The First Version of the OAEI Complex Alignment Benchmark

Elodie Thiéblin[1], Michelle Cheatham[2], Cassia Trojahn[1], Ondřej Zamazal[3] and Lu Zhou[2]

[1] IRIT & Université de Toulouse 2 Jean Jaurès, Toulouse, France
[2] Wright State University, Dayton, USA
[3] University of Economics, Prague, Czech Republic
elodie.thieblin@irit.fr, michelle.cheatham@gmail.com,
cassia.trojahn@irit.fr, ondrej.zamazal@vse.cz, zhou.34@wright.edu

Abstract. We present the first version of the complex benchmark of the Ontology Alignment Evaluation Initiative campaigns. This benchmark is composed of four datasets from different domains (conference, hydrology, geoscience and agronomy) and covers different evaluation strategies.

Keywords: complex ontology alignments, evaluation dataset, OAEI

1 Introduction

Complex correspondences involve transformation functions of literal values or logical constructors (e.g. $\forall x$, *ekaw:AcceptedPaper(x)* $\equiv \exists y$, *cmt:acceptedBy(x,y)*), which make them more expressive than simple correspondences. Complex alignments, composed of at least one complex correspondence, are therefore a complement to simple alignments. Different approaches for complex matching have emerged in the literature [2,4,5,8]. Most of them, however, have been evaluated on tailored datasets (e.g., targeting a specific correspondence pattern). Most efforts on systematic evaluation, in the context of the OAEI campaigns[1], are still dedicated to simple matchers.

This paper presents the first version of the OAEI complex track, composed of four datasets (Table 1) from different domains. This domain and correspondence variety allows for better covering different kinds of heterogeneity between ontologies. Different evaluation strategies aim at evaluating complex matchers under different perspectives. The evaluation will be supported by the SEALS platform and the output alignments must be in EDOAL. The detail of each dataset and evaluation process can be found on the OAEI's 2018 complex track webpage[2], and are introduced in the following.

[1] `http://oaei.ontologymatching.org/`
[2] `http://oaei.ontologymatching.org/2018/complex/index.html`

Dataset	Ontologies	(1:1)	(1:n)	(m:n)
Conference consensus	3	78	79	0
Hydrography	4	113	69	15
GeoLink	2	24	15	72
Taxon	4	6	17	3

Table 1. Number of ontologies and correspondences by kind in each dataset. (1:1) are simple correspondences, (1:n) and (m:n) are complex correspondences.

2 Conference consensual dataset

This dataset is based on the OntoFarm dataset [9], which is composed of 16 ontologies on the conference organisation domain and simple reference alignments between 7 of them. Here, we consider 3 out of the 7 ontologies from the reference alignments (*cmt*, *conference* and *ekaw*), resulting in 3 alignment pairs. The alignments involve both logical constructors (76 correspondences) and transformations (3 correspondences). Examples are given in the following :

1. $\forall$ *x, ekaw:AcceptedPaper(x)* $\equiv \exists$ *y, cmt:acceptedBy(x,y)* is a correspondence with the *existential* constructor.
2. $\forall$ *x,y, cmt:name(x,y)* $\equiv \exists$ y_1*,* y_2*, conference:has_the_first_name(x,*y_1*)* $\wedge$ *conference:has_the_last_name(x,*y_2*)* $\wedge$ *concatenation(y,*y_1*," ",* y_2*)*, where *concatenation(a,*b_1*,* b_2*, ...)* is a predicate ensuring that its first parameter *a* is equal to the string concatenation of the others $\{b_1, b_2, ...\}$. It uses a transformation function of the literal values.

The alignments have been manually created by three experts in the domain, following the methodology in [7]. Four experts assessed the generated correspondences to reach a consensus. The systems will be manually evaluated on their output alignments to produce precision and recall scores. Only the complex equivalence correspondences will be assessed. The systems can use a simple reference alignment as input. Confidence scores of correspondences will not be taken into account in the evaluation.

3 Hydrography dataset

The hydrography dataset is composed of 4 source ontologies (Hydro3, hydrOntology_native, hydrOntology_translated and Cree) that each should be aligned to a single target Surface Water Ontology (*swo*). The source ontologies vary in their similarity to the target ontology – Hydro3 is similar in both language and structure, hydrOntology_native and hydrOntology_translated are similar in structure but hydrOntology_translated is in Spanish rather than English, and Cree is very different in terms of both language and structure. The alignments were created by a geologist and an ontologist, in consultation with a native Spanish speaker regarding the hydrOntology_translated, and consist of logical relations such as the one shown below.

1. $\forall$*x, hydrOntology_translated:Aguas_Corrientes(x)* $\leq$ *swo:SurfaceFeature(x)* $\wedge$ *swo:Waterbody(x)* $\wedge$ $\exists$*y, swo:hasFlow(x,y)* $\wedge$ *swo:Flow(y)*

Performance on this dataset will be evaluated on three sub-tasks: 1) identifying the atoms (classes and properties) from the target ontology involved in the relations (e.g., *swo:SurfaceFeature*, *swo:Waterbody*, *swo:hasFlow* and *swo:Flow* from the correspondence above), 2) when given the atoms, identifying the logical relations that hold between them and 3) the full complex alignment task. Evaluation of the first sub-task will use traditional F-measure, while the remaining two subtasks will be evaluated on semantic F-measure [1].

4 GeoLink dataset

This dataset is from the GeoLink project[3], which was funded under the U.S. National Science Foundation's EarthCube initiative. It is composed of 2 populated ontologies: the GeoLink base ontology (*gbo*) and the GeoLink modular ontology (*gmo*). The GeoLink project is a real-world use case of ontologies. The alignment between the ontologies was developed in consultation with domain experts from several Geoscience research institutions. The complex correspondences include not only class and property subsumption and property chains (described in [5]), but also some that involve typecasting (c.f. [3]), for example:

1. Property Chain: $\forall x,z,\ gbo{:}Award(x) \wedge gbo{:}hasSponsor(x,z) \equiv$ $\exists y,\ gmo{:}FundingAward(x) \wedge gmo{:}providesAgentRole(x,y) \wedge$ $gmo{:}SponsorRole(y) \wedge gmo\ {:}performedBy(y,z)$
2. Class Typecasting: $\forall x,\ gbo{:}PlaceType(x) \equiv rdfs{:}subClassOf(x,\ gmo{:}Place)$

More information about this dataset can be found in [10] and the benchmark and alignment can be downloaded here[4]. The performance of alignment systems on this dataset will be evaluated in the same way as the hydrography dataset.

5 Taxon dataset

This dataset is composed of 4 populated ontologies whose common scope is plant taxonomy: AgronomicTaxon (*agtx*), Agrovoc (*agv* and *agronto*), DBpedia (*dbo*) and TaxRef-LD (*txr*). This dataset extends the one proposed in [6] by adding the TaxRef-LD ontology. The alignments were manually created with the help of one expert and involve only logical constructors, as for example:

1. $\forall x,\ agtx{:}GenusRank(x) \equiv agronto{:}hasTaxonomicRank(x,agv{:}c\_11125)$
2. $\forall x,\ agtx{:}GenusRank(x) \equiv \exists y,\ dbo{:}Species(y) \wedge dbo{:}genus(y,x) \wedge dbo{:}Species(x)$

The evaluation of this dataset is task-oriented. We will evaluate the generated correspondences using a SPARQL query rewriting system and manually measure their ability of answering a set of queries over each dataset. For example, a competency question could be "Retrieve all the genus taxa". For AgronomicTaxon, as source ontology, the corresponding SPARQL query is `SELECT ?x WHERE {?x a agtx:GenusRank.}` and the correspondences output by the systems with Agrovoc as target ontology, should be able to translate the query into: `SELECT ?x WHERE {?x agronto:hasTaxonomicRank agv:c_11125.}`

[3] `https://www.geolink.org/`

[4] `http://doi.org/10.6084/m9.figshare.5907172`

6 Conclusions

This paper has presented the first OAEI complex evaluation track, covering different kinds of complex correspondences, domains and evaluation strategies. For most datasets, the evaluation is still manually performed, opening directions on how complex alignments can be automatically generated and evaluated.

Acknowledgements. We thank Catherine Roussey (IRSTEA) and Nathalie Hernandez (IRIT) for their help on the Taxon dataset and Dalia Varanka (US Geological survey) for her work on the hydrography dataset. Ondřej Zamazal has been partially supported by the CSF grant no. 18-23964S. Creation of the GeoLink dataset was funded by NSF 1440202.

References

1. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 348–353 (2007)
2. Jiang, S., Lowd, D., Kafle, S., Dou, D.: Ontology matching with knowledge rules. In: Transactions on Large-Scale Data-and Knowledge-Centered Systems, pp. 75–95 (2016)
3. Krisnadhi, A.A., Hitzler, P., Janowicz, K.: On the capabilities and limitations of OWL regarding typecasting and ontology design pattern views. In: Ontology Engineering - 12th International Experiences and Directions Workshop on OWL, OWLED 2015, co-located with ISWC 2015, Bethlehem, PA, USA, October 9-10, 2015, Revised Selected Papers. pp. 105–116 (2015)
4. Parundekar, R., Knoblock, C.A., Ambite, J.L.: Discovering concept coverings in ontologies of linked data sources. In: ISWC. pp. 427–443 (2012)
5. Ritze, D., Meilicke, C., Šváb Zamazal, O., Stuckenschmidt, H.: A pattern-based ontology matching approach for detecting complex correspondences. In: 4th OM workshop. pp. 25–36 (2009)
6. Thiéblin, E., Amarger, F., Hernandez, N., Roussey, C., Trojahn, C.: Cross-querying lod datasets using complex alignments: an application to agronomic taxa. In: MTSR. pp. 25–37 (2017)
7. Thiéblin, E., Haemmerlé, O., Hernandez, N., Trojahn, C.: Task-oriented complex ontology alignment – two alignment evaluation sets. In: ESWC (2018), (to appear)
8. Walshe, B., Brennan, R., O'Sullivan, D.: Bayes-recce: A bayesian model for detecting restriction class correspondences in linked open data knowledge bases. International Journal on Semantic Web and Information Systems 12(2), 25–52 (2016)
9. Zamazal, O., Svátek, V.: The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere. Journal of Web Semantics 43, 46–53 (2017)
10. Zhou, L., Cheatham, M., Krisnadhi, A., Hitzler, P.: A complex alignment benchmark: Geolink dataset. In: ISWC. Springer (2018), (to appear)