Exploring the Semantic Web as Background Knowledge for Ontology Matching

Marta Sabou, Mathieu d'Aquin, and Enrico Motta

Knowledge Media Institute (KMi)
The Open University, Milton Keynes, United Kingdom
{r.m.sabou, m.daquin, e.motta}@open.ac.uk

Abstract. In this paper we propose an ontology matching paradigm based on the idea of harvesting the Semantic Web, i.e., *automatically* finding and exploring *multiple* and *heterogeneous* online knowledge sources to derive mappings. We adopt an experimental approach in the context of matching two real life, large-scale ontologies to investigate the potential of this paradigm, its limitations, and its relation to other techniques. Our experiments yielded a promising baseline precision of 70% and identified a set of critical issues that need to be considered to achieve the full potential of the paradigm. Besides providing a good performance as a stand-alone matcher, our paradigm is complementary to existing techniques and therefore could be used in hybrid tools that would further advance the state of the art in the ontology matching field.

1 Introduction

### 1.1 The Matching Problem in Ontologies and Databases

The issue of finding correspondences between heterogeneous conceptual structures is inherent to all systems that combine multiple information sources. The database community has identified schema matching as a core task in many application domains, such as integrating different databases (i.e., establishing mappings between their schemas), data warehousing and E-commerce (matching between different message schema) [39]. Matching also plays a major role in approaches that rely on ontologies to solve the semantic heterogeneity problem between information systems [25, 38, 52]. While both database schemas and ontologies provide a vocabulary of terms to describe a domain of interest, database schemas do not make explicit the semantics of their elements while ontologies, by definition (*"a formal, explicit, specification of a domain conceptualization"* [19]), do [44]. A direct implication is that matchers can try and exploit the explicit semantics of ontologies to improve their performance.

In this context, the appearance and growth of the Semantic Web, *"an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"* [4], marks an important stage in the evolution of the matching problem. Technologies such as RDF(S) and OWL, which allow to represent ontologies and information in a

formal, machine understandable way, have led to a rapid increase in the amount of online ontologies and semantic documents [26]. This online knowledge can be explored through novel infrastructures such as the Swoogle [11] semantic search engine or the Watson Semantic Web Gateway [10], which collect and index Semantic Web documents. These changes have important consequences for the design of Semantic Web applications. While early tools, resembling ontology-based information systems [12, 23, 29, 36], relied on a small number of ontologies selected and configured at design time, we are now witnessing the emergence of a *new generation of Semantic Web applications* [37], which aim to *dynamically* select, combine and exploit online ontologies [28]. Needless to say, matching is a key component of this new class of applications.

The approaches developed both for database schemas and ontologies follow two major paradigms depending on the types of information they use to derive mappings [25, 38, 39, 44]. *Internal* approaches typically explore information provided by the matched ontologies such as their labels, structure or instances [44]. Indeed, all the ontology matching tools evaluated within the Ontology Alignment Evaluation Initiative (OAEI'06)[1] primarily exploit label and structure similarity to derive correspondences associated to varying confidence values [14]. A limitation of such approaches is that they depend on the richness and the similarity of the internal information of the matched ontologies. For example, Aleksovski et al. [2] used two state of the art matchers, FOAM [13] and Falcon-AO [24], to match weakly structured medical vocabularies with a low overlap in their labels and obtained precision values of only 30% and 33%.

External (or background knowledge based) techniques aim to address this limitation by exploring an external resource to *bridge the semantic gap between the matched ontologies*. Indeed, continuing the example above, Aleksovski et al. obtained a precision value of 76% on the same dataset in the medical domain by exploring the DICE ontology as background knowledge [2]. As depicted in Figure 1, matchers from this category exploit an external resource by replacing the original matching problem (between concepts A and B) with two individual matching and an inference step: the two concepts are first matched to so called *anchor terms (A', B')* in the background source, and then mappings are deduced from the semantic relations of these anchors.
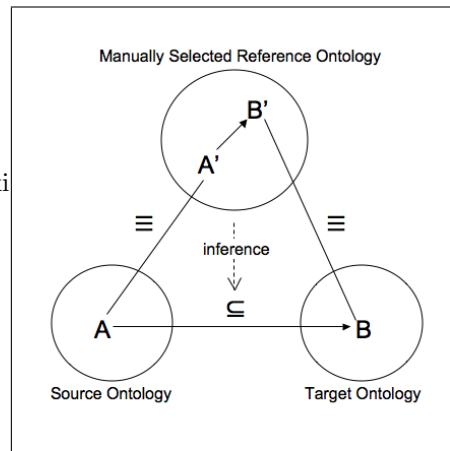
Fig. 1: Background knowledge based matching.

A pre-requisite for the success of such matchers is the *availability* of background knowledge sources with an *appropriate coverage* of the matched ontologies. Some approaches rely on readily available, large-scale, generic resources

---

[1] http://oaei.ontologymatching.org/2006/

such as WordNet or Cyc [9, 15, 22]. However, even if these resources cover a broad range of domains they might not cover specific domains (e.g., medicine, transportation planning) to the depth required by the matching task. In these cases, an appropriate domain ontology is either built manually (e.g., for the SIMS system [3]) or selected prior to the matching process [2]. As discussed in [3], the manual acquisition (or selection) of domain knowledge with appropriate coverage represents a considerable effort that should ideally be avoided.

1.2   Our Proposal: Exploiting the Semantic Web as Background Knowledge for Ontology Matching

We propose *a paradigm to ontology matching* based on the idea of harvesting the Semantic Web, i.e., *automatically* finding and exploring *multiple* and *heterogeneous* online knowledge sources. For example, when matching two concepts labeled *Researcher* and *AcademicStaff*, a matcher based on this paradigm would 1) identify (during matching) online ontologies that can provide information about how these two concepts inter-relate and then 2) combine this information to infer the mapping. The mapping can be either provided by a single ontology (e.g., stating that *Researcher ⊑ AcademicStaff*), or by reasoning over information spread in several ontologies (e.g., that *Researcher ⊑ ResearchStaff* in one ontology and that *ResearchStaff ⊑ AcademicStaff* in another).

While this approach enjoys the advantages derived from the use of background knowledge, it provides an elegant solution to the tradeoff between the availability and the coverage of background knowledge. First, instead of relying on a single (generic or domain) ontology, we *maximize the coverage of the background knowledge* by exploring multiple online ontologies. Second, instead of selecting or building a domain ontology prior to matching, we *minimize any knowledge acquisition effort prior to matching* through the automatic selection of the background knowledge. Such an approach can be particularly helpful when a large, domain ontology does not exist but, nevertheless, the required knowledge is potentially spread over multiple different ontologies, or when the matched ontologies spread over several domains, requiring the use of a variety of ontologies.

A small-scale, preliminary evaluation of this paradigm provided encouraging results but gave little insight in its strengths and weaknesses when faced with real life situations [40]. The objective of this paper is to report on an in-depth investigation of this paradigm along the line of three main research questions:

Does it work? The main research question focuses on the feasibility of this paradigm to be successfully applied in real life matching cases. In practice this means assessing the two core assumptions that underlie this work. First, that the amount and quality of online ontologies are sufficient to be used as a basis for matching and that an alignment can be obtained in a reasonable amount of time. Second, that it is possible to build algorithms that automatically discover and combine this knowledge in an intelligent (useful) way. A

proof that the use of such a paradigm is feasible and therefore worth pursuing further, would be to achieve a good performance in a real life matching experiment by using a simple, baseline implementation.

What are the limitations and how can they be avoided? Our preliminary experiments described in [40] provided limited insight 1) about the main steps needed for implementing a matcher based on this paradigm (e.g., ontology selection, knowledge combination), 2) about their relative influence on the quality of the derived mappings (e.g., are false mappings due more to the inability to select the right ontology or to the use of simple knowledge combination algorithms?), as well as 3) about typical problematic cases that need to be solved for each step (e.g., which ontology characteristics lead to false mappings and should be avoided through the selection process?). Gaining an understanding of all these issues is a pre-requisite for designing an improved technique based on the proposed paradigm.

How does it compare to other techniques? The third aim of this paper is to position the proposed paradigm in the landscape of the ontology matching field. On the one hand, we investigate the levels of performance that can be achieved with a stand-alone matcher based on this paradigm. On the other hand, since our goal is not to provide a stand-alone matcher but rather a complement to existing approaches, we analyze strengths and weaknesses with respect of other techniques in order to understand how the proposed paradigm would benefit from being combined with them, in hybrid approaches.

We rely on an experimental approach to answer these research questions. Our methodology consists of three major stages which are reflected in the structure of the paper. In the first stage, we propose two possible implementations of the paradigm and analyze the steps that are core to both (Section 3). In the second stage, we provide a baseline implementation based on the simplest solution for each of these steps (Section 4) and apply it on a real life, large-scale matching case using the experimental setup described in Section 5. The last and most important stage of our work consists in analyzing the results of the experiments (Sections 6 to 9). The performance of the implemented matcher represents a baseline that can be achieved with our paradigm and thus addresses the first question regarding its feasibility. The second research question, focusing on possible limitations, is answered by analyzing the evaluation results (Section 6 and 7). In Section 8 we assess our assumption that multiple online sources can be used for matching by providing some statistics about the number of ontologies explored to derive mappings. We address our final research question in Section 9, by comparing our results, strengths and weaknesses to those of other techniques. We conclude and point out future work in Section 10.

2   Related Work

A first body of related work consists of *approaches to matching that rely on the use of background knowledge*. We distinguish two categories of such matchers

depending on the type of the explored external resource, i.e., an ontology [2, 3, 6, 9, 48] or online textual sources [50].

Several ontology based matchers rely on a large-scale generic resource such as Cyc or WordNet. The Carnot system [9, 22] explores the Cyc knowledge base as a global context for achieving a semantic level integration of various information models (e.g., database schemas, knowledge bases). CTxMatch [6] (and its follow-up, SMatch [15]) translates ontology labels into logical formulae between their constituents, and maps them to the corresponding WordNet senses. A SAT solver is then used to derive mappings between the concepts. This approach has been extended to handle the problem of *missing background knowledge* [16]: if the simple techniques used to explore WordNet fail, then a second set of more complex and computationally expensive heuristics are applied to gain more knowledge.

While readily available, generic resources might fail to provide the appropriate coverage when matching is performed in a specific domain, such as medicine. In these cases, several matching approaches have opted for the use of a domain ontology. The SIMS system [3] relies on a manually built ontology about transportation planning for integrating several databases in this domain. In [2], the authors match two weakly structured vocabularies of medical terms by using the DICE ontology. Similarly, in [48] mappings between two medical ontologies (Galen and Tambis) are inferred from manually established mappings with a third medical ontology (UMLS), and by using the reasoning mechanisms permitted by the C-OWL language. Unfortunately, building (and even selecting) an appropriate domain ontology prior to matching is a considerable effort and represents a drawback of these techniques [3].

van Hage et. al [50] use the combination of two "linguistic ontology matching techniques" that exploit online texts to resolve mappings between two thesauri in the food domain. First, they rely on Google to determine subclass relations between pairs of concepts using the Hearst pattern based technique introduced by the PANKOW system [8]. Then, they exploit the regularities of an online cooking dictionary to learn hypernym relations between concepts of the matched ontologies. The strength of this approach is that, in principle, it is domain independent and therefore it does not require manual background knowledge selection. In reality, however, its precision dramatically decreases when relying on a corpus of general texts (50%), as opposed to a domain specific one (75%).

While the paradigm proposed in this paper explores ontologies as background knowledge, it differs from the above described matchers in several ways. First, we tackle the issue of *coverage* by exploring *multiple* rather than a single ontology. Second, we reduce the *knowledge acquisition effort prior to matching* by automatically selecting these ontologies. Finally, unlike some of the matchers which exploit the particularities of the background ontology [2, 15], our approach is entirely domain and ontology independent.

Besides matchers based on background knowledge, our work is also related to *approaches that explore multiple (online) ontologies*. The idea of finding mappings between two ontologies by exploring other ontologies as *semantic bridges* has been discussed in [46] where a finite set of small, independently developed

ontologies are interrelated by finding mappings between their concepts. These mappings are often discovered through a semantic bridge consisting of many other ontologies. Because the set of ontologies is finite, the technique can establish pairwise relations between the concepts of all ontologies (using a variety of matching techniques) and then rank and eliminate the redundant or useless ones. Our work is similar from the perspective that mappings are derived by exploring third party ontologies. However, a major difference is that we use a large set of heterogeneous ontologies where an exhaustive technique like the one of Stephens et al. cannot be applied.

The same paradigm of automatically selecting and exploring online ontologies has been proposed for solving other tasks than ontology matching. First, Alani proposes a method for ontology learning that relies on cutting and pasting ontology modules from online ontologies relevant to keywords from a user query [1]. Second, in [18] the authors describe a multi-ontology based method to disambiguate the senses of keywords that are given as a query to a search engine (e.g., *star* is used in its sense of celestial body in [*astronomy, start, planet*]). While the authors had previously relied on WordNet alone to collect possible senses for each keyword, now they exploit online ontologies to gather a larger set of senses and thus increase the quality of their method. Unfortunately, from these two methods, only the disambiguation process has been implemented and partially evaluated. Therefore, the contribution of our work to this line of research is to provide a first evaluation of automatically exploring online ontologies.

3   Proposed Paradigm

In the terminology of [44], we describe an element level matcher which relies on the use of external knowledge sources to derive mappings. In this section we investigate a set of issues that need to be considered when implementing this paradigm and conclude on a set of fine-grained research questions that should be experimentally investigated.

We describe two increasingly sophisticated strategies to discover and exploit online ontologies for matching. The first strategy derives a mapping between two concepts if this relation is defined within a single online ontology (Section 3.1). The second strategy (Section 3.2) addresses those cases when no single online ontology states the relation between the two concepts by combining relevant information which is spread over two or more ontologies. Both strategies need to address a set of tasks such as finding ontologies that contain equivalent concepts to those being matched (i.e., anchoring), selecting the appropriate ontologies, and using rules to derive mappings. We discuss all of these tasks in Sections 3.3 to 3.5. In Section 3.6 we discuss mechanisms for dealing with contradictory mappings derived from different sources.

Each strategy is presented as a procedure that takes two candidate concepts as an input and returns the discovered mapping between them. We use the letters A and B to refer to these candidate concepts. The corresponding concepts to A and B in an online ontology O_i are A'_i and B'_i ("anchor terms"). We rely on

the description logic syntax for semantic relations occurring between concepts in an online ontology O_i, e.g., $A'_i \sqsubseteq B'_i$ means that A'_i is a sub-concept of B'_i in O_i. The returned mappings are expressed using C-OWL like notations [5], e.g., $A \xrightarrow{\sqsubseteq} B$. Note that we are using the C-OWL notations without relying on the formalism itself and on its semantics.

3.1 Strategy S1: Mappings Within One Ontology

The first strategy consists of finding ontologies containing concepts similar with the candidate concepts (e.g., by relying on Swoogle) and then deriving mappings from their relations in the selected ontologies. Figure 2 (a) illustrates this strategy with an example where three ontologies are discovered (O_1, O_2, O_3) containing the concepts A' and B' corresponding to A and B. The first ontology contains no relation between the anchor concepts, while the other two ontologies declare a subsumption relation. The concrete steps of this strategy are:

1. Anchor A and B to corresponding concepts A' and B' in online ontologies;
2. Select ontologies containing A' and B';
3. For a given ontology (O_i) apply the following rules:
   - if $A'_i \equiv B'_i$ then derive $A \xrightarrow{\equiv} B$;
   - if $A'_i \sqsubseteq B'_i$ then derive $A \xrightarrow{\sqsubseteq} B$;
   - if $A'_i \sqsupseteq B'_i$ then derive $A \xrightarrow{\sqsupseteq} B$;
   - if $A'_i \perp B'_i$ then derive $A \xrightarrow{\perp} B$;
4. Combine all mappings derived from the considered ontologies.
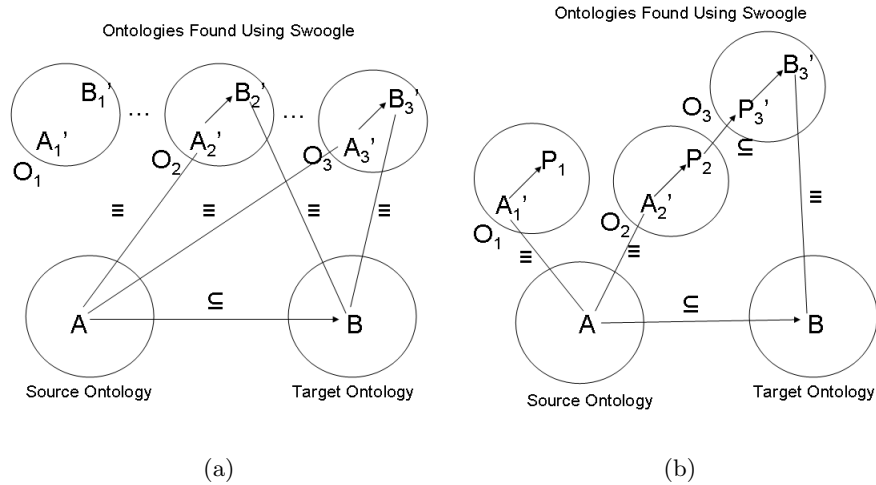
Fig. 2. Ontology matching (a) within one ontology (S1) and (b) across ontologies (S2).

For example, when matching two concepts labeled *Drinking Water* and *tap_water*, appropriate anchor terms are discovered in the TAP ontology and the following subsumption chain in the external ontology is used to deduce the mapping: *DrinkingWater* $\sqsubseteq$ *FlatDrinkingWater* $\sqsubseteq$ *TapWater*.

This strategy can be implemented in a multitude of ways depending on the type of anchoring mechanism applied in step 1, the criteria used to select the right ontologies in step 2, the complexity of the inferences employed by the derivation rules in step 3 or the strategy for integrating mappings originating from different sources in step 4. We discuss all these issues in the upcoming sections (Sections 3.3 to 3.6).

3.2   Strategy S2: Cross-Ontology Mapping Discovery

The previous strategy assumes that a relation between the candidate concepts can be discovered in a single ontology. However, some relations could be distributed over several ontologies. Therefore, if no ontology is found that relates both candidate concepts, then the mapping should be derived from two (or more) ontologies. In this strategy, matching is a recursive task where two concepts can be matched because the concepts they relate to in some ontologies are themselves matched. Figure 2 (b) illustrates this strategy where no ontology is available that contains anchor terms for both A and B, but where one of the parents (P_2) of the anchor term A_2' can be matched to B in the context of a third ontology (O_3). For example, a mapping between *Cabbage* and *Meat* can be derived by taking into account that *Cabbage* $\sqsubseteq$ *Vegetable*[2] and then discovering that *Vegetable* $\perp$ *Meat*[3] through another matching step. The concrete steps are:

1. Anchor A and B to corresponding concepts A' and B' in online ontologies;
2. If no ontologies are found that contain both A' and B' then select all ontologies containing A';
3. For a given ontology O_i apply the following rules:
    (a) for each P_i such that $A_i' \sqsubseteq P_i$, search for mappings between P_i and B;
    (b) for each C_i such that $A_i' \sqsupseteq C_i$, search for mappings between C_i and B;
    (c) derive mappings using the following rules:

    – (r1) if $A_i' \sqsubseteq P_i$ and $P_i \xrightarrow{\sqsubseteq} B$ then $A \xrightarrow{\sqsubseteq} B$
    – (r2) if $A_i' \sqsubseteq P_i$ and $P_i \xrightarrow{\equiv} B$ then $A \xrightarrow{\sqsubseteq} B$
    – (r3) if $A_i' \sqsubseteq P_i$ and $P_i \xrightarrow{\perp} B$ then $A \xrightarrow{\perp} B$
    – (r4) if $A_i' \sqsupseteq C_i$ and $C_i \xrightarrow{\sqsupseteq} B$ then $A \xrightarrow{\sqsupseteq} B$
    – (r5) if $A_i' \sqsupseteq C_i$ and $C_i \xrightarrow{\equiv} B$ then $A \xrightarrow{\sqsupseteq} B$
4. Combine all mappings derived from the considered ontologies.

The matching processes in steps 3(a) and 3(b) can be realized using either strategy S1 or S2. These two steps correspond to the recursive part of the algorithm and therefore a concrete implementation will need to avoid the exhaustive

---

[2] `http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf`
[3] `http://www.co-ode.org/resources/ontologies/Pizzademostep1.owl`

search of the semantic space. For example, mappings could be established only with the direct parents/children of A_i' (instead of *all*), the matching could stop as soon as a mapping is found or when a given amount of time has elapsed. As in the case of S1, strategy S2 can also be implemented differently depending on the chosen anchoring mechanism, ontology selection, the types of rules used and the way the final mappings are derived, as we discuss in the next sections.

3.3   Step1: Anchoring

Anchoring is a core part of all background knowledge based techniques: its role is to identify the appropriate part of the background knowledge that should be used for the matching (i.e., the part that refers to the two concepts being matched). Several anchoring mechanisms are reported in the literature.

The anchoring described in [2] is based on partial lexical matches between concept labels (i.e., it is sufficient that they share a subset of tokens) following the intuition that additional words added to a label denote a more specialized concept by constraining its meaning. For example, *"long brain tumor"* is anchored (as *narrower-than*) to *"long tumor"* because they share two tokens. Unfortunately, this strategy also introduces incorrect matches such as *"long brain tumor"* being anchored (i.e., as *narrower-than*) to *"brain"* [2].

The authors of [50] impose a strict string matching between concept labels and tokens in online texts (i.e., web pages) to establish equivalences between them. This stricter matching is likely to be more precise than the one in [2] but it still falls short of correctly anchoring polysemous words (e.g., *Squash* can be equally matched to words referring to a vegetable or a sport).

Unlike the previous two approaches which only exploit labels, S-Match [15] goes one step further and also relies on the structural information of a concept (i.e., its place in the concept hierarchy) when anchoring it into WordNet. First, the approach identifies all the WordNet senses relevant for the concept label. Then, the right sense is filtered out depending on the senses of the surrounding concepts in the hierarchy (using an algorithm presented in [30]). This approach ensures that concepts are anchored to concepts with the same sense in WordNet.

In the case of our technique, the anchoring is special because a concept is anchored to (possibly) many online ontologies with varying semantic richness. While anchoring should identify semantically (and not just syntactically) equivalent concepts (thus taking into account the semantic context of the concepts similarly to S-Match), it also needs to be light-weight enough to be usable during matching (in the case of S-Match, because a single background ontology is used, anchoring can be performed a priori). Before implementing a precise and optimal anchoring, we wished to find out:

RQ1: How well do simple anchoring techniques work? In our first implementation we use an anchoring technique similar to that of van Hage and we wish to assess the quality of the obtained results. If the results are reasonable, implementing a more complex and time consuming anchoring might not be worthwhile.

3.4   Step2: Ontology Selection

Anchoring identifies a set of ontologies that can lead to a mapping (e.g., in the example for S1, Figure 2, three such ontologies are identified). The choice of the ontologies that are used to derive mappings depends on two main design decisions: (1) the number of ontologies to be used and (2) the way they are selected. For the first design decision, we distinguish two situations:

Using a single ontology is the easiest way to deal with the multiple returned ontologies but it assumes that the discovered relation can be trusted and there is no need to inspect the other ontologies as well. In the example for S1, this would mean deriving the mapping from one of the three ontologies.

Using a subset (or all) of the returned ontologies is computationally more expensive but it has a higher accuracy by taking into account all the information that can be possibly derived from the returned ontologies. In this cases a mapping relation is derived from each ontology and then these are combined into a final mapping (see Section 3.6 for strategies about combining multiple, possibly inconsistent, mappings).

In both cases, whether using one or more ontologies, it is important to decide on some selection criteria. We distinguish two approaches to this issue:

Use the ranking mechanism of the underlying ontology search engine as the implicit selection mechanism. For example, in strategy S1 the mapping can be derived from the first ontology returned by Swoogle. Note that this ontology does not necessarily contain a relation between the candidate concepts (e.g., O_1 in Figure 2 (a)). In such cases, it could be considered that if an ontology covers the candidate concepts without relating them, then no mapping should be derived. Or, the algorithm could explore the remaining ontologies until a relation is provided by one of them (this will be the final mapping returned by the algorithm).

The selection criteria used by the search engine might not be appropriate for matching. For example, similarly to Web search engines such as Google, Swoogle ranks ontologies based on their popularity computed with a modified version of the PageRank algorithm which takes into account how many times an ontology is referenced by others [11]. Popularity, however, is not always a good indicator of an ontology's suitability for matching. Indeed, because it is frequently imported by other ontologies, FOAF is often ranked as the "best" ontology, even if this weakly structured vocabulary is of little help for deriving mappings.

Use predefined selection criteria to select the ontology (when using one ontology) or the ontologies (when using a subset of ontologies) that have the highest quality and can potentially lead to the best mapping. A pre-requisite to build a good selection mechanism that would identify the "good" ontologies is a better understanding of the ontology characteristics that typically result in good/false mappings. These could range from structural characteristics such as depth or width (i.e., deeper ontologies have a richer structure

thus they would lead to more mappings than shallow ones), to domain similarity with the matched ontologies or to qualitative characteristics such as the absence of certain modeling errors.

The need to better understand what constitutes a good ontology for matching leads us to the second research question:

RQ2: Which ontology characteristics lead to false mappings? One of the goals of our experiments is to determine some of these characteristics so that they can be used to build an appropriate selection mechanism.

3.5 Step 3: Derivation Rules

The derivation rules defined for both strategies can be implemented by considering different levels of inferences. In the simplest implementation, we can rely on *direct* and *declared* relations between A' and B' in the selected ontology. But, for better results, *indirect* and *inferred* relations should also be exploited (e.g., if $A' \sqsubseteq C$ and $C \perp B'$, then $A' \perp B'$). Different levels of inferences can be considered (no inference, basic transitivity, description logic reasoning), each of them representing a particular compromise between the performance of the matching process and the completeness of the obtained alignment.

3.6 Step 4: Combining Mappings

Unlike previous techniques where mappings were based on a single ontology ([2, 15]), our approach derives mappings from a variety of sources. However, mappings resulting from different sources can contradict each other.

At a simple level, different ontologies can lead to different and incoherent relations between the same pair of concepts. For example, $Seafood$ is subsumed by $Meat$ in one ontology[4], and disjoint with it in another[5], leading to two directly contradictory mappings. If the final mapping between a pair of concepts is derived from several ontologies (Section 3.4), situations when such contradictory relations are returned need to be considered. For example:

Keep all mappings. In the simplest case, all derived mappings can be returned, thus allowing the user to select the right mapping (favoring recall).
Keep mappings without contradiction. To favor precision, the algorithm could return a mapping between two concepts iff all the inferred intermediary mappings were the same (i.e., there was no contradiction).
**Keep the most frequent mapping.** Given a set of mappings, return the most frequent mapping (i.e., the mapping that was derived from most sources).
Keep the trusted mappings only. Return mappings derived from sources that satisfy certain trust criteria.

---

[4] `http://reliant.teknowledge.com/DAML/Economy.owl`
[5] `http://www.w3.org/2001/sw/WebOnt/guide-src/food`

At a more complex level, the combination of several mappings in the alignment can lead to intricate contradictions. Figure 3 provides an example of such a situation, where the concepts of $Tomatoes$ and $Vegetable$ can be related, directly or indirectly, on the basis of four different mappings, potentially derived from four different ontologies. The contradiction appears because $Tomatoes$ can be inferred to be at

Fig. 3:   Contradictory mappings.

the same time disjoint with $Vegetable$, and a sub-class of it. This situation can be described as an incoherence in the sense that the class $Tomatoes$ is *unsatisfiable*: there cannot be any instance of $Tomatoes$, since such an object would have to be an instance of two disjoint classes: $Vegetable$ and $Fruit\_(Food)$.
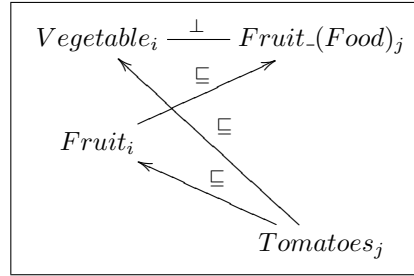
Generating such contradictions is a particularity of our technique, which combines information from different, heterogeneous knowledge sources. Incoherences are complex to detect as they require the use of reasoning mechanisms upon the source ontologies and the alignment. Handling these contradictions is a difficult task, requiring to select the appropriate strategy for removing the contradictory mappings. Therefore, an important research question is:

RQ3: How often do contradictions appear? The problem of dealing with contradictory mappings (both simple and complex) only needs to be addressed if such situations arise at all. We wish to get an insight in the scale of this phenomenon through experimental investigation.

4   Implementation Details

As described in Section 1, our methodology for exploring the proposed ontology matching paradigm consists in building and evaluating a baseline implementation. In this section we present the details of such a prototype which was built by using the simplest approach to implement all the tasks described in Section 3. We rely on Swoogle'05[6] which crawls and indexes a large amount of semantic metadata thus allowing access to a considerable part of the Semantic Web. We experiment with three implementations of the paradigm (these correspond to different configurations of the prototype):

Strategy S1, first variant stops as soon as one of the examined ontologies contains a relation between the matched concepts. We use this variant to evaluate the baseline performance of the paradigm for S1 (Section 6.1) and to understand the influence of anchoring and ontology selection (Section 6.2).

Strategy S1, second variant inspects *all* the ontologies that contain information about the two concepts to be matched and returns all the obtained

---

[6] At the time of the experiments, Swoogle'06 was too unstable to allow extensive experiments.

mappings. We use this implementation to investigate how often simple contradictory mappings are derived between two given concepts (Section 6.3).

Strategy S2 derives a mapping between two concepts by combining information spread over several ontologies. Given the recursive nature of S2 which can lead to long execution times, we took the following design decisions to limit the search space of the matcher. First, we avoid infinite recursion by using the non-recursive S1 strategy in steps 3(a) and 3(b). Indeed, this strategy will always investigate a restricted number of ontologies (those in which the concepts to match appear). Second, we restrict the recursive part (steps 3(a) and 3(b)) to find matches only between the direct parents and children (P and C) of the anchor terms corresponding to the source concept and the target concept (B). Finally, the matcher stops as soon as a mapping is found between A and B. We evaluate the baseline performance of the paradigm for S2 in Section 7.1 and investigate typical errors in Section 7.2.

We now discuss the details common to all these individual implementations.

4.1 Anchoring Mechanism

The anchoring mechanism (i.e., finding A', B') in the case of all implementations is based on strict string matching between concept labels, similar to that of van Hage [50]. We allow for variations in naming conventions and lexical form. For simple labels (made up of one word) we find anchors that match this word as well as it's lemma (i.e., base form): a label *Persons* will be anchored to concepts labeled either *Persons* or *person*. This is achieved by performing an exact search for each lexical form of the label. For compound labels (containing multiple words) we anchor to concept labels containing the same words, in the same order, but possibly written according to different naming conventions and having different lemmas: *TeaCups* $\simeq$ *Tea_Cup* $\simeq$ *tea cup*. Concretely, for each word in the label and its lexical variants we query Swoogle for the number of labels that contain the search string as a substring (fuzzy match). For the word that has the fewest appearances, we compare all its appearances to the compound label.

4.2 Ontology Selection

The first variant of S1 as well as S2 rely on the implicit ranking mechanism of Swoogle (based on popularity) to select the ontologies from which the mapping is derived. Both implementations inspect the first ontology returned by Swoogle and if no mapping can be derived from it, then the next ontology is considered until an ontology is found from which a mapping can be derived. The second variant of S1 simply inspects all ontologies returned by Swoogle and derives a mapping from each of them when possible. Note that the selected ontologies are *not* downloaded, parsed and interpreted locally (unlike envisioned in [1] and done in [18]). Instead, their inspection is performed through the Swoogle API by using a range of functionalities such as requesting the direct parent or the disjuncts of a given concept.

We adopt a broad notion of an ontology which is not limited to the physical file in which the content is stored, nor to a given namespace but which also considers imported knowledge. The side effect of this view is that our search for a mapping is also conducted in the ontologies imported (reused) by a given ontology. It is therefore possible that A' is identified in ontology O_i while B' is defined in an ontology O_j which is imported by O_i. For example, a mapping is derived between $Dredger$ and $Vehicle$ by identifying a subsumption chain that spans three ontologies importing each other, $(O_1{}^7, O_2{}^8, O_3{}^9)$:

$Dredger_1 \sqsubseteq Ship_2 \sqsubseteq DisplacementHullWatercraft_2 \sqsubseteq Watercraft_2 \sqsubseteq Vehicle_3$

Even if in such cases a mapping is derived by combining information from several ontologies, there is still a fundamental difference with respect to mappings derived using S2. Namely, the relations between concepts from different ontologies used in S1 have been declared by the ontology creator. On the contrary, when using S2, the correspondence between concepts in different ontologies is established automatically, by using the anchoring mechanism.

Technically, the ontology selection mechanisms provided by Swoogle do not suffice for implementing our broad view on ontologies (i.e., they cannot filter ontologies based on the content of the knowledge that they import). As a result, we used a technical artefact to implement the selection step: we select all ontologies that contain an anchor for A and then inspect its hierarchy until one of the concepts equals (or is disjoint with) B. We take advantage of the fact that the Swoogle function for inspecting the hierarchical context of a concept takes into account imported content. Compared to a previous implementation where we inspected ontologies containing anchors for both A and B, this implementation discovered more mappings without being noticeably slower than its predecessor.

4.3   Derivation Rules

We use the rules described in Section 3. In both strategies we have relied on the transitivity of the subsumption relations to take advantage of indirect relations between concepts. While Swoogle's API allows for retrieving direct subsumptions, indirect relations are explored by asking several queries about direct relations (i.e., asking for the parent of the parent). To reduce the time of our experiments, we implemented S2 in such a way that only the first direct parent of the discovered A' concept is considered (instead of exploring *all* parents).

4.4   Detecting Contradictions

As explained in Section 3.6, because it combines heterogeneous knowledge sources, our technique may result in contradictory mappings, leading to incoherences within the generated alignment. Simple contradictions, involving only two mappings between the same two terms, are easy to detect. However, as shown in

---

[7] http://reliant.teknowledge.com/DAML/Transportation.daml

[8] http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml

[9] http://reliant.teknowledge.com/DAML/SUMO.daml

Figure 3, contradictions and incoherences may appear because of the intricate combination of more than two mappings, and therefore, the use of reasoning mechanisms for detecting such situations is required.

Reasoning on mappings has received considerable attention lately with several papers reporting on the use of inferences on mappings in order to improve the quality of alignments [33, 34, 47, 49]. Among the diagnosis tasks defined in the literature, the detection of contradictions (called debugging in [34] and consistency checking in [47]) is recognized as being of particular importance. These studies rely on a rigorous formal framework, based on distributed description logics (DDLs). DDL is a formalism considering multiple ontologies, each of them with its own interpretation, interrelated through mappings (roughly sub-concept relations and equivalences), allowing distributed interpretations upon the ontologies globally, and upon the mappings [49]. However, relying on the DDL semantics introduces important constraints. In particular, the current implementation of DDL does not allow the use of disjoint relations in mappings. Moreover, mappings in DDLs are not transitive and are directional (e.g., $A \xrightarrow{\sqsubseteq} B$ is not equivalent to $B \xrightarrow{\sqsupseteq} A$ in DDLs), making this formalism inappropriate in our approach. Therefore, inspired by the previously mentioned work, we devised a simpler mechanism (not relying on DDLs) for detecting contradictions, using an ad-hoc reasoner (based on simple inference rules) for mappings, coupled with a classical DL reasoner (Pellet[10]) for reasoning upon the source ontologies.

We consider that the alignment contains a contradiction (incoherence) when it can be inferred, from the content of the alignment and from the source ontologies, that a concept is at the same time a sub-concept of and disjoint with another concept (e.g., $A \xrightarrow{\perp} B$ and $A \xrightarrow{\sqsubseteq} B$, or $A \xrightarrow{\perp} B$ and $A \sqsubseteq B$). According to this definition, the procedure for detecting incoherences is straightforward. For all the disjoint mappings that can be inferred from the alignment, we verify whether the involved concepts are sub-concepts of each other. Simple heuristics are used to avoid the detection of redundant contradictions. For example, if $A \xrightarrow{\perp} B$, $A \xrightarrow{\sqsubseteq} B$, and $C \xrightarrow{\sqsubseteq} A$, we only count one contradiction, even if it can be inferred that $C \xrightarrow{\sqsubseteq} B$ and that $C \xrightarrow{\perp} B$. This second contradiction is considered to be derived from the first one.

Note that our goal is not to provide a novel mechanism for incoherence detection in mappings. Indeed, the employed ad-hoc reasoner is only sufficient for detecting incoherences in an alignment. Handling these contradictions will require more advanced (and more complete) reasoning procedures (Section 6.3).

5   Experimental Setup

In this section we describe the experimental data sets and the real life scenario from where they originate, we provide an overview about how the experiments reported in the rest of the paper relate to the research questions identified in Section 3 and detail the methodology used for evaluating the alignments.

---

[10] http://www.mindswap.org/2003/pellet/

5.1   Experimental Scenario and Data

Our experimental data[11] is derived from a real life scenario, where two organizations wish to align their ontologies. These organizations are the UN's Food and Agriculture Organization (FAO) and the US's National Agricultural Library (NAL). Both organizations maintain large agricultural thesauri which they use for indexing their data. FAO's **AGROVOC** thesaurus, version May 2006, consists of 28.174 descriptor terms (i.e., preferred terms) and 10.028 non-descriptor terms (i.e., alternative terms). NAL's Agricultural Thesaurus **NALT**, version 2006, consists of 41.577 descriptor terms and 24.525 non-descriptor terms. Given their use to index data containing a vast amount of knowledge, these thesauri describe a broad range of domains, from animal species to chemical substances and information technology. Also, they use several technical terms (e.g., from chemistry) and a considerable amount of Latin terms (e.g., to describe animal species). There are several reasons for performing an alignment between these thesauri. First, such an alignment would facilitate data exchange between the two organizations. Second, the alignment process could identify concepts that are missing from one thesaurus but are covered by the other. Finally, an immediate benefit would be the enrichment of NALT, which currently contains only English and Spanish terms, with multilingual information contained in AGROVOC.

In our experiments we relied on both descriptor and non-descriptor terms, since the latter often describe synonyms of the preferred terms. There are several reasons behind choosing this data set as a basis for our experiments:

Large-Scale. Our hypothesis is that this large-scale, real life data set will allow us to evaluate the scalability of the proposed technique. Further, the large amount of data should provide a good test bed for all the research questions stated in Section 3.

Multi Domain. Because these thesauri contain information from a wide range of domains (and also because they are so large), it is virtually impossible to find a single ontology that could be used as a source of background knowledge to derive mappings (i.e., as current techniques do [2, 15]). Therefore this is an illustrative case where it is necessary to combine knowledge from multiple background ontologies, possibly selected automatically.

Useful benchmark. A further advantage is that five state of the art ontology matching tools have been already used to derive mappings between these ontologies. These results are important for understanding how the proposed technique can complement existing technology.

5.2   Overview of the Experiments

Table 1 summarizes the experiments reported in the rest of the paper and the corresponding research questions. To investigate the feasibility of the proposed paradigm i) we evaluate a baseline performance for the first variant of strategy S1 (Section 6) and for strategy S2 (Section 7) using the methodology described in

---

[11] This data was also used in the OAEI 2006 *food* Thesaurus Mapping Task.

Section 5.3 and ii) we assess the assumption that multiple online sources can be used to derive mappings by providing statistics about the number of ontologies used during the alignment process (Section 8).

Research Question	Experiment
Does it work?	Evaluate strategies S1 and S2 (Sect. 6.1 & 7.1)
	Number of explored ontologies (Sect. 8)
What are the limitations?	
Anchoring (RQ1)	Analyze results of S1 and S2 (Sect. 6.2 & 7.2)
Selection (RQ2)	Analyze results of S1 and S2 (Sect. 6.2 & 7.2)
Contradictions (RQ3)	Derive mappings from all ontologies (Sect. 6.3)
	Use incoherence detection mechanisms (Sect. 6.3)
How does it compare to other techniques?	Comparison with internal and external techniques (Sect. 9)

Table 1. Overview of the relation between research questions and experiments.

Another goal of this work is to understand in what ways the baseline performance can be improved. In Section 3 we stated a set of research questions about issues that might hamper performance. Some of these questions can be answered by analyzing the results of the performance evaluation (Sections 6.2 and 7.2). Indeed, by inspecting the causes of false mappings, we can get an insight into the influence of the anchoring (RQ1) and ontology selection methods (RQ2). In Section 6.3 we assess how often contradictory mappings appear: we use the second variant of S1 to identify simple contradictions and apply incoherence detection mechanisms to detect alignment level (i.e., complex) contradictions (RQ3).

In Section 9 we compare our paradigm, based on the obtained results, to both techniques relying on information internal to ontologies (by analyzing the outcome of the OAEI'06 contest) as well as to other background knowledge based approaches (by exploring results reported in the literature). Besides the simple performance based comparison, we also discuss the potential contribution of our approach when integrated with existing techniques.

5.3   Methodology for Evaluating Alignments

One of the expected benefits of working with the AGROVOC-NALT dataset was the reuse of the Gold Standards employed to evaluate the OAEI'06 contest results. However, because the participant tools only returned equivalences, the Gold Standards have been geared towards evaluating those and thus were unusable for our results, containing subclass, superclass and disjoint relations. Given the high number of the discovered mappings, as well as the lack of Gold Standards, we performed a manual assessment of a significant subset of the results (1000 mappings in the case of both strategies).

As evaluators, we relied on nine members of our lab working in the area of the Semantic Web, and thus familiar with ontologies and ontology modeling. We performed two parallel evaluations of the sample mappings (i.e., each mapping

has been evaluated by two different evaluators). The participants were asked to evaluate each mapping as *Correct*, *False* or *"I don't know"* for cases where they could not judge the correctness of the statement. They were allowed to use any kind of material (e.g., (web-)dictionaries, Google) in cases where they were not familiar with the domain and needed some more information for evaluating a given mapping (e.g., when judging that $Leukemia \sqsubseteq Neoplasm$). A specialized graphical interface has been developed to facilitate the task of the evaluators by displaying the mappings together with the context in which the mapped concepts appeared in the source ontologies (i.e., their neighborhood). We compute the precision of the alignment as the ratio of *Correct* mappings over all the evaluated mappings (i.e., those evaluated either as *Correct* or *False*). Formally:

$$Precision = \frac{Correct}{Correct + False}$$

6   Deriving Mappings from One Ontology (Strategy S1)

The matching process performed by using the first variant of S1 resulted in a total of 6687 mappings containing 2330 subclass, 3710 superclass and 647 disjoint relations. These mappings were derived during about two days by using an average laptop. Table 2 provides some examples of the derived mappings. For each mapping we present the source (AGROVOC) and target (NALT) concepts and their labels. Under each mapping we provide the URL(s) of the ontology(ies) from which the mapping was extracted, as well as the relations on which the mapping is based in these ontologies (i.e., it's explanation). For example, the first mapping was established between the AGROVOC concept c_6617 labeled with "Rivers, Streams, Brooks, Tributaries" and the NALT concept identified as *waterways* and labeled "waterways". The mapping was derived from ontology O_1[12] which declares that $river \sqsubseteq waterway$. O_1 has been used because the anchoring identified a correspondence between the "Rivers" label of c_6617 and O_1's *river* concept, as well as between *waterways* in NALT and *waterway* in O_1. This example illustrates how the anchoring mechanism is flexible with respect to different naming conventions (here, it matches capitalized vs. non-capitalized words) and lexical forms (here, a match is established between the plural and the base form, or lemma, of both anchored labels).

It is interesting to observe that the second mapping spans two ontologies, the first one (*Economy.owl*) importing the second (*Mid-level-ontology.owl*). As explained in Section 4.2, our implementation is capable of identifying such declared, cross-ontology relations and derive the corresponding mapping.

Another observation to be made is that, using an additional equivalence mapping, the mapping between c_10463 and *tap_water* could have been inferred, thanks to the structure of the matched ontologies. Indeed, in NALT it is declared that $tap\_water \sqsubseteq drinking\_water$. Therefore, by establishing an equivalence relation between c_10463 (having the label "Drinking water") and the NALT

---

[12] `http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/data/russia1a.rdf`

Mappings	Nr.	Examples			
		AGROVOC Concept	**Labels**	**NALT Concept**	**Labels**
Subclass $(\xrightarrow{\sqsubseteq})$	2330	*c_6617*	Rivers, Streams, Brooks, Tributaries	*waterways*	waterways
		O_1:river $\sqsubseteq$ O_1:waterway			
		$O_1 = $ `http://www.aifb.uni-karlsruhe.de/WBS/meh/` `mapping/data/russia1a.rdf`			
		c_25469	Cocaine	*narcotics*	narcotics, opioids
		O_1:Cocaine $\sqsubseteq$ O_2:Narcotic			
		$O_1 = $ `http://reliant.teknowledge.com/DAML/Economy.owl` $O_2 = $ `http://reliant.teknowledge.com/DAML/` `Mid-level-ontology.owl`			
SuperClass $(\xrightarrow{\sqsupseteq})$	3710	*c_10463*	Drinking water, Potable water	*tap_water*	tap_water, tap water
		O_1:DrinkingWater $\sqsupseteq$ O_1:FlatDrinkingWater $\sqsupseteq$ O_1:TapWater			
		$O_1 = $ `http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf`			
		c_1142	Buildings, Houses, Building structures	*supermarkets*	supermarkets
		O_1:Building $\sqsupseteq$ O_1:Public_Building $\sqsupseteq$ O_1:Shop $\sqsupseteq$ O_1:Supermarket			
		$O_1 = $ `http://frot.org/space/0.1/index.rdf`			
Disjoint $(\xrightarrow{\perp})$	647	*c_2761*	Exports	*imports*	imports
		$O_1 = $ `http://edge.mcs.drexel.edu/assemblies/` `ontologies/woolly/2003/02/functions.daml`			
		c_8309	Water	*solids*	solids
		O_1:Water $\sqsubseteq$ O_1:Fluid $\perp$ O_1:Solid			
		$O_1 = $ `http://www.lri.jur.uva.nl/~rinke/aargh.owl`			
Total	6687				

Table 2. Example mappings discovered between AGROVOC and NALT using S1.

drinking_water concept, the mapping $c\_10463 \xrightarrow{\sqsupseteq} tap\_water$ could be inferred. The fact that we have obtained the same result *without* relying on the structural information of the matched ontologies demonstrates the potential of our technique to derive rich mappings even in cases when a rich structure would not be provided by the source ontologies. Indeed, this shows that internal information can be replaced by external information drawn from online ontologies. Having said that, structural information should not be purposefully ignored for the sake of using online ontologies (we have only done so to get an insight in the functioning of our technique when employed stand-alone).

6.1   Evaluation Results

In order to evaluate the precision of the alignment obtained with the first variant of S1, we randomly selected 1000 mappings (i.e., 15% of the alignment) containing an appropriate proportion of different mapping relations, namely: 100 disjuncts, 350 subclass, 550 superclass relations. These mappings were then evaluated by two groups of evaluators as described in Section 5.3. Table 3 summarizes the number of *Correct*, *False* and unevaluated (*Don't know*) mappings for each group, as well as the number of these mappings agreed by both groups. The two groups agree on 742 mappings (we exclude the "Don't know" answers because there are no real agreements on those), and therefore have an agreement coefficient of 74%. Note that a similar agreement (72%) was observed between the two groups that evaluated equivalence mappings on this dataset during OAEI'06 [14].

	Group 1	Group 2	Agreed by All
Correct	586	666	525
False	346	299	217
Don't know	68	35	10
Precision	**63%**	**69%**	**70%**

Table 3. Evaluation of strategy S1 by both groups.

We obtained precision values of 63% and 69% for the two groups. The gap between these values is due to the variation in the way evaluators performed their task: some investigated each mapping thoroughly, while others simply provided no evaluation for the mappings they were not sure about. To level out these differences, we also computed the precision of the part of the alignment on which both groups agreed, as we think this better reflects the typical performance that can be achieved with our paradigm. In this case, the precision was equal to 70%. We consider this value as indicative for the baseline performance that can be obtained by harvesting online information. We compare it to typical performances of other matching approaches in Section 9.

6.2 Error Analysis

Besides getting an indication of the baseline precision that can be obtained with the proposed paradigm, we also wish to understand in which ways the performance can be improved, i.e., what are the major causes for errors and how could they be eliminated. To answer this question, we manually inspected the 217 false mappings on which both groups agreed. We observed two major causes for errors. On the one hand, 114 errors (i.e., 53%) are caused by the inherent limitations of the simplistic anchoring. On the other hand, 91 false mappings (i.e., 42%) are due to qualitatively inappropriate online ontologies. The rest of 12 (5%) false mappings are due to various smaller causes that are not significant in this analysis. Table 4 provides an overview of the type and number of identified errors as well as some illustrative examples.

Anchoring errors are a side-effect of the basic, string matching based anchoring and appear when a concept is related to an incorrect sense in online ontologies. For example, in the first entry in Table 4, concept c_3179 describing "Game" in the sense of a hunted animal is incorrectly anchored to the *Game* concept in SUMO which represents a physical activity. In the second example, both concepts are anchored incorrectly. First, c_6443 labeled with "Rams" and referring to an *"uncastrated adult male sheep"*[13] is put in correspondence with a similarly labeled concept ("ram"), but which refers to *Random Access Memory* in the context of the online ontology. In the same way, the *memory* concept of NALT refers to the term used as in psychology and thus has been incorrectly anchored to the identically labeled concept which refers to computer memory.

Because concept labels are ambiguous, anchoring errors are frequent and account for more than half of the false mappings (53%). Therefore, the current anchoring needs to be modified to take into account the context of the anchored concepts. Indeed, an anchoring mechanism that could prevent deriving these false mappings (thus reducing their number to 103) could potentially lead to an increase in precision from 70% to 87%.

We identified the following types of errors introduced by exploring low quality online ontologies:

Subsumption used to model generic relations. One of the most common errors in online ontologies was the use of subsumption as a way to model the fact that there exists some type of relation between two concepts, e.g., $Survey \sqsubseteq Marketing$, $Irrigation \sqsubseteq Agriculture$, $Biographies \sqsubseteq People$. This case leads to 40 false mappings (i.e., 18%).

Subsumption used to model part-whole relations. Subsumption is used in several ontologies to model part-whole relations. This resulted in incorrect mappings such as $Branch \sqsubseteq Tree$, $Leaf \sqsubseteq Plant$.

---

[13] Definition from WordNet2.1.

Error Type	Nr./%	AGROVOC Concept	Labels	Rel.	NALT Concept	Labels
Anchor	114, 53%	*c_3179*	Game, Hunted Animals	$\stackrel{\sqsupseteq}{\longrightarrow}$	*sports*	sports, ball games, athletics
		O_1:Game $\sqsupseteq$ O_1:Sport				
		$O_1 =$ `http://lists.w3.org/Archives/Public/` `www-rdf-logic/2003Apr/att-0009/SUMO.daml`				
		c_6443	Rams, Tups	$\stackrel{\sqsubseteq}{\longrightarrow}$	*memory*	memory
		O_1:ram $\sqsubseteq$ O_1:memory				
		$O_1 =$ `http://www.arches.uga.edu/~gonen/qos_bilal.owl`				
Subsumption as generic relation	40, 18%	*c_3954*	Irrigation	$\stackrel{\sqsubseteq}{\longrightarrow}$	*agriculture*	agriculture, agriculture (general)
		O_1:Irrigation $\sqsubseteq$ O_1:SoilCultivation $\sqsubseteq$ O_1:Agriculture				
		$O_1 =$ `http://sweet.jpl.nasa.gov/ontology/human_activities.owl`				
Subsumption as part-whole	16, 7%	*c_23995*	Branches	$\stackrel{\sqsubseteq}{\longrightarrow}$	*trees*	trees
		O_1:Branch $\sqsubseteq$ O_1:Tree				
		$O_1 =$ `http://www.site.uottawa.ca/~mkhedr/NewFuzzy.owl`				
Subsumption as role	11, 5%	*c_6211*	Products, Produce	$\stackrel{\sqsupseteq}{\longrightarrow}$	*wool*	wool
		O_2:Product $\sqsupseteq$ O_1:ManufacturedProduct $\sqsupseteq$ O_1:TextileProduct $\sqsupseteq$ O_2:Fabric $\sqsupseteq$ O_3:Wool				
		$O_1 =$ `http://reliant.teknowledge.com/DAML/Economy.owl` $O_2 =$ `http://reliant.teknowledge.com/DAML/SUMO.owl` $O_3 =$ `http://reliant.teknowledge.com/DAML/` `Mid-level-ontology.owl`				
Inaccurate labeling	12, 5%	*c_1693*	Coal	$\stackrel{\sqsubseteq}{\longrightarrow}$	*industry*	industry
		O_1:coal $\sqsubseteq$ O_1:industry				
		$O_1 =$ `http://www.aifb.uni-karlsruhe.de/WBS/meh/` `mapping/data/russia1a.rdf`				
		c_24833	Databases, Data bases, Databanks	$\stackrel{\sqsupseteq}{\longrightarrow}$	*enzymes*	enzymes
		O_1:Database $\sqsupseteq$ O_1:Enzyme				
		$O_1 =$ `http://mensa.sl.iupui.edu/ontology/Database.owl`				
Different View	12, 5%	*c_2943*	Fishes	$\stackrel{\sqsupseteq}{\longrightarrow}$	*lobsters*	lobsters
		O_1:Fish $\sqsupseteq$ O_1:MarineInvertebrate $\sqsupseteq$ O_1:Crustacean $\sqsupseteq$ O_1:Lobster				
		$O_1 =$ `http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf`				

Table 4. Examples of several types of false mappings.

Subsumption used to model roles. Roles are often modeled as subclass relations, for example, that $Aubergine, Leek \sqsubseteq Ingredient$ ($Leek$ is a $Vegetable$ but in some contexts it plays the role of an ingredient).

Inaccurate labeling. We also found cases of correct subclass relations which introduced errors due to the inaccurate labeling of their concepts. For example, O_1[14] states that $coal \sqsubseteq industry$, where $coal$ refers to *coal industry* rather than the concept of $Coal$ itself. Similarly, for $Database \sqsupseteq Enzime$ in O_1[15], $Enzyme$ refers to an *enzyme database* rather than describing the class of all enzymes. Note that this type of errors could be avoided by a semantic, context aware anchoring mechanism.

Different Views. Finally, some of the explored ontologies adopted views that were not in concordance with the context of the mapping and/or the perspective of the evaluators. For example, TAP considers *lobsters* kinds of *Fishes*, a perspective with which none of the evaluators agreed.

Because a high number of errors (42%) were caused by incorrectly designed ontologies, our implementation would benefit from a selection mechanism based on the *quality* rather than the popularity of ontologies. While some approaches exist to automatically assess the quality of the ontology modeling [51], this task remains an important and difficult research question to consider as future work.

6.3 Contradictory Mappings

Research question RQ3 refers to whether contradictory mappings can be derived from different ontologies. To assess if different ontologies can contradict each other concerning the relation between a single pair of concepts (simple contradiction), we ran the second variant of S1: for every pair of concepts we derive mappings from *all* the online ontologies that mention them. As it can be expected considering the relative simplicity of the detection method, the number of such contradictions is very low and accounts to only eight pairs of concept labels (Table 5). Three of the eight pairs also appear inverted because their labels exist both in AGROVOC and NALT. For the purposes of this analysis, we can regard them as redundant thus further reducing the number of problematic pairs to five.

This first experiment shows that direct contradictions on the relation derived between a single pair of concepts are rare. However, as shown in Section 3.6, detecting these simple cases is insufficient, since contradictions can appear because of the combination of several mappings, derived from more than two ontologies. We used the implementation of the incoherence detection process described in Section 4 on the 6687 mappings generated between AGROVOC and NALT with the first variant of S1 and obtained 306 base incoherences. This result shows that contradictions actually appear in an alignment derived from online ontologies and that it is important to define strategies to deal with them.

---

[14] http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/data/russia1a.rdf
[15] http://mensa.sl.iupui.edu/ontology/Database.owl

AGROVOC label	NALT label	Nr. Subclass relations ($\stackrel{\sqsubseteq}{\longrightarrow}$)	Nr. Superclass relations ($\stackrel{\sqsupseteq}{\longrightarrow}$)	Nr. Disjunct relations($\stackrel{\perp}{\longrightarrow}$)
fruit	tomato	0	3	1
sea	river	0	1	2
energy	light	0	1	1
meat	seafood	0	2	12
seafood	meat	2	0	12
mushroom	pizza	1	0	1
tomato	fruit	3	0	1
light	energy	1	0	1

Table 5. All the label level contradictions.

Analyzing these incoherences can help us to better understand some limitations of our matching technique, and can hint ways of improving it. For example, Table 6 lists the top ten mappings that are most frequently involved in incoherences as well as the number of incoherences that they cause. This data suggests that incoherences are caused by a restricted sub-set of the alignment, and that a small sub-set of these mappings are actually involved in a large proportion of the incoherences. In other terms, incoherences are *localized* in the mappings, and detecting them helps in pointing out particular "areas" of the alignment that have to be considered as problematic.

Mapping	Nr. of incoherences
$People \stackrel{\sqsubseteq}{\longrightarrow} Agents$	115
$Products \stackrel{\perp}{\longrightarrow} Environment$	82
$Products \stackrel{\perp}{\longrightarrow} People$	80
$Environment \stackrel{\sqsubseteq}{\longrightarrow} Agents$	69
$Foods \stackrel{\sqsubseteq}{\longrightarrow} Products$	66
$Organizations \stackrel{\sqsubseteq}{\longrightarrow} Agents$	58
$Organisms \stackrel{\sqsubseteq}{\longrightarrow} Individual$	50
$Industry \stackrel{\sqsubseteq}{\longrightarrow} Heaters$	50
$Heaters \stackrel{\sqsubseteq}{\longrightarrow} Organizations$	49
$Technology \stackrel{\sqsubseteq}{\longrightarrow} Science$	38

Table 6. The 10 mappings that are most involved in incoherences.

Most concepts in Table 6 correspond to rather generic concepts (e.g., *Agents*, *Products*) likely to have lots of subclasses, which would become incoherent through inheritance. Indeed, almost 50 000 incoherences can be derived from the set of 306 base incoherences that are detected through our mechanism. This shows that this small number of incoherences (306) and the small number of mappings associated to them (454) corrupt almost the entire alignment.

In conclusion, it appears that online ontologies actually contradict each other and that this has an important influence on the formal quality of the alignments generated using our technique. Ultimately, contradictory mappings should be removed. However, automatically identifying the mappings to be remove is not trivial. Indeed, as shown in Table 6 (and already observed in [35]), the mappings that are often involved in incoherences are not necessarily wrong. On a more positive tone, several studies have been targeted towards the management or the removal of incoherences in ontologies [20, 42, 43]. These techniques provide solutions to facilitate the detection of the problematic sub-part of the alignment and to resolve contradictions, thus improving the quality of the entire alignment.

7   Deriving Mappings Across Ontologies (Strategy S2)

The more complex S2 strategy lead to 6772 new mappings with respect to those derived with S1 (1966 subclass, 1568 superclass and 3238 disjoint relations) each obtained by combining information across ontologies. Interestingly, despite the fact that this strategy is more complex then S1 as it combines information from more ontologies, the time for deriving an alignment was roughly the same as for S1, i.e., around two days. In the case of the first mapping in Table 7, no ontology contains a relation between *BorealForest* and *Habitat*. However, because *BorealForest* $\sqsubseteq$ *Forest* in O_1[16] and *Forest* $\sqsubseteq$ *Habitat* in O_3,[17] the matcher derived that *BorealForest* $\xrightarrow{\sqsubseteq}$ *Habitat*.

One interesting observation to make is that almost half of the derived mappings are disjoint relations. These are obtained by combining relations between a concept and its generic type (e.g., *Cabbage* $\sqsubseteq$ *Vegetable*) with a relation between the generic concept and one of its disjoints (e.g., *Vegetable* $\perp$ *Meat*). This results in an explosion of new mappings since all the sub-concepts of the generic concept (here, *Vegetable*) are considered disjoint with all its disjoints (here, all the subclasses of *Vegetable*, like *Cabbage*, are disjoint with *Meat*). While these additional mappings are correct, their usefulness is questionable since they establish a relation between concepts at different level of abstraction and are redundant with respect to the original "top-level" disjoint relations.

7.1   Evaluation Results

For evaluating the precision of the alignment obtained with S2, we followed the methodology described in Section 5.3. Our evaluation sample of 1000 mappings (i.e., around 15% of the alignment) contained 478 disjunct, 290 subclass, 232 superclass relations and it was assessed by two groups of evaluators. Table 8 summarizes the results of the evaluation. The agreement coefficient between groups reached 79% (they agreed on 798 evaluations), a value which is close the one obtained for strategy S1 (i.e., 74%). The precision values obtained were 66%

---

[16] http://reliant.teknowledge.com/DAML/Geography.daml
[17] http://protege.stanford.edu/plugins/owl/owl-library/koala.owl

Mappings	Nr.	Examples			
		AGROVOC Concept	**Labels**	**NALT Concept**	**Labels**
Subclass ($\xrightarrow{\sqsubseteq}$)	1966	*c_1014*	Boreal forests, Taiga	*habitats*	habitats
		O_1:BorealForest $\sqsubseteq$ O_2:Forest $\cong$ O_3:Forest $\sqsubseteq$ O_3:Habitat			
		O_1=`http://reliant.teknowledge.com/DAML/Geography.daml` O_2=`http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml` O_3=`http://protege.stanford.edu/plugins/owl/owl-library/koala.owl`			
		c_1584	Cholesterol	*organic_compounds*	organic_compounds, organic compounds, organic chemicals
		O_1:Cholesterol $\sqsubseteq$ O_2:Steroid $\cong$ O_3:Steroid $\sqsubseteq$ O_3:Lipid $\sqsubseteq$ O_3:Organic_Chemical			
		O_1=`http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml` O_2=`http://reliant.teknowledge.com/DAML/SUMO.daml` O_3=`http://onto.cs.yale.edu:8080/umls/UMLSinDAML/NET/SRSTR.daml`			
SuperClass ($\xrightarrow{\sqsupseteq}$)	1568	*c_28628*	Age groups	*elderly*	elderly, aged (people), old age (humans), geriatric people, senior citizens, older people
		O_1:AgeGroup $\sqsupseteq$ O_1:Adult $\cong$ O_2:adult $\sqsupseteq$ O_2:elderly			
		O_1=`http://sweet.jpl.nasa.gov/ontology/human_activities.owl` O_2=`http://owl.man.ac.uk/2003/why/latest/ontology.rdf`			
		c_6170	Prepared foods, Convenience foods, Ready meals, Ready to cook foods	*iced_tea*	iced_tea, iced tea, tea, iced
		O_1:PreparedFood $\sqsupseteq$ O_1:Tea $\cong$ O_2:Tea $\sqsupseteq$ O_3:IcedTea			
		O_1=`http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml` O_2=`http://www.ksl.stanford.edu/projects/DAML/UNSPSC.daml` O_3=`http://www.wam.umd.edu/~katyn/CMSC828y/hw1/hw1.daml`			
Disjoint ($\xrightarrow{\perp}$)	3238	*c_1173*	Cabbages	*meat*	meat
		O_1:Cabbage $\sqsubseteq$ O_1:Vegetable $\cong$ O_2:Vegetable $\perp$ O_2:Meat			
		O_1=`http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf` O_2=`http://www.co-ode.org/resources/ontologies/Pizza_demo_step_1.owl`			
		c_935	Birds, Aves	*plants*	plants
		O_1:Bird $\sqsubseteq$ O_1:Animal $\cong$ O_2:Animal $\perp$ O_2:Plant			
		O_1=`http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf` O_2=`http://dannyayers.com/2003/08/udef.rdfs`			
Total	6772				

Table 7. Some of the mappings discovered with strategy S2.

for the first group, 63% for the second, and 70% when taking into account only the evaluations on which both groups agreed. Note that despite the increased complexity of this strategy, these values are similar to those obtained for S1: 63% and 69% per group, and 70% for the agreed mappings.

	Group 1	Group 2	Agreed by All
Correct	606	645	552
False	305	330	246
Don't know	89	25	7
Precision	**66%**	**63%**	**70%**

Table 8. Evaluation of strategy S2 by both groups.

7.2   Error Analysis

To understand the major causes for false mappings, we manually investigated all the 246 mappings that were rated as *False* by both groups of evaluators. While the same types of errors as in S1 were identified in S2 as well, false mappings obtained by S2 are sometimes caused by more than one error. Indeed, we found 285 causes for the 246 false mappings. This phenomenon is a direct consequence of the fact that S2 exploits more ontologies than S1 and relies on one extra anchoring step. Table 9 displays some examples of typical errors encountered when deriving mappings across ontologies.

Anchoring Errors. We identified 167 anchoring errors. In the case of S1 anchoring errors appear when the source concepts are anchored to semantically different concepts in online ontologies. In the case of S2, an additional anchoring process takes place for the intermediary concept that links the two concepts to be matched. This anchoring process is also prone to errors. For example, in the first mapping from Table 9, the intermediary concept is *Agent*. However, the senses of the concepts with this label in ontologies O_1[18] and O_2[19] are different: a participant in a chemical reaction in O_1 and a role played by a person in O_2.

Ontology Errors. S2 was also influenced by 118 errors specific to low quality online ontologies where subsumption is used incorrectly to model general relations (73 cases, e.g., between *Student* and *University*), part-whole relations (5 cases, e.g., between *Ohio* and *USA*) or roles (6 cases, e.g., between *Vegetable* and *Ingredient*). Some mappings were also derived due to inaccurate labeling (27 cases) or to incorrect views of the world modeled in the used ontologies (7 cases).

---

[18] http://mensa.sl.iupui.edu/ontology/BiologicalOntology.owl
[19] http://www.mindswap.org/2003/owl/swint/terrorism

| Error	Examples				
Type (Nr.)	**AGROVOC Concept**	**Labels**	**Rel.**	**NALT Concept**	**Labels**
Anchoring (167)	*c_5253*	Nucleic acids	$\overset{\sqsubseteq}{\longrightarrow}$	*people*	people, persons, mankind
	O_1:NucleicAcid $\sqsubseteq$ O_1:Agent $\cong$ O_2:Agent $\sqsubseteq$ O_3:Person				
	O_1=`http://mensa.sl.iupui.edu/ontology/BiologicalOntology.owl` O_2=`http://www.mindswap.org/2003/owl/swint/terrorism` O_3=`http://www.mindswap.org/2003/owl/swint/person`				
	c_802	Bamboos	$\overset{\perp}{\longrightarrow}$	*enterprises*	enterprises, businesses
	O_1:Bamboo $\sqsubseteq$ O_1:Plant $\cong$ O_2:Plant $\perp$ O_2:Enterprise				
	O_1=`http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf` O_2=`http://www.dannyayers.com/2003/08/udef.rdfs`				
Ontology Errors (118)	*c_139*	Adults	$\overset{\sqsubseteq}{\longrightarrow}$	*universities*	universities, colleges
	O_1:Adult $\sqsubseteq$ O_1:Student $\cong$ O_2:Student $\sqsubseteq$ O_2:Department $\sqsubseteq$ O_2:Faculty $\sqsubseteq$ O_2:University				
	O_1=`http://www710.univ-lyon1.fr/~s-suwa02/MSch/sc.owl` O_2=`http://www.srdc.metu.edu.tr/~yildiray/HW3.OWL`				
	c_5326	Ohio	$\overset{\sqsubseteq}{\longrightarrow}$	*North_America*	North_America, North America, America, North
	O_1:Ohio $\sqsubseteq$ O_1:USA $\cong$ O_2:USA $\sqsubseteq$ O_2:NorthAmerica				
	O_1=`http://www.cwi.nl/~media/ns/IWA/VideoGen.rdfs` O_2=`http://islab.hanyang.ac.kr/damls/Country.daml`				
	c_13735	Radishes	$\overset{\sqsubseteq}{\longrightarrow}$	*ingredients*	ingredients
	O_1:Radish $\sqsubseteq$ O_1:Vegetable $\cong$ O_2:vegetable $\sqsubseteq$ O_2:ingredient				
	O_1=`http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf` O_2=`http://cvs.sourceforge.net/viewcvs.py/instancestore/instancestore/ontologies/Attic/pizza9.daml?rev=1.2`				

Table 9. Examples of typical errors in compound mappings obtained with S2.

8   Harvesting the Semantic Web

A core assumption of our work is that matching can be performed by *harvesting the Semantic Web*, i.e., by combining appropriate background knowledge from *multiple*, automatically identified online ontologies. In this section we verify this assumption by investigating the number of ontologies that were employed during the matching process.

Strategy S1 explored 226 ontologies to derive 6687 mappings.[20] Figure 4 (a) shows the contribution of each ontology to the alignment in terms of the number of mappings to which it contributed and the percentage that this number represents. An analysis of this chart reveals that there is a high variation in the contribution of different ontologies: a few ontologies provide the majority of the mappings, while most ontologies lead to a small amount of mappings. Indeed, the 11 ontologies (Table 10) for which the percentages are shown in the chart (and which account to about 5% of all used ontologies) lead to deriving approximately 76% of the alignment.

Strategy S2 used 306 ontologies to obtain 6772 mappings (Figure 4 (b)). Given the nature of this technique, i.e., that of combining multiple ontologies, a higher number of ontologies (about 80 more) than in S1 were used to derive approximately the same number of mappings. As in S1, there are a few large ontologies that contribute most mappings, however, their level of contribution is more balanced. Indeed, instead of having a single ontology contributing 17% of the alignment as in S1, in S2, the top four ontologies provide about an equal percentage of the alignment (7%). This is a direct consequence of the fact that mappings are based on multiple rather than on a single ontology.
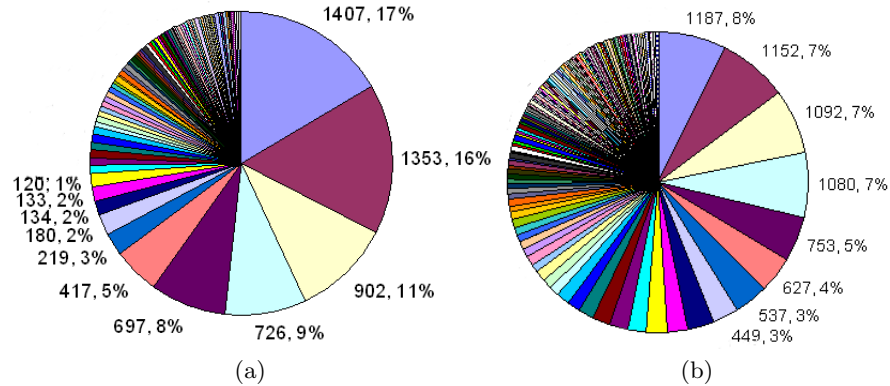
Fig. 4. Contribution to the alignment by ontologies used in (a) S1 and (b) S2.

---

[20] Note that these statistics were computed by considering an ontology to be equivalent to a namespace, independently of the actual, physical location of the concepts in files.

We observe a large overlap between the top contributor ontologies to S1 and S2 (Table 10). The same seven ontologies are used (although with different levels of contribution), with TAP and SUMO being the main contributors in both strategies. Such an overlap is not surprising since these large ontologies have a good coverage of the various topic domains of AGROVOC and NALT.

These statistics strengthen our hypothesis that harvesting the Semantic Web is feasible. Our findings suggest that the strength of the Semantic Web is not only in the use of single, isolated ontologies but also in reusing, combining and making sense of knowledge spread across a variety of different ontologies. Indeed, in such a scenario where large, multi-domain ontologies are matched, it would have been difficult and time-consuming (if not impossible) to manually identify appropriate ontologies in order to derive the same amount of mappings as our technique has done without requiring any a priori knowledge selection.

Ontologies/Ontology Namespaces	Contribution to (%)	
	S1	S2
`http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf`	17%	8%
`http://reliant.teknowledge.com/DAML/SUMO.daml`	16%	7%
`http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml`	11%	4%
`http://reliant.teknowledge.com/DAML/Economy.daml`	9%	3%
`http://gate.ac.uk/projects/htechsight/Technologies.daml`	8%	5%
`http://a.com/ontology`	5%	7%
`http://gate.ac.uk/projects/htechsight/Employment.daml`	3%	-
`http://reliant.teknowledge.com/DAML/WMD.daml`	2%	-
`http://sweet.jpl.nasa.gov/ontology/biosphere.owl`	2%	3%
`http://139.91.183.30:9090/RDF/VRP/Examples`	2%	-
`http://reliant.teknowledge.com/DAML/Geography.daml`	1%	-
`http://www.dannyayers.com/2003/08/udef.rdfs`	-	7%

Table 10. The top contributing ontologies to the alignments obtained with S1 and S2.

9   Comparison With Other Techniques

In this section we investigate how the proposed paradigm compares against and complements existing techniques. We describe our findings both for techniques relying on internal information (Section 9.1) and for background knowledge based techniques (Section 9.2).

9.1   Comparison with Techniques Relying on Internal Information

The five matching systems applied on this dataset during the OAEI'06 contest primarily exploit information that is internal to the two matched ontologies [7, 21, 27, 31, 32]. As such, their results could be used to investigate how our

paradigm relates to techniques from this category. Unfortunately, because all the tools provided exact matches, their evaluation was focused on such mappings: precision was assessed manually, while recall was approximated on a rather small set of 200 mappings containing only 30% of subsumption relations [14]. In addition, while these five systems are complete, self contained tools, our paradigm is intended to be used as a complement to other existing techniques. Indeed, the current implementation does not extract any mappings that lexical and structural techniques can discover (e.g., using basic string comparison on the labels). As a consequence, comparing the recall of this implementation to the one of complete, stand-alone tools would be misleading.

Nr.	Alignment	Nr. Of Mappings
1	COMA++ [32]	7626
2	FALCON-AO [21]	12900
3	PRIOR [31]	11504
4	HMATCH [7]	19924
5	RiMOM [27]	13966
	Union	25224
6	Using the SW	4464
	Union	29688
	Non-redundant	27083
	Non-redundant from 6	**1915**

Table 11. Identifying non redundant mappings.

We can nevertheless draw a set of conclusions which suggest that the proposed paradigm *complements* techniques exploring solely information internal to the matched ontologies. First, since our technique produces other relations than equivalences, a syntactic comparison with the alignments produced by the OAEI'06 tools yields that they are complementary (i.e., there is no overlap between them). Second, in order to *semantically* compare the matching techniques, we applied a redundancy detection mechanism on the union of their alignments[21]. We identified 1915 mappings discovered by our technique which were not redundant with the equivalence mappings identified by the OAEI'06 tools (Table 11). These mappings were obtained by exploring external sources and represent a net contribution to the alignments derived by exploring only information internal to the matched ontologies. Note that even if our technique performs anchoring using techniques similar to those employed by the OAEI'06 tools (i.e., string based comparison), it can identify additional mappings by exploring external sources. Similarly, Aleksovski et al. have shown that using syntactic technique for anchoring and then performing a deduction step using

---

[21] We assume that the relations extracted by the OAEI'06 tools correspond to equivalences and consider only mappings with a confidence value greater than 50%.

background knowledge leads to better performance than when these syntactic techniques are applied directly between the two source ontologies [2].

9.2   Comparison with Background Knowledge Based Techniques

The performances of background knowledge based techniques described in the literature were reported on different data sets, therefore we consider them only as indicative. Unfortunately, S-Match only reports on recall values [16]. The technique of Aleksovski et al. was evaluated on a Gold Standard of mappings for 200 concepts and produced a precision of 76% (compared to 30% and 33% achieved by two traditional techniques on the same dataset) [2]. The matching techniques proposed by van Hage et al. yield a range of precision values for a manually constructed Gold Standard: 17% - 30% when relying only on Google, 38% - 50% when taking into account the context given by the Google snippets, 53% - 75% when exploring a domain specific textual resource, and finally 94% when validating the results of the domain specific extraction with the Google based techniques [50]. We conclude that the 70% precision of our technique, which could eventually be improved through better anchoring to reach 87%, correlates with the performance of the other two techniques (75% - 76%). It is important to note, that the techniques in [2] and [50] reached a high precision when exploring a *single, high-quality, domain specific resource* (i.e., DICE [2], CooksRecipes.coms Cooking Dictionary [50]) while our technique achieves comparable results when *automatically combining multiple, heterogeneous and generic ontologies*. Indeed, we have shown in Section 8 that a high number of ontologies (200 to 300) are automatically discovered and combined.

This comparison indicates that the use of online ontologies leads to comparable performance as when exploring carefully selected, domain specific background knowledge. In addition, our hypothesis is that exploring multiple and dynamically selected ontologies gathers necessary knowlegde that cannot be found in a single, generic resource, even as broad as WordNet or Cyc. Indeed, Figure 4 in Section 8 supports this intuition by depicting that our alignments have been obtained by exploring a few large resources (namely, TAP and SUMO), complemented with a large number of smaller and more specific ontologies. In this line of idea, a set of experiments have been performed to assess the additional knowledge that online ontologies provide with respect to WordNet. We found that only 33% of the alignment obtained with S1[22] (2233 mappings) could have been obtained with WordNet. These findings illustrate that our method maximizes the coverage of background knowledge by exploring complementary, online sources ranging from large, generic resources to small, domain specific ontologies.

10   Conclusions and Future Work

---

[22] This alignment does not contain any WordNet based mappings because we could not explore this ontology through Swoogle due to parsing errors.

In this paper we describe and experimentally investigate an ontology matching paradigm based on the idea of harvesting the Semantic Web. Hereby we summarize our major conclusions and point out future work.

Two of our main findings suggest that *the proposed paradigm is feasible* (re the first research question in Section 1). First, a baseline implementation of the technique applied on a large-scale, real life data set has led to a precision value of 70% for both strategies (Sections 6.1 and 7.1) which correlates with the performance of other background knowledge based matchers (Section 9). Each alignment has been obtained within two days by using average equipment. Given the large size of the data set we consider this time performance reasonable and appropriate for the scenario in which the alignment process took place. Second, our core hypothesis that an alignment can be generated by exploring *multiple* ontologies has been verified since our prototype has automatically selected and reused between 200 and 300 online ontologies (Section 8). In a broader context, these encouraging results indicate the potential of the Semantic Web for solving real life problems [41].

We have experimentally assessed the core limitations of the current implementation (the second research question in Section 1) by investigating the fine-grained research questions stated in Section 3. *A first, major limitation of our prototype is its simple, string comparison based anchoring (RQ1)* which generated more than half of the false mappings for S1 (53%) and also had a significant negative influence on the precision of S2 (Sections 6.2 and 7.2). Indeed, if these mappings could be avoided the precision of S1 would increase from 70% to 87%. Therefore, a high priority task is the design and implementation of an anchoring mechanism that takes into account ontological context. Ongoing experiments with an adaptation of the semantic similarity technique employed in [18] have already lead to promising results [17].

Besides anchoring errors, another major source of false mappings (42% in the case of S1) is the exploitation of online ontologies that contain modeling errors, mostly related to an inaccurate use of subsumption relations to model generic relations, roles and part-whole relations (Sections 6.2 and 7.2). These findings indicate that *the ontology selection mechanism should focus on the quality of the selected ontologies (RQ2)* rather than on their popularity as in the case of Swoogle. Although already considered in the literature [51], the automatic evaluation of such qualitative features remains a challenging area of future work. When investigating the frequency of contradictory mappings (RQ3) we found *a low number of simple contradictions* (affecting only 8 out of 6425 pairs of labels – Section 6.3). This suggests that the implementation of a mechanism for combining mappings from different ontologies would not significantly improve results. At the same time, *complex (alignment level) contradictions are more frequent than expected*, with our automatic incoherence detection mechanism identifying 306 base incoherences that corrupted the entire alignment (since incoherences were inherited by several subclasses of generic concepts). Fortunately, these mappings can be isolated and disposed of automatically, thus leading to the improvement of the alignment (Section 6.3). We plan to integrate an incoherence detection step

into the matcher so that problematic mappings can be identified and excluded already *during* matching.

The third main research question stated in Section 1 refers to the relation of the proposed paradigm with other ontology matching approaches. Already when used as a stand alone matcher our prototype obtained precision values of 70% (and potentially even 87% given a more sophisticated implementation) comparable with the performance of state of the art matching tools (Section 9). *Besides a remarkable performance, the matcher is also complementary to existing techniques and could be more beneficial when used in a hybrid matcher.* Indeed, the obtained alignment is complementary with the results of existing tools (those used during the OAEI'06) i) by providing other relations than equivalences and ii) by identifying a set of mappings that are semantically non redundant with the union of all equivalence mappings obtained by the other tools. A hybrid matcher combining these two types of techniques would derive as many mappings as possible with traditional techniques and then it would explore external background knowledge for those entities about which not enough information exists to derive a mapping. Such a hybrid method has the potential to considerably advance the state of the art in ontology matching by exploring the Semantic Web.

Acknowledgements

References

1. H. Alani. Position Paper: Ontology Construction from Online Ontologies. In *Proc. of WWW*, 2006.
2. Z. Aleksovski, M. Klein, W. ten Katen, and F. van Harmelen. Matching Unstructured Vocabularies using a Background Ontology. In S. Staab and V. Svatek, editors, *Proc. of EKAW*, LNAI. Springer-Verlag, 2006.
3. Y. Arens, C.Y. Chee, C.N. Hsu, and C.A. Knoblock. Retrieving And Integrating Data From Multiple Information Sources . *International Journal of Cooperative Information Systems*, 2(2):127 – 158, June 1993.
4. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.
5. P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt. Contextualizing ontologies. *Journal of Web Semantics*, 1(4):24, 2004.
6. P. Bouquet, L. Serafini, and S. Zanobini. Peer-to-Peer Semantic Coordination. *Journal of Web Semantics*, 2(1), 2005.
7. S. Castano, A. Ferrara, and G. Messa. Results of the HMatch Ontology Matchmaker in OAEI 2006. In Shvaiko et al. [45].
8. P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In *Proc. of WWW*, 2004.

9. C. Collet, M.N. Huhns, and W. Shen. Resource Integration Using a Large Knowledge Base in Carnot. *IEEE Computer*, 24(12):55 – 62, Dec 1991.

10. M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. WATSON: A Gateway for the Semantic Web. In *Proc. of ESWC, Poster Session*, 2007.

11. L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng, and P. Kolari. Finding and Ranking Knowledge on the Semantic Web. In *Proc. of ISWC*, 2005.

12. M. Dzbor, J. Domingue, and E. Motta. Magpie - towards a semantic web browser. In *Proc. of ISWC*, 2003.

13. M. Ehrig and Y. Sure. FOAM - Framework for Ontology Alignment and Mapping; Results of the Ontology Alignment Initiative. In *Proc. of the Workshop on Integrating Ontologies*. 2005.

14. J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. R. van Hage, and M. Yatskevich. Results of the Ontology Alignment Evaluation Initiative 2006. In Shvaiko et al. [45].

15. F. Giunchiglia, P. Shvaiko, and M. Yatskevich:. Semantic Schema Matching. In *Proc. of CoopIS'05*, volume 3760 of *LNCS*, pages 347 – 360, 2005.

16. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Discovering Missing Background Knowledge in Ontology Matching. In *Proc. of ECAI*, 2006.

17. J. Gracia, V. Lopez, M. d'Aquin, M. Sabou, E. Motta, and E. Mena. Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching. ISWC Workshop on Ontology Matching (to appear), 2007.

18. J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the Web: A Multiontology Disambiguation Method. In *Proc. of ICWE*, 2006.

19. T. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199 – 220, 1993.

20. P. Haase and G. Qi. An analysis of approaches to resolving inconsistencies in DL-based ontologies. In *Proc. of Int. WS on Ontology Dynamics (IWOD)*, 2007.

21. W. Hu, G. Cheng, D. Zheng, X. Zhong, and Y. Qu. The Results of Falcon-AO in the OAEI 2006 Campaign. In Shvaiko et al. [45].

22. M.N. Huhns, N. Jacobs, T. Ksiezyk, W. Shen, M.P. Singh, and P.E. Cannata. Integrating enterprise information models in Carnot. In *Proc. of Intelligent and Cooperative Information Systems*, 1993.

23. E. Hyvonen, E. Makela, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, and S. Kettula. MuseumFinland – Finnish Museums on the Semantic Web. *Journal of Web Semantics*, 3(2), 2005.

24. N. Jian, W. Hu, C. Cheng, and Y. Qu. Falcon-AO: Aligning Ontologies with Falcon. In *Proc. of the Workshop on Integrating Ontologies*. 2005.

25. Y. Kalfoglou and M. Schorlemmer. Ontology Mapping: The State of the Art. *The Knowledge Engineering Review*, 18(1):1–31, 2003.

26. J. Lee and R. Goodwin. The Semantic Webscape: a View of the Semantic Web. IBM Research Report, 2004.

27. Y. Li, J. Li, D. Zhang, and J. Tang. Result of Ontology Alignment with RiMOM at OAEI06. In Shvaiko et al. [45].

28. V. Lopez, E. Motta, and V. Uren. PowerAqua: Fishing the Semantic Web. In *Proc. of ESWC*, 2006.

29. V. Lopez, M. Pasin, and E. Motta. AquaLog: An Ontology-portable Question Answering System for the Semantic Web. In *Proc. of ESWC*, 2005.

30. B. Magnini, L. Serafini, and M. Speranza. Making explicit the semantics hiden in schema models. In *Proc. of the Human Language Technology for the Semantic Web WS*, 2003.

31. M. Mao and Y. Peng. PRIOR System: Results for OAEI 2006. In Shvaiko et al. [45].

32. S. Massmann, D. Engmann, and E. Rahm. COMA++: Results for the Ontology Alignment Contest OAEI 2006. In Shvaiko et al. [45].

33. C. Meilicke. Reasoning about Mappings in Distributed Description Logics. Bachelor Thesis, University of Mannheim, August 2006.

34. C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Improving Automatically Created Mappings using Logical Reasoning. In Shvaiko et al. [45].

35. C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Repairing Ontology Mappings. In *Proc. of AAAI*, 2007.

36. P. Mika. Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. *Journal of Web Semantics*, 3(2), 2005.

37. E. Motta and M. Sabou. Next Generation Semantic Web Applications. In *Proc. of ASWC*, 2006.

38. N.F. Noy. Semantic Integration: a Survey Of Ontology-Based Approaches. *SIGMOD Record*, 33(4):65–70, December 2004.

39. E. Rahm and P. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334–350, 2001.

40. M. Sabou, M. d'Aquin, and E. Motta. Using the Semantic Web as Background Knowledge for Ontology Mapping. In Shvaiko et al. [45].

41. M. Sabou, J. Gracia, S. Angeletou, M. d'Aquin, and E. Motta. Evaluating the Semantic Web: A Task-based Approach. In *Proc. of ASWC/ISWC*, 2007.

42. S. Schlobach. Diagnosing Terminologies. In *In Proc. of AAAI*, 2005.

43. S. Schlobach, Z. Huang, R. Cornet, and F. van Harmelen. Debugging Incoherent Terminologies. *Journal of Automated Reasoning*, 2007.

44. P. Shvaiko and J. Euzenat. A Survey of Schema-based Matching Approaches. *Journal on Data Semantics*, IV, 2005.

45. P. Shvaiko, J. Euzenat, N. Noy, H. Stuckenschmidt, R. Benjamins, and M. Uschold, editors. *Proc. of the ISWC Ontology Matching WS*, November 2006.

46. L.M. Stephens, A.K. Gangam, and M.N. Huhns. Constructing Consensus Ontologies for the Semantic Web: A Conceptual Approach. *World Wide Web Journal*, 7(4):421–442, December 2004.

47. H. Stuckenschmidt, L. Serafini, and H. Wache. Reasoning about Ontology Mappings. In *Proc. of the ECAI WS on Context Representation and Reasoning*, 2006.

48. H. Stuckenschmidt, F. van Harmelen, L. Serafini, P. Bouquet, and F. Giunchiglia. Using C-OWL for the Alignment and Merging of Medical Ontologies. In *Proc. of the First Int. WS. on Formal Biomedical K. R. (KRMed)*, 2004.

49. A. Tamilin. Distributed Ontological Reasoning: Theory, Algorithms and Applications. PhD Thesis, University of Trento, February 2007.

50. W. van Hage, S. Katrenko, and G. Schreiber. A Method to Combine Linguistic Ontology-Mapping Techniques. In *Proc. of ISWC*, 2005.

51. J. Völker, D. Vrandecic, and Y. Sure. Automatic Evaluation of Ontologies (AEON). In *Proc. of ISWC*, 2005.

52. H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Huebner. Ontology-Based Integration of InformationA Survey of Existing Approaches. In *Proc. of the WS on Ontologies and Information Sharing*, 2001.