

Creating a new Ontology: a Modular Approach

Julia Dmitrieva and Fons J. Verbeek

Universiteit Leiden, Leiden Institute of Advanced Computer Science,
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
{jdmitrie,fverbeek}@liacs.nl
<http://bio-imaging.liacs.nl/>

Keywords: module extraction, ontology mapping, ontology integration

1 Introduction

Ontologies in life sciences, in particular, members of the OBO FOUNDRY [6], contain information about species, proteins, chemicals, genomes, pathways, diseases, etc. Information in these ontologies might overlap, and it is possible that a certain concept is defined in different ontologies from a different point of view and at different level of granularity. Therefore, the combination of information from different ontologies is useful to create a new ontology.

Case Study The integration will be illustrated with a case study on Toll-like receptors. If we want to investigate what kind of information about Toll-like receptors is available in MOLECULE ROLE ONTOLOGY (MoleculeRoleOntology) [6], then we will see that Toll-like receptors are defined as pattern recognition receptors. In the BIOLOGICAL PROCESS ONTOLOGY (GO) [6] the Toll-like receptors are described in the context of signaling pathway and are subsumed by the *pattern recognition receptor signaling pathway*. In the PROTEIN [6] ontology a Toll-like receptor is just a protein. In the NCI-THESAURUS [5] ontology Toll-like receptors are defined as *Cell Surface Receptors*. It follows from foregoing that multiple ontologies model different aspects of the same concept and the combination of the available information provides more knowledge about concepts where an ontology developer is interested in.

We introduce an approach for generating a new ontology in which ontologies from OBO FOUNDRY are reused. First, we extract modules from these ontologies, on the basis of the well defined modularity approach [2]. As a signature for the modules we are using the symbols that match the terms of interest as indicated by the user. In our case study we create an ontology about Toll-like receptors, therefore we use two *seed* terms (Toll, TLR). Subsequently, we create mappings between concepts in the modules. It has already been shown [1] that the simple similarity algorithms outperform structural similarity algorithms in biomedical ontologies. To this end, we have based our mappings on the similarity distance [4] between labels and synonyms of classes in the modules. Finally, a new ontology is created where the mappings are represented by means of OWL:EQUIVALENTCLASS axiom and small concise modules are imported.

2 Modules from Enriched Signature

In our case study we have used the following biomedical ontologies obtained from OBO FOUNDRY: National Cancer Institute Ontology (NCI_THESAURUS), GO Ontology (GO), Protein Ontology (PRO), Dendritic Cell ontology (DENDRITIC_CELL), Pathway ontology (PATHWAY), Molecule Role Ontology (MOLECULEROLEONTOLOGY), Gene Regulation Ontology (GENE_REGULATION), and finally, Medical Subject Heading ontology (MESH). All of these ontologies are in OBO format, except for the NCI_THESAURUS, which is in OWL format.

A module comprises knowledge of a part of the domain that is dedicated to a set of terms of user interest (*seed terms*). Let $T_1 = \{Toll, TLR\}$ be this set. Let S_1 be a set of terms (signature) from the ontology O_1 that represents the classes whose labels, descriptions, ID, or other annotation properties contain the symbols from T_1 . The first module that we have extracted is the module from NCI_THESAURUS M_1 . This is chosen because it is the largest ontology containing the most matches. In order to generate a signature for the next ontology O_2 , we are using not only the terms from T_1 but we enrich this set with the terms from the module M_1 . The same procedure is applied for the rest of the ontologies, namely module M_i is extracted on the basis of the terms $T_i = Sig(M_{i-1}) \cup T_{i-1}$. This method has two drawbacks. First, it depends on the order of ontologies. Second, with the generation of the new module M_i new symbols can be introduced that will match symbols from ontologies used in previous steps. These problems can be solved with the generation of a fixpoint.

Fixpoint Modules We have investigated whether or not we will find a fixpoint with our module extraction method. The fixpoint is reached at the moment the set of terms T which is used in order to generate modules during step t_i does not change any more after another run with all ontologies. This can be written as $\cup_{k=1}^n Sig(M_{k,i}) = \cup_{k=1}^n Sig(M_{k,i+1})$, where $M_{k,i}$ is the module k created during step t_i . It can be formulated in a "fixpoint-like" way $Match(T) = T$.

The fixpoint was reached with the following sizes of the modules, see Table 1.

3 Ontology Mapping

In this paper we use a more loosely definition of the concept *mapping* compared with the definition given in [3] in which mapping is a morphism. In our approach *mapping* is a partial function that maps from subset $S_1 \subseteq Sig(O_1)$ to subset $S_2 \subseteq Sig(O_2)$. We deliberately reject the morphism requirement, thus, the structural dependencies will not be preserved after mapping, because we are interested in consequents of this mapping to the original ontologies, namely, whether and how the structural dependences will be broken.

For our experimental prototype system we use our own mappings based on the syntactic similarity. It has been already shown [1] that in the case of biomedical ontologies the simple mappings methods are sufficient and outperform more complex methods.

Table 1: The size of the modules after reaching the fixpoint

module	size in KB
Toll_from_gene_regulation	88.7
Toll_from_protein	23.4
Toll_from_chebi	218.6
Toll_from_mesh	59.2
Toll_from_dendritic_cell	4.2
Toll_from_pathway	4.1
Toll_from_cellular_component	35.4
Toll_from_molecular_function	11.4
Toll_from_MoleculeRoleOntology	46.9
Toll_from_biological_process	221.1
Toll_from_Thesaurus	802.1

We compare characteristics (id, label, description) for all classes from ontology O_1 with the same characteristics for all classes from ontology O_2 . The comparison is based on the Levenshtein distance algorithm [4]. We have adapted the Levenshtein distance and introduce a metric *Lev* (in the range $[0 \dots 1]$). Two classes C_i and C_j are considered to be similar if they have the maximum value for *Lev* metric and if this value is also higher than the threshold $t = 0.95$ that was experimentally determined.

4 Integration Information from Ontologies

The final step of the ontology creation is the integration of the modules into one ontology. If there a mapping exists between two classes C_i and C_j from the modules M_i and M_j respectively we add the equivalence relation

OWL:EQUIVALENTCLASS between these classes in the new ontology. Besides the equivalence relationships the new ontology contains the OWL:IMPORTS axioms, where all the created modules are imported.

So far, this all seems rather straightforward. However, the problem with this integrated ontology $O_{1\dots n}$ is that it contains many unsatisfiable classes. In order to understand the reason of this unsatisfiability we have applied different experiments. First, we have merged all pairs of the modules, namely $\forall_{i \neq j} O_{i,j} \equiv M_i \cup M_j$. For each merged ontology $O_{i,j}$ we have checked for unsatisfiable classes. Already at this stage of integration different merged pairs contain unsatisfiable classes. We have used the Pellet [7] reasoner in order to reveal the explanations of unsatisfiability. After we have repaired unsatisfiable classes in the merged pairs of ontologies $O_{i,j}$ we have had to check satisfiability of the integrated ontology $O_{1\dots n}$. There were still 46 unsatisfiable classes. The unsatisfiabilities in the integrated ontology have also been solved by means of Pellet reasoner explanations.

5 Conclusion

We have described a method to generate a new ontology on the basis of the bio-ontologies most of which are available in OBO FOUNDRY. We have shown how to create modules on the basis of the terms of interest. The signature for the module extraction is enriched by the symbols from other modules with the fixpoint as a stop criterion. We have integrated modules on the basis of mappings created using Levenshtein distance similarity.

We have investigated how to solve unsatisfiable classes which appear after the integration of the modules. Although the number of unsatisfiable classes was high, it was possible to solve unsatisfiabilities with the help of explanations provided by the Pellet reasoner.

In this study we have shown that the modularity and simple mappings provide a good foundation for the creation of a new ontology in an pseudo-automated way. This method can be used when an ontology engineer does not want to create a new ontology from scratch, but rather wants to reuse knowledge already presented in other ontologies. Moreover, this is the strategy that should be preferred and has to be applied more often as ontologies gain importance in life sciences.

References

1. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: Simple methods work. In: AMIA 2009 Symposium Proceedings (2009)
2. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: Extracting modules from ontologies: A logic-based approach. In: Modular Ontologies, pp. 159–186 (2009)
3. Kalfoglou, Y., Schorlemmer, W.M.: Ontology mapping: The state of the art. In: Semantic Interoperability and Integration (2005)
4. Levenshtein, V.: Binary codes capable of correcting, deletions, insertions, and reversals. Soviet Physics-Doklady 10(8), 845–848 (August 1965)
5. NCI: Terminology resources: Nci enterprise vocabulary services (evs), dictionaries, fedmed, fda, cdisc, and ncpdp terminology. <http://www.cancer.gov/cancertopics/terminologyresources>
6. OBO: The open biomedical ontologies. <http://www.obofoundry.org/>
7. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. J. Web Sem. 5(2), 51–53 (2007), <http://www.informatik.uni-trier.de/~ley/db/journals/ws/ws5.html#SirinPGKK07>