Inferring logical definitions using compound ontology matching

Daniela Oliveira* and Catia Pesquita

LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Campo Grande 1749-016, Portugal

ABSTRACT

OBO logical definitions are a means to support the creation of integrated reference ontologies. In ontologies they exist for, logical definitions currently cover a small portion of classes, which limits the potential for integration.

We present a novel preliminary strategy to derive logical definition candidates based on an ontology compound matching algorithm. Preliminary results show that this strategy is able to increase the coverage of logical definitions between 2 and 19%.

1 INTRODUCTION

The Open Biological and Biomedical Ontologies (OBO) Foundry (Smith *et al.*, 2007) is a collaborative initiative for establishing a set of principles for ontology development in the biomedical domain. Its goal is to support the creation of orthogonal interoperable reference ontologies and OBO cross-products were created to provide computable logical definitions for classes.

Several of the current logical definitions present in the OBO Foundry were obtained with the Open Bio-Ontology Language (Obol) (Mungall, 2004). Obol has a fairly complex set of rules to define ontology-specific grammars and generate potential logical definitions, which have to be manually curated. It has been applied in the improvement of phenotype ontologies (Mungall *et al.*, 2010) and in the normalization of GO (Mungall *et al.*, 2011). A more recent approach, cross-products extension (CPE) (Quesada-Martínez *et al.*, 2014) has been applied to the GO.

However, adding and maintaining these definitions requires a significant amount of effort, which likely contributes to their incomplete coverage. For instance, the logical definitions of the three ontologies employed in this paper account for less than half of the classes in the ontology (see Table 1).

Ontology	Classes	Logical Definitions	Proportion
HP	28621	14059	49.1%
MP	28643	7694	26.9%
WBP	2290	957	41.7%

Table 1. Proportion of classes represented by logical definitions.

This paper describes a preliminary strategy to derive logical definitions candidates that is based on a novel algorithm used for the creation of compound alignments. Our algorithm is centered around a ternary compound mapping approach, which we define as a tuple <X, Y, Z, R, M>, where X, Y and Z are classes from three distinct ontologies, R is a relation established between Y and Z to generate a class expression that is mapped to X via a mapping relation M. Here, we consider the ontology to which X belongs to be the source ontology, and the ontologies that define Y and Z to be the target ontology 1 and 2, respectively. In this particular case the relation R is always an intersection and the mapping M an equivalence.

```
[Term]
id: HP:0000337 ! broad forehead
intersection_of: PATO:0000600 ! increased width
intersection_of: inheres_in FMA:63864 ! forehead
```

Fig. 1. Example of a possible ternary compound match in the HP logical definitions.

Due to the nature of the matching algorithm our strategy only finds logical definitions for classes which are composed of constructs from two different ontologies. This is the case of many of the classes in the Human Phenotype Ontology which have definitions that are composed of classes from the PATO and FMA ontologies (see Figure 1). Our goal is to investigate whether our proposed strategy is able to reliably find definitions which were not obtained through previous methodologies, and where thus not included in the available logical definitions.

2 MATERIALS AND METHODS

2.1 Ontologies

For creating and testing our algorithm we matched different combinations (see Table 2) of the following OBO ontologies: Cell Type (CL) (Bard *et al.*, 2005), Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003), Gene Ontology - Biological Process (GO) (Ashburner *et al.*, 2000), Human Phenotype Ontology (HP) (Köhler *et al.*, 2013), Mammalian Phenotype (MP) (Smith *et al.*, 2004), Neuro Behaviour Ontology (NBO) (Gkoutos *et al.*, 2012), Phenotypic quality (PATO) (Mungall *et al.*, 2010), Uber Anatomy Ontology (UBERON) (Haendel *et al.*, 2009) and *Caenorhabditis elegans* phenotype (WBP) (Schindelman *et al.*, 2011).

These ontologies were downloaded from the OBO Foundry (http://obo.sourceforge.net) in February 2015.

2.2 Algorithm

We developed a novel algorithm (Oliveira and Pesquita, 2015) to establish compound mappings integrated in AgreementMakerLight (AML) (Faria *et al.*, 2014) ontology matching system. We compute the confidence of the first step, based on the ratio of words of the first target ontology classes' labels that overlap with the words of the labels of the classes of the source ontology, weighted by their evidence content (i.e., the inverse log of their frequency in the

*To whom correspondence should be addressed: doliveira@lasige.di.fc.ul.pt

ontology's vocabulary). In the second step, we filter out source classes whose matches were below the threshold, and then match the remaining ones based on their unmatched words in step 1, to the second target ontology. To compute the confidence of this second step, if the number of words of a certain label is higher than the number of words of a target 2 ontology label we compare the unmatched words to the each word of the target 2 terms. Else, if the number of words of a certain label is lower than the number of words of a target 2 ontology label we compare the unmatched words to the each word of the source. Finally, the algorithm had a greedy selection step, which selects the mapping with the highest similarity, amongst the source classes with more than one mapping.

2.3 Evaluation

To evaluate our strategy we performed a manual analysis of the results, where we classified mappings into three possible categories: 'Correct', where the mapping is deemed correct and the source class has no mapping in the logical definitions; 'Conflict', where the mapping is potentially correct but the source class has a different mapping in the logical definitions; and 'Incorrect', where the mapping is deemed incorrect. We applied this to all mappings created by using 0.5 as a threshold for step 1 and 0.9 for step 2.

3 RESULTS AND DISCUSSION

The manual evaluations of the mappings (Table 2) reveals a very low proportion of incorrect mappings, and an intermediate proportion of conflicting mappings. Given the low error rate, we consider our strategy to be suitable to the identification of candidate logical definitions. However, we are also interested in ascertaining whether our strategy can contribute with a signficant number of novel definitions. In fact, the novel logical definitions represent a percentual increase between 2 and 19%, which corresponds to more than 800 new logical definitions for the three ontologies (see Table 3). This indicates that our strategy is able to find candidate logical definitions which are missed by the currently employed methods.

	Correct	Conflict	Incorrect
MP-CL-PATO	63.71 %	34.60 %	1.69 %
MP-GO-PATO	92.16 %	6.97 %	0.87 %
MP-NBO-PATO	72.46 %	26.09 %	1.45 %
MP-UBERON-PATO	91.33 %	7.96 %	0.70 %
WBP-GO-PATO	88.55 %	7.49 %	3.96 %
HP-FMA-PATO	77.82 %	15.56 %	6.61 %

Table 2. Manual evaluation of results.

Ontology	New Mappings	Logical Definitions	% of Growth
HP	259	14059	1.84
MP	422	7694	5.48
WBP	182	957	19.02

Table 3. Impact of the new mapping derived logical definitions.

However, for some ontologies, the number of conflicting mappings represents a greater proportion. Upon comparing the novel mapping with the conflicting logical definition we have found that in many cases this is due to similar PATO classes, whose synonyms are hard to distinguish.

4 CONCLUSION

Our proposed strategy was able to successfully identify a significant number of novel logical definitions candidates, with a low error rate. Therefore, this new methodology could help expert curators expand the current logical definitions. Although our current approach is limited to logical definitions established by the intersection of classes from two distinct external ontologies, we expect it can easily be adapted to logical definitions that employ classes from the source ontology and a single external ontology. In the future, we will also explore how different similarity thresholds can affect the accuracy and coverage of the obtained logical definitions.

ACKNOWLEDGEMENTS

The authors are grateful to Daniel Faria for his technical support. This work was supported by FCT through funding of LaSIGE Research Unit, ref.UID/CEC/00408/2013

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. *Genome biology*, **6**(2), R21.

Faria, D., Pesquita, C., Santos, E., Cruz, I. F., and Couto, F. M. (2014). AgreementMakerLight: A scalable automated ontology matching system. *10th International Conference on Data Integration in the Life Sciences 2014 (DILS)*, page 29.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2012). The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. *Int Rev Neurobiol*, **103**, 69–87.

Haendel, M. A., Gkoutos, G. G., Lewis, S. E., and Mungall, C. (2009). Uberon: towards a comprehensive multi-species anatomy ontology.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., *et al.* (2013). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, page gkt1026.

Mungall, C. J. (2004). Obol: integrating language and meaning in bio-ontologies. *Comparative and functional genomics*, **5**(6-7), 509–520.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, **11**(1), R2.

Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. (2011). Cross-product extensions of the Gene Ontology. *Journal of biomedical informatics*, **44**(1), 80–86.

Oliveira, D. and Pesquita, C. (2015). Compound matching of biomedical ontologies. In *International Conference on Biomedical Ontology (ICBO)* (to appear).

Quesada-Martínez, M., Mikroyannidi, E., Fernández-Breis, J. T., and Stevens, R. (2014). Approaching the axiomatic enrichment of the Gene Ontology from a lexical perspective. *Artificial intelligence in medicine*.

Rosse, C. and Mejino, J. L. (2003). A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of biomedical informatics*, **36**(6), 478–500.

Schindelman, G., Fernandes, J. S., Bastiani, C. A., Yook, K., and Sternberg, P. W. (2011). Worm Phenotype Ontology: integrating phenotype data within and beyond the c. elegans community. *BMC bioinformatics*, **12**(1), 32.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**(11), 1251–1255.

Smith, C. L., Goldsmith, C.-A. W., and Eppig, J. T. (2004). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, **6**(1), R7.