The Knowledge Engineering Review, Vol. 35, e21, 1 of 20.
doi:10.1017/S0269888920000168

Towards evaluating complex ontology alignments

LU ZHOU[1], ELODIE THIÉBLIN[2], MICHELLE CHEATHAM[3], DANIEL FARIA[4], CATIA PESQUITA[5], CASSIA TROJAHN[2] and ONDŘEJ ZAMAZAL[6]

[1]*Data Semantics Laboratory, Kansas State University, Manhattan, USA;*
e-mail: luzhou@ksu.edu
[2]*IRIT & Université de Toulouse 2 Jean Jaurès, Toulouse, France;*
e-mails: elodie.thieblin@irit.fr, cassia.trojahn@irit.fr
[3]*Wright State University, Dayton, USA;*
e-mail: michelle.cheatham@wright.edu
[4]*Instituto Gulbenkian de Ciência, Oeiras, Portugal;*
e-mail: dfaria@igc.gulbenkian.pt
[5]*Lasige, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal;*
e-mail: clpesquita@fc.ul.pt
[6]*Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic;*
e-mail: ondrej.zamazal@vse.cz

Abstract

The development of semi-automated and automated ontology alignment techniques is an important part of realizing the potential of the Semantic Web. Until very recently, most existing work in this area was focused on finding simple (1:1) equivalence correspondences between two ontologies. However, many real-world ontology pairs involve correspondences that contain multiple entities from each ontology. These 'complex' alignments pose a challenge for existing evaluation approaches, which hinders the development of new systems capable of finding such correspondences. This position paper surveys and analyzes the requirements for effective evaluation of complex ontology alignments and assesses the degree to which these requirements are met by existing approaches. It also provides a roadmap for future work on this topic taking into consideration emerging community initiatives and major challenges that need to be addressed.

1 Introduction

Ontology alignments specify the relations that hold between entities in two or more ontologies. Identifying these relations is critical for integrating data across the Semantic Web. The development of automated and semi-automated techniques to establish alignments between ontologies has been an active area of research since at least 2004; however, the vast majority of existing alignment systems seek to identify relatively simple (1:1) equivalence and (more rarely) subsumption relationships. While simple (1:1) relationships are limited in expressiveness by linking single entities, complex matching approaches are able to generate correspondences which better express the relationships between entities of different ontologies. Earlier works have introduced the need for complex alignments (Maedche *et al*., 2002; Visser *et al*., 1997).

Recent work has shown that alignments between pairs of real-world ontologies contain many relations that are more complex than those targeted by current systems. These relations may involve set operations such as union, intersection, disjunction, cardinality restrictions, and other constraints. For example, two ontologies representing the domain of conference organization may have the following relationship between their entities, which states that the class *AcceptedPaper* in the source ontology is equivalent to

the intersection of the class *Paper* with entities that appear in the domain of the *acceptedBy* property: ⟨o1:AcceptedPaper, *intersectionOf*(o2:Paper, *minCardinality*(1, o2:acceptedBy)), ≡, 1.0⟩.

These more complex relationships often make up half or more of the relations within an alignment, as discussed in Zhou *et al*. (2018). It is therefore an important research area for developers of alignment systems to consider. Unfortunately, the topic of complex ontology alignment has received relatively little attention thus far. Different complex matching approaches have emerged in the literature (Ritze *et al*., 2009, 2010; Parundekar *et al*., 2010, 2012; Jiang *et al*., 2016; Walshe *et al*., 2016); however, most efforts on evaluation are still dedicated to the matching approaches dealing with simple alignments.

We posit that part of the reason for the lack of research on complex alignment systems is a lack of benchmarks that contain complex relations and a lack of appropriate metrics with which to evaluate the performance of systems on such benchmarks. The issue of the lack of ontology alignment benchmarks involving complex relationships is being addressed with the introduction of a new complex alignment track within the Ontology Alignment Evaluation Initiative[1] (OAEI), as described in Thiéblin *et al*. (2018a). This paper begins work on the second issue.

The most common evaluation approach for ontology alignments is to perform an exact match comparison between the correspondences suggested by an alignment system and those in a reference alignment and to compute precision and recall based on this. This is a somewhat unforgiving approach. For example, in the case of the aforementioned conference ontologies, if an alignment system identified a relation between the ontologies of the form ⟨o1:AcceptedPaper, *unionOf*(o2:Paper, *minCardinality*(1, o2:acceptedBy)), ≡, 1.0⟩, that is, with union instead of intersection, it would be considered a false positive, and the correct relation would be considered a false negative. While the system should clearly be penalized for not producing the correct relation, considering this as completely incorrect lacks important nuance. For instance, this relation could be relatively easily corrected by a user in a semi-automated alignment system. Moreover, the alignment system developer would likely benefit from knowing how close the system came to generating the correct output in this case.

The primary goal of this position paper is to survey and analyze the requirements for effective evaluation of complex ontology alignments, assess the degree to which these requirements are met by existing approaches, and provide a roadmap for future work on this topic. We begin by discussing related work in Section 2. Section 3 presents the relevant background information, including a formal definition of complex alignments and their representation formats. A generic model of the ontology alignment evaluation process that highlights the choices implicit in implementing a complete evaluation strategy based on reference alignments is then presented in Section 4. The paper then surveys existing ontology alignment evaluation metrics and analyzes their strengths and weaknesses with respect to evaluation of complex alignments when a reference alignment is available. Section 5 overviews the alternative evaluation measures applicable in the absence of reference alignments. The paper continues with a discussion of the gaps that exist between the current state of the art and what is needed for effective evaluation of complex alignments following with its feasibility analysis in Section 6 and then argues about necessary future work to fill in those gaps (Section 7).

2 Related work

Early studies have introduced the need for complex ontology alignments (Visser *et al*., 1997; Maedche *et al*., 2002), and different approaches for generating such alignments have been proposed in the literature since. These approaches rely on diverse methods such as correspondence patterns (Ritze *et al*., 2009, 2010), knowledge-rules (Jiang *et al*., 2016), statistical methods (Parundekar *et al*., 2010, 2012; Walshe *et al*., 2016), or genetic programming (Nunes *et al*., 2011) and path-finding algorithms (Qin *et al*., 2007). While most work on complex ontology matching has been dedicated to the development of complex matching approaches, automatic support for evaluating complex approaches has still not been extensively addressed in the literature.

[1] http://oaei.ontologymatching.org/2018/complex/.

The evaluation of most existing approaches has been done by manually calculating the precision of the alignments generated by the systems (Ritze *et al*., 2009, 2010; Parundekar *et al*., 2012; Walshe *et al*., 2016). In order to be able to measure recall, specific datasets have been constructed. The approach of Parundekar *et al*. (2012) estimated their recall based on the recurring pattern between DBpedia and Geonames:$\exists$ *dbpedia:country.*{`theCountryInstance`} $\equiv \exists$ *geo-names:countryCode.*{`theCountryCode`} where `theCountryInstance` is a country instance of DBpedia such as *dbpedia:Spain* and `theCountryCode` is a country code such as *'ES'*. They estimated the number of occurrences of this pattern between these ontologies and calculated the recall based on this estimation. In Qin *et al*. (2007), a set of reference correspondences between two ontologies was manually created, involving nine reference correspondences from which only two cannot be expressed with simple correspondences. In Walshe *et al*. (2016), the authors proposed an algorithm to create an evaluation dataset that is composed of a synthetic ontology containing 50 classes with known *Class-by-attribute-value* (a correspondence pattern) correspondences with DBpedia and 50 classes with no known correspondences with DBpedia. Both ontologies are populated with the same instances.

As described by Thiéblin *et al*. (2018b), the metrics of *accuracy* and *top-x accuracy* have been also applied in evaluation settings in which the number of correspondences is predefined, for example, there is one correspondence for each entity of the target schema/ontology. The accuracy is then the percentage of predefined questions having a correct answer. A 'question' in this context could be a source entity to be matched and the 'answers' the correspondences having this entity as source member. Some approaches output various answers for each question, for example, a ranked list of correspondences for each source entity. In this case, the top-*x* accuracy is the percentage of questions whose correct answer is in the top-*x* answers to the question. For example, top-3 accuracy is the fraction of source entities for which the correct correspondence is in the three best correspondences output by the system. Alternatively, the approach in Thiéblin *et al*. (2017) to evaluate complex correspondences between agronomic ontologies is based on manually comparing the results of the reference queries and queries automatically rewritten with the help of the complex alignments.

More recently, complex evaluation was introduced in the 2018 OAEI (Thiéblin *et al*., 2018a). The track consisted of four datasets from a variety of domains: conference organization, hydrography, geoscience, and plant taxonomies. Each dataset was evaluated in a different way. For the conference dataset, precision and recall of the system's alignment were manually calculated based on exact match with respect to the reference alignment. For the plant taxonomy dataset, the evaluation was twofold. First, the precision of the output alignment with respect to exact match against the reference was manually assessed. Then, a set of source queries was rewritten using the output alignment. Each rewritten target was then manually classified as correct or incorrect. A source was considered successfully rewritten if at least one of the target queries was semantically equivalent to it. Finally, for the hydrography and geoscience datasets, the evaluation plan was to divide the alignment task into three subtasks and assess performance on each one separately: (1) given an entity from the source ontology, identify all related entities in the source and target ontology; (2) given an entity in the source ontology and the set of related entities, identify the logical relation that holds between them; (3) identify the full complex correspondences. The first subtask was evaluated based on precision and recall with respect to exact match against the reference alignment and the latter two were evaluated using semantic precision and recall.

The evaluation plan for the hydrography and geoscience datasets was not really put to the test in 2018, however, because no alignment systems were capable of finding complex correspondences across these ontologies. The manual nature of the evaluation for the conference organization and plant taxonomy datasets was feasible because only two alignment systems, AMLC and CANARD, were able to generate any complex relations for those datasets; however, there are obvious limitations to a manual approach, both during system development (system developers cannot quickly test modifications to their system to assess whether or not they improve the performance) and evaluation (the time taken is prohibitive for the OAEI track organizers if many systems participate). Additionally, manual evaluation might introduce bias or inconsistencies into the performance assessment.

3 Background

The examples of both simple and complex correspondences provided throughout this paper are based on the OntoFarm ontologies from the conference domain (Šváb *et al.*, 2005; Zamazal & Svátek, 2017). Complex examples are based on the complex version of this dataset, which consists of alignments between all combinations of three of the OntoFarm ontologies, *ekaw*, *cmt*, and *conference*, created by domain experts from three universities who were all familiar with ontology alignment (Thiéblin *et al.* 2018a).

3.1 Complex ontology alignment

We define ontology matching as the process of generating an alignment A between two ontologies: a source ontology O and a target ontology O', as in Euzenat and Shvaiko (2013). A is directional, denoted $A_{O \to O'}$, and is a set of correspondences $\langle e, e', r, s \rangle$. Each correspondence contains a relation r (e.g., equivalence ($\equiv$), subsumption ($\leq$, $\geq$)) between two members e and e', and s expresses the strength or confidence (in [0;1]) of this correspondence. Each member can be a single ontology entity (class, object property, data property, individual, value) of respectively O and O' or a more complex construction that is composed of some entities using constructors or transformation functions.

We consider two types of correspondences depending on the type of their members (Thiéblin *et al.*, 2018; Zhou *et al.*, 2018).

- a correspondence is simple if both e and e' are single entities (represented as IRIs):
  $\langle$*ekaw:Paper, cmt:Paper,* $\equiv$*, 1*$\rangle$
- a correspondence is complex if at least one of e or e' involves a constructor or a transformation function:
  $\langle$*ekaw:AcceptedPaper, someValuesFrom(cmt:hasDecision, cmt:Acceptance),* $\equiv$*, 1.0*$\rangle$
  $\langle$*concatenation(edas:hasFirstName, " ", edas:hasLastName), cmt:name,* $\rightarrow$*, 1*$\rangle$

A simple correspondence is usually noted (1:1), and a complex correspondence can be (1:n) if its source member is a single entity, (m:1) if its target member is a single entity or (m:n) if neither of the members are single entities. Note that these cardinalities refer to the number of entities from the source and target ontologies *in a single correspondence*, not across all correspondences within the alignment. For example, a (1:n) correspondence means that one source entity is related to n target entities via a relationship expressed in a single complex correspondence, not that the same source entity is mapped in a (1:1) manner to n different target entities.

Because relations between instances are generally (1:1) in nature (e.g., *sameAs*, *differentFrom*), complex correspondences predominantly involve entities from the TBox of the ontologies rather than the ABox.

3.2 Representation formats

A general understanding of formats used to express complex correspondences between entities is necessary to comprehend some of the metrics designed to measure the similarity between such correspondences. This section provides an overview of common approaches.

The $\langle e, e', r, n \rangle$ tuples making up a simple alignment are most often encoded using Resource Description Framework (RDF) in a representation format commonly referred to as the Alignment API format, which was introduced in Euzenat (2004). This API is used by the OAEI and has wide adoption within the ontology alignment research community. Version 4 of the Alignment API, described in David *et al.* (2011), also contains a representation format for complex correspondences, known as the Expressive and Declarative Ontology Alignment Language (EDOAL) (Euzenat *et al.*, 2007). While in the simple alignment format e and e' are single Internationalized Resource Identifiers (IRIs), in EDOAL these are expressions involving classes and properties that can be combined using intersection, union, disjunction, and composition operators and/or restricted using constraints on attributes, such as domain, range, cardinality, or value restrictions. The EDOAL representation for

the correspondence ⟨cmt:ProgramCommitteeMember, *someValuesFrom*(conference:was_a_member_of, conference:Program_committee), ≡, 1.0⟩ is shown below.

```
<map>
    <Cell>
        <entity1>
            <edoal:Class rdf:about="&cmt;ProgramCommitteeMember"/>
        </entity1>
        <entity2>
            <edoal:AttributeDomainRestriction>
                <edoal:onAttribute>
                    <edoal:Relation rdf:about="&conference;was_a_member_of"/>
                </edoal:onAttribute>
                <edoal:exists>
                    <edoal:Class rdf:about="&conference;Program_committee"/>
                </edoal:exists>
            </edoal:AttributeDomainRestriction>
        </entity2>
        <measure rdf:datatype="&xsd;float">1.</measure>
        <relation>Equivalence</relation>
    </Cell>
</map>
```

While its general acceptance and associated toolset make EDOAL, a convenient choice for representing complex relationships between ontologies, there are some limitations to this approach. For instance, while EDOAL supports a limited set of transformations, this aspect of the language is somewhat immature. Another issue is that in some cases a concept that is represented as a class in one ontology is modeled as an instance in another ontology (or, one may need to restrict a set of possible instance values involved in a relationship based on their type). This is similar to the OWL concept known as punning, but it is not currently possible in EDOAL. Finally, some relations may be modeled as object properties in one ontology and data properties in another. This occurs frequently when one ontology author has used a 'strings as things' approach while the other has instead created instances. EDOAL does not allow one to specify relationships between object and data properties. Indeed, this is not possible in Web Ontology Language Description Logics (OWL DL) either, though it is permissible in OWL Full.

EDOAL is the most common representation format for complex alignments, but they can be represented in a variety of different ways. For example, OWL can be used directly. This has the benefit of existing tool support for creating, modifying, and reasoning with the alignment, as well as merging ontologies based on it, but it limits the possible complex correspondences to those expressible in OWL (or OWL DL if reasoning is desired), which in particular makes it difficult to encode relationships that involve transformation functions. Another option is to use logical rules following one of a range of different syntaxes, which has the benefit of being generally easier for humans to parse from text than either EDOAL or OWL, but there is a lack of tool support for direct use of alignments expressed in this way. Other possibilities for complex correspondence representation include using a dedicated vocabulary or representing them as queries. As described in Xiao *et al.* (2018), in the area of OBDA (Ontology-Based Data Access) the R2RML format, a W3C standard, has been extended in many different ways, including for this purpose. For a more complete survey on the representation of ontology alignments, we refer the reader to the one presented in Scharffe (2009).

4 Evaluation with a reference alignment

The evaluation of ontology alignments is often performed with respect to a reference alignment, as is the case in most of the OAEI tracks. Usually, this evaluation relies on the traditional information retrieval

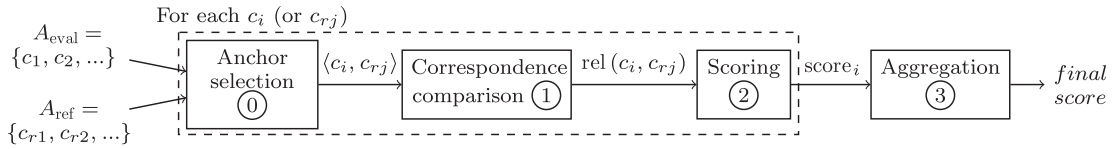

Figure 1 Evaluation process of the alignment A_{eval} with the reference alignment A_{ref}

evaluation metrics of precision and recall, and only contemplates correspondences that are exactly equal between the evaluated and reference alignments. However, as we will overview in this section, several alternative approaches to score inexact matches between the evaluated and reference alignments have been proposed.

4.1 Generic evaluation process

The generic process of evaluating an ontology alignment A_{eval} using a reference alignment A_{ref} can be decomposed into four steps, as schematized in Figure 1: anchor selection, correspondence comparison, scoring, and aggregation. Note that these steps are not independent, and in fact, much existing work on the topic of ontology alignment evaluation conflates the latter three steps (Ehrig & Euzenat, 2005; Euzenat, 2007). In practice, the correspondence comparison approach selected and corresponding scoring scheme have ramifications throughout the evaluation process.

In the anchor selection step, the set of correspondences c_{rj} from the reference alignment A_{ref} that have to be compared with each correspondence c_i from the evaluated alignment A_{eval} (or vice versa) is computed. This selection depends on the correspondence comparison approach adopted. In the traditional evaluation where only exactly matching correspondences are to be scored, only these need be selected in this step. But if related correspondences are also contemplated, then each evaluated correspondence may have several such correspondences in the reference alignment, and all of them will need to be compared unless it is evident *a priori* which is the most similar (e.g., if there is an equivalent correspondence).

In the correspondence comparison step, for each pair of correspondences $\langle c_i, c_{\text{rj}} \rangle$, where $c_i = \langle e_i, e_i', r_i, s_i \rangle$ and $c_{\text{rj}} = \langle e_{\text{rj}}, e'_{\text{rj}}, r_{\text{rj}}, s_{\text{rj}} \rangle$, a relation $\text{rel}(c_i, c_{\text{rj}})$ between c_i and c_{rj} is computed. $\text{rel}(c_i, c_{\text{rj}})$ can be decomposed into the relations between the elements of c_i and c_{rj}:

$$\text{rel}(c_i, c_{\text{rj}}) = \begin{cases} \text{rel}(e_i, e_{\text{rj}}) \\ \text{rel}(e_i', e'_{\text{rj}}) \\ \text{rel}(r_i, r_{\text{rj}}) \\ \text{rel}(s_i, s_{\text{rj}}) \end{cases}$$

As we will overview in Section 4.2, relations between entities include syntactic equivalence, semantic equivalence, and semantic relatedness; the relation between the correspondence relations includes equivalence and relatedness; and the relation between the confidence scores, when considered, is typically numerical similarity. In the traditional evaluation, the relation between the confidence scores is ignored, and correspondences are considered equivalent if both the entities and the correspondence relation are syntactically equivalent.

In the scoring step, a scoring function is applied to the relation $\text{rel}(c_i, c_{\text{rj}})$ between c_i and c_{rj}. This is usually done by applying the scoring scheme associated with the correspondence comparison approach to score the relations between each element in the correspondences, then multiplying these scores. score_i is the result of this scoring function. In the traditional evaluation, equivalent correspondences are treated as true positives and scored 1, and no other correspondences are scored.

In the aggregation step, the scores are aggregated over the whole alignment to produce the *final score*. In the traditional evaluation, this aggregation means computing precision and recall by tallying the true positives and dividing by the number of correspondences in the evaluated and reference alignments, respectively. Correspondences in the evaluated alignment that are not in the reference alignment are false positives, and those in the latter and not in the former are false negatives. In cases where inexact

correspondence matches are contemplated, then the aggregation must also include the selection of which correspondence pairs to score, as each evaluated correspondence may have a non-zero score when compared with several reference correspondences. Intuitively, it makes sense to select only the most similar reference correspondence for each evaluated correspondence (which in the trivial case would be an exact match). However, the fact that multiple evaluated correspondences may have the same reference correspondence as the most similar makes this selection less straightforward. There is some argument to enforcing that each correspondence from both the reference and evaluated alignments be considered only once in the aggregation, but this may not make sense when neither evaluated correspondence is related to any other reference correspondence.

While this generic evaluation workflow is valid for both simple and complex alignments, it is challenging to apply it to complex alignments due to the fact that complex correspondences feature expressions of arbitrary complexity with a wide range of constructs, rather than singular entities. Thus, one cannot simply compare Uniform Resource Identifiers (URIs) of the mapped entities between two correspondences and check for identity or a semantic relation between them, as there are additional layers to contemplate when comparing correspondences. This affects the anchor selection step as it may not be trivial to determine that two complex correspondences are related in a manner that is computationally more efficient than the worse–case scenario of skipping anchor selection and making the full pairwise comparison of all correspondences in the subsequent step. It also affects the correspondence comparison step, as determining the relation between complex entities requires comparing all the singular entities they list, as well as the expressions in which they are listed, likely in recursive fashion, as there is no theoretical limit to the nesting of expressions within expressions. Furthermore, there are cases where one might want to consider making a joint evaluation of two or more correspondences against a single reference correspondence, which complicates both the correspondence comparison and the aggregation step. For example, consider this reference correspondence from the Conference test set in the complex alignment track of the OAEI:

⟨*intersectionOf*(ekaw:Paper_Author, *complementOf*(*someValuesFrom*(ekaw:reviewerOfPaper, ekaw: Paper))), *intersectionOf*(conference:Regular_Author, *complementOf*(conference:Review- er)), ≡,1.0 ⟩

Consider also the following two correspondences produced by an alignment system:

⟨ekaw:Paper_Author, conference:Regular_Author, ≡, 1.0 ⟩
⟨*someValuesFrom*(ekaw:reviewerOfPaper, ekaw:Paper), conference:Reviewer, ≡, 1.0 ⟩

In this scenario, if neither of the system correspondences were in the reference alignment, it is arguable that both should be scored against the reference correspondence together, as the latter can be logically derived from them ($\{A \equiv A'; B \equiv B'\} \Rightarrow A \cap !B \equiv A' \cap !B'$).

An additional challenge to the evaluation of complex alignments is that, in practice, there is a greater variety of correspondence relationships, since most simple ontology alignment benchmarks consist entirely of equivalence relations. This aggravates the difficulty in comparing correspondences, as the relation may factor into how two correspondences are related. Picking up on our example above, consider the following correspondence produced by a matching system:

⟨*intersectionOf*(ekaw:Paper_Author, *complementOf*(*someValuesFrom*(ekaw:reviewerOfPaper, ekaw: Paper))), conference:Regular_Author, ≤, 1.0 ⟩

This correspondence is logically derived from the reference correspondence, and thus formally correct (if less specific than desired), whereas it would not be correct if the relation were equivalence.

Finally, the several layers involved in comparing complex correspondences make it desirable to use comparison approaches that generate more nuanced similarity scores than the simple all-or-nothing approach traditionally used in alignment evaluation. This means that there will likely be more correspondence comparisons involved in evaluating complex alignments, and the aggregation step will be less straightforward.

4.2 Existing approaches for correspondence comparison

As we detailed in the previous section, the correspondence comparison approach affects the whole evaluation workflow, as it determines which correspondences are selected as anchors, how they are compared and how they should be scored, as well as how they can be aggregated. Due to this central importance, and to the fact that they are the characterizing factor of different forms of alignment evaluation, this section is devoted to surveying existing approaches for correspondence comparison and discussing their application to complex alignments. This is not an exhaustive survey, but rather an attempt to provide insights on the strengths and weaknesses of each type of approach when used to evaluate complex ontology alignments.

The following example will be used throughout this section. Correspondences in the reference alignment (R):

1. ⟨cmt:Author, conference:Regular_Author, ≡, 1.0⟩
2. ⟨cmt:ProgramCommitteeMember, *someValuesFrom*(conference:was_a_member_of, conference:Program_committee), ≡, 1.0⟩
3. ⟨cmt:User, *unionOf*(conference:Regular_Author, conference:Reviewer), ≥, 1.0⟩
4. ⟨cmt:AuthorNotReviewer,*intersectionOf*(conference:Regular_Author, *complementOf*(conference:Reviewer)), ≡, 1.0⟩

Correspondences generated by alignment system 1 (S1):

1. ⟨cmt:Author, conference:Regular_Author, ≤, 1.0⟩
2. ⟨cmt:ProgramCommitteeMember, *minCardinality*(1, conference:was_a_member_of, conference:Program_committee), ≡, 1.0⟩
3. ⟨cmt:User, conference:Regular_Author, ≥, 1.0⟩
4. ⟨cmt:User, conference:Reviewer, ≥, 1.0⟩
5. ⟨cmt:AuthorNotReviewer, *unionOf*(conference:Regular_Author, conference:Reviewer), ≡, 1.0⟩

Correspondences generated by alignment system 2 (S2):

1. ⟨cmt:Author, conference:Contribution_1th-Author, ≡, 1.0⟩ (Note that conference:Contribution_1th-Author is a subclass of conference:Regular_Author)
2. ⟨cmt:AuthorNotReviewer,*intersectionOf*(conference:Conference_participant, *complementOf*(conference:Committee_member)), ≡, 1.0⟩

We will use the notation $c_i =< e_i, e'_i, r_i, s_i >$ to refer to any correspondence generated by an alignment system, and $c_{\mathrm{rj}} =< e_{\mathrm{rj}}, e'_{\mathrm{rj}}, r_{\mathrm{rj}}, s_{\mathrm{rj}} >$ to refer to any reference correspondence.

4.2.1 Syntactic

Syntactic approaches to alignment evaluation compare the elements of two correspondences based on their syntactic description (i.e., the URIs of entities, or the identifiers of correspondence relations or complex expressions). This includes the traditional evaluation approach of scoring only exact matches, where a correspondence is scored 1 if both of its entities and its relation are syntactically equivalent to the reference correspondence (i.e., $e_i \equiv e_{\mathrm{rj}}$, $e_i' \equiv e'_{\mathrm{rj}}$, and $r_i \equiv r_{\mathrm{rj}}$) and scored 0 otherwise.

This exact match approach is used to compare correspondences in most existing work on ontology alignment, including in most OAEI tracks and in the majority of ontology alignment papers. In fact, if papers do not explicitly state what evaluation approach they are using, it is assumed to be exact match. Thus, this approach has the advantage of being both simple and widely used. It is often possible to compare the results of an alignment system to previous work based on this approach by referring to the original papers rather than re-running the experiments. Furthermore, available computational tools for handling ontology alignments, such as the Alignment API, usually contain evaluation facilities based on exact match and do not require users to write additional code.

However, this approach is unforgiving in that it treats as incorrect correspondences that, while not listed in the reference alignment, can be logically derived from it (or even equivalent to it), and thus are formally correct. Furthermore, it does not distinguish between correspondences that are formally incorrect but closely related to correct correspondences, and those that are completely incorrect. Referring to the example above, the first correspondence in the reference alignment is an equality, but the first alignment system identifies the relation between the same entities as subsumption. This is considered completely incorrect under the exact match approach, even though it is formally correct (if imprecise) and may be a useful result in some applications of the alignment, such as query answering. Meanwhile, the second alignment system correctly identified the equality relationship for cmt:Author, but rather than conference:Regular_Author, it specified conference:Contribution_1th-Author, a subclass of conference:Regular_Author, as the equivalent entity. This is formally incorrect, but the correct correspondence can be inferred from it, so it is only partially incorrect, as the correspondence holds true for a subset of cmt:Author. If the alignment system had specified conference:Chair, which has no relation at all to conference:Regular_Author, then the correspondence would be fully incorrect. The case of the second reference correspondence is even more grave, as the first system identified a correspondence that is syntactically different but logically equivalent and thus formally correct. Under a syntactic approach, this correspondence would result in both a false positive and a false negative (as the syntactically correct correspondence is missing), whereas it should clearly result in a true positive. Regarding the third correspondence from the reference alignment, the first alignment system states that cmt:User is related to both conference:Regular_Author and conference:Reviewer, yet this is also treated as incorrect (specifically, as one false negative and two false positives), because the system specified each relation separately instead of as a union. Finally, both alignment systems generate relations that are somewhat similar to the fourth one from the reference alignment. The first system has the correct entities but incorrect expressions while the second has the expressions correct but incorrect entities. Both of these are treated as completely incorrect.

An alternative to the traditional binary syntactic evaluation is the weighted syntactic evaluation, where the confidence scores of the alignment to evaluate and those of the reference alignment are taken into consideration. This is particularly relevant when the reference alignment is not considered ground truth and has similarity scores other than 1, such as in the approach proposed by Cheatham & Hitzler (2014). In this approach, which is also implemented in the Alignment, the true-positive count is replaced by the sum of the products of confidence scores $s_i * s_{\text{rj}}$, and the false positive and false-negative counts were replaced by the sum of differences of confidence scores $|s_i - s_{\text{rj}}|$, respectively, for $s_i < s_{\text{rj}}$ and $s_i > s_{\text{rj}}$. This penalizes an alignment system more if it fails to identify a strong correspondence than a weak one and rewards the alignment system if its scoring scheme approximates the confidence scores of the reference alignment. A similar methodology, albeit relying on a vector representation of the ontology alignments, was also proposed by Sagi & Gal (2018).

4.2.2 Rule-based semantic and reasoning-based semantic

Semantic approaches compare correspondences based on their semantic meaning rather than their syntactic representation. This is done by looking at the correspondence within the context of the ontologies and determining whether they are semantically related. If they are 'closely' related, but not equivalent, they are typically scored in (0; 1], depending on the scoring scheme of the approach.

An example of such an approach is the relaxed precision and recall metric proposed by Ehrig & Euzenat (2005), which defines different similarity functions for the various elements of a correspondence, depending on whether precision or recall is to be computed. It scores the similarity between two entities, e_i and e_{rj}, according to:

$$\text{entity prec similarity} = \begin{cases} 1 & \text{if } e_i \leq e_{\text{rj}} \\ 0.5 & \text{if } e_i > e_{\text{rj}} \\ 0 & \text{otherwise} \end{cases} \qquad \text{entity rec similarity} = \begin{cases} 1 & \text{if } e_i \geq e_{\text{rj}} \\ 0.5 & \text{if } e_i < e_{\text{rj}} \\ 0 & \text{otherwise} \end{cases}$$

where $>$ and $<$ stand for direct sub- or super-classes/properties only. The similarity between two relations is defined only for the case where r_i is $\equiv$ (as most matching systems tend to produce only equivalence correspondences) and depends only on r_{rj} according to:

$$\text{relation prec similarity} = \begin{cases} 1 & \text{if } \equiv \text{ or } < \\ 0.5 & \text{if } > \\ 0 & \text{otherwise} \end{cases} \qquad \text{relation rec similarity} = \begin{cases} 1 & \text{if } \equiv \text{ or } > \\ 0.5 & \text{if } < \\ 0 & \text{otherwise} \end{cases}$$

Finally, the similarity between confidence scores, s_i and s_{rj}, is scored according to:

$$\text{score similarity} = 1 - |s_i - s_{\text{rj}}|$$

This approach thus aims to reward correspondences that are semantically close to the correct correspondence from the perspective of query answering. Namely, in the case of precision, it does not penalize at all correspondences that are narrower than the correct correspondence (but implied by it) since these would result in missing but only correct query results (full precision). Likewise, in the case of recall, it does not penalize correspondences that are broader than the correct correspondence (and imply it) since these would result in no missing results but some incorrect ones (full recall).

Another semantic approach also proposed by Ehrig & Euzenat (2005) focuses on the perspective of alignment validation rather than query answering and seeks to account for the effort it would take a human reviewer to correct an erroneous correspondence that is semantically close to the correct one. This approach can be considered a simple edit-distance approach, as it attributes a cost to each edition necessary for converting an incorrect correspondence to a correct one. Under this approach, the similarity between entities is given by:

$$\text{entity effort similarity} = \begin{cases} 1 & \text{if } e_i \equiv e_{\text{rj}} \\ 0.6 & \text{if } e_i < e_{\text{rj}} \\ 0.4 & \text{if } e_i > e_{\text{rj}} \\ 0 & \text{otherwise} \end{cases}$$

where again, $>$ and $<$ stand for direct sub- or super-classes/properties only. The rationale behind attributing a different similarity to sub- and super-entities is that typically ontology entities are expected to have more sub- than super-entities, and thus correcting to a broader entity requires less effort than correcting to a narrower entity. The similarity between relations is 1 if the relations are the same and 0.5 if they are different, under the rationale that correcting the relation is fairly trivial even if the relation predicted by the matching system is completely off.

Another semantic approach is the semantic precision and recall proposed by Euzenat (2007). Under this approach, a reasoner is employed to count the number of correspondences suggested by the alignment system that are entailed by a merged ontology consisting of the source and target ontologies and the reference alignment. This count is then divided by the number of correspondences in the proposed alignment to produce the system's precision. Analogously, recall is computed by counting the number of relations in the reference alignment that are entailed by a merged ontology consisting of the source and target ontologies and the proposed alignment, then dividing this by the number of correspondences of the reference alignment.

Regardless of their scope and implementation, semantic approaches tend to mitigate some of the issues we reported for syntactic approaches, since they account for correspondences that are semantically close to the correct ones. Here we discuss the score produced by each metric for the example alignments presented at the start of Section 4.2. This information is summarized in Table 1. The first alignment system identified the relation as subsumption rather than equivalence. This would score 0 under a syntactic approach, but would score, respectively, 1 and 0.5 in relaxed precision and recall, 0.5 in effort similarity, and 1 and 0 in semantic precision and recall (as subsumption is entailed by but does not entail equivalence). Likewise, the correspondence proposed by the second system, in which the entity from the target ontology was a subclass of the correct entity, would score 0 under a syntactic approach, but, respectively,

Table 1 Scores of the surveyed metrics on the sample alignments

Issue	Related items	Exact match	Relaxed prec.	Relaxed rec.	Effort	Sem. prec.	Sem. rec.
Mismatched relation	R1,$S1_1$	0	1	0.5	0.5	1	0
Subclass rather than exact class match	R1,$S2_1$	0	1	0.5	0.6	0	0
Logically equivalent	R2,$S1_2$	0	0	0	0	1	1
Correct but expressed as multiple correspondences	$S1_3$, $S1_4$	0	0	0	0	1	0
Correct entities; incorrect construction	R4,$S1_5$	0	0	0	0	1	0
Correct construction; incorrect entities	R4,$S2_2$	0	0	0	0	0	0

1 and 0.5 in relaxed precision and recall, and 0.6 in effort similarity. In this case, it would also score 0 under semantic precision and recall, as equivalence to a class neither entails nor is entailed by equivalence to its superclass. The contrary happens in the second reference correspondence in the example, in which the first system produced a logically equivalent correspondence. In this case, the system's correspondence would score 1 under semantic precision and recall, but would still be scored 0 under relaxed precision and recall as well as effort similarity, as these rule-based approaches have no provisions for complex alignment expressions and thus cannot detect that these correspondences are logically equivalent. Similarly, the fifth correspondence of the first system, which differs from the fourth correspondence of the reference alignment only in that a union was used instead of an intersection, would score, respectively, 1 and 0 in semantic precision and recall (as equivalence to the union entails equivalence to the intersection but not the other way around) but also be scored 0 under the other approaches, for the same reason as in the previous case. From the perspective of alignment validation, such a correspondence should be fairly trivial to correct, and thus should have a non-zero score. In the case of the third reference correspondence, for which system one predicts two related correspondences (3 and 4), these would also be scored, respectively, 1 and 0 under semantic precision and recall (as superclass of the union entails superclass of each element in the union but not the other way around), but again 0 under the other approaches, as they have no provision for comparing correspondences other than on a one-to-one basis.

In summary, the main limitations of rule-based semantic approaches with respect to complex alignments are that no such approach has been proposed that encompasses the range of expressions possible in these alignments, and that they do not contemplate joint correspondence evaluation in the cases where a correspondence is decomposed into several related ones. Furthermore, proposed approaches are coarse in granularity and only distinguish between identical entities, direct sub-/super-entities, and all other cases. They do not account for cases of other relations, such as indirect sub-/super-entities, even though Ehrig & Euzenat (2005) did suggest that more a granular approach could do so by explicitly taking into the account the edge-distance between entities in the similarity function.

By contrast, reasoning-based semantic approaches do account for all complex expressions that can be encoded in OWL and also handle cases of correspondence decomposition well. However, they only score correspondences that are logically entailed, ignoring those that are semantically related but not entailed. Thus, in cases that can be handled by both reasoning-based and rule-based approaches, reasoning-based approaches are stricter in their assessment of performance for purposes such as query answering or alignment validation. Furthermore, reasoning-based approaches are computationally more complex than rule-based approaches and may not be applicable in practice to very large ontologies, as reasoning over these is still a computational challenge. Finally, reasoning is only possible if the merged ontology is in OWL DL, which may not be the case in complex alignments even if the original ontologies are (for example, if a correspondence is made between an object property and a datatype property).

4.2.3 Instance based

Instance-based approaches compare two correspondences between ontology classes based on the overlap between their sets of instances. In Isaac *et al*. (2007), instance-based similarity measures are divided into two primary categories: traditional set similarity metrics and information-theoretic measures. Set similarity is most often computed based on the Jaccard index, which is the ratio of instances that belong to both the source and target classes to the number of instances belonging to either the source or target classes. Information-theory measures reflect the degree to which knowledge of an instance's categorization via e of a correspondence provides knowledge about the appropriateness of the e' categorization. Examples include point-wise mutual information, log likelihood ratio, and information gain. More recent work has proposed instance-based metrics based on locality-sensitive hashing (Duan *et al*., 2012) an on Cohen's kappa coefficient (Kirsten *et al*., 2007).

Instance-based correspondence comparisons are powerful in that they directly correspond to the underlying definition of ontological entities as sets of instances that are related in some way. However, the applicability of such metrics is limited to the evaluation of class correspondences, and only in cases in which common instances exist in both ontologies. These common instances can either be the same individuals (with identical URIs) or individuals with different URIs that have been declared identical through the use of a co-reference resolution procedure (though this procedure can of course introduce errors that would negatively impact the alignment evaluation). Furthermore, even if dual-typed instance data exists, there may be particular valid complex correspondences for which few or no instances are available, which can compromise the evaluation (even though some, metrics such as the log likelihood ratio and the modified version of the Jaccard metric described in Isaac *et al*. (2007), handle sparse data better than others). A solution for handling sparse data is to synthetically generate additional instance data, as described in Schopman *et al*. (2012), but this has the potential of biasing the evaluation and no assurance of covering particular complex correspondences better.

5 Evaluation without a reference alignment

Constructing reference alignments is a time-consuming task that requires the involvement of domain experts. In the absence of time, an alternative evaluation strategy can be the manual validation of sample alignments, as detailed in Van Hage *et al*. (2007), although this still requires significant involvement of domain experts. Alternative approaches consider the generation of natural language questions to support end-users in the validation task (Abacha & Zweigenbaum, 2014) or validation of correspondences in a semi-automatic way (Serpeloni *et al*., 2011).

In the absence of both reference alignments and domain experts, there are two families of approaches to ontology alignment evaluation: one that uses quality metrics to assess the logical soundness of the alignment (Meilicke & Stuckenschmidt, 2008; Solimando *et al*., 2017), and another that focuses on the suitability of the alignment for a specific task or application (Isaac *et al*., 2008; Hollink *et al*., 2008; Solimando *et al*., 2014). In this section, we discuss how complex alignments can be evaluated using these strategies.

5.1 Alignment quality metrics

The union of two ontologies through an alignment can lead to logical errors such as unsatisfiable classes (i.e., classes than can only be interpreted as empty sets) even if both ontologies were originally logically sound. In such cases, the merged ontology is said to be incoherent, and by extension, so is the ontology alignment. Since, for many applications, incoherence would cause problems, there are several approaches to measure ontology incoherence (Qi & Hunter, 2007). Derived from these Meilicke & Stuckenschmidt (2008) proposed two measures to assess an alignment's quality based on its logical coherence: one based on counting unsatisfiable classes; and another, named maximum cardinality measure (degree of incoherence), based on the minimum number of correspondences that must be removed to obtain a coherent merged ontology. Additionally, the authors proposed a variant of the latter measure that considers the confidence scores of the correspondences and measures the minimum loss of total confidence required for

coherence, called the maximum trust measure. Interestingly, they reported that the maximum cardinality measure can be used to compute a strict upper bound of precision (Meilicke & Stuckenschmidt, 2008).

Also on the topic of logical soundness, Jiménez-Ruiz *et al.* (2011) proposed three principles for ontology alignments: consistency, conservativity, and locality. The consistency principle states that correspondences should not lead to unsatisfiable classes in the merged ontology. This is a bit of a misnomer, as the principle pertains to ontology coherence (all classes are realizable) rather than ontology consistency (there are no contradicting axioms). Compliance with this principle can thus be assessed by using the metrics described above. The conservativity principle states that correspondences should not introduce, in the merged ontology, new semantic relationships between any two entities that were originally from the same input ontology. Compliance with this principle can be assessed by counting the number of violation to it, as proposed by Solimando *et al.* (2017). Finally, the locality principle states that correspondences tend not to be semantically isolated in the ontologies, which is to say, two semantically related concepts from one of the input ontologies are more likely to be aligned to two concepts from the other input ontology that are themselves semantically related, than to unrelated concepts. This principle is more a guideline for identifying potential false correspondences than a basis for assessing alignment quality, which is why no metric to assess its compliance has been proposed.

We must note that coherence and conservativity may sometimes be at odds with alignment completeness, as ontologies have different modeling views of their domain, which have to be reconciled when two ontologies are merged, possibly leading to new semantic relations between entities of one or both of them, as well as to logical conflicts (Pesquita *et al.*, 2013). Thus, it may very well be that the complete and correct alignment between two ontologies is incoherent and/or unconservative. Nevertheless, alignment coherence is critical for several applications, such as ontology merging and query answering, and therefore is commonly used as an evaluation criterion in the OAEI, in tracks such as *Anatomy*, *Conference*, *Large Biomedical Ontologies*, *Disease and Phenotype*, and *Ontology Alignment for Query Answering (OA4QA)*. Evaluation modalities include binary assessment of coherence, the maximum cardinality measure, and the number or fraction of unsatisfiable classes. The relevance of alignment conservativity is more debatable, as it is not strictly required for any application, but it has also been used as an evaluation criterion in the OAEI *Conference* and *OA4QA* tracks. Note also that neither coherence nor conservativity evaluations are a substitute for an evaluation of alignment completeness and correctness, and they have always been used in complement of the latter in the OAEI. In the extreme case, an empty alignment is fully coherent and conservative, but utterly useless.

In complex alignments, assessing coherence is particularly desirable, as the very interest in making a complex alignment is underpinned by a concern with semantic precision beyond what simple alignments allow. However, assessing coherence requires reasoning and is computationally challenging, particularly for large and/or semantically complex ontologies, and even more so if the alignment itself is large and/or complex. Even more important, assessing coherence requires that the merged ontology be expressible in OWL DL, which may not be the case in complex alignments, even if the input ontologies are. Some complex correspondences are not expressible in OWL at all, while others are expressible in OWL but not OWL DL.

Assessing conservativity of complex alignments makes less sense than doing so for simple alignments, as complex alignments tend to contribute substantially to the semantics of both input ontologies by design (e.g., by defining ontology restrictions) and thus it is not at all unexpected that they lead to conservativity violations. That said, assessing conservativity violations in complex alignments should be little harder than doing so for simple alignments, assuming the correspondences can be encoded in OWL DL.

5.2 Task-based evaluation

The quality of an alignment can also be assessed regarding its suitability for a specific task or application. Considering that ontology alignments are, in practice, constructed for a given application or with a given task in mind, it would be useful to set up experiments that do not stop at the delivery of the alignment but carry on to the application or task for which the alignment was constructed. This is especially true when

there is a clear measure of success for the overall task or application, but even when there is not, it can be useful to share corresponding aggregate measures associated with a task or application profile.

With respect to application-oriented evaluation, Isaac *et al*. (2008) proposed ontology alignment evaluation methods for the specific scenarios of thesaurus merging and data translation. They defined sets of tasks which need an alignment or part of it, then evaluated the alignment on how well it fulfilled these tasks in terms of quality (for each task how good is the answer) and quantity (how many tasks were fulfilled by the alignment).

Regarding task-oriented evaluation, Euzenat and Shvaiko (2013) argued that different task profiles can be established to explicitly compare matching systems for certain tasks, such as ontology evolution or query answering, that have different constraints in terms of coverage and run time. One such task-oriented evaluation approach was introduced in the OAEI in 2015 at the *OA4QA* track[2] (Solimando *et al*., 2014), which focused on the task of query answering. This track used a synthetically populated version of the *Conference* dataset and a set of manually constructed queries over these ABoxes. A given query, such as $Q(x)$:=Author(x) expressed using the vocabulary of the Cmt ontology, was executed over the merged ontology $Cmt \cup Ekaw \cup A$, where A is an alignment between *Cmt* and *Ekaw*. Precision and recall were calculated with respect to model answer sets, that is, for each ontology pair and query $Q(x)$, and for each alignment A computed by each matching system. An alternative approach for evaluating query answering without using instances was proposed by David *et al*. (2018), where queries are compared without instance data, by grounding the evaluation on query containment.

While task-based evaluation is equally valid for both simple and complex alignments, some tasks tend to have higher expressiveness requirements, and thus to more often involve complex alignments, such as query answering/rewriting and ontology merging (Thiéblin *et al*., 2018). Query answering in particular has already been a subject of focus for complex alignments, with Makris *et al*. (2012) presenting a set of complex correspondences used for query rewriting[3] for a few pairs of ontologies. More recently, complex correspondences have been exploited for the task of query rewriting for federating agronomic taxonomy knowledge on the LOD cloud (Thiéblin *et al*., 2017). This (Taxon) dataset was also used in the *Complex* track of the OAEI 2018 campaign, with the aim of assessing the performance of matching systems over large knowledge bases. The evaluation was performed based on the quality of the generated alignments (in terms of precision) and on the ability to rewrite SPARQL queries using these alignments. In particular, a manual analysis of the number of queries satisfyingly rewritten based on the alignments was carried out. The queries written for the source ontology were rewritten automatically when dealing with (1:1) or (1:n) correspondences, using the system described by Thiéblin *et al*. (2016) and manually when dealing with (m:n) correspondences.

Given the relevance of complex alignments for query answering, and the fact that this task is one of the main applications of these alignments, evaluation approaches based on this task would be highly relevant. One of the main challenges in implementing such approaches lies in establishing a query rewriting scheme that encompasses the expressivity and cardinality of complex correspondences. In the case of simple alignments, a naive approach for rewriting SPARQL queries can be to simply replace the IRI of an entity of the initial query by the IRI of the corresponding entity in the alignment, as described in David *et al*. (2011). For complex alignments, such a naive approach is obviously not possible, as the semantics of the alignment itself has to be taken under consideration. Euzenat *et al*. (2008) proposed an approach for writing specific SPARQL *construct* queries, but most query rewriting systems still rely on simple or (1:n) complex correspondence and fail in covering highly expressive (m:n) complex correspondences.

6 Discussion

The nature of complex ontology alignments presents unique evaluation challenges that were not considered when existing evaluation techniques were developed. This section outlines those challenges and analyzes the areas in which current approaches are lacking.

[2] http://www.cs.ox.ac.uk/isg/projects/Optique/oaei/oa4qa/index.html.
[3] http://www.music.tuc.gr/projects/sw/sparql-rw/.

6.1 Challenges

Regarding evaluation using reference alignments, challenges exist at each stage of the evaluation process:

Anchor Selection: Given the less bounded nature of complex matching, it is to be expected that systems will produce a large number of correspondences.

Challenge 1: Selecting which candidates will be compared to which reference correspondences in order to avoid the necessity of a full pairwise comparison of all candidates in the comparison step

Correspondence Comparison: Complex correspondences can consist of entities of arbitrary complexity and be expressed in a multitude of semantically equivalent or nearly-equivalent ways.

Challenge 2: Determining the relation between a candidate correspondence and a reference correspondence, which requires comparing all of the singular entities and the expressions in which they are listed for both correspondences

Challenge 3: Handling correspondence decomposition, which involves comparing sets of correspondences to a single correspondence, since the combination of several correspondences (simple or complex) can be equivalent or related to a single complex correspondence

Challenge 4: Comparing correspondences whose relation differs (e.g., a subsumption to an equivalence)

Scoring: Complex correspondences contain more axes than simple correspondences because e and e' are not single entities but rather (potentially nested) combinations of entities, constructors, and transformation functions. This necessitates more nuanced scoring metrics, which can be used to determine how close a correspondence is to a reference correspondence. This allows for measuring the effort required of a human validator or by matching approaches creators to understand the limitations of their approaches and thus drive development.

Challenge 5: Accurately reflecting the quality of a correspondence, especially considering that in complex matching, a correspondence is still useful even if only partially correct.

Aggregation: Existing aggregation approaches for alignment evaluation with a reference alignment were designed with simple alignments in mind. They are tightly coupled to particular correspondence comparison and scoring methods and tend to take an all-or-nothing, or at best all, half, or nothing, approach.

Challenge 6: Factoring correspondences that are partially correct into the scoring process

Challenge 7: Considering a set of candidate correspondences in conjunction as related to a single reference correspondence (and vice-versa)

Challenge 8: Handling the occurrence of multiple correct candidate correspondences that are implied by a single reference correspondence (as is the case in correspondence 3 from the reference alignment and 3 and 4 from the first alignment system)

Evaluation when no reference alignment is available presents an orthogonal set of challenges. Task-based evaluations require a well-defined task, for which a quality metric is definable. The quality of the alignment is measured by proxy through the quality of the task results, which results in a narrow scope for the evaluation. Furthermore, for tasks where the output needs to be manually evaluated (e.g., query rewriting), the manual effort required presents an additional challenge.

Challenge 9: Developing generalizable quality metrics for task-based complex alignment evaluation

Challenge 10: Automating the query rewriting process based on a set of complex correspondences

In addition to being able to handle the aforementioned challenges, evaluation metrics for complex alignments should also be fully automated and independent of manual input, even if the alignment is intended to be manually validated post hoc. This is a crucial feature to further promote the development of complex matching approaches, by shortening the time between development cycles. Consequently,

techniques for the evaluation of complex alignments need to be able to handle the computational complexity the challenges pose, both at the correspondence level and at the alignment level.

6.2 Gap analysis

We now turn our attention to assessing the degree to which existing alignment evaluation approaches address the challenges above. This analysis begins with approaches focused on cases in which a reference alignment is available (i.e., those relevant to challenges 1 through 8).

Syntactic approaches are unsuited to address challenges 3 and 4, since they do not employ reasoning and consider correspondences that are logically equivalent or can be derived as incorrect. They are also unable to address challenges 5, 6, and 7, given that they do not consider closely related correspondences. By virtue of their simplicity, they struggle less with challenges 1, 2, and 8, which are related to the computational complexity of the approach.

Rule-based semantic approaches provide strategies that can partially address challenges 5 and 6, since they are able to account for closely related correspondences involving direct super/subclasses. However, they are unable to handle the full gamut of expressions required by complex matching. Furthermore, they do not address the remaining challenges. Edit-distance metrics, which assess the number of modifications that must be made to a candidate correspondence in order to arrive at reference correspondence, can be considered a type of rule-based semantic approach in the context of complex alignment evaluation. Examples of edit-distance metrics for strings include Levenstein and Smith-Waterman. These metrics are potentially able to handle challenges 4, 5, and 6, while not specifically addressing the remaining challenges. However, we are not aware of any existing edit-distance metrics for any of the common complex alignment representation languages discussed in Section 3.

Reasoning-based semantic approaches are better suited to answer challenges 2, 3, and 7, since they can cover the semantic complexity of complex expressions, and also handle correspondence decomposition. However, this is restricted to the cases where the merged ontologies are in OWL-DL. Furthermore, they are unable to handle challenges 5 and 6, since they only cover correspondences that can be logically derived, they consider closely related correspondences as incorrect. They also do not offer any specific features to address challenges 1 and 8.

Instance-based approaches circumvent many of the outlined challenges, by simplifying correspondence evaluation to a measure of the overlap between sets of instances. However, they are applicable to class correspondences and transformations, but it is not straightforward to apply them to property correspondences. Furthermore, they require that all classes in the alignment be populated with instances.

With respect to evaluation approaches that do not require a reference alignment, the existing work primarily consists of manually intensive evaluation strategies that were uniquely developed for particular cases. There is significant room for future work on the challenges relevant to these metrics (i.e., challenges 9 and 10).

6.3 Feasibility analysis

The task of evaluating complex correspondences is inherently expensive computationally, due to the syntactical and semantic complexity of these correspondences.

From a syntactic perspective, there is no theoretical limit to the complexity of the expressions that can be constructed through nesting. This is not a challenge for the use of the traditional *syntactic* evaluation metric, which can still be implemented with $\mathcal{O}(n)$ time complexity. It is, however, a substantial challenge for the implementation of more sophisticated and promising approaches, such as *rule-based* and particularly edit-distance metrics, which have to cope with a potentially endless search space of possible combinations of constructions and transformations. This means that, in all likelihood, such evaluation

approaches would have to adopt non-naive techniques to reduce the search space and contemplate only the more plausible combinations of constructions in order to ensure efficiency.

From a semantic perspective, the more expressive complex correspondences go beyond OWL DL, and thus may not be decidable, while transformations cannot be expressed in OWL at all. This means that *semantic* approaches relying on existing OWL reasoners would only be able to evaluate correspondences with constructions supported by those reasoners, which would limit their applicability.

By contrast, instance-based approaches are largely unaffected by the complexity of the correspondences and could be the most realistic way to address the complex alignment evaluation problem, by shifting from the comparison of correspondences into the comparison of sets of instances. One approach for this would be to determine, for each correspondence c_i in the evaluated alignment, the relation between the sets of instances I_s and I_t, belonging to the source and target members of the correspondence, respectively. Each correspondence could then be classified as *equivalent*, *subsumed*, *overlapping*, or *disjoint*, given the relation between I_s and I_t, or *empty* if $I_s = I_t = \emptyset$ (i.e., if both members are either unsatisfiable or non-populated entities). Having a reference alignment, one would know what are the sets of expected instances to be compared. Different precision scores could then be computed for each type of correspondence member relation: the *equivalent* precision would measure the percentage of correspondences whose members are exactly populated with the same instances, and likewise, the *subsumed*, *overlapping* and *not disjoint* precision would measure the percentage of correspondences whose members subsume one another, overlap, or are not *disjoint*, respectively.

Such a strategy could rely on expressing complex correspondences as SPARQL queries, which would cover also transformation functions. As we discussed previously, it is limited in coverage, since it can be applied to the evaluation of class (expression) correspondences or transformations, but it is not straightforward to apply to the evaluation of property (expression) correspondences. Furthermore, it requires the knowledge bases to be consistently populated (i.e., complete population of all entities the complex correspondences are supposed to cover). However, the cost of creating such a knowledge base (e.g., with artificially populated data) is smaller than the cost of creating reference alignments or applying evaluation strategies such as query rewriting.

7 Conclusions and future work

In this paper, we have defined complex ontology alignments and shown that the few systems that have attempted to generate such alignments have been evaluated using methods that are difficult to generalize and/or labor intensive. A survey of existing evaluation approaches, which were developed with simple alignments in mind, has shown that they are insufficient in several ways. In particular, the most common evaluation approach, based on exact syntactic match, lacks the nuance necessary to distinguish between completely unhelpful correspondence suggestions and those that are 'almost correct'. Other existing evaluation techniques are not scalable to the complex case, can only be used under certain conditions (e.g., when dually annotated instance data is present or when the alignment is expressible in OWL DL), or have other drawbacks. We have enumerated what we view as the most pressing gaps between current techniques and what are needed for complex alignment evaluation. In the remainder of this section, we propose future work that can potentially bridge these gaps.

Evaluating complex ontology alignments is too broad a challenge to tackle with a single approach, as there are multiple aspects to take into account, and different tasks will likely merit different evaluation paradigms. Considering that the two main applications of complex alignments are ontology/linked-data integration, and query answering/rewriting, it stands to reason to focus our efforts in developing evaluation approaches with these applications in mind.

With respect to ontology/linked-data integration, it is unlikely that the state of the art in ontology alignment ever reaches a point where human validation is unnecessary. This is true even for simple alignments, but particularly so for complex alignments, given the inherent difficulty in generating them automatically with reasonable precision or recall. Under this premise, we believe that the most adequate approach to evaluate complex alignments in the context of this application would be an edit-distance approach that reflected the effort involved in human validation, in the same spirit as the effort similarity approach we

reviewed in Section 4.2.2. Therefore, our future work will concentrate on developing an analogous edit-distance approach that encompasses all the requirements and nuances of complex alignment evaluation. Concretely, the approach must explicitly contemplate all complex expressions in use and define costs for inter-converting them and must adequately handle cases of correspondence decomposition, where a reference correspondence should be compared with two or more system correspondences that cover it partially (or vice versa). Greater granularity with respect to the edit-costs between semantically related classes would also be desirable. Last, but not least, the approach should be scalable and avoid the need to do all-vs.-all correspondence comparisons. Given these constraints, we believe that a deterministic rule-based edit-distance approach that covers all the key complex correspondences constructions explicitly, in a way that reflects the effort required to correct them, would be the best candidate.

With respect to query answering/rewriting, we believe that there are two major hurdles to be tackled: developing an automated converter for transforming any complex alignment into a query rewriting scheme and developing a query generating algorithm that can automatically generate queries adequate in coverage and scope to the complex alignment to evaluate. The primary focus of our future work will be the first hurdle, as only by overcoming it can we use query-based approaches to fully evaluate complex alignments automatically. Overcoming the second hurdle will be essential to enable the widespread use of query-based evaluation and will also contribute to make query-based evaluation efforts more comprehensive and comparable, as otherwise queries have to be manually defined for each test case.

We will also explore instance-based evaluation approaches, such as the one delineated in the previous section. This approach can complement or even replace the edit-distance approach in a linked-data integration scenario and can be a computationally efficient and labor-friendly alternative to query answering.

Acknowledgments

DF was funded by the ELIXIR-EXCELERATE project (INFRADEV-3-2015). CP was funded by the Fundação para a Ciência e Tecnologia through the funding of the LaSIGE research unit (ref.UID/CEC/00408/2013) and project PTDC/EEI-ESS/4633/2014. OZ was supported by the CSF grant no. 18-23964S and by long-term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

Abacha, A. B. & Zweigenbaum, P. 2014. Means: une approche sémantique pour la recherche de réponses aux questions médicales. *TAL* **55**(1), 71–104.

Cheatham, M. & Hitzler, P. 2014. Conference v2.0: an uncertain version of the OAEI conference benchmark. In *International Semantic Web Conference*, 33–48. Springer.

David, J., Euzenat, J., Genevès, P. & Layada, N. 2018. Evaluation of query transformations without data: short paper. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23–27, 2018*, 1599–1602.

David, J., Euzenat, J., Scharffe, F. & Trojahn dos Santos, C. 2011. The alignment 4.0. *Semantic Web* **2**(1), 3–10.

Duan, S., Fokoue, A., Hassanzadeh, O., Kementsietsidis, A., Srinivas, K. & Ward, M. J. 2012. Instance-based matching of large ontologies using locality-sensitive hashing. In *International Semantic Web Conference*, 49–64. Springer.

Ehrig, M. & Euzenat, J. 2005. Relaxed precision and recall for ontology matching. In *Integrating Ontologies'05, Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies, Banff, Canada, October 2, 2005.*

Euzenat, J. 2004. An API for ontology alignment. In *International Semantic Web Conference*, 698–712. Springer.

Euzenat, J. 2007. Semantic precision and recall for ontology alignment evaluation. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6–12, 2007,* 348–353.

Euzenat, J., Polleres, A. & Scharffe, F. 2008. Processing ontology alignments with SPARQL. In *2008 International Conference on Complex, Intelligent and Software Intensive Systems*, 913–917.

Euzenat, J., Scharffe, F. & Zimmermann, A. 2007. Expressive alignment language and implementation. https://hal.inria.fr/file/index/docid/822892/filename/kweb-2210.pdf

Euzenat, J. & Shvaiko, P. 2013. *Ontology Matching*. Springer.

Hollink, L., Van Assem, M., Wang, S., Isaac, A. & Schreiber, G. 2008. Two variations on ontology alignment evaluation: methodological issues. In *5th European Semantic Web Conference*, 388–401.

Isaac, A., Matthezing, H., van der Meij, L., Schlobach, S., Wang, S. & Zinn, C. 2008. Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In *5th European Semantic Web Conference*, 402–417.

Isaac, A., Van Der Meij, L., Schlobach, S. & Wang, S. 2007. An empirical study of instance-based ontology matching. In *The Semantic Web*, 253–266. Springer.

Jiang, S., Lowd, D., Kafle, S. & Dou, D. 2016. Ontology matching with knowledge rules. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVIII*, 75–95. Springer.

Jiménez-Ruiz, E., Grau, B. C., Horrocks, I. & Berlanga, R. 2011. Logic-based assessment of the compatibility of UMLS ontology sources. *Journal of Biomedical Semantics* **2**(1), S2.

Kirsten, T., Thor, A. & Rahm, E. 2007. Instance-based matching of large life science ontologies. In *International Conference on Data Integration in the Life Sciences*, 172–187. Springer.

Maedche, A., Motik, B., Silva, N. & Volz, R. 2002. Mafra—a mapping framework for distributed ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, 235–250. Springer.

Makris, K., Bikakis, N., Gioldasis, N. & Christodoulakis, S. 2012. SPARQL-RW: transparent query access over mapped RDF data sources. In *15th International Conference on Extending Database Technology*, 610–613. ACM.

Meilicke, C. & Stuckenschmidt, H. 2008. Incoherence as a basis for measuring the quality of ontology mappings. In *3rd International Conference on Ontology Matching*, **431**, 1–12.

Nunes, B. P., Mera, A., Casanova, M. A., Breitman, K. K. & Leme, L. A. P. 2011. Complex matching of RDF datatype properties. In *Proceedings of the 6th International Conference on Ontology Matching*, **814**, 254–255. CEUR-WS.org.

Parundekar, R., Knoblock, C. A. & Ambite, J. L. 2010. Linking and building ontologies of linked data. In *ISWC*, 598–614. Springer.

Parundekar, R., Knoblock, C. A. & Ambite, J. L. 2012. Discovering concept coverings in ontologies of linked data sources. In *ISWC*, 427–443. Springer.

Pesquita, C., Faria, D., Santos, E. & Couto, F. M. 2013. To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *OM*, 13–24.

Qi, G. & Hunter, A. 2007. Measuring incoherence in description logic-based ontologies. In *The Semantic Web*, 381–394. Springer.

Qin, H., Dou, D. & LePendu, P. 2007. Discovering executable semantic mappings between ontologies. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, 832–849. Springer.

Ritze, D., Meilicke, C., Šváb Zamazal, O. & Stuckenschmidt, H. 2009. A pattern-based ontology matching approach for detecting complex correspondences. In *4th ISWC Workshop on Ontology Matching*, 25–36.

Ritze, D., Völker, J., Meilicke, C. & Šváb Zamazal, O. 2010. Linguistic analysis for complex ontology matching. In *5th Workshop on Ontology Matching*, 1–12.

Sagi, T. & Gal, A. 2018. Non-binary evaluation measures for big data integration. *The VLDB Journal* **27**(1), 105–126.

Scharffe, F. 2009. *Correspondence Patterns Representation*. PhD thesis, Faculty of Mathematics, Computer Science and University of Innsbruck.

Schopman, B., Wang, S., Isaac, A. & Schlobach, S. 2012. Instance-based ontology matching by instance enrichment. *Journal on Data Semantics* **1**(4), 219–236.

Serpeloni, F., Moraes, R. & Bonacin, R. 2011. Ontology mapping validation. *International Journal of Web Portals* **3**(3), 1–11.

Solimando, A., Jimenez-Ruiz, E. & Guerrini, G. 2017. Minimizing conservativity violations in ontology alignments: algorithms and evaluation. *Knowledge and Information Systems* **51**(3), 775–819.

Solimando, A., Jimenez-Ruiz, E. & Pinkel, C. 2014. Evaluating ontology alignment systems in query answering tasks. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, **1272**, 301–304. CEUR-WS.org.

Šváb, O., Svátek, V., Berka, P., Rak, D. & Tomášek, P. 2005. Ontofarm: towards an experimental collection of parallel ontologies. In *Poster Track of ISWC*, *2005*.

Thiéblin, E., Amarger, F., Haemmerlé, O., Hernandez, N. & dos Santos, C. T. (2016). Rewriting SELECT SPARQL queries from 1:n complex correspondences. In *Proceedings of the 11th International Workshop on Ontology Matching Co-Located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016*, 49–60.

Thiéblin, E., Amarger, F., Hernandez, N., Roussey, C. & Trojahn, C. 2017. Cross-querying lod datasets using complex alignments: an application to agronomic taxa. In *Research Conference on Metadata and Semantics Research*, 25–37. Springer.

Thiéblin, E., Cheatham, M., Trojahn, C., Zamazal, O. & Zhou, L. 2018a. The first version of the OAEI complex alignment benchmark. In *ISWC Posters and Demos*. Springer.

Thiéblin, E., Haemmerlé, O., Hernandez, N. & Trojahn, C. 2018. Task-oriented complex ontology alignment: two alignment evaluation sets. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings,* 655–670.

Thiéblin, E., Haemmerlé, O. & Trojahn, C. 2018b. Complex matching based on competency questions for alignment: a first sketch. In *OM 2018 - 13th ISWC Workshop on Ontology Matching*.

Van Hage, W. R., Isaac, A. & Aleksovski, Z. 2007. Sample evaluation of ontology-matching systems. In *EON*, **2007**, 41–50.

Visser, P. R., Jones, D. M., Bench-Capon, T. J. & Shave, M. 1997. An analysis of ontology mismatches; heterogeneity versus interoperability. In *AAAI 1997 Spring Symposium on Ontological Engineering, Stanford CA, USA*, 164–72.

Walshe, B., Brennan, R. & O'Sullivan, D. 2016. Bayes-recce: a bayesian model for detecting restriction class correspondences in linked open data knowledge bases. *International Journal on Semantic Web and Information Systems (IJSWIS)* **12**(2), 25–52.

Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R. & Zakharyaschev, M. 2018. Ontology-based data access: a survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5511–5519. International Joint Conferences on Artificial Intelligence Organization.

Zamazal, O. & Svátek, V. 2017. The ten-year ontofarm and its fertilization within the onto-sphere. *Web Semantics: Science, Services and Agents on the World Wide Web* **43**, 46–53.

Zhou, L., Cheatham, M., Krisnadhi, A. & Hitzler, P. 2018. A complex alignment benchmark: geolink dataset. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II*, 273–288.