Ontology Alignment in Ecotoxicological Effect Prediction*

Erik B. Myklebust[1,2], Ernesto Jiménez-Ruiz[2,3], Jiaoyan Chen[4],
Raoul Wolf[1], and Knut Erik Tollefsen[1,5]

[1] Norwegian Institute for Water Research, Oslo, Norway
[2] SIRIUS, University of Oslo, Oslo, Norway
[3] City, University of London, London, United Kingdom
[4] University of Oxford, Oxford, United Kingdom
[5] Norwegian University of Life Sciences, Ås, Norway

1 Introduction

The Toxicological and Risk Assessment Knowledge Graph (TERA) [1] integrates several disparate datasets relevant to ecological risk assessment and effect prediction. TERA is being used in conjunction with knowledge graph embedding models to improve the extrapolation of chemical effect data in the Norwegian Institute for Water Research (Norsk institutt for vannforskning, NIVA) [1].[1]

The largest publicly available repository of effect data is the ECOTOXicology knowledge base (ECOTOX) developed by the US Environmental Protection Agency [2]. The dataset consists of $940k$ experiments using $12k$ compounds and $13k$ species. ECOTOX contains a taxonomy (of species), however, this only considers the species represented in the ECOTOX effect data. Hence, to enable extrapolation of effects across a larger taxonomic domain, an alignment to the NCBI taxonomy have to be established. However, there does not exist a complete and public mapping set between the 47,785 ECOTOX taxa and the 2,140,344 NCBI taxa. In this paper we present the ECOTOX-NCBI alignment results of three ontology matching algorithms.

2 Methods and Evaluation

Although there does not exist a complete and public alignment between the ECOTOX and NCBI, a partial mapping curated by experts can be obtained through the ECOTOX Web.[2] We have gathered a total of 2,321 mappings for validation purposes. We have used three methods to align the two vocabularies: *(i)* LogMap system [3]. *(ii)* AgreementMakerLight (AML) , and *(iii)* a baseline string matching algorithm based on Levenshtein distance [4].

Table 1 shows the alignment results over the ground truth samples. Note that the results represent 1-to-1 alignments as, in our setting, it is expected an entity from

[1] Knowledge Graphs at NIVA: `https://github.com/NIVA-Knowledge-Graph/`

[2] ECOTOX search interface: `https://cfpub.epa.gov/ecotox/search.cfm`

Algorithm	# mappings	Recall	Precision (*)
LogMap	$32,726$	0.81	0.88
AML	$31,659$	0.80	0.87
String distance (> 0.8)	$33,554$	0.38	0.70
Union all	$57,511$	0.72	0.73
Consensus (LogMap $\cap$ AML)	$20,217$	0.78	0.95
LogMap $\cup$ AML	$39,985$	0.83	0.85

Table 1. Alignment results for ECOTOX-NCBI. (*) Estimated precision with respect to the known entities in the incomplete reference alignment, assuming only 1-1 mappings are valid.

ECOTOX to match to a single entity in NCBI, and vice-versa. Hence, 1-to-N (respectively N-to-1) alignments were filtered according to the system computed confidence. LogMap and AML produce mapping sets with similar recall and (estimated) precision, with LogMap producing a larger number of mappings. The baseline matcher, as expected, achieves both a lower recall and (estimated) precision. This shows that a simple string matching solution may not be enough in this setting. Table 1 also shows the results of the consensus alignment between AML and LogMap and the union of different mapping sets. Note that the lower recall of the union is down to overconfidence in the string distance method when 1-to-1 filtering.

3 Conclusions

The used alignment techniques achieve relatively good scores for recall over the available (incomplete) reference mappings. However, aligning such large and challenging datasets required some preprocessing before ontology alignment systems could cope with them. The preprocessing involved to split NCBI into manageable fragments, leading to a set of matching subtasks instead of a single task. Thus, the alignment of ECOTOX and NCBI has the potential of becoming a new track of the Ontology Alignment Evaluation Initiative (OAEI)[3] [5] to push the limits of state-of-the-art systems. The output of the different OAEI participants could be merged into a rich consensus alignment that could become the reference to integrate ECOTOX and NCBI. At the same time, as the alignment between ECOTOX and NCBI is not public nor complete, the consensus mappings could also be seen as a very relevant resource to the ecotoxicology community.

References

1. Myklebust, E.B., Jiménez-Ruiz, E., Chen, J., Wolf, R., Tollefsen, K.E.: Knowledge Graph Embedding for Ecotoxicological Effect Prediction. In: ISWC. (2019)
2. U.S. EPA: ECOTOXicology knowledgebase (ECOTOX) (2019)
3. Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: ECAI. (2012)
4. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady **10** (1966)
5. Algergawy, A., et al.: Results of the Ontology Alignment Evaluation Initiative 2019. In: 14th International Workshop on Ontology Matching. (2019) 46–85

[3] OAEI: `http://oaei.ontologymatching.org/`