CT(c_i, \lambda)$ in both source and target ontologies on the mapping refinement effectiveness?

We consider that \mathcal{O}_T has not evolved (thus \mathcal{O}_T^j and \mathcal{O}_T^{j+1} are the same version of the ontology \mathcal{O}_T). \mathcal{O}_S^j and \mathcal{O}_S^{j+1} are two distinct versions of the same ontology

\mathcal{O}_S . At time $j + 1$, newly added concepts appear in \mathcal{O}_S^{j+1} and we attempt to refine the original mapping set M_{ST}^j to provide a set of valid mappings M_{ST}^{j+1} .

4 Mapping Refinement under Concept Addition Changes

Our goal is to propose adequate correspondences for each newly added concept at time $j + 1$. In the first step, our approach identifies all newly added concepts using the *Conto-Diff* tool [7]. This tool allows to identify atomic and complex ontology changes. Next, we extract the contextual information, *i.e.*, super, sub and sibling concepts of those newly added concepts (*cf.* Formula 1). We then examine the existing mappings between the source concept in the context of the newly added concept and the corresponding target concepts. The idea behind the context-oriented technique is that the candidate mapping is established between a newly added concept and a target concept of an existing mapping at time t .

Figure 2 A illustrates a situation where there are two ontologies that have an alignment in time j . Each circle represents a concept of an ontology. Light blue circles represent concepts of source ontology. Yellow circles represent concepts of target ontology. Continuous lines represent mappings between concepts from source ontology and target ontology.

Figure 2 B illustrates a situation where source ontology evolve and change to time $j + 1$. The algorithm find newly added concepts and explore the context of each newly added concepts. In this case, we are exploring the context of the right concept using source level 1. Purple circles represent newly added concepts. Dark blue circles represent concepts of a context with a certain source level; the number inside the circle represents the source level needed to access this concept.

After finding some concepts inside the newly added concepts' context that have an alignment in previous time, the concepts from target ontologies that have an alignment in previous time are added as candidate concepts. The context of each candidate concepts is explored and added as candidate concepts. Figure 3 illustrate this situation using target level 1. Red circles represent candidate concepts for a new concept from source ontology; the number inside the circles represent the target level needed to access this concept. Dashed lines represent a possible alignment between a new concept (in \mathcal{O}_S^{j+1}) and candidate concepts (in \mathcal{O}_T^j).

Algorithm 1 computes the *diff* between two given versions of the source ontology (line 1). For each newly added concept c_i^1 , the algorithm considers a candidate concept namely c_i^0 in the target ontology by exploiting already existing mappings related to $CT(c_i^1, \gamma)$ (lines 4-8). Note that we recover the before evolution version (c_k^0) of the concept c_k^1 found in the context of c_i^1 .

For each c_i^0 , the algorithm obtains a set of concepts from $CT(c_i^0, \lambda)$ (line 11). We determine a new refined mapping by calculating the similarity between a new concept c_i^1 of \mathcal{O}_S^{j+1} and a candidate $c_n \in \mathcal{C}_t$. If the maximum similarity (among the concept attributes) is greater than or equal to a threshold τ , the algorithm establishes a mapping between the newly added concept and the candidate target

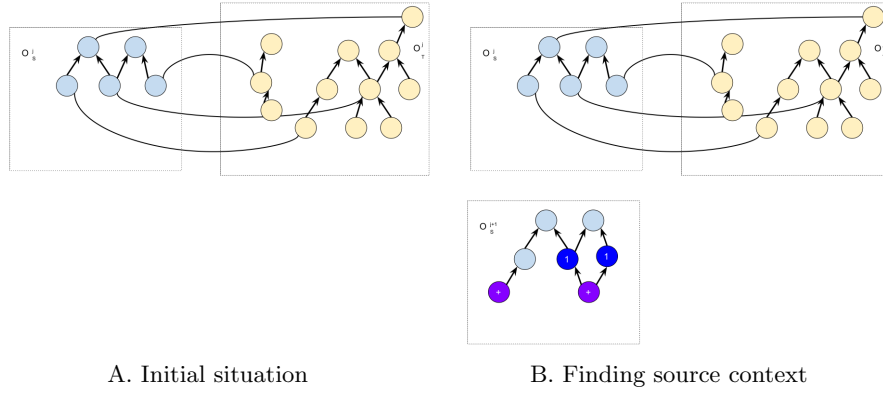


Fig. 2: Representation of situations before applying alignment algorithm

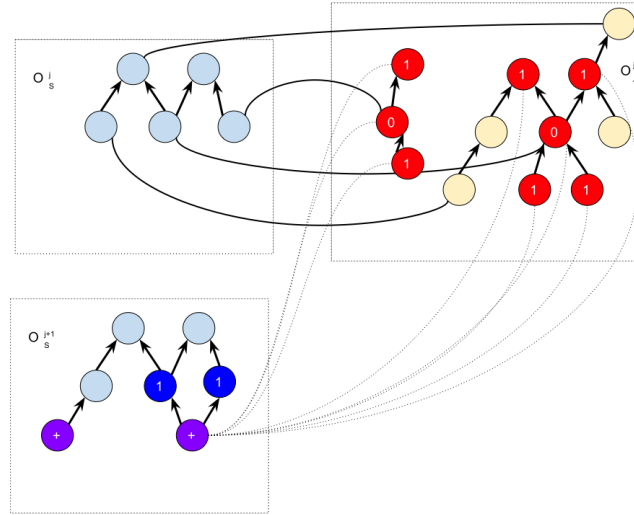


Fig. 3: Calculating similarity with candidate concepts

concept. Algorithm 1 searches for the candidate c_t that yields the maximum similarity value.

In order to compare with the results obtained by our approach (*cf.* Section 5), we propose another algorithm that ignores new concepts' context to calculate similarity. It means that the algorithm computes the similarity between each newly added concepts with all concepts in the target ontology. More specifically, these algorithm computes the *diff* between two given versions of the source ontology. For each newly added concepts, it calculates similarity between all concepts with the target ontology. If there are any similarity greater than a threshold,

Algorithm 1 Contextual approach to mapping refinement

Require: $\mathcal{O}_S^j, \mathcal{O}_S^{j+1}, \mathcal{O}_T^j, \mathcal{O}_T^{j+1}, \mathcal{M}_{ST}^j, \lambda, \gamma, \tau \in \mathbb{R}$
Ensure: $M_A = \{m_1, m_2, \dots, m_N\}$

- 1: $\mathcal{C}_{add} \leftarrow \text{diff}_{add}(\mathcal{O}_S^j, \mathcal{O}_S^{j+1})$ {newly added concepts}
- 2: $\mathcal{C}_t \leftarrow \emptyset$ {initialize target concepts of candidate mappings}
- 3: **for all** $c_i^1 \in \mathcal{C}_{add}$ **do**
- 4: **for all** $c_k^1 \in CT(c_i^1, \gamma)$ **do**
- 5: **if** $\exists c_t^0 \in \mathcal{C}_{\mathcal{O}_T^j}, \exists m(c_k^0, c_t^0) \in \mathcal{M}_{ST}^0$ **then**
- 6: $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \{c_t^0\}$
- 7: **end if**
- 8: **end for**
- 9: $m_{it} \leftarrow \emptyset$
- 10: **for all** $c_t \in \mathcal{C}_t$ **do**
- 11: **for all** $c_n \in CT(c_t, \lambda)$ **do**
- 12: $m_{cand} \leftarrow \text{argmax}_{sim(c_i^1, c_n)}$ {Create a mapping between concepts c_i^1 and c_n }
- 13: **if** $\max(sim(c_i^1, c_n)) \geq \tau$ **then**
- 14: $m_{it} \leftarrow m_{cand}$
- 15: $\tau \leftarrow \max(sim(c_i^1, c_n))$
- 16: **end if**
- 17: **end for**
- 18: **end for**
- 19: $\mathcal{M}_A \leftarrow \mathcal{M}_A \cup \{m_{it}\}$
- 20: **end for**
- 21: **return** \mathcal{M}_A

the algorithm creates a new mapping between the newly added concept and a target ontology's concept with the greatest similarity.

In our algorithms, source concept's attributes are compared with all target concept's attributes to obtain similarity value between concepts. The value of similarity between two concepts is the maximum value of similarity from their attributes. The method used to calculate similarity affects the precision and recall. In this work, we explored Bi-gram Dice to calculate similarity. Bi-gram is a sequence of two adjacent letters of a word. Dice's coefficient is defined as twice the number of common elements divided by sum of each elements. Formula 5 shows the application of Bi-gram Dice to strings X and Y. The n-gram's strength is in the fact it has context sensitivity, but it not have good resolution when gram size is increased [8]. For the data set used in this work Bi-gram Dice have better results than Levenshtein distance, Cosine distance and Jaccard distance.

$$\text{Similarity} = \frac{2 \times (\text{Bi-gram}(X) \cap \text{Bi-gram}(Y))}{\text{Bi-gram}(X) + \text{Bi-gram}(Y)} \quad (5)$$

5 Experimental Evaluation

We aim to validate the quality of the refined set of mappings as the outcome of our approach. Data used in this evaluation come from five biomedical ontologies: SNOMED-CT (SCT), MeSH, ICD-9-CM, ICD10-CM and NCI Thesaurus. SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) is an ontology which objective is to create a taxonomy of terms referring to the medical environment and a framework of rules guaranteeing that each term is used with exactly one meaning [2]. MeSH Thesaurus is a controlled vocabulary produced by the National Library of Medicine and used to index, catalogue and search information and documents related to biomedicine and health <https://www.nlm.nih.gov/> ICD-9-CM and ICD-10-CM are a formalization in OWL-DL of International Classification of Diseases published by World Health Organization² NCI Thesaurus³ contains terminologies used in the National Cancer Institute’s semantic infrastructure and information systems Table 1 shows the statistics of source and target ontologies for each of the considered versions.

Table 1: Statistics of ontologies

Ontology	Release	#Concepts	#Attributes	#Subsumptions	#New Concepts
ICD9	2009	12,734	34,065	11,619	325
	2011	13,059	34,963	11,962	
ICD10	2011	43,351	87,354	40,330	0
SCT	2010	386,965	1,531,288	523,958	8,381
	2012	395,346	1,570,504	539,245	
NCI	2009	77,448	282,434	86,822	17,284
	2012	94,732	365,515	105,406	
MeSH	2012	50,367	259,565	59,191	604
	2013	50,971	264,783	59,844	

The mappings obtained by the proposed Algorithm 1 are compared with the official mappings (their new official release). Mappings between SNOMEDCT and ICD9CM is offered by the International Health Terminology Standards Development Organisation (IHTSDO)⁴. Mappings between MeSH and ICD-10-CM were offered by the *Catalogue et Indexation des Sites Mdicaux de langue Franais* (CISMeF)⁵. Table 2 shows the quantity of each mapping set between the ontologies used in this experiment.

To analyze results obtained experimentally, it was necessary to compare our obtained mappings with mappings created only for newly added concepts in the new version of the considered ontologies. Table 3 shows the quantity of mappings really considered in the metrics.

² <http://www.who.int/classifications/icd/en/>

³ <https://ncit.nci.nih.gov/ncitbrowser/>

⁴ https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

⁵ <http://www.chu-rouen.fr/cismef>

Table 2: Statistics of the studied mappings

SCT-ICD9 #Mappings		SCT-NCI #Mappings		MeSH-ICD10CM #Mappings	
2010-2009	84,519	2009-2009	19,971	2012-2011	4,631
2012-2011	86,638	2012-2012	22,732	2013-2011	5,378

Table 3: Number of new mappings created and associated to newly added concepts in the new ontology version (considered gold standard)

Mappings	#official mappings created after newly added concepts
SNOMEDCT-ICD9CM	1,583
SNOMEDCT-NCI	158
MeSH-ICD10CM	21

The experiments were performed for the three datasets (SCT-NCI, SCT-ICD9 and MeSH-ICD10) considering SCT and MeSH as source ontologies. As assessed configurations, we considered three source levels, three threshold values (0.5, 0.75 and 0.9), and four target levels. For each dataset, we fixed source level and threshold to verify the results for each target level. After examining all target levels, we changed the threshold and repeated for each target level. After examining all threshold values, we changed source level and repeated the whole procedure for all thresholds and target levels.

We used three metrics to evaluate the results: Precision, Recall and F-Measure. These metrics were used comparing results obtained by our approach and expected results from the official mappings.

Precision is defined as the relation between correctly identified mappings and identified mapping (Formula 6).

$$Precision = \frac{\#IdentifiedAndCorrectMappings}{\#IdentifiedMappings} \quad (6)$$

Recall is defined as the relation between correctly identified mappings and those expected new official release of mappings (Formula 7).

$$Recall = \frac{\#IdentifiedAndCorrectMappings}{\#CorrectMappings} \quad (7)$$

F-measure is the harmonic mean of precision and recall (Formula 8).

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

Tables 4 (SNOMED-CT and NCI Thesaurus), 5 (SNOMED-CT and ICD-9) and 6 (MeSH and ICD-10) show the obtained results in terms of precision, recall and f-measure in applying our Algorithm 1 for the studied datasets.

Results in Table 4 reveal a decrease in precision and f-Measure for threshold set as 0.5 when it increases source level. The results increases in terms of precision, recall and F-measure for other thresholds. We found that the best results are obtained when increasing the level of the context in the source concept.

Table 4: Mapping derivation results for SNOMED-CT and NCI

Source level	Threshold	Target level	Precision	Recall	F-Measure
1	0.5	0	0.009	0.018	0.012
		1	0.042	0.101	0.060
		2	0.048	0.120	0.068
		3	0.048	0.127	0.069
	0.75	0	0	0	0
		1	0.088	0.082	0.085
		2	0.086	0.082	0.0841
		3	0.101	0.108	0.104
	0.9	0	0	0	0
		1	0.344	0.070	0.116
		2	0.378	0.089	0.144
		3	0.341	0.089	0.140
2	0.5	0	0.006	0.025	0.010
		1	0.031	0.139	0.051
		2	0.035	0.171	0.058
		3	0.034	0.184	0.058
	0.75	0	0.006	0.006	0.006
		1	0.089	0.120	0.102
		2	0.105	0.152	0.124
		3	0.108	0.165	0.130
	0.9	0	0.071	0.006	0.012
		1	0.357	0.095	0.012
		2	0.423	0.127	0.195
		3	0.407	0.139	0.208
3	0.5	0	0.004	0.019	0.006
		1	0.023	0.139	0.039
		2	0.029	0.190	0.050
		3	0.028	0.190	0.048
	0.75	0	0.005	0.006	0.006
		1	0.08	0.127	0.097
		2	0.104	0.177	0.125
		3	0.097	0.177	0.125
	0.9	0	0.071	0.006	0.012
		1	0.356	0.101	0.158
		2	0.434	0.146	0.218
		3	0.418	0.146	0.216

Results presented in Table 5 (concerning the mapping between SNOMED-CT and ICD-9) are different from those of SNOMED-CT and NCI. We observe an increase in recall when it increases source level, but it has lower precision. Table 5 presents the best results for the first level in the source concept and with lower thresholds. We could not observe huge differences in the results when increasing the context level of the target concept.

Table 6 presents the results for the refinement for MeSH and ICD-10. We observed an overall improvement of results when increasing the level of the source concept.

We evaluated our proposal in considering the neighbourhood for the derivation of new mappings associated to new concepts (Algorithm 1) with the approach in applying the matching with the whole target ontology. To this end, we applied the non-context approach in the datasets considering the threshold τ yielding the best results in Algorithm 1 obtained for each dataset. Table 7 shows the results concerning precision, recall and f-measure obtained for each dataset using the matching with all concepts in the target ontology. The comparison of

Table 5: Mapping derivation results for SNOMED-CT and ICD-9

Source level	Threshold	Target level	Precision	Recall	F-Measure
1	0.5	0	0.535	0.186	0.276
		1	0.340	0.163	0.220
		2	0.310	0.152	0.204
		3	0.296	0.145	0.196
	0.75	0	0.630	0.037	0.069
		1	0.461	0.044	0.081
		2	0.449	0.041	0.075
		3	0.439	0.041	0.075
	0.9	0	0.778	0.004	0.009
		1	0.692	0.006	0.011
		2	0.727	0.005	0.10
		3	0.75	0.006	0.011
2	0.5	0	0.325	0.181	0.233
		1	0.233	0.159	0.189
		2	0.241	0.167	0.198
		3	0.230	0.160	0.230
	0.75	0	0.487	0.046	0.084
		1	0.349	0.0455	0.080
		2	0.449	0.042	0.076
		3	0.382	0.048	0.085
	0.9	0	0.8	0.005	0.010
		1	0.687	0.007	0.014
		2	0.615	0.005	0.010
		3	0.615	0.005	0.010
3	0.5	0	0.256	0.177	0.209
		1	0.190	0.158	0.166
		2	0.200	0.159	0.177
		3	0.199	0.157	0.175
	0.75	0	0.444	0.051	0.091
		1	0.342	0.049	0.085
		2	0.360	0.050	0.089
		3	0.348	0.049	0.085
	0.9	0	0.833	0.006	0.013
		1	0.687	0.007	0.014
		2	0.687	0.007	0.014
		3	0.687	0.007	0.014

results reveals that for the dataset SCT-NCI the results using all target concept as candidates were better. For the dataset SCT-ICD9, our context-approach is better; concerning the dataset MeSH-ICD10, the approaches obtained similar results. However, we need to consider that applying mapping candidates with the whole target ontology have a worst run-time complexity than our contextual approach.

6 Discussion

This investigation aimed to create mappings to update ontology alignments based on new concepts added in novel ontology releases. Our approach have three variables affecting mapping quality: threshold, target level and source level. Threshold increases precision, but decreases recall. It is caused by the fact that high threshold can remove false positive mappings, but as an effect removes correct mappings. For two datasets (SCT-NCI and MeSH-ICD10) the increasing in

Table 6: Mapping refinement results for MeSH and ICD-10

Source level	Threshold	Target level	Precision	Recall	F-Measure
1	0.5	0	0.059	0.048	0.053
		1	0.059	0.048	0.053
		2	0.059	0.048	0.053
		3	0.0625	0.048	0.054
	0.75	0	0.250	0.048	0.08
		1	0.250	0.048	0.08
		2	0.250	0.048	0.08
		3	0.250	0.048	0.08
	0.9	0	0	0	0
		1	0	0	0
		2	0	0	0
		3	0	0	0
2	0.5	0	0.067	0.095	0.078
		1	0.079	0.143	0.102
		2	0.0714	0.143	0.095
		3	0.0714	0.143	0.095
	0.75	0	0.250	0.048	0.08
		1	0.429	0.143	0.214
		2	0.429	0.143	0.214
		3	0.429	0.143	0.214
	0.9	0	0	0	0
		1	1.000	0.095	0.174
		2	1.000	0.095	0.174
		3	1.000	0.095	0.174
3	0.5	0	0.036	0.095	0.052
		1	0.044	0.143	0.067
		2	0.043	0.143	0.067
		3	0.043	0.143	0.066
	0.75	0	0.125	0.048	0.069
		1	0.333	0.143	0.2
		2	0.333	0.143	0.2
		3	0.333	0.143	0.2
	0.9	0	0	0	0
		1	1.000	0.095	0.174
		2	1.000	0.095	0.174
		3	1.000	0.095	0.174

Table 7: Mapping derivation results exploring the matching with all concepts of the target ontology

Data set	Threshold	Precision	Recall	F-Measure
SCT-NCI	0.9	0.593	0.525	0.557
SCT-ICD9	0.5	0.264	0.042	0.072
MeSH-ICD10	0.75	0.312	0.238	0.270

precision compensated decreasing in recall. However, we observed for one dataset (SCT-ICD9) that a high threshold implied bad effects.

Target level increases candidate concepts for mapping by increasing the context in the target ontology. It means that each new added concept has more options to compare. The increasing in candidate concepts means more chances to find a correct mappings, but it can cause finding a wrong mapping when a wrong concept have a better results in terms of similarity value than the expected concept. For two datasets (SCT-NCI and MeSH-ICD10), precision increased between target level 0 and 1 and recall improved when target level increases. For

one dataset (SCT-ICD9), precision decreased between target level 0 and 1 and recall had only minor effects caused by changes in target level. We found that the results were very dependent on the characteristics of the datasets.

Source level increases source context to find candidate concepts from the target ontology. Our approach depends if in the neighbourhood of a new concept there are concepts presenting a mapping in prior version. If the source level is low, new concepts have less chances to find concepts mapped in prior version. In the worst case, if there is no concept mapped in a prior version, the new concepts are not analyzed to find new mappings. Therefore, in our approach, the derivation of mappings related to new concepts depends directly on the source level. We found better results in improving the level of the source concept.

The analysis of results obtained by all concepts approach indicates that SCT-NCI got better results using such approach. Whereas precision presents very similar results, the recall is very low using contextual approach. SCT-ICD9 presented better results using contextual approach. In this case, precision had good values using contextual approach, but f-measure suffered with low recall. Findings on the dataset MeSH-ICD10 presented similar results for both approaches. In summary, contextual approach implies in a better precision, but all concepts approach obtains a better recall.

This research found that is possible to exploit existing mappings for mapping refinement based on new concepts added. Our findings indicated the possibility of reaching mapping refinement for alignment of new concepts in ontology evolution without applying a matching operation in the whole target ontology. We found further impact in considering the level of the source concept than in the target ontology for the effectiveness of the ontology alignment refinement.

7 Conclusion

Ontology mappings play a central role for semantic data integration in the life sciences. However, domain knowledge update leads to new concepts in ontology versions. This requires to maintain mapping sets up-to-date according to the knowledge dynamics. We proposed a technique to refine ontology alignments based on evolving ontologies. Our constructed algorithm considered the context of concepts in both ontologies as a way to find the matching between concepts. Experimental evaluation with aligned ontologies in the life sciences demonstrated the effectiveness of our approach. Future work involves further investigating heuristics to update the type of semantic relation in the refinement procedure.

Acknowledgements

This work was financially supported by the São Paulo Research Foundation (FAPESP) (grant #2017/02325-5)⁶. We also thank the Scientific Initiation Program (PIBIC) from UNICAMP for the scholarship grant.

⁶ The opinions expressed in here are not necessarily shared by the financial support agency.

References

1. Dos Reis, J.C., Pruski, C., Da Silveira, M., Reynaud-Delaitre, C.: Dykosmap: A framework for mapping adaptation between biomedical knowledge organization systems. *Journal of Biomedical Informatics* **55**, 153 – 173 (2015)
2. El-Sappagh, S., Franda, F., Ali, F., Kwak, K.: Snomed ct standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making* **18** (12 2018)
3. Euzenat, J., Shvaiko, P.: *Ontology matching*. Springer (2007)
4. Groß, A., Reis, J.C.D., Hartung, M., Pruski, C., Rahm, E.: Semi-Automatic Adaptation of Mappings between Life Science Ontologies. In: *Proceedings The 9th International Conference on Data Integration in the Life Sciences*. pp. 90–104 (2013)
5. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowl. Acquis.* **5**(2), 199–220 (Jun 1993)
6. Hamdi, F., Safar, B., Niraula, N.B., Reynaud, C.: Taxomap alignment and refinement modules: Results for oaei 2010. In: *Proceedings of the 5th International Workshop on Ontology Matching*. vol. 689, pp. 212–219 (2010)
7. Hartung, M., Gross, A., Rahm, E.: COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Biomedical Informatics* **46**, 15–32 (2013)
8. Kondrak, G.: N-gram similarity and distance. In: Consens, M., Navarro, G. (eds.) *String Processing and Information Retrieval*. pp. 115–126. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
9. Noy, N.F., Musen, M.A.: Anchor-prompt: Using non-local context for semantic matching. In: *Workshop on ontologies and information sharing*. pp. 63–70 (2001)
10. Otero-Cerdeira, L., Rodriguez-Martinez, F.J., Gmez-Rodriguez, A.: Ontology matching: A literature review. *Expert Systems with Applications* **42**(2), 949 – 971 (2015)
11. Pruski, C., Dos Reis, J.C., Da Silveira, M.: Capturing the relationship between evolving biomedical concepts via background knowledge. In: *Proceedings of the 9th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS16)* (2016)
12. Seddiqui, M.H., Aono, M.: Anchor-flood: Results for oaei 2009. In: *Proceedings of the 4th International Conference on Ontology Matching - Volume 551*. pp. 127–134. OM’09, CEUR-WS.org, Aachen, Germany, Germany (2009)
13. Sekhavat, Y.A., Parsons, J.: Sesm: Semantic enrichment of schema mappings. In: *Proceedings of the 29th International Conference on Data Engineering Workshops (ICDEW 2013)*. pp. 7–12. IEEE (2013)
14. Shvaiko, P., Euzenat, J.: *Ontology Matching: State of the Art and Future Challenges*. *IEEE Trans. Knowl. Data Eng.* **25**(1), 158–176 (2013)
15. Stoutenburg, S.K.: Acquiring advanced properties in ontology mapping. In: *Proceedings of the 2nd PhD Workshop on Information and Knowledge Management (PIKM 2008)*. pp. 9–16. ACM (2008)
16. Zhang, S., Bodenreider, O.: Experience in aligning anatomical ontologies. *International journal on Semantic Web and information systems* **3**(2), 1 (2007)