Selection and Combination of Heterogeneous Mappings to Enhance Biomedical Ontology Matching

Amina Annane, Zohra Bellahsene, Faiçal Azouaou, Clement Jonquet

Selection and Combination of Heterogeneous Mappings to Enhance Biomedical Ontology Matching

Amina Annane[1,2], Zohra Bellahsene[1], Faiçal Azouaou[2], and Clement Jonquet[1,3]

[1] Université de Montpellier, Laboratoire d'Informatique, de Robotique et de Microélectronique (LIRMM), France,
{amina.annane@lirmm.fr}
[2] Ecole nationale Supérieure en Informatique (ESI), Algérie
[3] Center for Biomedical Informatics Research, Stanford University, USA

Abstract. This paper presents a novel background knowledge approach which selects and combines existing mappings from a given biomedical ontology repository to improve ontology alignment. Current background knowledge approaches usually select either manually or automatically a limited number of different ontologies and use them as a whole for background knowledge. Whereas in our approach, we propose to pick up only relevant concepts and relevant existing mappings linking these concepts all together in a specific and customized background knowledge graph. Paths within this graph will help to discover new mappings. We have implemented and evaluated our approach using the content of the NCBO BioPortal repository and the Anatomy benchmark from the Ontology Alignment Evaluation Initiative. We used the mapping gain measure to assess how much our final background knowledge graph improves results of state-of-the-art alignment systems. Furthermore, the evaluation shows that our approach produces a high quality alignment and discovers mappings that have not been found by state-of-the-art systems.

Keywords: ontology matching, background knowledge, repository of ontologies, biomedical ontologies, BioPortal

1 Introduction

Ontology alignment is recognized by the scientific community as an important area of research because of its multiple applications in different domains [7]: ontology engineering, data integration, information sharing, etc. Especially in the biomedical domain that generates and manipulates a big volume of data. Ontology matching plays a key role in the development of biomedical research by facilitating the development of data warehouses articulated around common ontologies. Many works have been made to extract mappings automatically, mainly using lexical and structural matchers, but these matchers often fail when the ontologies to align have different structures and do not use the same vocabulary

(different terms to describe the same concepts) [21]. In the recent years, the community has started to consider an alternative solution for automatic approaches in the use of *background knowledge* as a semantic mediator to discover mappings between ontologies. These background knowledge resources span from thesaurus, lexical resources, linked open data, one or several ontologies or a full repository of ontologies [20, 19, 18] and in our case, already existing mappings. The use of background knowledge has raised the following challenges: (1) selection: How to select the most useful background to align ontologies? (2) usage: How to use such knowledge in order to enhance alignment results? In all proposed approaches, the use of background knowledge was a complementary solution to traditional automatic approaches. In this paper, we propose a novel approach to align ontologies using only a background knowledge built from heterogeneous mappings, the main idea is to combine the knowledge formalized in mappings produced manually by human experts, to mappings produced automatically by simple lexical matching to discover new mappings between the ontologies to be aligned. The main contributions of this paper are:

- A novel approach to align ontologies using a background knowledge graph automatically built from existing mappings
- A novel measure called *Path Confidence Measure* to select the most accurate from several candidates mappings derived from the previously built background knowledge graph.

We have implemented and evaluated our approach using the content of the NCBO BioPortal[4] repository and the Anatomy benchmark[5] from the Ontology Alignment Evaluation Initiative. The obtained results show that our approach produces a high quality alignment, and discovers mappings not found by state-of-the-art alignment systems.

The rest of this paper is organized as follows. Section 2 defines ontology matching and common biomedical ontology mappings. Section 3 describes our novel approach exploiting existing mappings extracted from a given repository to align biomedical ontologies. Section 4 presents the proposed Path Confidence Measure. Section 5 describes the implementation of our approach. Section 6 provides the evaluation results of our approach. Section 7 discusses related work. Finally, Section 8 concludes our paper and points out future work.

2 Preliminaries

2.1 Ontology Matching

Ontology matching is the process of finding correspondences between two given ontologies O_1 and O_2. Each correspondence can be formalized by a quadruplet $\prec e_1, e_2, r, n \succ$ with $e_1 \in O_1$ and $e_2 \in O_2$, r is a relationship between two

[4] http://bioportal.bioontology.org/
[5] http://oaei.ontologymatching.org/2015/anatomy/index.html

given entities e_1 and e_2, and n is the confidence value of this relationship (generally, a value between 0 and 1) [7]. In this paper, we deal only with equivalence relationship between entities.

We distinguish the direct matching which has only the two ontologies to be aligned as an input, from the indirect matching which uses external resources, that we call Background Knowledge (BK), to enhance the quality of direct matching. These resources may be one mediator ontology, a set of ontologies, an existing alignment. The common schema to perform an alignment using a BK is composed of two steps: anchoring and deriving relations [20, 19]. Anchoring consists in finding for source and target entities their equivalent entities in the BK. This step is generally done by using a lexical matcher. The second step consists in deriving relations between the entities of ontologies to align according to the relations between the anchored entities in the BK.

2.2 Biomedical Ontologies Mapping

The number of biomedical ontologies is too big to allow manual alignment of all of them (the repository NCBO BioPortal stores more than 500 biomedical ontologies). In addition, their size is also very large (e.g, SNOMEDCT, Gene Ontology). Therefore, interconnecting manually all biomedical ontologies is not feasible. However, we can find some reliable manually produced mappings in several resources such as UMLS[6] [3], the OBO Foundry [6] and the NCBO BioPortal[7] [11]. For instance, the OBO Foundry ontology developers produce Xref relations between the concepts of their ontologies(more than 141 ontologies) that can be considered mappings (latter called OBO mappings). As another example CUI (Concept Unique Identifier) mappings that are produced by the US National Library of Medicine team. When an ontology or a terminology is integrated in the UMLS Meta-Thesaurus, a CUI is manually assigned to each concept, grouping concepts together. These manually produced mappings are the formalization of human experts knowledge that we aim to exploit to enhance biomedical ontology matching.

3 Overview of our Approach

Our approach aims to reuse mappings that can be extracted from a repository of ontologies to discover new ones, especially by combining manually and automatically produced mappings. Indeed, we hypothesis that manual mappings may be the bridge that overcomes the limitations of automatic matchers. As we can see in Fig. 1, our approach involves five steps: (1) Extraction of different kinds of mappings between all ontologies stored in the repository to construct the *Global Mapping Graph*, (2) Anchoring the concepts of the source ontology on the resulted graph, (3) Selection of mappings that may help to discover new

[6] Unified Medical Language System

[7] Not all mappings in BioPortal are manually produced, see section 5.1 for more information about NCBO BioPortal mappings

ones using resulted anchors. The selected mappings are organized in the form of a graph called the *Specific Mapping Graph*, (4) Anchoring the concepts of the target ontology on the *Specific Mapping Graph* and extract all paths between the source and target anchors (candidate mappings. Finally (5) Filtering discovered candidates mappings to keep only the most reliable ones according to a given aggregation strategy.

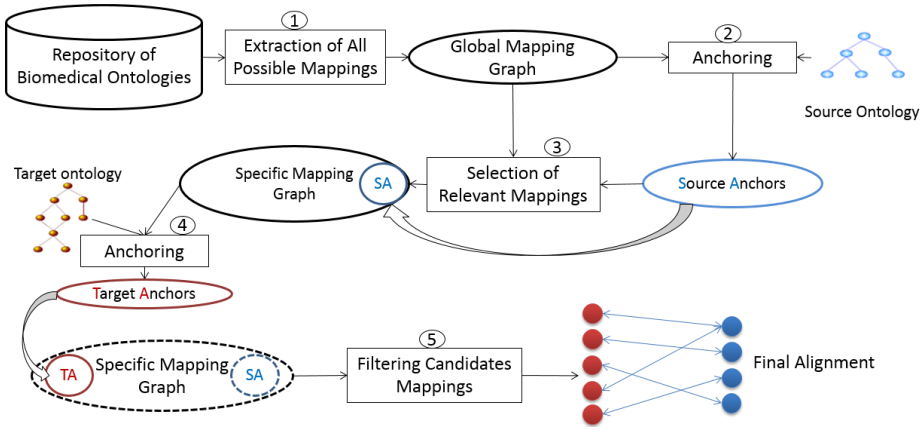

Fig. 1. Overview of the proposed approach

3.1 Building the Global Mapping Graph

In the biomedical domain the most known resources of manually produced mappings are: (i)ontologies produced by the OBO Foundry, (ii)ontologies integrated in UMLS. For a given repository of ontologies, to build the *Global Mapping Graph* we start by checking for each ontology if it is an OBO ontology, or if it is integrated in UMLS. Then, we extract from each one its manually produced mappings (OBO from the first category and CUI mappings from the second one). After that, we use a lexical matcher or any other efficient matcher to match each ontology with all others ontologies in the repository. We add these mappings produced automatically to those produced manually. For each extracted mapping we keep the source and the target concepts, the ontology of each concept, the set of labels of each concept and the provenance of this mapping (OBO, CUI, etc.). We can add any other sets of relevant mappings to enrich the final set of extracted mappings. At the end of the mappings extraction step we obtain a large set of mappings. We merge these mappings to obtain the *Global Mapping Graph* (naturally some mappings have common concepts). We note that this step is done just for once; the *Global Mapping Graph* is an independent resource that can be exploited to match any couple of ontologies. In case of enriching the

repository with a new ontology, we will only extract its related mappings with other ontologies, and adding them to the resulted graph.

3.2 Anchoring Source Concepts

The second step consists in anchoring source concepts on the *Global Mapping Graph*. If the source ontology is stored in the repository, the anchors are the source concepts themselves. Otherwise, the anchors can be found using a lexical matcher on the concept labels between the source ontology and all concepts of the *Global Mapping Graph*. In this case, the mappings returned by the lexical matcher will be the first selected mappings in the *Specific Mapping Graph*. The use of a lexical matcher offers the advantages of being fast (anchoring is a preprocessing stage) and effective in aligning biomedical ontologies [10]. For a given source concept we can get wrong anchors, for that we can imagine to use more sophisticated matchers but this choice could entail higher costs in terms of resources (time and memory). In our approach we propose to let the filter at the end (see section 3.5).

3.3 Selection of the Specific Mapping Graph

This step allows selecting the appropriate fragment from the *Global Mapping Graph* for a given input ontology (Algorithm. 1). For each concept in the list of source anchors, we select its direct mappings in the *Global Mapping Graph* (mappings of different provenance). For each new concept in the *Specific Mapping Graph* , we search for their direct mappings and so on, until no new concept is found. Indeed if a concept A is mapped directly to B, the concept B may be automatically or manually mapped to another concept C that has no mapping with A. Finally, we obtain the *Specific Mapping Graph* which is composed of all concepts related to the source ontology interconnected via selected mappings. It is interesting to note that this *Specific Mapping Graph* is not limited in number of used ontologies, our units are concepts, not ontologies.

3.4 Anchoring Target Concepts

This step is necessary only if the target ontology is not in the initial repository. Otherwise, the anchors are the target concepts themselves. Indeed, if a target concept belongs to a mapping related to the source ontology, this target concept should be already in the resulted *Specific Mapping Graph*. In the same manner (see section 3.2), we can use any efficient lexical matcher to anchor target concepts on *Specific Mapping graph* concepts and add the returned alignment in it.

3.5 Filtering Candidates Mappings

To derive mappings between the source and the target ontologies, we search for all paths between the source anchors and the target anchors in the *Specific*

Algorithm 1 Specific Mapping Graph Selection

Input: $GlobalMappingGraph$, $sourceAnchors$,$MappingsResultedFromAnchoring$
Output: $SpecificMappingGraph$

```
if sourceOntology ∉ BiomedicalOntologyRepository then
  SpecificMappingGraph=MappingsResultedFromAnchoring
end if
for each c ∈ sourceAnchors do
  listConcepts.add(c)
end for
next ← 0
while next < listConcepts.size() do
  x ← listConcepts.get(next)
  Extract S from GlobalMappingGraph: all direct mappings of x
  for each m ∈ S do
    if m ∉ SpecificMappingGraph then
      SpecificMappingGraph.add(m)
    end if
    if m.targetConcept ∉ listConcepts then
      listConcepts.add(m.targetConcept)
    end if
  end for
  next++
end while
return SpecificMappingGraph
```

Mapping Graph. In Fig. 2 we can find an example of paths between the concept (MA:1012) and the concept (NCIT:C32337). One source concept may have several target concepts (several mapping candidates). Indeed, mappings composing the *Specific Mapping Graph*, in particular automatically produced ones, may be not precise (or wrong) which lead to derive wrong mappings. The challenge is to select the most accurate candidate target concept, especially if we deal with 1:1 mappings (searching only for equivalence relationship). In our case, a candidate mapping corresponds to one or several paths linking the same source concept to the same target concept. Paths in Fig. 2 represents a candidate mapping between the concept (MA:1012) and the concept (NCIT:C32337). We have experimented different aggregation strategies (see section 6.2) to select one mapping from several candidates for a given source concept, but these strategies produced a low recall. To improve the quality of the final alignment, we propose a novel measure to select for a given source concept the best mapping from several candidates. This measure is described in the next Section.

4 Path Confidence Measure

We define the type of a given path as a distinct sequence of provenances that forms this path, independently from intermediate concepts. For example, the type of path linking the concept (MA:1012) to the concept (MeSH:D17626) in

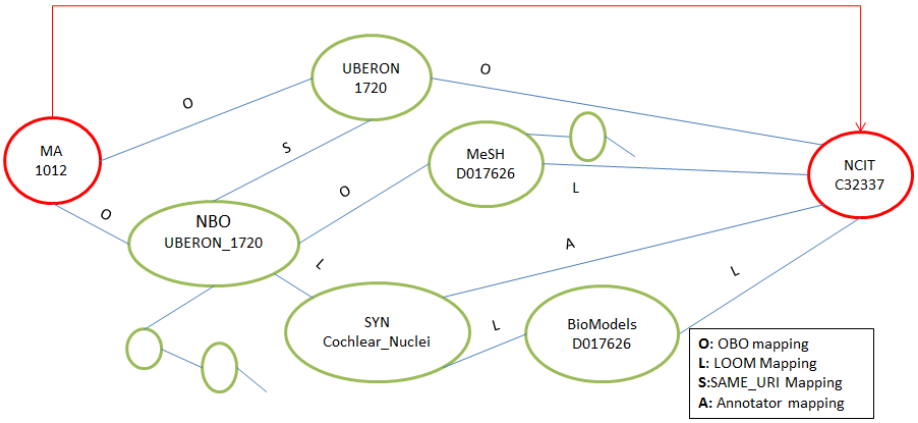

Fig. 2. Extracted mappings from BioPortal for the mouse anatomy concept 1012 (each concept is represented by the acronym of its ontology and its code within BioPortal)

Fig. 2 is OO (OBO_OBO). The types of path linking the concept (MA:1012) to the concept (NCIT:C32337) are: OO, OSO, OLLL, etc.

To enhance the selection of the final mappings, we propose the novel *Path Confidence Measure*(PCM) that takes the confidence value of given path type into account. The confidence value is a score assigned to each path type according to its ability to discover correct mappings. This measure is inspired from the most frequent aggregation strategy (also called popularity in [16]) based on the hypothesis: for a given source concept, the most accurate target concept is the concept that has the highest number of paths linking it to this source concept. In this hypothesis we assume that all path types has the same confidence value. However, the quality of discovered mappings is different from one path type to another. Indeed, some types give better results than the others (see Section. 5). For this purpose, we introduce the confidence value of a given path type as a coefficient to be multiplied by the number of paths of this type. The Path Confidence Measure for a given candidate mapping (C_s, C_t) is defined as the sum of the number of each path type linking C_s to C_t multiplied by its confidence value. We use the log function to avoid the over-estimation of a given candidate mapping due to a large number of a given path type. We add 1 to avoid $log(0)$ and we divide by the max sum to normalize values between 0 and 1. For a given candidate mapping (C_s, C_t), we compute the PCM value of the target concept C_t as follows:

$$PCM(C_s, C_t) = \frac{\sum_{i=1}^{n} log(1 + NP_i * CV_i)}{\max_{j=1}^{m} \sum_{i=1}^{n} log(1 + NP_{ji} * CV_i)}$$

Where n is the number of different types of paths that lead to the target concept C_t from the source concept C_s; NP_i is the number of paths of type i linking C_s to C_t; CV_i is the confidence value of the path of type i; m is the number of

concepts of the source ontology. This measure is proposed only to select for a given source concept, one target concept from several candidates.

5 Implementation

To evaluate our approach, we have implemented it using the reference repository of biomedical ontologies NCBO BioPortal and the ontologies of the Anatomy track from Ontology Alignment Evaluation Initiative 2015[8].

5.1 NCBO BioPortal

NCBO BioPortal is a community based repository. Currently, it is one of the richest repository in the biomedical domain with more than 500 biomedical ontologies. The repository offers a REST web services API.[9] In particular, mappings of different provenances[10] between stored ontologies. In addition of OBO and CUI mappings that we have previously explained, the repository generates automatically other mappings such as LOOM [10], SAME_URI and REST mappings. LOOM mappings are based on close lexical match between preferred names of concepts or a preferred name and a synonym. The lexical match involves removing white-space and punctuation from labels. SAME_URI mappings are based on exact match between the URI of concepts. Finally, REST mappings that are mappings uploaded manually by users of the portal, they represent the minority. In addition, the portal integrates an efficient Annotator [15] which can be used as a lexical matcher. For a given concept label, the Annotator returns a list of concepts that have the same label.

5.2 Anatomy track

The Anatomy track consists in finding an alignment of 1516 mappings between the Adult Mouse Anatomy ontology (2738 concepts)and a part of the NCI Thesaurus (describing the human anatomy 3298 concepts). The task has a good share of non-trivial mappings.

Instead of creating a local repository of biomedical ontologies, we have chosen to use the NCBO BioPortal. Another factor that motivates our choice is the mappings of different provenances that are stored and accessible through its REST API. Consequently, BioPortal can be considered as a huge graph where nodes are concepts and edges are mappings with different provenances. With this vision, BioPortal can play the role of the *Global Mapping Graph* in our approach. Also,the source and the target ontologies of the Anatomy track are already stored in BioPortal, we do not need to anchor concepts (see section 3.2 and 3.4), we can access directly to them using their URI. Consequently, to run our approach, we need just to execute the steps 3 and 5 of the proposed approach to produce the final alignment.

[8] http://oaei.ontologymatching.org/2015/
[9] http://data.bioontology.org/documentation
[10] http://www.bioontology.org/wiki/index.php/BioPortal_Mappings

6 Evaluation

The selection of the *Specific Mapping Graph* step with the mouse anatomy (MA) as a source ontology and the NCBO BioPortal as *Global Mapping Graph* has produced a graph[11] combining 85192 concepts and 368371 mappings of different provenance (see Fig. 2). We have extracted the preferred label of each concept and annotate it using the BioPortal Annotator, because it works with a richest synonym dictionary which allows to discover mappings that the LOOM algorithm does not discover. Indeed, the LOOM algorithm is based only on close lexical match without using any complementary resources. Mappings are extracted in JSON format as we can see in [2], we note that no score is assigned to these mappings, we have just the information about their provenance. It is important to keep this information to be able to explain the provenance of a given derived mapping by the end. The distribution of extracted mappings per provenance is presented in Table.1. As we can see, the number of the annotator mappings is greater than the number of LOOM mappings, this can be explained by the fact that the annotator works only with exact string match whereas LOOM involves some pretreatment such as removing white-space and punctuation from labels.

Provenance of mappings	Number of mappings
LOOM	196225
Annotator	78446
OBO	65305
CUI	17551
SAME_URI	10488
REST	356

Table 1. Number of extracted mappings per provenance

6.1 Evaluation of Paths Types Quality

From the resulted *Specific Mapping Graph*, we have extracted all possible paths between the concepts of the source ontology *MA* and the concepts of the target ontology *NCIt*. Each path represents a candidate mapping that may be true or false according to the reference alignment provided by OAEI2015. We have computed the true positive mappings (mappings present in the reference alignment) and the false positive mappings (mappings absent in the reference alignment) for each type of path. Using these parameters, we have computed the precision, recall and F-Score for each type of path. Fig. 3 represents the top 50 path types ranked according to the F-Score measure. Based on the obtained results, we can

[11] We have created the graph using the graph database Neo4J (https://neo4j.com/)

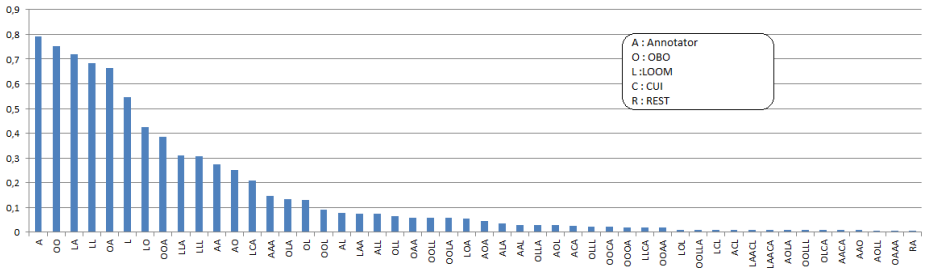

Fig. 3. Fscore per type of path

conclude that the best paths are the shortest ones: direct matching (paths of type A and L), paths with two steps; one mediator concept (OO,LA,LL,OA,LA) and paths with three steps; two mediator concepts (OOA,LLA,LLL). We note that the combination of manually and automatically produced mappings provides a good results(e.g., LA,OOA,LA). The longest paths return a few mappings candidates, and generally wrong ones (see Fig. 4).

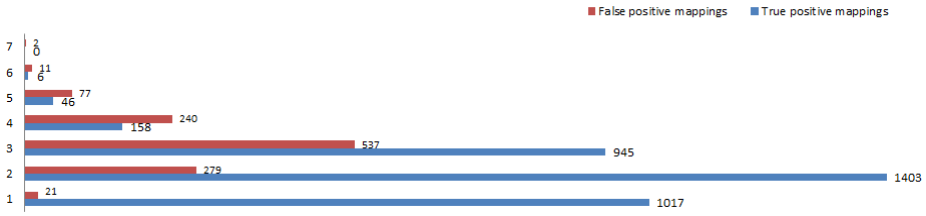

Fig. 4. True positive/False positive mappings per length of paths (number of steps)

According to this study, we have chosen to use the F-Score of each path's type as its confidence value to asses its ability to discover true positive mappings.

6.2 Evaluation of Final Alignment Quality

In order to evaluate the quality of the *Specific Mapping Graph*, we have compared mappings derived from it (mappings linking MA concepts to NCIt concepts) to the reference mappings of the Anatomy track. First of all, we have evaluated all mappings derived from the *Specific Mapping Graph* without any aggregation strategy. Then, we have experimented three strategies to select only one target concept for each source concept: (i) the first found; i.e. the final node of the shortest path leading to the target ontology (ii) the most visited target concept; it is the concept of the target ontology that has the highest number of paths from a given source concept and (iii) the target concept that has the greatest PCM score (path's type F-Score as confidence value). Then, we have compared

the alignment produced by our approach to the final alignments of the four top systems in OAEI 2015 [4] for the Anatomy track. The results presented in Table 2 show that our final alignment is competitive with top alignment systems. Without any strategy of aggregation, our final alignment has the best precision but relatively a low recall, what gives it the worst F-Score. However, the use of any aggregation strategy improve the recall, and lets our final alignment having the second position after AML system. We note that AML and LogMapBio [14] systems use already biomedical ontologies as BK. Also, AML implement several features that help improving the final alignment. The best F-Score is obtained using the PCM measure for the selection of final mappings. Indeed, the proposed measure promotes paths with high confidence.

Systems		Mappings	Correct	Incorrect	Precision	Recall	F-Score
Resulted BK	All mappings	2247	1416	831	**0,934**	0,630	0,753
	First found	1504	1366	138	0,901	0,909	0,905
	Most frequent	1504	1372	132	0,905	0,912	0,909
	PCM	1503	1395	108	0,920	**0,928**	**0,924**
AML		1477	1412	66	**0,931**	**0,956**	**0,944**
LogMapBio		1549	1366	183	0,901	0,882	0,891
LogMap		1397	1282	115	0.846	0,918	0,88
XMAP		1414	1312	102	0,865	0,928	0,896

Table 2. Quality evaluation of the discovered mappings

6.3 Specific Mapping Graph: Usefulness Evaluation

The mapping gain [8] is a measure proposed to asses the usefulness of a BK for a given task of alignment. It measures how many new mappings have been found in an alignment A thanks to a given BK comparing to another alignment B. For clarity, we recall here the formula of this measure. Given two alignments A and B between ontologies S and T, the mapping gain between A and B is defined as the fraction of mappings in A that are not in B.

$$MG(A,B) = Min(\frac{C_s(A \cap \neg B)}{C_s(B)}, \frac{C_t(A \cap \neg B)}{C_t(B)})$$

Where C_s and C_t denote respectively the sets of concepts in the alignments (A and B) and belong respectively to the source and the target ontologies.

To evaluate the usefulness of the *Specific Mapping Graph* as a BK, we have computed the mapping gain using the previous formula replacing A by our final derived alignment (with PCM) and B by one of alignments produced by the four top systems in the OAEI 2015[12] (see Table 3).

[12] http://oaei.ontologymatching.org/2015/results/anatomy/index.html

Systems	# Absent concepts of MA	# Absent concepts of NCIT	Mapping Gain
AML	77	195	5%
LogMapBio	134	247	9%
XMAP	188	302	13%
LogMap	218	337	16%

Table 3. Mapping gain using resulted BK

Based on analysis done in [8], the authors conclude that if the use of a BK provides a mapping gain greater than 2%, the BK could be considered as useful. According to that, the *Specific Mapping Graph* is useful for all these systems (state-of-the-art alignment systems). We can observe that the resulted BK is significantly useful for XMAP and LogMap because they do not use any biomedical ontologies as a BK. The other systems already use biomedical ontologies as a BK. AML uses three ontologies (Uberon, DOID and Mesh) which represents 292 591 concepts. LogMap uses top ten ontologies returned by the algorithm presented in [5]. The first ontology returned by this algorithm is SNOMEDCT which contains 324129 concepts. In the last both cases we observe the large number of concepts comparing to the *Specific Mapping Graph*'s concepts number (85192 concepts). We observe also that even if AML and LogMap use a biomedical BK, the *Specific Mapping Graph* allows to enhance their results. Table 4 presents the number of reference mappings found by our approach, missed by the other systems.

AML	LogMapBio	XMAP	LogMap
20	87	161	133

Table 4. Mappings found by our approach, missed by top alignment systems

7 Related Work

The selection of the appropriate BK to enhance biomedical ontology matching is an active research issue. Several approaches have been proposed to address it. To avoid the complexity of an automatic selection, many approaches usually manually select the relevant BK. For examples, WordNet is used in [20], DOLCE in [17]. The manual selection does not guarantee the enhancement of a given task of alignment, and requires a wide range of knowledge. For this purpose, several automatic approaches have been defined to select the appropriate BK as those described in [19, 18]. The most similar work to this paper is done in [12]. Their approach consists in aligning the source and the target ontologies with each ontology in a set of intermediate ontologies. Then, compose the different produced alignments to derive mappings between source and target ontologies. The authors do not extract manually produced mappings and they do not extract

mappings between intermediates ontologies. Using their approach, one can derive only mappings with one mediator concept (paths of two steps only). In the same manner [5] propose to compose mappings after selecting dynamically five ontologies from BioPortal. However, and as we can see in Fig. 4, paths of length three (two mediator concepts) and four (three mediator concepts) return many reference mappings. For example, 945 reference mappings are returned by three-step-paths. This can be the explanation of the high F-Score obtained by our approach (0.928) comparing to the F-Score obtained in their experimentation (0.847 and 0.913 respectively).

Recently, other measures have been proposed to select the most appropriate set of ontologies (which represents the BK) as the effectiveness [13] and the mapping gain [8] measures. The drawback of the proposed measures resides in the fact that they select the whole ontologies (many thousands of concepts) even if we need just for a fragment from these ones. Furthermore, dealing with whole ontologies makes it necessary to limit the number of selected ontologies. In our approach, there is no limitation of the number of selected ontologies, our units are concepts. We select only concepts that may help us to discover new mappings without considering the number of used ontologies. In [8] the selection is based on the mapping gain score. The ontologies with a low mapping gain (less then the defined threshold) are eliminated even if they contain some concepts that may help to discover reliable mappings. In our case, we do not select specific ontologies but we work with all ontologies in the repository at the same time. We propose to follow mappings of different provenances, and select progressively potential useful concepts. Therefore, we combine the lexical overlapping with the human knowledge from mappings produced manually without eliminating any candidate mediator concept.

Furthermore, in all other approaches, the selection and the combination of different ontologies is based only on mappings produced automatically, they do not distinguish different types of mappings (different provenances). They are based mainly on the lexical overlapping between the BK and ontologies to be aligned. This criteria does not guarantee the selection of the best BK. For instance, the huge biomedical ontology SNOMED-CT with its rich lexical content may always be ranked first to match biomedical ontologies, even if more appropriate BK are available as Uberon for Anatomy in [5]. The use of SNOMED-CT needs more resources, memory to manage the whole ontology and time to anchor concepts on it.

Moreover, the *Specific Mapping Graph* could be reused as a resource to map the source ontology with any other ontology. If a new ontology is added to the initial repository, we just need to extract its related mappings with the concepts in the *Specific Mapping Graph* and integrate them. In the previous approaches, one will need to restart the selection process from scratch. The probability of not finding an anchor for a given concept in a rich repository of biomedical ontologies as NCBO BioPrtal (8150126 concepts) is very low. In this case, we can search on the web for ontologies that may contain this concept as proposed in [1] and [18].

8 Conclusion and Future Work

This paper deals with the selection and the combination of heterogeneous existing mappings, produced manually and automatically, stored in a biomedical repository, to discover new ones. Our approach is based on building the *Specific Mapping Graph* as a BK. Such graph allows to get an alignment of high quality between ontologies to be aligned without using complex lexical and structural measures. One source concept may have several candidates target concepts. To select the most accurate one, we have proposed the *Path Confidence Measure* that takes the confidence of a given path type into account.

The presented evaluation shows that our approach provides good results, competitive to those of state-of-the-art systems. Also, that the reuse of existing mappings allows discovering mappings missed by the previous approaches.

The explanation of final mappings is one of challenges of ontology matching [21]. Indeed, it is very important to be able to justify the provenance of a given mapping instead of a simple score. In our approach, each found mapping is deducted from one or several paths. The edges of paths are tagged with their provenance. Consequently, all found mappings are explained.

Moreover, we have evaluated our approach using one benchmark (Anatomy benchmark). For a better evaluation, we will evaluate it on other OAEI biomedical benchmarks. Also to improve the quality of the final alignment, we plan to study the impact of the variation of the PMC threshold on the F-Score, currently no threshold is applied. Also, the coherence of automatically produced BioPortal mappings has been critiqued in [9]. For this purpose, we plan to integrate a semantic verification into our approach to improve the quality of produced alignment. Currently our approach is used to derive only 1:1 mappings. We will experiment the usefulness of our method to derive n:m mappings. This will be possible if we extract not only mappings but also fragments of ontologies (sequence of concepts linked with is_a relationship) that connect two concepts in the *Specific Mapping Graph* if they belong to the same ontology.

9 Acknowledgment

This work was achieved during a LIRMM-ESI collaboration within the SIFR project funded in part by the French National Research Agency (grant ANR-12-JS02-01001), as well as by University of Montpellier, the CNRS and the EU H2020 MSCA program.

References

1. Zharko Aleksovski, Warner Ten Kate, and Frank Van Harmelen. Exploiting the structure of background knowledge used. In *1st International Conference on Ontology Matching-Volume 225*, pages 13–24, 2006.

2. Amina Annane, Vincent Emonet, Faiçal Azouaou, and Clement Jonquet. Multilingual mapping reconciliation between english-french biomedical ontologies. In *6th International Conference on Web Intelligence, Mining and Semantics, WIMS*, pages 13:1–13:12, 2016.
3. Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.
4. Michelle et al. Cheatham. Results of the ontology alignment evaluation initiative 2015. In *10th ISWC workshop on ontology matching*, pages 60–115, 2015.
5. Xi Chen, Weiguo Xia, Ernesto Jiménez-Ruiz, and Valerie V. Cross. Extending an ontology alignment system with bioportal: a preliminary analysis. In *ISWC*, pages 313–316, 2014.
6. Barry Smith et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.
7. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer, 2013.
8. Daniel et al. Faria. Automatic background knowledge selection for matching biomedical ontologies. *PloS one*, 9(11):e111226, 2014.
9. Daniel et al. Faria. Towards annotating potential incoherences in bioportal mappings. In *ISWC*, pages 17–32. Springer, 2014.
10. Amir Ghazvinian, Natalya Fridman Noy, Mark A Musen, et al. Creating mappings for ontologies in biomedicine: simple methods work. In *AMIA*, pages 198–202, 2009.
11. Amir et al. Ghazvinian. What four million mappings can tell you about two hundred ontologies. In *International Semantic Web Conference*, pages 229–242. Springer, 2009.
12. Anika Gross, Michael Hartung, Toralf Kirsten, and Erhard Rahm. Mapping composition for matching large life science ontologies. In *ICBO*, pages 109–116, 2011.
13. Michael Hartung, Anika Gross, Toralf Kirsten, and Erhard Rahm. Effective mapping composition for biomedical ontologies. In *ESWC*, pages 176–190, 2012.
14. Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Alessandro Solimando, and Valerie V. Cross. Logmap family results for OAEI 2015. In *10th International Workshop on Ontology Matching*, pages 171–175, 2015.
15. Clement Jonquet, Nigam Shah, and Mark Musen. The open biomedical annotator. In *AMIA summit on translational bioinformatics*, pages 56–60, 2009.
16. Angela Locoro, Jérôme David, and Jérôme Euzenat. Context-based matching: design of a flexible framework and experiment. *Journal on data semantics*, 3(1):25–46, 2014.
17. Viviana Mascardi, Angela Locoro, and Paolo Rosso. Automatic ontology matching via upper ontologies: A systematic evaluation. *IEEE Transactions on knowledge and data engineering*, 22(5):609–623, 2010.
18. Christoph Quix, Pratanu Roy, and David Kensche. Automatic selection of background knowledge for ontology matching. In *International Workshop on Semantic Web Information Management*, page 5. ACM, 2011.
19. Marta Sabou, Mathieu dAquin, and Enrico Motta. Exploring the semantic web as background knowledge for ontology matching. In *Journal on data semantics XI*, pages 156–190. Springer, 2008.
20. Brigitte Safar, Chantal Reynaud, and François Calvier. Techniques d'alignement d'ontologies basées sur la structure dune ressource complémentaire. *1ères Journées Francophones sur les Ontologies*, pages 21–35, 2007.
21. Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.