Conference v3.0: A Populated Version of the Conference Dataset

Elodie Thiéblin, Cassia Trojahn

Institut de Recherche Informatique de Toulouse, France
{firstname.lastname}@irit.fr

Abstract. The Conference dataset consists of independently designed ontologies in the domain of conference organisation, together with a subset of reference alignments between these ontologies. It has been widely used in ontology matching evaluation, in particular in the context of the Ontology Alignment Evaluation Initiative (OAEI). This dataset however is not equipped with instances, limiting its exploitation by matchers. This paper describes the methodology followed to populate a subset of the Conference dataset, both synthetically and with real data.

1 Introduction

Ontology matching is the task of generating alignments between the entities of different ontologies. Several matching approaches have been proposed in the literature [2] and systematic evaluation of them has been carried out over the last fifteen years in the context of the Ontology Alignment Evaluation Initiative (OAEI)[1]. One OAEI track offering expressive and real-world ontologies is the Conference track, whose dataset has been proposed in [8]. This dataset consists of 16 independently designed ontologies of the conference organisation domain, together with a subset of 21 reference alignments between 7 of these ontologies. This dataset has became one of the most used ones in matching evaluation [7] and has been extended in different proposals [1, 3]. Recently, it has been extended with complex alignments [5].

The Conference dataset however is not equipped with instances, limiting the evaluation of matching approaches relying on them. While in [4], a partially populated version of the dataset has been used to evaluate alignments on the query rewriting task, the resulting dataset is limited to the scope of the queries used in the evaluation (only ontology concepts corresponding to 18 queries). In this paper, a fully populated version of the dataset is proposed. We present the methodology that has been followed to populate a subset of 5 Conference ontologies, both synthetically and with real data. It has been based on the notion of *competence questions for alignment* (CQA) which define the knowledge needs to be covered (at best) by the ontologies and the alignment between them [6]. The use of CQAs ensures that the populating is homogeneous across ontologies. Thanks to this dataset, it will be possible to automatise the evaluation process of complex matchers using an evaluation strategy based on the comparison of instances in a query rewriting setting rather than comparing syntactically complex correspondences to references ones.

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

2 Overall methodology

The methodology followed for populating the dataset has the following main steps:

1. Create a set of CQAs based on an application scenario in order to guide the ontology interpretation by the experts. Examples of CQAs include: “What are the accepted papers?” (unary CQA) or “Which are the authors of accepted papers” (binary CQA).
2. Create a pivot format (e.g., JSON schema) for covering the CQAs from step 1 (e.g. covering attributes describing specific types of objects, such as papers or people):

```
{ "id": "10",
  "title": "User-Centric Ontology Population",
  "authors": ["K. Clarkson", ...],
  "type": "Research track",
  "decision":"accept" }
```

3. For each ontology of the dataset, create SPARQL INSERT queries from the pivot format (here, an ontology may not cover the whole pivot format).

```
INSERT DATA {
      {{pap}} a :Camera_ready_contribution.
      {{pap}} rdfs:label {{paptitle}}.
      {{pap}} :is_submitted_at {{conf}}.
      {{pap}} :has_authors {{auth}}.
      ...  }
```

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 exception occurs, try to change the interpretation of the ontology and iterate over steps 3 to 5.

3 Populated dataset

The methodology above has been followed to populate 5 ontologies from the Conference data: *cmt*, *conference* (Sofsem), *confOf* (confTool), *edas* and *ekaw* (Table 1). This choice is motivated by the fact that these ontologies have been also the ones used in the complex version of this dataset. A total of 152 CQAs have been created by an expert using as basis the ESWC 2018 conference scenario (whose data were fully open) and expanded by ontology exploration. The pivot format was first instantiated with data from the ESWC 2018 website and an automatic instantiation script of the pivot format was developed taking into account some statistics (e.g, proportion of members of the program committee author of articles, etc.). The dataset and instantiations of the pivot format have been made available[2].

In addition to the ESWC 2018 dataset, 6 other datasets (with 25 artificial conferenes) have been generated in order to cover the cases where ontologies share common

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

instances. In these artificial datasets, each ontology has been populated with 5 pivot instantiation data. In the “dataset 0%” all ontologies were populated with 5 different pivot format instantiations; in the “dataset 20%”, the ontologies were populated with 1 identical and 4 different instantiations; the other datasets (40%, 60%, 80%, and 100%) followed the same strategy. Since the size of each instantiation may differ, the percentage of common instances between two ontologies varies. For example, in the dataset 20%, the instances *Papers* common to the ontologies represent between 7% instances of *Papers* of *ekaw* and 11% of instances of *Papers* of *cmt*.

Table 1. Populated entities/total entities by ontology. Number of CQAs covered by each ontology.

	cmt	conference	confOf	edas	ekaw
Classes	26 / 30	51 / 60	29 / 39	42 / 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
CQAs	46	90	67	60	84

4 Discussion

Running the Hermit reasoner (step 6 of the methodology), several incoherences were encountered. For most of them, the problem was with the interpretation of the ontology. For example, in *cmt*, *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*. Hence, the SPARQL INSERT queries have been modified accordingly. We have also detected exceptions that could not be resolved by changing the interpretation. In that case, the original ontologies have been slightly modified. For instance, in *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.

With respect to the CQAs, if a given CQA is not fully covered by an ontology, it would not be populated with the corresponding instances. This results in an uneven population of equivalent concepts. For example, considering *ekaw* and *cmt*, which both contain a *Document* class. However, *“What are the documents?”* are rather covered by *paper, review, web site* and *proceedings* instantiations, as *ekaw:Document* has four subclasses (*ekaw:Paper*, *ekaw:Review*, *ekaw:Web_Site* and *ekaw:Conference_Proceedings*) and *cmt:Document* has only two subclasses (*cmt:Paper* and *cmt:Review*). We could also have considered each class with exactly the same instances, e.g., populating *cmt:Document* with all the *Paper*, *Review*, *Web site* and *Conference proceedings* instances. Therefore, *cmt:Document* and *ekaw:Document* would share exactly the same instances. However, we chose to remain closer to the original ontologies as possible (the lack of a class in an ontology is due to the requirements of its creators). The instances also reflects the conceptual mismatches between the ontologies.

In order to evaluate the dataset itself, we verified that two equivalent classes would not obtain a disjoint relation on the populated dataset. For that, we used the reference alignment *ra1* from the original Conference dataset and modified it in order to take into account our interpretations. Then, the instances of the source and target member of each correspondence of the modified *ra1* were compared, resulting in non disjoint correspondences. We have also calculated the intrinsic precision of reference alignments such as the simple alignment *ra1* and the two complex ones (rewriting and merging) of [5] (Table 2). In a dataset where two common classes are either populated with the same instances, not populated or share at least a subclass with the same instances, this metric gives a lower and upper bound for the precision of the alignment. The lower bound is given by the *classical* score in which only equivalent members are considered as correct. The upper bound is given by the *not disjoint* score in which all correspondences with overlapping or empty members are considered correct.

Table 2. Results on comparing the instances related to the entities in the correspondences.

	classical	recall oriented	precision oriented	overlap	not disjoint
ra1	0.563	0.763	0.763	0.923	0.990
Ontology merging	0.445	0.724	0.724	0.880	0.955
Query rewriting	0.429	0.719	0.719	0.911	0.976

5 Conclusion

This paper has presented a populated version of a subset of the Conference ontologies. This dataset will contribute to automatising the evaluation of complex matchers and expanding the scope of its use in ontology matching in general. We plan to evaluate the behaviour of the approaches generating complex alignments under the different percentages of instances overlap, and to integrate this dataset in the evaluation of complex matchers in OAEI 2019.

References

1. M. Cheatham and P. Hitzler. Conference v2. 0: An uncertain version of the OAEI Conference benchmark. In *ISWC*, pages 33–48, 2014.
2. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer Berlin Heidelberg, 2013.
3. C. Meilicke, R. Garcia-Castro, F. Freitas, W. R. van Hage, E. Montiel-Ponsoda, R. R. de Azevedo, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, A. Tamilin, C. Trojahn, and S. Wang. Multifarm: A benchmark for multilingual matching. *JWS*, 15:62–68, 2012.
4. A. Solimando, E. Jimnez-Ruiz, and C. Pinkel. Evaluating ontology alignment systems in query answering tasks. In *ISWC Poster Track*, pages 301–304, 2014.
5. É. Thiéblin, O. Haemmerlé, N. Hernandez, and C. Trojahn. Task-oriented complex ontology alignment: Two alignment evaluation sets. In *ESWC*, pages 655–670, 2018.
6. E. Thiéblin, O. Haemmerlé, and C. Trojahn. Complex matching based on competency questions for alignment: a first sketch. In *OM@ISWC*, 2018.
7. O. Zamazal and V. Svatek. The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere. *Web Semantics*, 43:46–53, Mar. 2017.
8. O. Zamazal, V. Svatek, P. Berka, D. Rak, and P. Tomasek. Ontofarm: Towards an experimental collection of parallel ontologies. *ISWC Poster Track*, 2005, 2005.