Exploiting the structure of background knowledge used in ontology matching

Zharko Aleksovski[1,2], Warner ten Kate[1], and Frank van Harmelen[2]

[1] Philips Research, Eindhoven,
[2] Vrije Universiteit, Amsterdam

Abstract. We investigate the use of a background knowledge ontology in ontology matching. We conducted experiments on matching two medical ontologies using a third extensive one as background knowledge, and compare the results with directly matching the two ontologies. Our results indicate that using background knowledge, in particular the exploitation of its structure, has enormous benefits on the matching. The structure of the background ontology needs closer examination to determine how to use it in order to obtain maximal benefit.

1 Introduction

The problem of ontology alignment (also known as ontology integration, semantic integration, ontology mapping, etc.) plays a central role in the development of knowledge based systems. New technologies such as Semantic Web make it easier to use ontologies in the information systems. These trends have driven the development of new ontologies, which in turn has resulted in an increasing amount of ontologies becoming available in the recent years. Essential to an ontology is its reusability, which implies one needs to integrate it into the system using it. Problem arises if the ontology to be integrated uses a different vocabulary from the system using it.

This problem initiated a lot of research on ontology matching lately, see [1–3]. Various approaches have been developed. They mainly focus on two aspects: lexically matching the elements of the ontologies, and using the structure of the ontologies. The first uses string-based and linguistic methods to detect relatedness between elements based on string similarity of their labels, and the second uses the relations within the ontologies to detect similarities. Elements in the ontologies that are related but have neither lexical nor structural similarity remain undetected. Motivated by this issue, we focused on using background knowledge. We followed the intuition that a background ontology which comprehensively describes the domain of the source and target ontologies will provide a way to find matches missed by other approaches.

In earlier papers, we showed that the use of background ontology can compensate for lack of structure and lexical overlap, and increasing the amount of background knowledge (multiple ontologies) improves the matching result, see [4, 5]. In this work we investigate the benefits and problems of using a *comprehensive* domain ontology as background knowledge. We conducted experiments of matching one medical ontology to another, while using a much larger and detailed ontology of the same domain as background knowledge. The results of our experiments confirmed that the background knowledge can significantly boost the performance of the matching process. In particular, maximal benefit is achieved when combining different pieces of knowledge within the background knowledge. However, these pieces need careful consideration when combining them.

The paper is organized as follows: in Section 2 we describe the general scheme of our approach, that is how we use an ontology as background knowledge in ontology integration. In Sections 3 and 4 we describe a case study and a set of experiments to test our expectations. In Section 5 we report on the results of these experiments with evaluation on validity. In Section 6 we discuss representative matches of the different experiments, and Sections 7 and 8 conclude the paper with discussion on related and future work.

2 Our approach: Using a background knowledge ontology in ontology matching

In our approach we match two ontologies using a third as background knowledge. We call the ontologies being matched the source and target, see Figure 1. We make use of the background knowledge by first relating the concepts from the source and target ontology to the background knowledge, and then checking if these concepts are related. Hence, this process proceeds in two steps: anchoring and deriving relations.

Anchoring is matching the source and target concepts to the background knowledge. In general, this process can be performed by using an existing ontology matching technique. Besides the concept's labels one can also use the structure of the ontologies. In the anchoring we are not only interested in finding the corresponding equivalent concepts. As we will see in our experiments later, other kind of relatedness with the concepts in the background knowledge can be useful as well.

Deriving relations is the process of discovering relations between source and target concepts by looking for relations between their anchored concepts in the background knowledge. Both the source and target concept's anchors are part of the background knowledge, and checking if they are related means using the reasoning service in the background knowledge ontology. Combining the anchor relations with the relations between the background knowledge concepts derives the relation between source and target concepts, which is what we are looking for.

To explain this process in the context of medical ontologies, a realistic example is the following: the source concept $\mathcal{SRC}$*:Brain* is anchored to background knowledge concept $\mathcal{BK}$*:Brain*, and the target concept $\mathcal{TAR}$*:Head* is anchored to a background knowledge concept $\mathcal{BK}$*:Head*. The background knowledge reveals a relation $\mathcal{BK}$*:Brain part-of* $\mathcal{BK}$*:Head*, and we derive a relation that source concept $\mathcal{SRC}$*:Brain* has a narrower meaning than the target concept $\mathcal{TAR}$*:Head*. Using background knowledge was crucial in this case; the match was not found by directly matching the source to the target ontology, $\mathcal{SRC}$*:Brain* is classified under $\mathcal{SRC}$*:Central nervous system* which is in no way related to the concept $\mathcal{TAR}$*:Head*.

As the example suggests, of particular interest in our approach is exploiting the structure of the background knowledge ontology. It is done in the deriving relations step, when checking for relatedness between the anchored concepts in the background

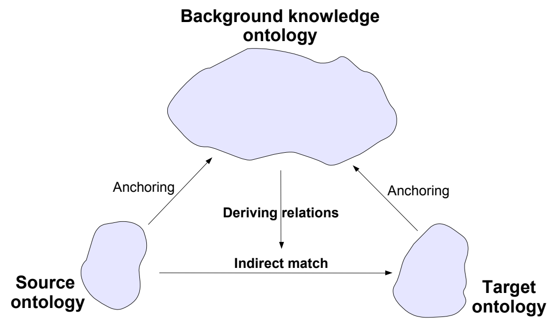

Fig. 1. Matching source to target ontology using background knowledge.

knowledge ontology. Now we introduce the formal definitions of all the components in this scheme, which we will use in the rest of the paper.

Concept is a class of things grouped together due to some shared property. It is named with one or more labels which are synonymous to each other. Besides the labels, concept is also determined by its relations to other concepts. We will refer to a concept in two ways: with capital italic letters X, Y... when referring to an arbitrary concept, $X^{\mathcal{O}}$ or $X^{\mathcal{ONT}}$ when referring to a concept from a specific ontology, and by using its label, for example *Temporal lobe*, and $\mathcal{ONT}$:*Temporal lobe* or $\mathcal{O}$:*Temporal lobe* when referring to a concept from a specific ontology.

Relation is a triple $(X, \mathit{relation}, Y)$, where X and Y are concepts, and $\mathit{relation} \in T$, where T is the set of all types of relations. We will also write it as $X \sim Y$ with "$\sim$" being the symbol of the relation. Examples of relations used in this paper are: $X \equiv Y$, - the two concepts have the same meaning, and $X \preceq Y$ (with inverse: $\succeq$), also written as X *is-narrower-than* Y - the first has narrower meaning than the second. Other relations are used in the existing ontologies as well, see Section 4. Relations can be established between concepts from the same and also from different ontologies.

Ontology is a pair of sets: $\mathcal{ONT}(C, R)$. C is a set of concepts, R is the set of relations among these concepts. We will refer to an ontology using shortened form of its name written in calligraphic letters, like $\mathcal{ONT}$. When referring to C or R of a specific ontology, we will write them as $C^{\mathcal{ONT}}$ and $R^{\mathcal{ONT}}$

Ontology match is a function of two ontologies that returns a set of relations between their concepts:

$$f : (\mathcal{SRC}, \mathcal{TAR}) \to \{(X, \mathit{relation}, Y) | X \in C^{\mathcal{SRC}}, \mathit{relation} \in T, Y \in C^{\mathcal{TAR}}\} \quad (1)$$

Specific types of ontology matches of interest to our approach are the following two: **Anchorings** are two ontology matches from the source and target ontology to the background knowledge respectively, and **Deriving relations** is an ontology match between the source and the target ontology which is an indirect matching that uses their anchors to the background knowledge, and the background knowledge itself.

3 Our case study

Hypothesis: Using comprehensive domain ontology as background knowledge can significantly boost the performance of an ontology matching process.

To test this hypothesis and investigate the problems that occur when our matching scheme is used in practice, we conducted a set of experiments matching existing ontologies available on the Semantic Web. We matched the anatomy parts of CRISP and MeSH using the FMA ontology as a background knowledge. CRISP and MeSH were choosen randomly, and FMA because it extensively covers the anatomy domain.

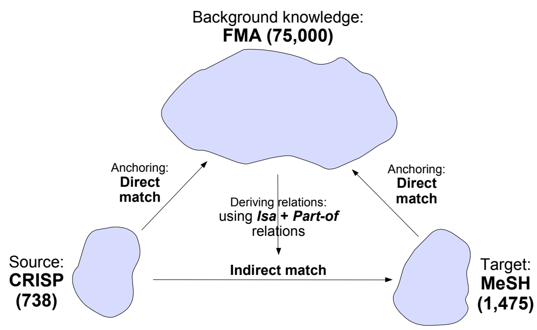

Fig. 2. Matching CRISP to MeSH ontology using FMA as background knowledge.

The test data used in the experiments.

Source ontology: **CRISP**[3] (Computer Retrieval of Information on Scientific Projects) is a biomedical document classification system. It contains 738 concepts organized in a hierarchy. The relations in its hierarchy are established based on the classified document sets. The hierarchy contains two relations: *broader-than* and its inverse *narrower-than*, meaning superset and subset between the corresponding document sets respectively. In our experiments we used the part of CRISP describing anatomy.

Target ontology: **MeSH**[4] is the National Library of Medicine's controlled vocabulary thesaurus intended for classification of documents. The part which we used in the experiments is the anatomy sub hierarchy. It contains is 1475 concepts, and is based on *broader-than* and its inverse *narrower-than* relations, the same as CRISP.

Background knowledge ontology: **FMA**, as stated in its description[5]: "The Foundational Model of Anatomy is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy." The version of FMA used in our experiments dates from the end of 2005. It contains 75000 concepts interconnected with around 160 different relation types. We used the main two hierarchies: *isa* and *part-of*.

[3] http://crisp.cit.nih.gov/
[4] http://www.nlm.nih.gov/mesh/
[5] http://sig.biostr.washington.edu/projects/fm/AboutFM.html

Testing our hypothesis by conducting experiments.
We performed five experiments matching CRISP to MeSH. In the first we matched them directly, and in the other four we matched them indirectly using the FMA ontology as background knowledge. The direct matching served as a baseline, against which we compared each of the indirect matchings. With the intention to distile the added value of using background knowledge, we analyzed the additional matches discovered by indirect matching. Furthermore, there were some cases of matches found in the direct and not in the indirect matching. We discuss these in Section 6.

4 The experiments

We performed five experiments. In Experiment 1 we matched CRISP to MeSH directly and in the other four we matched them indirectly using FMA as background knowledge. In Experiment 2 we exploited each *isa* and *part-of* relation in FMA separately; in Experiment 3 we used *isa* and *part-of* relations with their transitive closures; in Experiment 4 we used *isa* and *part-of* combined, and in Experiment 5 we restricted to specific combinations of them to induce the matches. The result of each experiment was an ontology match between CRISP and MeSH concepts, using one of the three relations: $\equiv$ (*equivalent*), $\preceq$ (*narrower-than*), $\succeq$ (*broader-than*). Now, we explain the direct and then the indirect matchings in detail.

Experiment 1: Direct matching was performed in two steps: lexical and structural. In the lexical phase we matched CRISP to MeSH using the concept's labels. We cleaned the labels of interpunction, general words like *the*, *of*, *and* etc., we accounted for word order and singular/plural forms of the same words. When matching two concepts X and Y, we concluded $X \equiv^d Y$ if a pair of their labels matches[6]. Further, we used partial matches as well, if X has a label consisting of a superset of words of a label of Y we concluded $X \preceq^d Y$, and analogously $X \succeq^d Y$ if X has label of subset of the words of a label of Y. In other words, we used the partial lexical matches following the intuition that additional words in a label additionally constrain the meaning of that concept. This way, for example, we concluded that $\mathcal{CSP}$*: Mesenteric artery* $\preceq^d$ $\mathcal{MSH}$*: Artery*.

In the structural phase of direct matching we used the structure of CRISP and MeSH to further induce matches by combining the relations from CRISP and MeSH and the lexical matches. For example, from the two relations:

- $\mathcal{CSP}$: *Brain* $\equiv^d$ $\mathcal{MSH}$: *Brain*
- $\mathcal{MSH}$: *Brain* $\succeq$ $\mathcal{MSH}$: *Temporal lobes*

we can induce the relation

- $\mathcal{CSP}$: *Brain* $\succeq^d$ $\mathcal{MSH}$: *Temporal lobes*

We extended the set of lexical matches with the matches implied by the structure of CRISP and MeSH. The following rules were used to extend the result set[7]:

[6] The small d letter in the right upper corner means that the relation is a direct match, letter a means it is an anchor relation, and letter i means that it is an indirect match.
[7] $X^{\mathcal{C}}$, $X^{\mathcal{M}}$, $X^{\mathcal{F}}$ stand for an arbitrary concept from CRISP, MeSH and FMA respectively

- if $(X^{\mathcal{C}} \preceq^d Y^{\mathcal{M}}) \wedge (Y^{\mathcal{M}} \preceq Z^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \preceq^d Z^{\mathcal{M}})$
- if $(X^{\mathcal{C}} \preceq Y^{\mathcal{C}}) \wedge (Y^{\mathcal{C}} \preceq^d Z^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \preceq^d Z^{\mathcal{M}})$
- if $(X^{\mathcal{C}} \succeq^d Y^{\mathcal{M}}) \wedge (Y^{\mathcal{M}} \succeq Z^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \succeq^d Z^{\mathcal{M}})$
- if $(X^{\mathcal{C}} \succeq Y^{\mathcal{C}}) \wedge (Y^{\mathcal{C}} \succeq^d Z^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \succeq^d Z^{\mathcal{M}})$

These rules also used $\equiv$ relations, where $X \equiv Y$ was considered as $X \preceq Y$ and $X \succeq Y$. The rules were exhaustively applied on the result set.

Indirect matching followed the scheme that we described in Section 2. It was performed in two steps: first anchoring CRISP and MeSH to FMA, and then deriving relations between CRISP and MeSH using FMA as background knowledge, Figure 2.

In the anchoring we used the same direct matching technique described for matching CRISP to MeSH directly. Both CRISP and MeSH were anchored to FMA. The result was set of matches with three different kinds of relations: $X \equiv^a Y$, $X \preceq^a Y$, $X \succeq^a Y$, where Y is a concept from FMA, and X is in CRISP or MeSH.

When deriving the relations we used the following rules:

- if $(X^{\mathcal{C}} \preceq^a Y^{\mathcal{F}}) \wedge (Y^{\mathcal{F}} \preceq Z^{\mathcal{F}}) \wedge (Z^{\mathcal{F}} \preceq^a Q^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \preceq^i Q^{\mathcal{M}})$
- if $(X^{\mathcal{C}} \succeq^a Y^{\mathcal{F}}) \wedge (Y^{\mathcal{F}} \succeq Z^{\mathcal{F}}) \wedge (Z^{\mathcal{F}} \succeq^a Q^{\mathcal{M}})$ induce $(X^{\mathcal{C}} \succeq^i Q^{\mathcal{M}})$

where we used the relations *isa* and *part-of* for $\succeq$, and their inverse *has-kind* and *has-part* for $\preceq$. However, in FMA there are no *broader-than* and *narrower-than* relations, but their specializations: *isa* and *part-of* with their inverses *has-kind* and *has-part*. We conducted four different experiments of indirect matching while using FMA as a background knowledge. The experiments differ in the way *isa* and *part-of* relations from FMA were used and combined when deriving *broader-than* and *narrower-than* relations which are then used in the two rules stated above to derive the indirect matches between CRISP and MeSH concepts.

Experiment 2: Indirect matching by using FMA *isa* and *part-of* relations without transitive closure. We induced a relation between the FMA concepts if they were directly related with *isa* or *part-of* relation. We used the following rules:

- $(X^{\mathcal{F}}\ \textit{isa}\ Y^{\mathcal{F}})$ induce $(X^{\mathcal{F}} \preceq Y^{\mathcal{F}})$
- $(X^{\mathcal{F}}\ \textit{part-of}\ Y^{\mathcal{F}})$ induce $(X^{\mathcal{F}} \preceq Y^{\mathcal{F}})$

When a relation $X^{\mathcal{F}} \preceq Y^{\mathcal{F}}$ was induced, we added its semantic equivalent $Y^{\mathcal{F}} \succeq X^{\mathcal{F}}$ as well. We did this in all the indirect matching experiments.

Experiment 3: Indirect matching by using FMA *isa* and *part-of* relations with their transitive closures. Relation between two FMA concepts was induced when they were related with the transitive closure of *isa* or *part-of* relations. We used the following rules:

- $(X_1^{\mathcal{F}}\ \textit{isa}\ X_2^{\mathcal{F}}\ \textit{isa} \ldots \textit{isa}\ X_n^{\mathcal{F}})$ induce $(X_1^{\mathcal{F}} \preceq X_n^{\mathcal{F}})$
- $(X_1^{\mathcal{F}}\ \textit{part-of}\ X_2^{\mathcal{F}}\ \textit{part-of} \ldots \textit{part-of}\ X_n^{\mathcal{F}})$ induce $(X_1^{\mathcal{F}} \preceq X_n^{\mathcal{F}})$

Experiment 4: Indirect matching by using the transitive closure of FMA *isa* and *part-of* relations combined. In this experiment we completely merged *isa* and *part-of* relations and then used the transitive closure of the resulting relation. We used one single inference rule:

– $(X_1^{\mathcal{F}}\ \mathit{rel}^1\ X_2^{\mathcal{F}}\ \mathit{rel}^2 \ldots \mathit{rel}^{n-1}\ X_n^{\mathcal{F}})$ where $\mathit{rel}^i \in \{\mathit{isa}, \mathit{part\text{-}of}\}$ induce $(X_1^{\mathcal{F}} \preceq X_n^{\mathcal{F}})$

After analyzing the results of Experiment 4 it appeared that false positive matches were created due to using *isa* relation before *part-of* in the process of inducing matches, see Section 6 for clarification. To overcome this negative effect we conducted the next experiment.

Experiment 5: Indirect matching by using the transitive closure of FMA *isa* and *part-of* relations without using *isa* before *part-of*. What we did in this experiment was avoiding the use of *isa* relation before *part-of*. We used one single inference rule:

– $(X_1^{\mathcal{F}}\ \mathit{part\text{-}of}\ X_2^{\mathcal{F}} \ldots X_{k-1}^{\mathcal{F}}\ \mathit{part\text{-}of}\ X_k^{\mathcal{F}}\ \mathit{isa}\ X_{k+1}^{\mathcal{F}} \ldots X_{n-1}^{\mathcal{F}}\ \mathit{isa}\ X_n^{\mathcal{F}})$ induce $(X_1^{\mathcal{F}} \preceq X_n^{\mathcal{F}})$

5 Results and evaluation

We present now the results of the experiments. First we explain the numbers presented in the tables, then we interpret and explain their meaning, and finally we provide evaluation on the results.

An important issue in presenting the matching results is that in one set of matches many of them may be implied by the others, in combination with the structure of the ontologies. For example, all the concepts in CRISP are found more specific than the root concept in MeSH, whereas having equivalence between the two root concepts already implies all those matches. Similarly, having a match between two concepts contains implicit knowledge about their sub and super-concepts. To make a fair trade-off between the two cases of having all the possible matches and having only the minimal set of matches that implies all the rest, we decided on a result set that is in between.

In each matching experiment we did the following: We started from the set of all matches, including the implied. For each source concept we took the set of all its matches, and then minimized that set by discarding the matches which are implied by the rest of the set. The minimal set is not sensitive to the order of discarding the implied matches. The union of these minimized sets was the final result. This trade-off matching set extracts the minimal knowledge from the matching result for each of the source concepts separately.

5.1 Results of direct and indirect matchings

In the anchoring phase we matched CRISP and MeSH to FMA directly. The results are shown in Figure 3. The equivalence relations were established as 1-1 matches, while narrower-than and broader-than as many to many. Looking for equivalences only already produced successful anchoring: 65.5% of CRISP and 70.6% of MeSH concepts were anchored to their equivalent concepts in FMA. This success comes from the richness of FMA. On the other hand, for many there were no equivalent concepts in FMA because of disagreement on the coverage of anatomy domain. In CRISP there is a concept $\mathcal{CSP}$*: Muscle movement* which is not an anatomical part of the human body, and as such does not exist in FMA. Still, as shown in the last column on Figure 3, nearly 99% of the concepts from both CRISP and MeSH were anchored due to using the structure

	Anchoring concepts	$\equiv$	$\preceq$	$\succeq$	Anchored concepts
Anchoring CRISP to FMA	738	483 (65.5%)	607	1,474	730 (98.9%)
Anchoring MeSH to FMA	1,475	1,042 (70.6%)	1,545	2,227	1,462 (99.1%)

Fig. 3. Anchoring CRISP and MeSH to FMA

of CRISP and MeSH. For example, $\mathcal{CSP}$*: Muscle movement* was anchored as *narrower-than* $\mathcal{FMA}$*: Muscle* because within CRISP it is *narrower-than* $\mathcal{CSP}$*: Muscle*.

Figure 4 summarizes the results of the five experiments. Comparing the indirect to the direct matching, the indirect matchings found many more *narrower-than* and *broader-than* relations than the direct matching. It appeared that the concepts in CRISP and MeSH can be related in many more ways which can not be found by using only the structure of these ontologies alone. In our case FMA contributed the missing knowledge which resulted in such an improvement over the direct matching.

Matches of CRISP to MeSH	$\equiv$	$\preceq$	$\succeq$	$\equiv + \preceq + \succeq$	increase
Exper. 1: Direct	448	417	156	1,021	
Exper. 2: Indir. *isa* and *part-of*	395	516	405	1,316	29%
Exper. 3: Indir. *isa* and *part-of* closure	395	933	1,402	2,730	167%
Exper. 4: Indir. *isa* and *part-of* mixed and closure	395	1,511	2,228	4,143	306%
Exper. 5: Indir. *isa* and *part-of* isa only after part-of	395	972	1,800	3,167	210%

Fig. 4. Matching CRISP to MeSH directly and indirectly

The last column in Figure 4 shows the increase of amount of matches of the indirect matching when compared to the direct matching. The indirect matching of Experiment 2 produced 29% more matches than the direct matching. So, using only the direct *isa* and *part-of* relations between the concepts in FMA already outperformed the direct matching. When using the transitive closure of *isa* and *part-of* (Experiment 3) we obtained increase of 167%, or nearly 2.7 times more matches than the direct matching. When arbitrarialy mixing *isa* and *part-of* with their transitive closure we got increase of 306%, or 4 times more matches than the direct. The fifth experiment, when combining the *isa* and *part-of* in a resticted way, there was an increase of 210% which is 3.1 times more matches than the direct. It produced 26% less matches than the fourth, and 19% more than the third experiment.

These numbers show that using background knowledge produces enormously more matches than direct matching. Without combining the relations within the background knowledge it is already better than the direct matching, then combining the relations in the background knowledge produces much more matches, and combining different relations within background knowledge produces the maximal number. Of course, these numbers do not say anything about the quality of these matches. This will be discussed in the next section. In particular, if the relations are combined arbitrarily then there is big increase in the amount of matches but also false positive matches are created, but when combining them in a specific way we retain the precision of the matches while again considerably increasing the recall.

When looking at *Equivalent* ($\equiv$), the indirect matching found slightly less relations. All the indirect matchings discovered the same amount of equivalences because the only way to find equivalence indirectly is to have both concepts anchored to the same concept in background knowledge. The equivalences found directly and not indirectly were caused by concepts which existed in CRISP and MeSH but not in FMA. In the next section we take a closer look at such a case. In few cases equivalences were detected indirectly and not directly because their labels were found as synonymous only through the background knowledge.

5.2 Evaluation of results

To test for correctness of the matches that we produced with the different experiments, we randomly choose 30 CRISP concepts, and inspected their matches by manually browsing the Wikipedia[8] pages describing these concepts. The evaluation is presented in Figure 5.

	$\equiv$	$\preceq$	$\succeq$	Total	Correct(%)
Exper. 1: Direct	17	18	3	38	38 (100%)
Exper. 4: Indir. *isa* and *part-of* mixed and closure	14	39	59	112	105 (93.7%)
Exper. 5: Indir. *isa* and *part-of* not isa after part-of	14	37	50	101	101 (100%)

Fig. 5. Evaluation of the matchings CRISP to MeSH directly and indirectly on 30 random CRISP concepts

In the last column of Figure 5 is shown the correctness of the matches produced by the different experiments. Only in Experiment 4 there were wrong matches found where the precision dropped to 93.7%, and in the other experiments it was 100% meaning that all the discovered matches were correct. The evaluation of Experiment 2 and 3 was left out because they produced subsets of Experiment 5 which already produced 100% correct matches.

Clearly, 30 concepts is not a sufficient number to get to a full evaluation. However, all these cases were closely examined and gave us the impression that this evaluation depicts the correctness of the experiments. The number of matches on these 30 concepts resembles the ratio as found on the whole test set.

In lack of gold standard, the evaluation phase turned out not to be straight forward. We had to make a choice what to consider correct and what not. Namely, some matches are arguably correct because of the nature of the relation *constitutional-part-of*. For example, the ulnar artery is constitutional part of the elbow, but it also stretches through the whole arm, and therefore it is not part of the elbow only. We call these matches shared.

[8] http://wikipedia.org/

Yet, having the relation between ulnar artery and elbow is a useful one, somebody looking for medical resources about an elbow is interested in the arteries passing through the elbow as well, see Wikipedia for more details on this example. We explored the matches $\preceq$ or $\succeq$ produced in Experiment 5, and found out that 30 matches are shared, while the other 57 are not. This means that even if inspecting the matches rigidly by discarding shared matches, the background knowledge still produces a large gain in the matching results.

6 Analysis of discovered matches

We selected three representative cases of matches from the result sets. We will discuss matches found by the indirect matchings and not by the direct, then take a look at the causes for finding incorrect matches when arbitrary mixing the *isa* and *part-of* relations in the background knowledge, and finally we discuss matches found by the direct but not by the indirect matching.

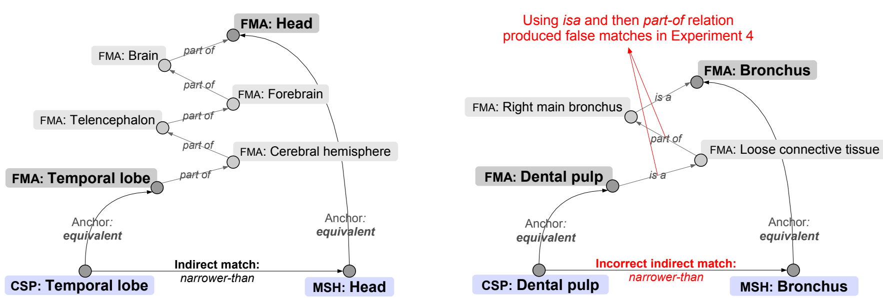

Fig. 6. Two indirect matches: on the left the correct match *Temporal lobe* $\preceq^i$ *Head*, and on the right incorrect match *Dental pulp* $\preceq^i$ *Bronchus*.

Case 1: Matches found by indirect and not by the direct matching. $\mathcal{CSP}$*: Temporal lobe* $\preceq^i$ $\mathcal{MSH}$*: Head* is a representative case of these matches, and is shown on the left hand side in Figure 6. Temporal lobes are parts of the brain, and consequently parts of the human's head. In the structure of MeSH and CRISP they are classified under the Brain which is classified under Central Nervous System, and are not connected in any way with the Head. Therefore relation to the Head is impossible to establish using direct matching, and the background knowledge is crucial in discovering the match.

Case 2: Incorrect match produced by arbitrary mixing of *isa* and *part-of* hierarchies. When using *isa* and then *part-of* in the inference, some of the matches were incorrect. An example is shown on the right hand side in Figure 6 finding $\mathcal{CSP}$:*Dental pulp* $\preceq^i$ $\mathcal{MSH}$:*Bronchus*. Each of the two relations *Dental pulp isa Loose connective tissue part-of Right main bronchus* remain correct when generalizing to *Dental pulp* $\preceq$ *Loose connective tissue*, and *Loose connective tissue* $\preceq$ *Right main bronchus*, but their transitive closure does not hold any more: *Dental pulp* $\preceq$ *Right main bronchus* is incorrect, which then implies the incorrect match $\mathcal{CSP}$:*Dental pulp* $\preceq^i$ $\mathcal{MSH}$:*Bronchus* is found.

Case 3: Matches found by direct but not by indirect matching. An example of such a match is $\mathcal{CSP}$:*Mesenteric artery* $\equiv^d$ $\mathcal{MSH}$:*Mesenteric Arteries*. The relation was not found indirectly because a concept *Mesenteric artery* does not exist in FMA, instead there are two more specific $\mathcal{FMA}$:*Inferior Mesenteric Artery* and $\mathcal{FMA}$:*Superior Mesenteric Artery* and one broader concept $\mathcal{FMA}$:*Artery*. Using FMA the following indirect relations were discovered for $\mathcal{CSP}$:*Mesenteric artery*:

- $\mathcal{CSP}$:*Mesenteric artery* $\preceq^i$ $\mathcal{MSH}$:*Arteries*
- $\mathcal{CSP}$:*Mesenteric artery* $\succeq^i$ $\mathcal{MSH}$:*Mesenteric Artery, Inferior*
- $\mathcal{CSP}$:*Mesenteric artery* $\succeq^i$ $\mathcal{MSH}$:*Mesenteric Artery, Superior*

Combining the results of the direct and indirect matching will improve the result of the both. We showed that in an earlier study in [4].

7 Related work

The research topic of semantic integration is a very active one, yet we encountered two major difficulties when comparing our approach with the others. First is the objective in matching: some target at finding pairs of most corresponding concepts, others aim at 1-1 mappings only, etc.; and second is the different way we use background knowledge.

Existing approaches mostly use background knowledge in the form of lexicons for discovering synonyms, see [6, 2, 7]. S-Match, [8], is example where background knowledge is exploited in doing the mapping, which is very similar to the approach we used in this work. In the current state S-Match uses a predefined set of background knowledge sources, such as Wordnet and UMLS. Moreover, it uses the class hierarchy of the background knowledge ontologies.

Related case-study was performed in [9], where the authors investigated enriching user search queries for image retrieval, by using relations from Wordnet. This follows closely our own scheme: the user search query is a source ontology of one single concept with one single label, Wordnet is the background knowledge, and the classification of images is the target ontology. They used *isa* and *part-of* relations, and they arrived to conclusion comparable with ours about using and combining these relations.

The work we present here is a sequel of another case study that followed similar scheme of using background knowledge in ontology integration, see [4]. In contrast to this work, in the previous studies we used source and target ontologies without structure, the relations we were chasing for were semantic closeness rather than relations with exact semantics such as *broader-than* and *narrower-than*. The background knowledge contained only *broader-than* and *narrower-than* relations, which prohibited us in investigating the impact of combining different relations.

8 Conclusions

Based on the results produced in the experiments, we draw the following conclusions from our study:

(i) Using comprehensive background knowledge in form of ontology can boost the ontology matching process as compared to a direct matching of the two ontologies.

(ii) Most of the value in using background knowledge comes from combining different pieces of knowledge within the background knowledge.

(iii) Different pieces of knowledge within the background knowledge need careful combination in order to gain maximal benefit.

A crucial requirement in using background knowledge is the existence of extensive reference ontologies in different domains at hand. Therefore, the development of such ontologies and subsequent publication on the Semantic Web will make the problem of integration easier.

Currently we are expanding the reported experiments further, we are looking for approximation schemes when deriving the relations within FMA, and we are investigating the usefulness of other relations in FMA like: X *is-attached-to* Y, X *sends-output-to* Y, etc. These relations will produce more matches different from those we present here.

Our findings were concluded from experiments conducted on medical test data. Therefore, we are conducting similar experiments in music domain on matching styles and genres from different music providers. In contrast to the medical domain, the knowledge in music is much weakly structured as different music content providers largely disagree on the meaning of music terms. In this direction, we took the effort to extract relatively extensive music ontology from Wikipedia which will serve as background knowledge in our experiments.

References

1. Jian, N., Hu, W., Cheng, G., Qu, Y.: Falcon-ao: Aligning ontologies with falcon. In: K-Cap 2005 Workshop on Integrating Ontologies. (2005)
2. Rahm, E., Bernstein, P.A.: A Survey of Approaches to Automatic Schema Matching. VLDB Journal **10**(4) (2001)
3. Ehrig, M., Sure, Y.: Foam - framework for ontology alignment and mapping; results of the ontology alignment initiative. In Ashpole, B., Ehrig, M., Euzenat, J., Stuckenschmidt, H., eds.: Proceedings of the Workshop on Integrating Ontologies. Volume 156., CEUR-WS.org (2005) 72–76
4. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: Proceedings of Knowledge Engineering and Knowledge Management (EKAW). (2006)
5. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Improving ontology mapping by using web ontologies as background knowledge. In: Submitted for acceptance at International Semantic Web Conference (ISWC). (2006)
6. de Bruijn, J., Martin-Recuerda, F., Manov, D., Ehrig, M.: D4.2.1 state-of-the-art-survey on ontology merging and aligning v1. SEKT Project deliverable D4.2.1 (2004)
7. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal on Data Semantics **IV** (2005) 146–171
8. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-match: an algorithm and an implementation of semantic matching. In: Proceedings of the European Semantic Web Symposium (ESWC). (2004) 61–75
9. Hollink, L.: Enriching search queries for image retreival using wordnet. In: Submitted for acceptance at ISWC'06. (2006)