Automatic evaluation of complex alignments: an instance-based approach

Elodie Thiéblin, Ollivier Haemmerlé, Cássia Trojahn [*]
Institut de Recherche en Informatique de Toulouse, Toulouse, France
E-mails: elodie@thieblin.fr, ollivier.haemmerle@irit.fr, cassia.trojahn@irit.fr

Editor: Jens Lehmann, University of Bonn, Germany
Solicited reviews: Jérôme Euzenat, INRIA, France; Shu-Chuan Chu, Shandong University of Science and Technology, China; Pavel Shvaiko, Trentino Digitale SpA, , Italy; Seven anonymous reviewers

Abstract. Ontology matching is the task of generating a set of correspondences (i.e., an alignment) between the entities of different ontologies. While most efforts on alignment evaluation have been dedicated to the evaluation of simple alignments (i.e., those linking one single entity of a source ontology to one single entity of a target ontology), the emergence of matchers providing complex alignments (i.e., those composed of correspondences involving logical constructors or transformation functions) requires new strategies for addressing the problem of automatically evaluating complex alignments. This paper proposes i) a benchmark for complex alignment evaluation composed of an automatic evaluation system that relies on queries and instances, and ii) a dataset about conference organisation. This dataset is composed of populated ontologies and a set of competency questions for alignment as SPARQL queries. State-of-the-art alignments are evaluated and a discussion on the difficulties of the evaluation task is provided.

Keywords: ontology matching, complex alignment, evaluation, benchmark

1. Introduction

Ontology matching is the task of generating a set of correspondences (i.e., an alignment) between the entities of different ontologies. This is the basis for other tasks, such as data integration, ontology evolution, and query rewriting. While the field has fully developed in the last decades, most works are still dedicated to the generation of simple correspondences (i.e., those linking one single entity of a source ontology to one single entity of a target ontology). However, simple correspondences are insufficient for covering the different kinds of heterogeneities (lexical, semantic, conceptual) in the ontologies to be matched. More expressiveness is achieved by complex correspondences, which can better express the relationships between entities of different ontologies. For example, the piece of knowledge that a conference paper has been accepted can be represented as a class IRI *ekaw:Accepted_Paper* in a source ontology, or as a class expression representing the papers (the range of *cmt:hasDecision* is *cmt:Paper*) having a decision of type *cmt:Acceptance* in a target ontology. The correspondence $\langle$*ekaw:Accepted_Paper, $\exists$cmt:hasDecision.cmt:Acceptance, $\equiv$, 1*$\rangle$ expresses an equivalence between the two representations of "accepted paper", with a confidence value of 1.

Earlier works in the field have introduced the need for complex ontology alignments [1, 2], and different approaches for generating them have been proposed in the literature afterwards. These approaches rely on diverse methods, such as correspondence patterns [3–5], knowledge-rules [6], statistical methods [7–9], competency questions for alignment [10, 11], genetic programming [12] or still path-finding algorithms [13]. In others fields, such as relational databases, different approaches have been proposed so far [14, 15], however,

[*] Corresponding author. E-mail: cassia.trojahn@irit.fr.

covering less expressive knowledge representation languages and models. The reader can refer to [16] for a survey on complex matching. While works on complex ontology matching have been mostly dedicated to the development of approaches able to generate complex alignments, benchmarks[1] on which the approaches can be systematically evaluated are still lacking. On the one hand, most existing matching proposals have been manually evaluated [3], usually in terms of precision, or on approach-tailored datasets [9] on which recall is calculated. On the other hand, most efforts on systematic evaluation are still dedicated to matching approaches dealing with simple alignments. Although a large spectrum of matching cases has been proposed so far in the Ontology Alignment Evaluation Initiative campaigns (OAEI)[2] e.g. involving synthetically generated or real world datasets with large and domain-specific ontologies, these cases are mostly limited to simple alignments. Recently, the first OAEI complex track was proposed [17] opening new perspectives for the automatic evaluation in the field.

In this paper, a benchmark for evaluating complex alignments is proposed. This benchmark is composed of a dataset involving ontologies, populated with controlled and shared instances, reference competency question queries, and an automatic evaluation system. “Controlled” or “regularly” populated instances mean that every entity (class or property) concerned by the alignment should have at least one instance in both ontologies. While classical benchmarks in the field [18, 19] rely on reference alignments and measurements of compliance between the generated and reference alignments (usually using classical precision and recall as evaluation metrics), here we propose a set of competency questions for alignment (CQA) as reference. A competency question expresses, through a SPARQL query, the knowledge an alignment should cover between the source and target ontologies [20]. In particular, we propose two evaluation measures. While the *CQA coverage* measure relies on pairs of equivalent SPARQL queries (source and target queries) and measures how well an evaluated alignment covers these queries, the *intrinsic precision* compares the instances of the correspondences members. Intrinsic precision balances the CQA coverage like precision balances recall in information retrieval.

The contribution of this paper is manifold:

- we discuss the challenges of automatic evaluation of complex alignments with respect to classical evaluation in the literature;
- we propose an automatic approach for evaluating complex alignments, which is based on competency questions for alignment in the form of SPARQL queries as references, and comparison of instances;
- we propose a dataset with controlled instance population and competency questions for alignment on which the alignments are evaluated;
- we evaluate state-of-the-art complex alignments on the proposed dataset and discuss their main strengths and weaknesses.

The automatic evaluation system and the populated datasets (and the scripts to generate them) are published under LGPL license[3].

The rest of this paper is organised as follows. The background on complex ontology matching and competency question for alignment are introduced in Section 2. Related works are discussed in Section 3. Then, the proposed evaluation system is presented in Section 4. Next, the methodology followed to create the dataset and the dataset itself are detailed in Section 5. Evaluation of existing complex alignments over the benchmark is discussed in Section 6. Finally, conclusions and future work are presented in Section 7.

2. Background

Before introducing the notions of complex alignment and competency questions, the ontologies and their instances that will be used in the rest of this paper are introduced. The ontologies *cmt* and *ekaw* come from the Conference dataset [19]. Their fragments are depicted in Figures 1 and 2 using the format proposed in [21].

2.1. Complex ontology alignment

Ontology matching (as in [22]) is defined as the process of generating an alignment *A* between two ontologies: a source ontology *o* and a target ontology

[1] Following the definition of “benchmark” as a standard by which something can be measured or judged (from the American Heritage® Dictionary of the English Language, Fifth Edition. S.v. “benchmark.” Retrieved January 7 2019 from https://www.thefreedictionary.com/benchmark), an alignment **benchmark** is considered composed of a dataset and an evaluation system.

[2] http://oaei.ontologymatching.org/

[3] https://framagit.org/IRIT_UT2J/conference-dataset-population

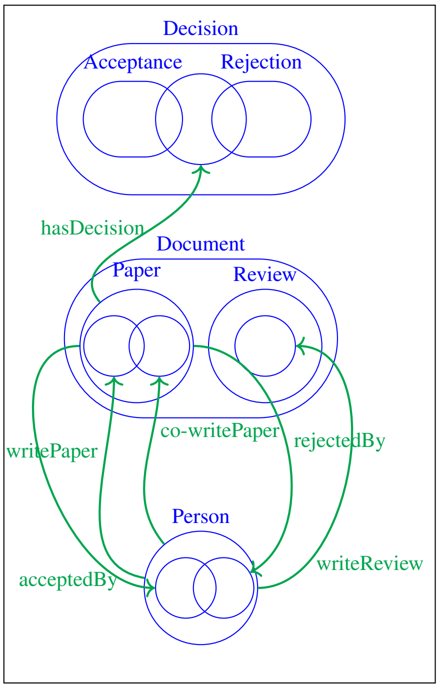

Fig. 1. Fragment of the *cmt* ontology used in the running examples.

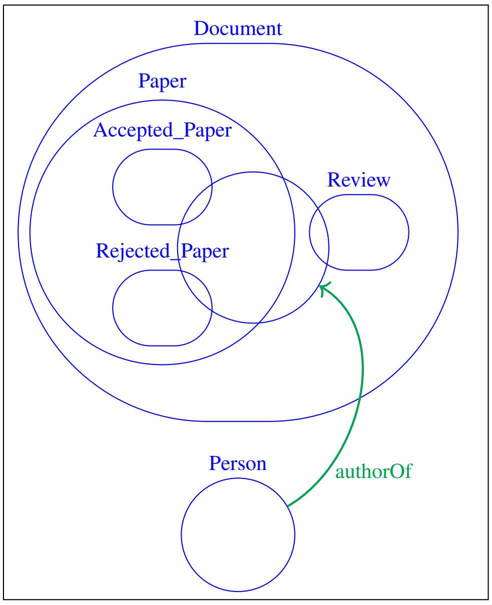

Fig. 2. Fragment of the *ekaw* ontology used in the running examples.

o'. A is directional, denoted $A_{o\to o'}$. $A_{o\to o'}$ is a set of correspondences $\langle e, e', r, n\rangle$. Each correspondence expresses a relation r (e.g., equivalence ($\equiv$), subsumption ($\sqsupseteq$, $\sqsubseteq$)) between two members e and e', and n expresses the level of confidence $[0..1]$ in this correspondence. A member can be a single ontology entity (class, object property, data property, individual) of respectively o and o' or a more complex construction which is composed of some entities using constructors or transformation functions (as in the examples in the following). From that, two types of correspondences are considered depending on the type of their members [23]:

- 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, respectively: $\langle$*ekaw:Accepted_Paper,* $\exists$ *cmt:hasDecision.cmt:Acceptance,* $\equiv$, *1*$\rangle$ and $\langle$*concatenation(edas:hasFirstName, " ", edas:hasLastName), cmt:name,* $\rightarrow$, *1*$\rangle$

A simple correspondence is usually noted (s:s), and a complex correspondence can be (s:c) if its source member is a single entity, (c:s) if its target member is a single entity or (c:c) if both members are complex entities. An approach which generates a complex alignment will be referred as a "complex matching approach", a "complex matching system" or a "complex matcher" in the rest of this paper.

2.2. *Competency questions for alignment (CQAs)*

In ontology authoring, in order to formalise the knowledge needs of an ontology, competency questions (CQs) have been introduced as *ontology's requirements in the form of questions the ontology must be able to answer* [24]. As defined in [10, 20], a competency question for alignment (CQA) is a competency question which should be covered by two or more ontologies, i.e., it expresses the knowledge that an alignment should cover (if both ontologies' scopes can answer the CQA). The first difference between CQA and CQ is that the scope of the CQA is limited by the intersection of its source and target ontologies' scopes. The second difference is that this maximal and ideal alignment's scope is not known *a priori* (as it is the purpose of the alignment). As the ontology authoring competency questions (CQs) [25], a CQA can be expressed in natural language or as SPARQL SELECT queries.

Inspired from the predicate arity in [25], the notion of **question arity**, which represents the arity of the expected answers to a CQA was introduced in [10]:

- A *unary* question expects a set of instances or values, e.g., "Which are the accepted papers?" *(paper1), (paper2).*
- A *binary* question expects a set of instances or value pairs, e.g., "What is the decision on a paper?" *(paper1, accept), (paper2, reject).*
- A *n-ary* question expects a tuple of size *n*, e.g., "What is the decision associated with the review of a given paper?" *(paper1, review1, weak accept), (paper1, review2, reject).*

3. Related work

Evaluation of matching systems is carried out over an **evaluation dataset**, usually composed of a set of ontologies, a reference alignment, and potentially different inputs (e.g., queries, instances, partial alignment). The generated alignment is then evaluated by an **evaluation system** which gives a score to the alignment produced by the system. Different evaluation dimensions can be considered in the process (that applies for both simple and complex evaluation):

Resource-oriented This dimension refers to the evaluation of the system performance in terms of runtime and memory usage. It is often performed over ontologies of different sizes and levels of expressiveness. Most OAEI tracks adopt this kind of evaluation.

Controlled input Evaluation of the generated alignment given different (and controlled) inputs. Such an evaluation was proposed for the GeoLink and Hydrography datasets of the OAEI Complex track [17]. Given a list of entities, the system should be able to find the correct (complex) construction involving these entities.

Output-oriented Evaluation of the output alignment itself over a dataset. This evaluation can be intrinsic or extrinsic. With the former, the quality of an alignment can be measured based on its intrinsic characteristics, as in [26] who evaluates the quality of an alignment over its logical coherence or in [27] where a good alignment should not violate the conservativity principle. With the latter, the evaluation is usually based on the compliance of the generated alignment with respect to a reference one (i.e., applying precision and recall metrics).

Task-oriented The quality of an alignment can also be assessed regarding its suitability for a specific task or application [28, 29]. While current evaluation settings have not been set-up for evaluating matchers specifically designed for a given application or with a given task in mind, alignments generated by general purpose matchers are rather evaluated with respect to its suitability to a given task.

In the following, the main related works considering these evaluation dimensions are discussed.

3.1. Complex alignment evaluation metrics

Most works on alignment evaluation address the evaluation of simple alignments using a reference alignment or a sample of it. This is what has been done in the context of the OAEI campaigns. With respect to the evaluation of complex alignments, they have been evaluated manually, usually in terms of precision [3, 4, 8, 9], or on specific datasets in order to compute recall. In particular, the approach adopted in [8, 9] estimated their recall based on a recurring pattern (*Class by attribute-value*) between DBpedia and Geonames. They estimated the number of occurrences of this pattern between these ontologies and calculated the recall based on this estimation. In [13] a set of reference correspondences between two ontologies was manually created, involving few reference correspondences from which only two could not be expressed with simple correspondences. In [9] the authors proposed an algorithm to create an evaluation dataset that is composed of a synthetic ontology containing 50 classes with *Class-by-attribute-value* correspondences with DBpedia and 50 classes with no known correspondences with DBpedia. Both ontologies are populated with the same instances. In [30], inspired from [15], the approach for discovering complex attribute correspondences (i.e., {First Name, Last Name} = {Author}) between web interfaces is evaluated using *target accuracy* (that includes target precision and target recall) as metric. It evaluates how similar the generated alignment is with respect to a set of manually collected ones, using the notion of synonym attribute sets.

As discussed in [10] (inspired from [31]), alternative metrics of *accuracy* and *top−x accuracy* have been also applied in evaluation settings in which the number of correspondences is predefined, e.g., there is one correspondence for each entity of the target schema/ontology. The accuracy is calculated as 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, e.g., 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 generated by the system. Alternatively, the approach in [32], 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.

3.2. Complex alignment benchmarks

As discussed above, complex matchers are usually evaluated on custom evaluation alignment sets, usually covering the specificities of the approach to be evaluated. Recently, the first complex benchmark has been introduced in the OAEI campaigns [17]. The track consists of four datasets from different domains and considering different evaluation strategies:

Complex conference a consensual complex alignment was created using the query rewriting methodology from [23]. Each generated correspondence is manually classified as true positive or false positive, with respect to a reference alignment. The evaluated and reference correspondences are (s:c). In 2019, the benchmark presented in this paper has been used to automatically evaluate complex alignments.

Hydrography and GeoLink a set of ontologies on the hydrography domain and a pair of ontologies from GeoScience (more details bout the GeoLink dataset are provided in [33]). The matchers are evaluated following three subtasks: i) finding all entities which appear in a given correspondence, ii) finding the right construction involving those entities, and iii) finding the complex correspondences from scratch. Only the first subtask was implemented in the OAEI 2018 campaign [34], and the evaluation was automatically carried out using classical precision and recall (all alignments were simple equivalences). In 2019, a close metric to relaxed precision and recall [35] has been applied to entity identification and relationship identification tasks.

Taxon a set of CQAs over agronomic knowledge bases is rewritten with the evaluated alignments. Each rewritten query is manually classified as semantically equivalent to the source query or not. A "Query Well Rewritten" metric measures the percentage of CQA which had a semantically equivalent query after the rewriting process. Each correspondence of the evaluated alignment is also manually classified as true positive or false positive without a reference.

In 2018, only two systems, AMLC [5] and CANARD [36], were able to generate complex correspondences for those datasets. In 2019, a new system has been proposed, AROA[4].

3.3. Task-oriented benchmarks

Regarding task-oriented evaluation, [22] 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 runtime. One such task-oriented evaluation approach was introduced in the OAEI in 2015 at the *OA4QA* track[5] [37], 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 $\mathcal{A}boxes$. A given query, such as *Q(x):=Author(x)* expressed using the vocabulary of the *cmt* ontology, is executed over the merged ontology *cmt* ∪ *ekaw* ∪ *A*, where *A* is an alignment between *cmt* and *ekaw*. The evaluation metrics were precision and recall on the result sets of the query evaluation. A reference or model answer set for the query results was computed using the reference alignment (RA1) of the Conference track. The answer set of *Q(x)* executed over *O1* ∪ *O2* ∪ *A* was compared with respect to the result sets of running the same query *Q(x)* over *O1* ∪ *O2* ∪ *RA1*. An alternative approach for evaluating query answering without using instances was proposed by [38], where queries are compared without instance data, by grounding the evaluation on query containment.

[4] http://oaei.ontologymatching.org/2019/results/complex/index.html

[5] http://www.cs.ox.ac.uk/isg/projects/Optique/oaei/oa4qa/index.html

In [39], an "end-to-end" evaluation in which a set of queries are rewritten using an evaluated alignment is proposed. The results of the queries are manually classified by relevance for a user on a 6-point scale. This evaluation was performed with two rewriting systems. If a source member *e* does not appear in any correspondence of the alignment, the *upwards* rewriting system will use super-classes of *e* which appear as source member in the alignment's correspondences and the *downwards* system will use subclasses of *e*. Three alignments were evaluated. For each alignment, 20 concepts were randomly selected to be queried and evaluated.

While the task-based evaluation is relevant for both simple and complex alignments, some tasks tend to have higher expressiveness requirements, such as query rewriting and ontology merging, as discussed in [23]. Complex alignments for query rewriting have been the focus of the work of [40][6], applied to 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 [32] cloud. This dataset is the one used in the OAEI *Complex* track, on the ability to rewrite SPARQL queries using these alignments. The queries written for the source ontology were rewritten automatically using (s:s) or (s:c) correspondences and the system described by [41], and manually for (c:c) correspondences.

In fact, the query rewriting task can be seen as one of the main applications for complex alignments, and evaluation approaches based on this task are highly relevant. 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 [42]. For complex alignments, such a naive approach is not enough, as the semantics of the alignment itself has to be taken under consideration. [43] proposed an approach for writing specific SPARQL CONSTRUCT queries, but most query rewriting systems still rely on simple or (s:c) complex correspondence and fail in covering highly expressive (c:c) correspondences.

3.4. Positioning with respect to existing benchmarks

With respect to the evaluation of complex alignments, several works focus on manually evaluating alignments, in terms of precision as in [3, 4], calculating recall on recurring patterns as in [8, 9], or relying on a sample of reference correspondences [13]. While most of these approaches focus on the comparison of correspondences, we shift the problem to the comparison of instances. We propose an evaluation benchmark that considers queries as references and relies on metrics based on query coverage (as for recall) and intrinsic precision (as for precision without a reference alignment). Our approach requires, however, datasets populated in a controlled manner, differently from the datasets in [33].

As [37], we have queries as references instead of reference alignments. Close to ours, the evaluation in [37] relies on a synthetically populated version of the *Conference* dataset. However, their queries are executed over a merged ontology and alignments are limited to simple correspondences. Here, the queries are executed over different populated ontologies. As [39], here a set of queries are rewritten using an evaluated alignment. However, their evaluation process relies on manually classifying the query results.

Table 1 summarizes the existing alignment evaluation benchmarks that are close to our proposal (**CQA** benchmark, marked in bold in Table 1). Automation for (c:c) correspondences is still an open issue in the field. The proposal here is to automatise the evaluation process by shifting the problem to the comparison of instances, as detailed in the following sections.

4. Automatic evaluation of complex alignments

As discussed above, evaluation of simple alignments have been largely exploited in the literature and in particular in OAEI campaigns. Automatic evaluation of complex alignments being addressed to a lesser extent [17]. In terms of evaluation metrics, most of the solutions so far are based on the comparison of alignments using syntactic or semantic approaches leaving under-exploited the comparison at instance-level. This is the proposal of this paper.

With respect to a **syntactic** comparison of alignments, it can measure how much effort should be done to transform an evaluated correspondence into the reference one. However correspondences which use different constructors, or different levels of factorisation can express the same meaning. A syntactic comparison also depends on the language in which the correspondences are expressed. Such a comparison strongly depends on the way the reference correspondences,

[6] http://www.music.tuc.gr/projects/sw/sparql-rw/

Table 1

Comparison of ontology alignment evaluation benchmarks. The *Type of corresp.* column represents the form of the most expressive correspondences dealt with by the benchmarks – (c:c) is more complex than (s:c), which is more complex than (s:s).

Benchmark	*Type of evaluation*	*Type of reference*	*Type of corresp.*
OA4QA [37]	Automatic (precision/recall)	Query	(s:s)
Query rewrite [39]	Manual	Query	(s:s)
Patterns evaluation [9]	Manual	Alignment	(s:c)
Patterns evaluation [8]	Manual	Alignment	(s:c)
Thieblin 2018 [23]	Manual	Alignment	(s:c)
GeoLink 2018 [33]	Automatic (precision/recall) /Manual	Alignment	(c:c)
Hydrography 2018 [33]	Automatic (precision/recall)/Manual	Alignment	(c:c)
GeoLink 2019	Automatic (relaxed precision/recall)	Alignment	(c:c)
Hydrography 2019	Automatic (relaxed precision/recall)	Alignment	(c:c)
Taxon [32]	Manual	Query	(c:c)
CQA benchmark	**Automatic instance-based (CQA coverage/intrinsic precision)**	**Query**	**(c:c)**

queries, *etc.* are expressed. For example, ⟨ *o:Author* , $\exists$*o′:authorOf*.$\top$, $\equiv$ ⟩ is semantically equivalent to the correspondence ⟨ *o:Author* , $\exists$*o′:writtenBy*$^-$.$\top$, $\equiv$ ⟩. However, these two correspondences use different URIs in their constructors and thus are syntactically different. The correspondences ⟨ *o:AcceptedPaper* , $\exists$*o′:acceptedBy*.$\top$, $\equiv$ ⟩ and ⟨ *o:AcceptedPaper* , $\geqslant 1$ *o′:acceptedBy*.$\top$, $\equiv$ ⟩ are equivalent but expressed using different constructors (respectively an existential restriction or a cardinality restriction over the *o′:acceptedBy* property). They are also syntactically different. A factorisation problem would consist in verifying that ⟨ *o:paperWrittenBy* , *dom(o′:Paper)* $\sqcap$ *o′:writes*$^-$, $\equiv$ ⟩ and ⟨ *o:paperWrittenBy* , *(o′:writes* $\sqcap$ *range(o′:Paper))*$^-$, $\equiv$ ⟩ are equivalent correspondences. The *inverse* constructor is factorised in the second correspondence. A syntactic comparison of queries is faced with the same problems: syntactically different SPARQL queries can share the same semantics.

A **semantic** comparison would be an alternative solution. Semantic precision and recall perform evaluation against references without needing to decide what to compare to what. It is only necessary to evaluate if the reference entails each correspondence of the result (precision) or if the result entails each correspondence of the reference (recall). Concerning complex alignments, it has the advantage that any item that can be converted into OWL can be practically evaluated in this way. However, the expressiveness of the evaluated alignment with a semantic comparison is limited to $\mathcal{SROIQ}$ (the decidable fragment of OWL [44]). Correspondences with transformation functions could not be compared with such a comparison. Alternatively, the semantic query comparison proposed by [38] is based on query containment which can be based on inferences. However, it is also limited with regard to queries with transformation functions. While all semantic approaches (as in semantic precision and recall or in query containment) hold for any ontology population, the notion of entailment of constructs involving transformation functions need to be defined and implemented.

An **instance-based** comparison (of correspondences or query results) is an alternative comparison method to automatize. However, it has the drawback of requiring the knowledge bases to be regularly populated. Hence, for this kind of comparison, the desiderata for instance data is that the ontologies to be matched have ideally to be regularly and consistently populated with common instances, and a complex alignment dataset fulfilling such requirements does not exist. Here, a benchmark to evaluate complex alignments is proposed, including i) an evaluation system implementing instance-based comparison and using equivalent queries as references and ii) a dataset with controlled instances. Using equivalent SPARQL CQA as reference would ensure that the two compared objects are equivalent because they model the same piece of knowledge.

With respect to i), we propose two evaluation measures. While the *CQA coverage* measure relies on pairs of equivalent SPARQL queries (source and target queries) and measures how well an evaluated alignment covers these queries, the *intrinsic precision* compares the instances of the correspondences members. Intrinsic precision balances the CQA coverage like precision balances recall in information retrieval. With respect to ii) a methodology based on CQAs, as introduced in [10], is proposed to synthetically populate ontologies. This methodology was applied to five ontologies of the well-known Conference dataset [19].

In the following, before detailing the CQA coverage metric (Section 4.2), the overall evaluation workflow adopted in the approach is presented (Section 4.1). Then, the description of the intrinsic metric is presented (Section 4.3).

4.1. Overall workflow

Figure 3 presents the overall workflow adopted in the proposed approach. The steps followed in the evaluation process are:

① **Anchor selection** The anchor selection step consists of outputting a pair of comparable objects $\langle x_i, x_{rj} \rangle$. x_i is an object related to the evaluated alignment A_{eval} and x_{rj} is an object related to the reference $reference$. In the case the reference is equivalent queries, x_i can be a query derived from A_{eval} and x_{rj} a reference query.

② **Comparison** The purpose of the comparison step is to output a relation $rel(x_i, x_{rj})$ for each pair previously obtained $\langle x_i, x_{rj} \rangle$. The relation can be an equivalence (*i.e.*, $x_i \equiv x_{rj}$), a subsumption, an overlap, a disjoint, *etc.* (this list can be extended according to the type of comparison performed). A similarity value can be associated with the relation. The comparison here is instance-based.

$$rel(x_i, x_{rj}) = \begin{cases} rel(e_i, e_{rj}) \\ rel(e'_i, e'_{rj}) \\ rel(r_i, r_{rj}) \\ rel(n_i, n_{rj}) \end{cases} \tag{1}$$

③ **Scoring** The scoring step associates a score with each relation found in the previous step. Thus, the scoring functions are directly impacted by the relation $rel(x_i, x_{rj})$ found between the objects. Different scoring metrics have been proposed in the literature (classical score, used in the classical precision and recall metrics, or relaxed precision and recall measures were defined which replace the set intersection by a distance [35]:

$$classical\ score = \begin{cases} 1 & \text{if } x_i = x_{rj} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$relaxed\ prec\ score = \begin{cases} 1 & \text{if } x_i \leqslant x_{rj} \\ 0.5 & \text{if } x_i > x_{rj} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$relaxed\ rec\ score = \begin{cases} 1 & \text{if } x_i \geqslant x_{rj} \\ 0.5 & \text{if } x_i < x_{rj} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

④ **Aggregation** The scores are locally and globally aggregated to give the $final\ score$. The aggregations can be performed with different functions: best match, average, weighted average, *etc.* The local aggregation aggregates all scores for a given object. There can be different local aggregations. For example, there can be an aggregation over the evaluated object and one over the reference object. The global aggregation aggregates all the locally-aggregated scores. For example, if the local aggregation was performed over the reference object, all the reference objects were given a score. The reference object scores can be aggregated into a final score. A final score locally aggregated over the evaluated objects is often referred to as the *precision* score. A final score locally aggregated over the reference objects is often referred to as the *recall* score.

4.2. CQA coverage metric

With this evaluation strategy, the reference is a set of equivalent CQAs in the form of SPARQL queries. An evaluated alignment A_{eval} will be used to rewrite each source CQA. The rewritten queries will then be compared to the reference target CQA. The comparison of the queries is instance-based and a value is associated with each query relation based on the common part of the evaluated query and target CQA instances.

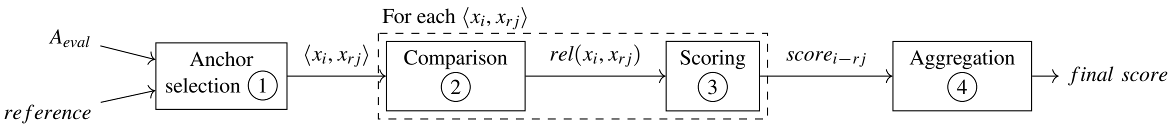

Fig. 3. Evaluation process of the alignment A_{eval} with a generic $reference$.

The scoring metric chosen is the one keeping the comparison relation value. A best-match aggregation is locally performed over the reference queries. The locally aggregated scores are then aggregated by an average. In the following, each step of the proposed evaluation process is described.

4.2.1. Source CQA anchoring

As stated above, the reference in this kind of evaluation is a set of equivalent CQAs as SPARQL queries. Each source CQA cqa_s has an equivalent target CQA cqa_t. In the anchoring step, each source cqa_s is rewritten using the generated alignment A_{eval}. The rewriting phase outputs all the possible rewritten target queries from the rewriting systems as the set Q_t. For each query q_t in Q_t, a pair (q_t, cqa_t) is formed.

Two rewriting systems have been considered. None of these systems consider the correspondence relation or correspondence value. The first system is the one from [41]. Each triple of cqa_s is rewritten using A_{eval}. When the predicate or object of the triple appears as the source member of a correspondence in A_{eval}, the target member of this correspondence is transformed into a SPARQL subgraph and put in the triple's place in the query. This system only deals with (s:c) correspondences. If a triple can be rewritten with different correspondences, all the possible combinations are added into Q_t. For example, consider the CQA:

```
SELECT ?s WHERE{
?s a ekaw:Accepted_Paper.}
```

which contains *ekaw:Accepted_Paper* which is the source member of the correspondences $c_{1k}, k \in [1..5]$.

The rewritten query using the c_{11} correspondence is:

```
SELECT ?s WHERE{
?s cmt:hasDecision ?o.
?o a cmt:Acceptance.}
```

This rewriting system cannot however work the other way around. For example, the CQA

```
SELECT ?s WHERE{
?s cmt:hasDecision ?o.
?o a cmt:Acceptance.}
```

cannot be rewritten with c_{11}.

The second system is based on instances and has been developed in the context of this paper. The instances I_s^{cqa} of cqa_s are retrieved from the source ontology. For each correspondence c of A_{eval}, the source member is transformed into a query and which retrieves the set of instances I_s over the source ontology. If $I_s \equiv I_s^{cqa}$, then, the target member of c is transformed into a query and added to Q_t. For example the CQA:

```
SELECT ?s WHERE{
?s a ekaw:Accepted_Paper.}
```

retrieves a set of accepted paper instances in the *ekaw* ontology. This set of instances is then compared to the set of instances described by the source member of each correspondence. In this case, *ekaw:Accepted_Paper* describes the same instances as the source member of all the $c_{1k}, k \in [1..5]$. Therefore, the target member of each correspondence can be transformed into a query. For c_{11}, the output query is

```
SELECT ?s WHERE{
?s cmt:hasDecision ?o.
?o a cmt:Acceptance.}
```

This rewriting system allows queries such as

```
SELECT ?s WHERE{
?s cmt:hasDecision ?o.
?o a cmt:Acceptance.}
```

to be rewritten too using the inverse of c_{11} for example (the inverse of a correspondence is its equivalent except that the source member becomes the target member and vice-versa).

Out of the existing rewriting systems dealing with complex correspondences, the one described in [41] deals with the most types of constructions. So far, the proposed instance-based rewriting system is one of the few systems able to deal with (c:c) correspondences. However, it is a feature of the system that (c:c) cannot be combined together.

4.2.2. Comparison

The instances I_t^{cqa} of cqa_t are retrieved over the target ontology. The instances I_t of q_t are retrieved over the target ontology. I_t and I_t^{cqa} are compared and the query precision (QP) and query recall (QR) are associated as value with the relation $rel(q_t, cqa_t)$ (subsumption, overlap, equivalence, etc.) between the two queries.

$$QP = \frac{|I_t \cap I_t^{cqa}|}{|I_t|} \qquad QR = \frac{|I_t \cap I_t^{cqa}|}{|I_t^{cqa}|}$$

$$rel(q_t, cqa_t) = \begin{cases} \equiv \text{ if } QR = 1 \text{ and } QP = 1 \\ \sqsubseteq \text{ if } QR \leqslant 1 \text{ and } QP = 1 \\ \sqsupseteq \text{ if } QR = 1 \text{ and } QP \leqslant 1 \\ overlap \text{ if } 0 < QR \leqslant 1 \text{ and} \\ 0 < QP \leqslant 1 \\ \perp \text{ if } QR = 0 \text{ and } QP = 0 \end{cases}$$

4.2.3. Scoring

The relation (associated with the query precision and query recall values) between cqa_t and q_t is transformed by an harmonic mean into a query F-measure score:

$$Fmeasure = 2 \times \frac{QR \times QP}{QR + QP}$$

The query F-measure (equally balancing precision and recall) was preferred over other metrics to be the scoring function as it is commonly used in alignment evaluation to aggregate the results of precision and recall. However, users may prefer one score to another, depending on alignment usage or manipulation. This was an implementation choice, as a matter of facilitating the comparison of the evaluated alignments.

4.2.4. Aggregation

As the rewriting phase outputs all the possible queries regardless of the correspondence relation, a lot of noise can be introduced. Moreover, the same query can be output by both rewriting systems. Therefore, for each cqa_t, the query q_t with the best query F-measure score is kept. The best-match aggregation prevents the final score to suffer from the noise introduced by the query rewriting systems. If a cqa_s could not be rewritten by the alignment, its query precision, query recall and query F-measure scores are 0.0. Here we make the decision of scoring query precision and recall to 0 if a CQA cannot be rewritten. However, if these have to be evaluated then precision could be set to 1 as long no mistake has been made.

The global aggregation method is the average function. The final output of the evaluation system is an average query precision, query recall and query F-measure score for the evaluated alignment.

4.3. Intrinsic precision

The CQA coverage evaluation locally aggregates the results over the CQA and not the rewritten queries because of the noise added by the rewriting systems. In return, an alignment with all the possible correspondences (correct and erroneous) between the source and target ontologies would obtain a good CQA coverage score. To counterbalance the CQA coverage score, we propose to measure the **intrinsic instance-based Precision** of an alignment.

For each correspondence c_i in the evaluated alignment, the instances I_s represented by the source member are compared to the instance I_t represented by the target member. Each correspondence is then classified as an *equivalent*, *subsumed*, *overlapping*, or *disjoint*, given the relation between I_s and I_t, or *empty* if $I_s = I_t = \emptyset$. Therefore, a correspondence can be *empty* if both its members are either unsatisfiable entities or non populated entities.

Different precision scores are given for each type of correspondence member relation: the *equivalent* precision measures the percentage of correspondences whose members are exactly populated with the same instances, the *subsumed* precision measures the percentage of correspondences whose members subsume one another, the same goes for *overlapping* and *not disjoint* which consider correct all correspondences except the *disjoint* ones.

5. CQA-based dataset

In this section, first the methodology followed to create the evaluation dataset (populated ontologies and associated CQAs) is presented (Section 5.1). Then, the OAEI Conference dataset (Section 5.2) is described, followed by the population of its ontologies from real-life data (Section 5.3). Finally, the set of evaluation CQAs extracted from the CQAs used for the dataset population is discussed (Section 5.4).

5.1. Dataset creation methodology

The purpose here is to create a dataset on which ontology matchers can be run and on which the evaluation described in the previous section can be performed. Therefore, the dataset must contain populated ontologies and a set of CQAs expressed as SPARQL queries over these ontologies. The population step is very important as the chosen instances may influence the result of the evaluation.

The proposed methodology has the following main steps:

1. Create a set of CQAs based on an application scenario. Only unary and binary CQAs were considered in this work.
2. Create a pivot format (i.e., the bridge format used for representing in a uniform way the data extracted from the data sources) which covers all the CQAs from step 1.
3. For each ontology of the dataset, create SPARQL INSERT queries corresponding to the pivot format.
4. Instantiate the pivot format with real-life or synthetic data.
5. Populate the ontologies with the instantiated pivot format using the SPARQL INSERT queries.
6. Run a reasoner to verify the consistency of the populated ontologies. If an inconsistency is detected, try to change the interpretation (i.e., add, suppress or modify axioms) of the ontology and iterate over steps 3 to 5.
7. Based on SPARQL INSERT queries, translate the CQAs covered by two or more ontologies as SPARQL queries.

In this methodology, the interpretation of the ontologies is the same for ontology population and CQA creation. The creation of CQAs can be done by interviewing users and domain experts, as recommended in the NeOn methodology [45] for competency question authoring. The CQAs can also derive from the competency questions which were used to design the ontologies of the dataset. In this implementation, however, one expert created the CQAs. This set has been discussed with a second expert who judged the set exhaustive enough for covering the conference organisation scenario.

In [23], (c:c) correspondences were not included in the dataset hence no exhaustive coverage could be guaranteed. However, as CQAs represent basic pieces of knowledge, they can be exhaustively covered by an alignment regardless of the shape of the correspondences. Using the same list of CQAs for ontology population and evaluation also insures the consistency of the answers of the evaluation CQAs.

5.2. Conference dataset

The dataset used here is the Conference dataset[7] proposed in [46]. It has been widely used [19], especially in the OAEI campaigns where it is a reference evaluation track. It is composed of 16 ontologies on the conference organisation domain and simple reference alignments between 7 of these ontologies. These ontologies were developed individually. The motivation for the extension of this dataset is that the ontologies are real ontologies (as opposed to synthetic ones), they are expressive and largely used for evaluation in the field. The query-oriented evaluation benchmark OA4QA was also based on this dataset [37]. Furthermore, reference complex alignments for query rewriting and ontology merging tasks have been proposed over five ontologies of this dataset [23].

In the first OAEI complex track, an evaluation was proposed over a consensual complex alignment between three ontologies (*cmt*, *conference*, *ekaw*) [17]. Here, the five ontologies covered by [23] have been populated: *cmt*, *conference* (Sofsem), *confOf* (confTool), *edas* and *ekaw* (Table 2).

Table 2

Number of entities by type of each ontology.

	cmt	conference	confOf	edas	ekaw
Classes	30	60	39	104	74
Obj. prop.	49	46	13	30	33
Data prop.	10	18	23	20	0

Even though this dataset has been largely used, it has only been partially populated. In the OA4QA track, only the classes covered by the 18 queries were populated and the creation of the synthetic $\mathcal{A}box$ has not been documented.

5.3. Populating the conference ontologies

In order to create the CQAs and re-interpret the Conference ontologies, the conference organisation scenario has been considered. First, the list of CQA has been established by examining a real-life use case: the Extended Semantic Web Conference 2018 edition. Second, the list of CQAs created from this use case has been extended by exploring the conference ontologies scope. The Extended Semantic Web Conference[8]

[7] http://oaei.ontologymatching.org/2018/conference/index.html
http://owl.vse.cz:8080/ontofarm/

[8] https://2018.eswc-conferences.org/

(ESWC) is open review and its website provided a good base to analyse which information is needed for conference organisation. In order to create the artificial instances of the pivot format, the ESWC 2018 use case as well as data from Scholarly Data [47] were considered.

5.3.1. Re-interpreting the ontologies with real-life data

As mentioned before, the first step of the process was to create a list of CQAs and re-interpret the ontologies under the perspective of a conference organisation application. By analysing the ESWC 2018 website, a first list of CQAs was created. The methodology was followed based on this first list of CQAs. The pivot format was instantiated with the website data.

While running the Hermit [48] reasoner in step 6 of the methodology, several exceptions were encountered. For most of them, the problem was with the interpretation of the ontology. For example, in the *cmt* ontology, *cmt:hasAuthor* is functional. Unlike primarily interpreted, this means that *cmt:hasAuthor* represents a “is first author of” relationship between a *cmt:Paper* and a *cmt:Author*. Then, the SPARQL INSERT queries have been modified in order to fit the new interpretation of the ontology.

Two exceptions have been detected, which could not be resolved by a change of interpretation. In that case, the original ontologies have been slightly modified:

- *cmt*: the relation *cmt:acceptPaper* between an *Administrator* and a *Paper* was defined as functional and inverse functional. This leads to an inconsistency when a conference administrator accepts more than one paper. *cmt:acceptPaper* has been changed to be only inverse functional.
- *conference*: *conference:Contribution_1st_author* was disjoint with *conference:Contribution_co-author*, which lead to an inconsistency when a person was at the same time the first author of a paper and the co-author of another paper. The disjunction axiom from the ontology has been then removed.

If a CQA was not exactly covered by an ontology, the ontology would not be populated with its associated instances. This results in an uneven population of equivalent concepts in the ontologies. For example, considering the *ekaw* and *cmt* ontologies, which both contain a *Document* class. *“What are the documents?”* was not a CQA whereas *paper, review, web site* and *proceedings* were the focus of CQAs. While *ekaw:Document* class has for subclasses *ekaw:Paper*, *ekaw:Review*, *ekaw:Web_Site* and *ekaw:Conference_Proceedings*, *cmt:Document* has only two subclasses *cmt:Paper* and *cmt:Review*. *ekaw:Document* will, by consequence of its subclasses, be populated with paper, review, website and proceedings instances whereas *cmt:Document* will be populated with paper and review instances only.

5.3.2. Conference data analysis

In order to populate the conference ontologies and make it close to real scenarios, some figures from past conferences have been analysed. The information from ISWC 2018 and ESWC 2017 from Scholarly Data[9] complemented the ESWC 2018 website data for this analysis. Indeed, some information such as which program committee member reviewed which paper does not appear in Scholarly Data and the ESWC 2018 website did not show which person is affiliated to which organisation. Some points could be observed:

- percentage of accepted papers having at least a program committee member as author: 44% for ESWC 2017 and 59% for ISWC 2018
- distribution of the number of authors per submitted papers (ESWC 2018): 1 (6%), 2 (17%), 3 (29%), 4 (26%), 5 (9%), 6 (8%) ou 7-10 (2%)
- distribution of the number of collaborating institutions per accepted papers over scholarly data (global represents the statistics over all data from the scholarly data endpoint):

nb inst.	global	ESWC 2017	ISWC 2018
1	56%	40%	40%
2	18%	16 %	30%
3	10 %	10 %	17%
4	6%	7 %	7%
5	5%	6%	5%
6+	between 0 and 2 %		

- distribution of the number of authors per accepted papers over scholarly data:

nb auth.	global	ESWC 2017	ISWC 2018
1	12%	7%	13 %
2	21%	11%	14%
3	27%	28%	24%
4	19%	25%	23%
5	17%	17%	14%
6	5%	5%	6%
7+	between 0 and 4 %		

[9] http://www.scholarlydata.org/

5.3.3. Population of conference ontologies

The first population of the ontologies with the ESWC 2018 data left some important knowledge unrepresented. For example, the concepts of external reviewer, presenter of a paper, and person affiliation, which appeared important for a conference organisation were not available on the website. Always in the perspective of conference organisation, the conference ontologies were browsed to complete the list of CQAs with useful concepts. The pivot format and associated SPARQL INSERT queries were also extended to cover the new list of CQAs. Then, the next step was to artificially generate the pivot format instantiation. For that, a score between 1 and 10 is given to each conference. This score determines the number of submitted papers, program committee members, etc. as shown in Table 3.

Table 3

Number of submitted papers, pc members, etc. for a conference of size 1 and 10 (min – max values).

Number of	Size 1	Size 10
submitted papers	40 – 45	940 – 990
people	300 – 330	1830 – 2130
pc members	50 – 52	500 – 530
oc members	20 – 22	110 – 140
sc members	15 – 17	60 – 90
institutions	30 – 32	210 – 240
tutorials	1 – 2	10 – 11
workshops	1 – 2	19 – 20
tracks	1	6

The statistics from the ESWC 2018, ISWC 2018, ESWC 2017 datasets were globally reproduced: 50% of papers have at least a program committee member as author, the number of authors per paper is 1 (6%), 2 (17%), 3 (29%), 4 (26%), 5 (9%), 6 (8%) or 7-10 (2%), the number of collaborating institutions is around 1 (40%), 2(30%), 3 (17%), 4 (7%), 5 (5%) 6(2%). These statistics are pointers, as the generation process is pseudo-random, these figures may vary in practice. Some proportions were arbitrarily chosen: 20% of the submitted papers are poster papers, and 20% are demo papers, the regular paper acceptance rate is in [0.1 – 0.7] and a poster/demo paper acceptance rate is in [0.4 – 1.0], 20% of the reviews are done by an external reviewer.

In order to evaluate statistics-based matchers on the benchmark, different sets of population were considered for the ontologies. The idea is to provide the same conference ontologies but with partially overlapping set of instances (instances linked with *owl:sameAs*). To do so, 6 sets of instance population with a more or less important overlapping parts were created. Each ontology is populated with different conferences[10] (with absolutely no common instance between the conferences –no common person, no common paper, etc.). This ensures that there is a quantifiable common part and that the ontologies are consistent. As a result, 6 artificial datasets were created with 25 artificial conferences:

- 0 %: 5 different conferences per ontology
- 20 %: 1 common conference for all ontologies and 4 different conferences per ontology
- 40 %: 2 common and 3 different conferences
- 60 %: 3 common and 2 different conferences
- 80 %: 4 common and 1 different conference
- 100 %: 5 common conferences for all ontologies

Note that the percentage given in the name of the datasets is the percentage of common conference event instances per ontology. As the size of each conference is different, the percentage of common instances (papers, authors, etc.) will not be same. In Table 4, the minimum and maximum percentage of the common paper instances is given for each dataset.

Table 4

Percentage (min, max) of common submitted papers in the different datasets. The second line reads *"In the 20% dataset, the proportion of common paper instances is between 7 and 11 %"*. Which means that for one of the ontologies, the common part of paper instances represents 7% of all its paper instances. For another ontology, the common part of paper instances represents 11% of all its paper instances.

Dataset	Min	Max
0%	0%	0%
20 %	7%	11 %
40 %	29%	51%
60 %	40 %	57%
80 %	57%	84 %
100 %	100 %	100 %

Not all the ontology concepts were covered by the pivot CQAs. Table 5 shows the number of entities covered by the CQAs, *i.e.*, instantiated after the CQA-based population, in each ontology.

[10] A *conference* here refers to the data related to a conference event.

Table 5

Number of populated entities by ontology. Number of populated entities / number of entities in the original ontology.

	cmt	conference	confOf	edas	ekaw
Classes	26 / 30	51 / 60	29 / 39	43 / 104	57 / 74
Obj. prop.	43 / 49	37 / 46	10 / 13	17 / 30	26 / 33
Data prop.	7 / 10	13 / 18	10 / 23	11 / 20	0 / 0

5.4. CQA for evaluation creation

For the evaluation, the focus is on CQAs which can actually be covered by two or more ontologies. To write the CQAs which will be used in the dataset, the list of CQAs used for the population was trimmed:

- the CQAs which were only covered by one ontology
- some CQAs which were not considered relevant such as “What is the name of a reception?”, the answer being an *rdfs:label* “Reception” for all reception instances.

The remaining CQAs were then written as SPARQL SELECT queries by adapting the SPARQL INSERT queries. Table 6 shows the number of CQAs which were covered by the pivot format, by each ontology (in the SPARQL INSERT queries) and which were transformed into SPARQL SELECT queries for the evaluation dataset. 278 SPARQL SELECT queries result from this process.

Table 6

Number of initial (pivot) CQAs covered by each ontology and number of evaluation (eval) CQAs covered by each ontology.

	cmt	conference	confOf	edas	ekaw	total
pivot	46	90	67	60	84	152
eval	34	73	54	52	65	100

6. Evaluation

Existing alignments over the conference dataset were evaluated with the proposed evaluation system. The dataset used for the evaluation is the 100 % dataset so that instance-based precision can be measured.

6.1. Evaluated alignments

Existing alignments between the Conference ontologies in EDOAL format[11] [42] have been evaluated. The EDOAL format was necessary so that the alignments could be processed by the rewriting systems. Five alignments have been evaluated. The number of ontology pairs (out of 10 pairs) that these alignments cover are indicated in the following.

Query_rewriting the query rewriting oriented alignment set[12] from [23]. It has been manually generated and is composed of 431 correspondences with 191 complex correspondences from 17 different patterns (some patterns are composite) - 10 pairs of ontologies

Ontology_merging the ontology merging oriented alignment set[12] from [23]. It has been manually generated and is composed of 313 correspondences with 54 complex correspondences from 9 different patterns (some patterns are composite) - 10 pairs of ontologies.

ra1 the reference simple alignment[13] from the conference dataset [19]. This dataset is limited to simple alignments between 7 ontologies - 10 pairs of ontologies.

Ritze_2010 the output alignment[12] from [4] (automatically generated) - complex correspondences found on 4 pairs of ontologies. This alignment is the smallest one as only one correspondence has been found for each pair.

Faria_2018 the output alignment from [5] (automatically generated) - alignments between 3 pairs publicly available. It is composed of two types of complex equivalence correspondences: those with attribute occurrence restriction and those with attribute domain restriction. These are the alignments available in the context of the OAEI 2018 campaign[14].

The ra1 alignment had been used as input by the systems of Ritze_2010 and Faria_2018. Ra1 has been added to these two alignments for the CQA coverage evaluation. The precision evaluation was made only on the complex correspondences (the output of the original approaches).

6.2. CQA coverage

The CQA coverage evaluation was run over all datasets in order to measure the standard deviation of

[11] http://alignapi.gforge.inria.fr/edoal.html

[12] https://doi.org/10.6084/m9.figshare.4986368.v7

[13] http://oaei.ontologymatching.org/2018/conference/

[14] http://oaei.ontologymatching.org/2018/results/complex/conference/index.html

Table 7
Standard deviation and average of the query precision, query f-measure and query recall scores over the 6 datasets.

		Query_rewriting	Ontology_merging	ra1	Faria_2018	Ritze_2010
Standard deviation	Precision	1.45×10^{-3}	1.48×10^{-3}	6.75×10^{-4}	2.74×10^{-3}	1.64×10^{-3}
	F-measure	5.55×10^{-4}	7.95×10^{-4}	6.87×10^{-4}	2.65×10^{-3}	1.76×10^{-3}
	Recall	3.89×10^{-4}	1.17×10^{-3}	7.26×10^{-4}	2.63×10^{-3}	1.91×10^{-3}
Average	Precision	0.69	0.63	0.42	0.42	0.48
	F-measure	0.68	0.63	0.42	0.41	0.47
	Recall	0.70	0.65	0.42	0.41	0.47

Table 8
Average of CQA f-measure for each pair of ontologies for each alignment on the 100% dataset.

pair	Query_rewriting	Ontology_merging	ra1	Faria_2018	Ritze_2010
cmt-conference	0.70	0.57	0.31	0.45	
cmt-confOf	0.69	0.69	0.69		
cmt-edas	0.65	0.65	0.41		0.53
cmt-ekaw	0.65	0.64	0.25	0.42	0.34
conference-cmt	0.69	0.59	0.28	0.41	
conference-confOf	0.50	0.48	0.43		
conference-edas	0.66	0.52	0.48		0.48
conference-ekaw	0.48	0.45	0.33	0.36	
confOf-cmt	0.77	0.71	0.72		
confOf-conference	0.73	0.56	0.45		
confOf-edas	0.87	0.74	0.28		
confOf-ekaw	0.83	0.72	0.51		0.54
edas-cmt	0.73	0.67	0.43		0.54
edas-conference	0.63	0.52	0.50		0.50
edas-confOf	0.56	0.70	0.30		
edas-ekaw	0.92	0.83	0.50		
ekaw-cmt	0.66	0.65	0.27	0.46	0.36
ekaw-conference	0.51	0.46	0.34	0.38	
ekaw-confOf	0.74	0.74	0.45		0.52
ekaw-edas	0.77	0.77	0.50		
Average	**0.69**	**0.63**	**0.42**	**0.41**	**0.48**

the query precision, recall and f-measure between the datasets, as shown in Table 7. The standard deviation is maximal for Faria_2018 and Ritze_2010, but is still rather low (10^{-3}). As the standard deviation is low, the CQA coverage evaluation was performed over the 100% dataset so that the same dataset could be used for CQA coverage and instance-based precision evaluation (Table 8). Ritze_2010 and Faria_2018 both have better coverage than ra1 that they include. It means that the complex correspondences in these alignments are indeed a complement to the simple ones.

Globally, as shown in Table 8, the Query_rewriting alignments have a better coverage than the others. An exception for the edas-confOf pair could be noted. The Ontology_merging alignment outperforms the Query_rewriting one. This is explained by the choice made in the methodology for the creation of both alignments combined with the rewriting systems. In the Ontology_merging alignments, unions of properties were separated into individual subsumptions which were usable by the rewriting system, whereas in the Query_rewriting one, the subsumptions were unions. For example:

Query_rewriting correspondence:

$\langle$ *confOf:starts_on, edas:startDate* $\sqcup$ *edas:hasStartDate,* $\sqsupseteq$*, 1.0*$\rangle$

$\langle$ *confOf:Conference.confOf:starts_on.*$\top$*, edas:startDate,* $\equiv$*, 1.0*$\rangle$

Ontology_merging correspondences:

⟨*confOf:starts_on, edas:startDate,* $\sqsupseteq$*, 1.0*⟩
⟨*confOf:starts_on, edas:hasStartDate,*$\sqsupseteq$*,1.0* ⟩

Therefore, when a query contained the *edas:hasStartDate* relation, the Ontology_merging correspondence could be used, but the Query_rewriting ones could not. The precision-oriented methodology prevented the addition of the two Ontology_merging correspondences to the Query_rewriting alignment.

When closely looking at the results, many CQAs retrieving literals (titles, names, etc.) were not rewritten by the alignments. This is mainly explained because the *rdfs:label* property was introduced in the population phase when no labelling property was included in the original ontologies. The CQAs which needed (c:c) correspondences to be rewritten were not covered by the evaluated alignments. Indeed, these alignments are restricted to (s:s), (s:c) and (c:s) correspondences.

6.3. *Intrinsic precision*

Table 9 shows the precision of the alignments considering different sets of correspondences as correct. The *equivalent* precision is calculated by considering that only the correspondences whose members are *equivalent* are correct. The *subsumed* precision considers correct the correspondences whose members subsume one another (this includes the equivalent ones). The *overlapping* precision considers correct the correspondences with equivalent, subsumed or overlapping members. The *not disjoint* precision considers all correspondences whose members are not disjoint correct. The difference with the *overlapping* one is that an empty correspondence is correct in this case.

The real precision of the alignments is considered to be between the *equivalent* and the *not disjoint* values. The Query_rewriting, Ontology_merging alignments do not have a very good equivalent precision score (0.42 and 0.43). Indeed, their correspondences include a lot of subsumptions. For the subsumed, overlapping and not disjoint scores, their scores are much higher (0.94 and 0.91). ra1 has a better equivalence score (0.56) than the other two manually created alignments because it originally contains only correspondence with an equivalence relation. However, given this score seems low for a reference alignment. This low score is partly due to the different CQA coverage of the ontologies in the population phase.

For example, for the pair *cmt-edas*, the ra1 correspondence ⟨*cmt:Document, edas:Document,* $\equiv$*, 1.0* ⟩ is a subsumption in the ontology population. *cmt:Document* has for subclasses *cmt:Paper* and *cmt:Review*, whereas *edas:Document* has for subclasses *edas:Paper*, *edas:Review*, *edas:Programme* and *edas:SlideSet* which were all populated. Therefore, even if the correspondence is correct with an equivalence relation, its instance interpretation is a subsumption. Note that the instance interpretation could also be an overlap if *cmt* had another subclass (e.g., Website) which did not appear in *edas*.

The low *equivalence* score of ra1 is also due to the different interpretation of the ontologies. For example, in the pair *cmt-confOf*, the ra1 correspondence ⟨ *cmt:hasAuthor, confOf:writtenBy,* $\equiv$*, 1.0* ⟩ is a subsumption in the ontology population. *cmt:hasAuthor* was interpreted as the *"has 1st author"* relationship because of its functional property (Section 5.3.1).

Ritze_2010 has only equivalent or disjoint correspondences, therefore its precision scores are the same for all metrics. Faria_2018 achieves a good precision score overall (between 0.65 and 0.71).

Given the different population issues, the overlapping and not disjoint scores give a good representation of the alignment precision.

6.4. *Discussion*

Table 10 shows the results of the evaluation over the alignments. The CQA coverage and precision scores have been aggregated in an harmonic mean (called HM in Table 10). Overall, the Query_rewriting and Ontology_merging alignments have the better results. This is satisfactory given that these two alignments are complex reference alignments on this dataset. Even if ra1 has the best precision, its low CQA coverage (0.42) shows that a lot of CQAs from the benchmark need complex alignments to be covered. Faria_2018 and Ritze_2010 are compared to the other even if they do not contain the same number of pairs. Therefore, these numbers cannot be exactly compared to the others.

In the results of the OAEI 2018 [34], the precision measured for the Faria_2018 alignment was 0.54 (cf. Table 11). The instance-based precision gives the same result as the manual evaluation for the *cmt-ekaw* pair. For the other pairs, the gap is quite important. For the *cmt-conference* pair, this is probably due to a difference of interpretation of the ontologies. The *conference:Written_contribution* being considered as a superclass of *cmt:Paper* in the OAEI 2018 evaluation, but equivalent classes in the ontology population.

In the *conference-ekaw* pair, the ⟨$\exists$*conference:was_a_track-workshop_chair_of. conference:Tutorial, ekaw:Tutorial_Chair,* $\equiv$*, 0.369*⟩

Table 9

Different precision metrics over the alignments. The name of the precision metric is the relation between a correspondence member which is considered correct. For example, in the *equivalent* precision, the correspondences whose members were found equivalent is considered correct, the other correspondences not correct.

Average Precision	Query_rewriting	Ontology_merging	ra1	Faria_2018	Ritze_2010
equivalent	0.42	0.43	0.56	0.65	0.75
subsumed	0.80	0.80	0.83	0.71	0.75
overlapping	0.90	0.86	0.92	0.71	0.75
not disjoint	0.94	0.91	0.96	0.71	0.75

Table 10

CQA coverage and equivalence, overlapping and not disjoint precision of the alignments, harmonic mean (HM) of the two scores.

Metric	Query_rewriting	Ontology_merging	ra1	Faria_2018	Ritze_2010
CQA Coverage	**0.69**	0.63	0.42	0.41	0.48
Precision overlapping	0.90	0.86	**0.92**	0.71	0.75
Precision not disjoint	0.94	0.91	**0.96**	0.71	0.75
HM overlap	**0.78**	0.73	0.58	0.52	0.59
HM not disjoint	**0.80**	0.74	0.58	0.52	0.59

was considered correct in the OAEI 2018 evaluation. However, an axiom of the *conference* ontology restrains the domain of *conference:was_a_track-workshop_chair_of* to *conference:Track* ⊔ *conference:Workshop*. This has been taken into account in the ontology population and the correspondence was evaluated as disjoint for the evaluation system.

Table 11

Comparison of the OAEI 2019 and instance-based precision metrics over the Faria_2018 alignment. The not disjoint, subsumed and overlap precision scores are the same for this alignment.

pair	OAEI 2018	equivalent	not disjoint
cmt-conference	0.4	1.00	1.00
cmt-ekaw	0.86	0.86	0.86
conference-ekaw	0.36	0.09	0.27
Average	0.54	0.65	0.71

7. Conclusions and future work

This paper has presented an evaluation benchmark on which complex correspondences can be evaluated. In general, alignment evaluation is often performed by comparing a generated alignment to a reference one. It involves comparing the members of the correspondences generated by the systems to the members of the correspondences in the reference alignment. While this comparison is straightforward for simple alignments, this step becomes harder when dealing with complex correspondences. For example, these three correspondences can be considered as true positive: $(o\text{:AcceptedPaper},\exists o'\text{:hasDecision}.o'\text{:Acceptance},\equiv)$, $(\exists o'\text{:accepted}.\{\text{true}\},\exists o''\text{:hasDecision}.o''\text{:Acceptance},\equiv)$, or $(o\text{:AcceptedPaper},\exists o'\text{:acceptedBy}.\top,\equiv)$.

While syntactic-oriented evaluation metrics (measuring the effort to transform a correspondence into another) would fail in covering the high space of possible combinations between constructors, semantic-oriented approaches would restrict the expressiveness of correspondences to those supported by current reasoners, leaving aside for instance, transformation functions. Hence, comparison of instance sets seems to be reasonable. Our proposal shifts the problem to the comparison of instances in a task of query rewriting targeting user needs. We proposed two evaluation measures. While the *CQA coverage* measure relies on pairs of equivalent SPARQL queries (source and target queries) and measures how well an evaluated alignment covers these queries, the *intrinsic precision* compares the instances of the correspondences members.

CQA coverage, in particular, requires a way for rewriting the source query into the target query, in terms of the evaluated alignment. Such an evaluation however requires that the ontologies of the evaluation dataset are consistently populated and a system for rewriting the queries. With respect to the former, this problem has been addressed here by proposing an artificially and regularly populated dataset, as datasets with cross-ontology consistency may not be easy to find. The population process was guided by CQAs. We argue that the synthetic population ensures that each CQA is consistently populated across the ontologies. However, one can argue that in case the CQAs have different coverage for correspondences achieved through different patterns, this may have an impact on evaluation. As our evaluation is instance-based, two correspondences that do not exactly follow the same pattern but that represent the same piece of knowledge, will be considered to be comparable.

With respect to the query rewriting systems, most existing SPARQL rewriting systems are limited to (s:c) correspondences and dealing with (c:c) correspondences is still a challenge. A rewriting system which deals with such correspondences has been proposed here. However, it can not combine several (c:c) correspondence together. Instance-based rewriting could, however, be a new lead for this challenge. While the two systems have been manually evaluated in the task of rewriting queries, in the way discussed in [41], we did not evaluate the impact of each of the systems in the evaluation task. While this has to be done, we reduced their potential impact by choosing the best rewriting query, by selecting the one with the best f-measure. Another point is that these systems do not take into account correspondence relation and confidence within the rewrite process, what has to be addressed in the future.

The proposed approach has been applied for evaluating existing alignments. This system has also been applied for automating the evaluation of complex alignments in the OAEI 2019 campaign. The evaluation reported here shows that the reference alignments all have a good precision score and that complex alignments provide a better coverage of the CQAs than simple alignments. The evaluation of the alignments from two complex matchers shows that, even though both achieve a rather good precision, their CQA coverage is below 0.5. However, these results are far from the ones obtained with the original dataset and reported in OAEI campaigns, leaving a large room for improvements in the field. As our approach requires the alignments to be a priori known, it is suitable for scenarios such as the ones in OAEI. In that sense, as for the largely used artificial datasets, as the OAEI Benchmark, our dataset covers a lack of complex datasets under which an automatic evaluation can be carried in a controlled manner.

Evaluating complex ontology alignments, however, is a too broad challenge to be tackled with a single approach, as there are multiple aspects to take into account. A complementary approach to the instance-based one proposed in this paper could be an edit-distance approach that would reflect the effort involved in human validation. The approach should be also scalable, and avoid the need to do all correspondence comparisons. This could also be achieved by considering the possibility of computing minimal complex correspondences (or key complex correspondences, which can be used for computing all the other ones), in line with the work of [49]. In order to cover ontologies of various sizes and domains, developing a query generation system able to automatically generate queries adequate in coverage and scope to the evaluation of complex alignments could also help in the evaluation task.

References

[1] P.R.S. Visser, D.M. Jones, B.T.J.M. Capon and M.J.R. Shave, An Analysis of Ontological Mismatches: Heterogeneity versus Interoperability, in: *AAAI 1997 Spring Symposium on Ontological Engineering*, Stanford, USA, 1997, pp. 164–72.

[2] A. Maedche, B. Motik, N. Silva and R. Volz, MAFRA — A MApping FRAmework for Distributed Ontologies, in: *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, A. Gómez-Pérez and V.R. Benjamins, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 235–250. ISBN 978-3-540-45810-4.

[3] D. Ritze, C. Meilicke, O. Sváb-Zamazal and H. Stuckenschmidt, A Pattern-based Ontology Matching Approach for Detecting Complex Correspondences, in: *Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009*, P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, N.F. Noy and A. Rosenthal, eds, CEUR Workshop Proceedings, Vol. 551, CEUR-WS.org, 2009.

[4] D. Ritze, J. Völker, C. Meilicke and O. Sváb-Zamazal, Linguistic analysis for complex ontology matching, in: *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010*, P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, M. Mao and I.F. Cruz, eds, CEUR Workshop Proceedings, Vol. 689, CEUR-WS.org, 2010.

[5] D. Faria, C. Pesquita, B.S. Balasubramani, T. Tervo, D. Carriço, R. Garrilha, F.M. Couto and I.F. Cruz, Results of AML participation in OAEI 2018, in: *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham and O. Hassanzadeh, eds, CEUR Workshop Proceedings, Vol. 2288, CEUR-WS.org, 2018, pp. 125–131.

[6] S. Jiang, D. Lowd, S. Kafle and D. Dou, *Ontology Matching with Knowledge Rules*, in: *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVIII: Special Issue on Database- and Expert-Systems Applications*, A. Hameurlain,

J. Küng, R. Wagner and Q. Chen, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016, pp. 75–95.

[7] R. Parundekar, C.A. Knoblock and J.L. Ambite, Linking and Building Ontologies of Linked Data, in: *The Semantic Web – ISWC 2010*, P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks and B. Glimm, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 598–614. ISBN 978-3-642-17746-0.

[8] R. Parundekar, C.A. Knoblock and J.L. Ambite, Discovering Concept Coverings in Ontologies of Linked Data Sources, in: *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J.X. Parreira, J. Hendler, G. Schreiber, A. Bernstein and E. Blomqvist, eds, Lecture Notes in Computer Science, Vol. 7649, Springer, 2012, pp. 427–443, 10.1007/978-3-642-35176-1_27.

[9] B. Walshe, R. Brennan and D. O'Sullivan, Bayes-ReCCE: A Bayesian Model for Detecting Restriction Class Correspondences in Linked Open Data Knowledge Bases **12**(2) (2016), 25–52–.

[10] É. Thiéblin, O. Haemmerlé and C. Trojahn, Complex matching based on competency questions for alignment: a first sketch, in: *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham and O. Hassanzadeh, eds, CEUR Workshop Proceedings, Vol. 2288, CEUR-WS.org, 2018, pp. 66–70.

[11] É. Thiéblin, O. Haemmerlé and C. Trojahn, Generating Expressive Correspondences: An Approach Based on User Knowledge Needs and A-Box Relation Discovery, in: *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part I*, J.Z. Pan, V.A.M. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Lecture Notes in Computer Science, Vol. 12506, Springer, 2020, pp. 565–583, 10.1007/978-3-030-62419-4_32.

[12] B.P. Nunes, A.A.M. Caraballo, M.A. Casanova, K.K. Breitman and L.A.P.P. Leme, Complex matching of RDF datatype properties, in: *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, P. Shvaiko, J. Euzenat, T. Heath, C. Quix, M. Mao and I.F. Cruz, eds, CEUR Workshop Proceedings, Vol. 814, CEUR-WS.org, 2011.

[13] H. Qin, D. Dou and P. LePendu, Discovering Executable Semantic Mappings Between Ontologies, in: *On the Move to Meaningful Internet Systems 2007: CoopIS, DOA, ODBASE, GADA, and IS, OTM Confederated International Conferences CoopIS, DOA, ODBASE, GADA, and IS 2007, Vilamoura, Portugal, November 25-30, 2007, Proceedings, Part I*, R. Meersman and Z. Tari, eds, Lecture Notes in Computer Science, Vol. 4803, Springer, 2007, pp. 832–849.

[14] R. Dhamankar, Y. Lee, A. Doan, A. Halevy and P. Domingos, IMAP: Discovering Complex Semantic Matches between Database Schemas, in: *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, Association for Computing Machinery, New York, NY, USA, 2004, pp. 383–394–, 10.1145/1007568.1007612. ISBN 1581138598.

[15] B. He, K.C. Chang and J. Han, Discovering complex matchings across web query interfaces: a correlation mining approach, in: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, W. Kim, R. Kohavi, J. Gehrke and W. DuMouchel, eds, ACM, 2004, pp. 148–157, 10.1145/1014052.1014071.

[16] É. Thiéblin, O. Haemmerlé, N. Hernandez and C. Trojahn, Survey on complex ontology matching, *Semantic Web* **11**(4) (2020), 689–727, 10.3233/SW-190366.

[17] É. Thiéblin, M. Cheatham, C.T. dos Santos, O. Zamazal and L. Zhou, The First Version of the OAEI Complex Alignment Benchmark, in: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*, M. van Erp, M. Atre, V. López, K. Srinivas and C. Fortuna, eds, CEUR Workshop Proceedings, Vol. 2180, CEUR-WS.org, 2018.

[18] J. Euzenat, M. Rosoiu and C. Trojahn, Ontology matching benchmarks: Generation, stability, and discriminability, *Journal of Web Semantics* **21** (2013), 30–48, 10.1016/j.websem.2013.05.002.

[19] O. Zamazal and V. Svátek, The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere, *Journal of Web Semantics* **43** (2017), 46–53, 10.1016/j.websem.2017.01.001.

[20] É. Thiéblin, Do Competency Questions for Alignment Help Fostering Complex Correspondences?, in: *Proceedings of the EKAW Doctoral Consortium 2018 co-located with the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018), Nancy, France, November 13, 2018*, L. Hollink and F. Osborne, eds, CEUR Workshop Proceedings, Vol. 2306, CEUR-WS.org, 2018.

[21] G. Stapleton, J. Howse, A. Bonnington and J. Burton, A Vision for Diagrammatic Ontology Engineering, in: *Proceedings of the International Workshop on Visualizations and User Interfaces for Knowledge Engineering and Linked Data Analytics co-located with 19th International Conference on Knowledge Engineering and Knowledge Management, VISUAL@EKAW 2014, Linköping, Sweden, November 24, 2014*, V. Ivanova, T. Kauppinen, S. Lohmann, S. Mazumdar, C. Pesquita and K. Xu, eds, CEUR Workshop Proceedings, Vol. 1299, CEUR-WS.org, 2014, pp. 1–13.

[22] J. Euzenat and P. Shvaiko, *Ontology Matching*, 2nd edn, Springer Berlin Heidelberg, 2013. ISBN 3642387209.

[23] É. Thiéblin, O. Haemmerlé, N. Hernandez and C. Trojahn, 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*, A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Lecture Notes in Computer Science, Vol. 10843, Springer, 2018, pp. 655–670, 10.1007/978-3-319-93417-4_42.

[24] M. Grüninger and M.S. Fox, Methodology for the Design and Evaluation of Ontologies, in: *Workshop on Basic Ontological Issues in Knowledge Sharing*, Vol. 15, 1995.

[25] Y. Ren, A. Parvizi, C. Mellish, J.Z. Pan, K. van Deemter and R. Stevens, Towards Competency Question-Driven Ontology Authoring, in: *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, V. Presutti,

C. d'Amato, F. Gandon, M. d'Aquin, S. Staab and A. Tordai, eds, Lecture Notes in Computer Science, Vol. 8465, Springer, 2014, pp. 752–767, 10.1007/978-3-319-07443-6_50.

[26] C. Meilicke and H. Stuckenschmidt, Incoherence as a Basis for Measuring the Quality of Ontology Mappings, in: *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008) Collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, October 26, 2008*, P. Shvaiko, J. Euzenat, F. Giunchiglia and H. Stuckenschmidt, eds, CEUR Workshop Proceedings, Vol. 431, CEUR-WS.org, 2008.

[27] A. Solimando, E. Jiménez-Ruiz and G. Guerrini, Detecting and Correcting Conservativity Principle Violations in Ontology-to-Ontology Mappings, in: *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C.A. Knoblock, D. Vrandecic, P. Groth, N.F. Noy, K. Janowicz and C.A. Goble, eds, Lecture Notes in Computer Science, Vol. 8797, Springer, 2014, pp. 1–16, 10.1007/978-3-319-11915-1_1.

[28] A. Isaac, H. Matthezing, L. van der Meij, S. Schlobach, S. Wang and C. Zinn, Putting Ontology Alignment in Context: Usage Scenarios, Deployment and Evaluation in a Library Case, in: *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, S. Bechhofer, M. Hauswirth, J. Hoffmann and M. Koubarakis, eds, Lecture Notes in Computer Science, Vol. 5021, Springer, 2008, pp. 402–417, 10.1007/978-3-540-68234-9_31.

[29] A. Isaac, D. Kramer, L. van der Meij, S. Wang, S. Schlobach and J. Stapel, Vocabulary Matching for Book Indexing Suggestion in Linked Libraries - A Prototype Implementation and Evaluation, in: *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, A. Bernstein, D.R. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta and K. Thirunarayan, eds, Lecture Notes in Computer Science, Vol. 5823, Springer, 2009, pp. 843–859, 10.1007/978-3-642-04930-9_53.

[30] W. Su, J. Wang and F.H. Lochovsky, Holistic Schema Matching for Web Query Interfaces, in: *Advances in Database Technology - EDBT 2006, 10th International Conference on Extending Database Technology, Munich, Germany, March 26-31, 2006, Proceedings*, Y.E. Ioannidis, M.H. Scholl, J.W. Schmidt, F. Matthes, M. Hatzopoulos, K. Böhm, A. Kemper, T. Grust and C. Böhm, eds, Lecture Notes in Computer Science, Vol. 3896, Springer, 2006, pp. 77–94, 10.1007/11687238_8.

[31] Y. An, X. Hu and I. Song, Learning to discover complex mappings from web forms to ontologies, in: *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, X. Chen, G. Lebanon, H. Wang and M.J. Zaki, eds, ACM, 2012, pp. 1253–1262, 10.1145/2396761.2398427.

[32] É. Thiéblin, F. Amarger, N. Hernandez, C. Roussey and C.T. dos Santos, Cross-Querying LOD Datasets Using Complex Alignments: An Application to Agronomic Taxa, in: *Metadata and Semantic Research - 11th International Conference, MTSR 2017 Tallinn, Estonia, November 28 - December 1, 2017, Proceedings*, E. Garoufallou, S. Virkus, R. Siatri and D. Koutsomiha, eds, Communications in Computer and Information Science, Vol. 755, Springer, 2017, pp. 25–37, 10.1007/978-3-319-70863-8_3.

[33] L. Zhou, M. Cheatham, A. Krisnadhi and P. Hitzler, 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*, D. Vrandecic, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L. Kaffee and E. Simperl, eds, Lecture Notes in Computer Science, Vol. 11137, Springer, 2018, pp. 273–288, 10.1007/978-3-030-00668-6_17. https://doi.org/10.1007/978-3-030-00668-6_17.

[34] A. Algergawy, M. Cheatham, D. Faria, A. Ferrara, I. Fundulaki, I. Harrow, S. Hertling, E. Jiménez-Ruiz, N. Karam, A. Khiat, P. Lambrix, H. Li, S. Montanelli, H. Paulheim, C. Pesquita, T. Saveta, D. Schmidt, P. Shvaiko, A. Splendiani, É. Thiéblin, C. Trojahn, J. Vatascinová, O. Zamazal and L. Zhou, Results of the Ontology Alignment Evaluation Initiative 2018, in: *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham and O. Hassanzadeh, eds, CEUR Workshop Proceedings, Vol. 2288, CEUR-WS.org, 2018, pp. 76–116.

[35] M. Ehrig and J. Euzenat, 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*, B. Ashpole, M. Ehrig, J. Euzenat and H. Stuckenschmidt, eds, CEUR Workshop Proceedings, Vol. 156, CEUR-WS.org, 2005.

[36] É. Thiéblin, O. Haemmerlé and C. Trojahn, CANARD complex matching system: results of the 2018 OAEI evaluation campaign, in: *Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018*, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham and O. Hassanzadeh, eds, CEUR Workshop Proceedings, Vol. 2288, CEUR-WS.org, 2018, pp. 138–143.

[37] A. Solimando, E. Jiménez-Ruiz and C. Pinkel, Evaluating Ontology Alignment Systems in Query Answering Tasks, in: *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014*, M. Horridge, M. Rospocher and J. van Ossenbruggen, eds, CEUR Workshop Proceedings, Vol. 1272, CEUR-WS.org, 2014, pp. 301–304.

[38] J. David, J. Euzenat, P. Genevès and N. Layaïda, 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*, P. Champin, F. Gandon, M. Lalmas and P.G. Ipeirotis, eds, ACM, 2018, pp. 1599–1602, 10.1145/3184558.3191617.

[39] L. Hollink, M. van Assem, S. Wang, A. Isaac and G. Schreiber, Two Variations on Ontology Alignment Evaluation: Methodological Issues, in: *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, S. Bechhofer, M. Hauswirth, J. Hoffmann and M. Koubarakis, eds, Lecture Notes in Computer Science, Vol. 5021, Springer, 2008, pp. 388–401, 10.1007/978-3-540-68234-9_30.

[40] K. Makris, N. Bikakis, N. Gioldasis and S. Christodoulakis, SPARQL-RW: transparent query access over mapped RDF data sources, in: *15th International Conference on Extending Database Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings*, E.A. Rundensteiner, V. Markl, I. Manolescu, S. Amer-Yahia, F. Naumann and I. Ari, eds, ACM, 2012, pp. 610–613, 10.1145/2247596.2247678.

[41] É. Thiéblin, F. Amarger, O. Haemmerlé, N. Hernandez and C.T. dos Santos, 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*, P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh and R. Ichise, eds, CEUR Workshop Proceedings, Vol. 1766, CEUR-WS.org, 2016, pp. 49–60.

[42] J. David, J. Euzenat, F. Scharffe and C. Trojahn, The Alignment API 4.0, *Semantic Web* **2**(1) (2011), 3–10, 10.3233/SW-2011-0028.

[43] J. Euzenat, A. Polleres and F. Scharffe, Processing Ontology Alignments with SPARQL, in: *Second International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2008), March 4th-7th, 2008, Technical University of Catalonia, Barcelona, Spain*, F. Xhafa and L. Barolli, eds, IEEE Computer Society, 2008, pp. 913–917, 10.1109/CISIS.2008.126.

[44] I. Horrocks, O. Kutz and U. Sattler, The Even More Irresistible SROIQ, in: *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2-5, 2006*, P. Doherty, J. Mylopoulos and C.A. Welty, eds, AAAI Press, 2006, pp. 57–67.

[45] M.C. Suárez-Figueroa, A. Gómez-Pérez and M. Fernández-López, The NeOn Methodology for Ontology Engineering, in: *Ontology Engineering in a Networked World*, M.C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta and A. Gangemi, eds, Springer, 2012, pp. 9–34, 10.1007/978-3-642-24794-1_2.

[46] O. Šváb, V. Svátek, P. Berka, D. Rak and P. Tomášek, Ontofarm: Towards an experimental collection of parallel ontologies, in: *Proceedings of the 4th International Semantic Web Conference (ISWC). Poster*, 2005.

[47] A.G. Nuzzolese, A.L. Gentile, V. Presutti and A. Gangemi, Conference Linked Data: The ScholarlyData Project, in: *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, P. Groth, E. Simperl, A.J.G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck and Y. Gil, eds, Lecture Notes in Computer Science, Vol. 9982, 2016, pp. 150–158, 10.1007/978-3-319-46547-0_16.

[48] R. Shearer, B. Motik and I. Horrocks, HermiT: A Highly-Efficient OWL Reasoner, in: *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, October 26-27, 2008*, C. Dolbear, A. Ruttenberg and U. Sattler, eds, CEUR Workshop Proceedings, Vol. 432, CEUR-WS.org, 2008.

[49] V. Maltese, F. Giunchiglia and A. Autayeu, Save Up to 99% of Your Time in Mapping Validation, in: *On the Move to Meaningful Internet Systems, OTM 2010 - Confederated International Conferences: CoopIS, IS, DOA and ODBASE, Hersonissos, Crete, Greece, October 25-29, 2010, Proceedings, Part II*, R. Meersman, T.S. Dillon and P. Herrero, eds, Lecture Notes in Computer Science, Vol. 6427, Springer, 2010, pp. 1044–1060. doi:10.1007/978-3-642-16949-6_28.