Extending Logic Programming for Life Sciences Applications

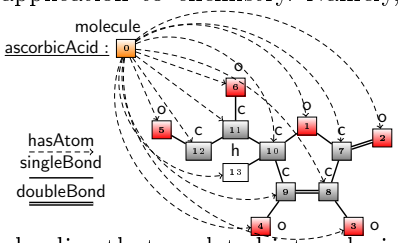
Despoina Magka

Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford, OX1 3QD, UK

The volume of bioinformatics data produced by research laboratories worldwide is increasing at an astonishing rate turning the need to adequately catalogue, represent and index the vast amounts of produced data sources into a pressing challenge. Semantic Web technologies have achieved significant progress towards the federation of biochemical information via the definition and use of domain vocabularies with formal semantics, also known as *ontologies*. OWL [4], a family of logic-based knowledge representation formalisms, has played a pivotal role in the advent of Semantic Web due to its significant ability to reason over ontologies by means of logical inference. A core reasoning task in biochemical domains is classification; with the help of performant OWL reasoners life scientists can employ OWL to drive fast, automatic and repeatable classification processes. As a consequence, OWL bio- and chemo-ontologies are widely used for the modelling of life sciences knowledge.

A prominent Semantic Web ontology is ChEBI, an open-access dictionary that provides taxonomical information for molecular entities with numerous applications such as drug discovery and study of disease pathways. Inspite of ChEBI being available in OWL format, classifying ChEBI with an OWL reasoner produces only few of the desired subsumptions because OWL is not expressive enough to capture the structure and the properties of chemical entities. For instance, the tree-model property of OWL prevents one from faithfully representing *cyclic* molecules: OWL axioms can state that butane molecules have four carbon atoms, but they cannot state that the four atoms in a cyclobutane molecule are arranged in a *ring*. Additionally, due to the first-order logic semantics of OWL it is difficult to describe classes based on the *absence* of certain characteristics: if an inorganic molecule is defined as 'a molecule *not* containing a carbon atom', then in order to derive that water is inorganic one needs to also specify that water contains at most three atoms and that both hydrogens and oxygens are disjoint from carbons. As a consequence of these inadequacies, ChEBI is manually curated by human experts who determine the chemical classes of new molecular entries. Currently, ChEBI describes nearly 30,000 entities and grows at a rate of 3,500 entities per year. Given the size of other publicly available chemical databases with millions of molecules, there is clearly a strong potential for ChEBI to expand by speeding up the curating tasks through automation of chemical classification.

Towards that direction, we developed a theoretical and practical framework for the representation of graph-shaped objects, with a particular application to chemistry. Namely, we introduced Description Graph Logic Programs (DGLP) [6] a logic-based language that addresses the previously outlined expressivity limitations of OWL and admits a sound, complete and terminating reasoning procedure for the automatic classification of structured objects. The main modelling artefact of our formalism is the *description graph*, a directed labeled graph that abstracts out the structure of objects and can be directly translated into a logic program. E.g., the description graph of the ascorbic acid molecule is depicted above.



From a technical point of view, DGLP is a language that is based on logic programming rules with function symbols in the head and adopts the stable model semantics. Thus, DGLP is directly related to extensions of datalog with existential rules, a formalism that has been extensively studied in various areas such as theory of databases, knowledge representation and answer set programming [1]. Existential rules may incur non-termination of the reasoning procedure due to the creation of fresh terms that can be infinitely many; so, the main focus of research is to guarantee termination by formulating suitable restrictions, the so-called *acyclicity* conditions. In order to ensure decidability of our formalism we proposed a number of novel acyclicity criteria, explored their computational properties and proved that they are strictly more general than previously suggested analogous conditions; reasoning over acyclic DGLPs was shown to be 2EXPTIME-complete [2].

In terms of applicability, we implemented LoPStER, a prototype that performs logic-based chemical classification and draws upon DLV [5], a state of the art logic programming reasoner. In order to assess the feasibility of our approach, we empirically evaluated our implementation using data extracted from the ChEBI ontology. Our software classified 500 molecules under 51 chemical classes in 40 secs, which exhibits a significant improvement in comparison with our previous (450 seconds to classify 70 molecules [6]) and related (Hastings et al. report a total of 4 hours to compute the superclasses of 140 molecules [3]) work.

Furthermore, while conducting the experiments we discovered a number of missing and inconsistent axioms from the manually curated ChEBI ontology. As one can infer from the molecular structure of ascorbic acid, ascorbic acid is a carboxylic ester (i.e. a molecule containing (C=O)O) as well as a polyatomic cyclic entity. Inspite of the fact that these superclasses were exposed by our classification methodology, we were not able to identify them in the ChEBI hierarchy. Moreover, ascorbic acid is asserted as a carboxylic acid (i.e. a molecule with a carboxy group, which has formula C(=O)OH) which is not the case as it can be deduced by the lack of a carboxy group in its molecular graph. We interpret the revealing of these discrepancies as an indication of the practical relevance of our contribution.

Concerning future benefits, our prototype could form the basis of an application to assist biocurators towards a more rapid development of the ChEBI ontology. From a modelling point of view, our approach could stimulate the adoption of a different and expressive reasoning paradigm based on logic programming for which highly optimised reasoners are available. For the future, we plan to design a surface syntax that will enable life scientists to represent knowledge without the need to script intricate logic programs; from a theoretical point of view, it would be interesting to investigate extensions of DGLP with numerical values that would allow for more expressive modelling such as molecular weights.

References

1. Calì, A., Gottlob, G., Lukasiewicz, T., Marnette, B., and Pieris, A. Datalog+/-: A family of logical knowledge representation and query languages for new applications. In *LICS'10*.
2. Cuenca Grau, B., Horrocks, I., Krötzsch, M., Kupke, C., Magka, D., Motik, B., and Wang, Z. Acyclicity Conditions and their Application to Query Answering in Description Logics. In *KR 2012* (2012), AAAI Press.
3. Hastings, J., Dumontier, M., Hull, D., Horridge, M., Steinbeck, C., Stevens, R., Sattler, U., Hörne, T., and Britz, K. Representing Chemicals Using OWL, Description Graphs and Rules. In *OWLED* (2010), vol. 614.
4. Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. From SHIQ and RDF to OWL: the making of a Web Ontology Language. *J. Web Sem. 1*, 1 (2003), 7–26.
5. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., and Scarcello, F. The DLV system for knowledge representation and reasoning. *ACM TOCL 7*, 3 (2006).
6. Magka, D., Motik, B., and Horrocks, I. Modelling Structured Domains Using Description Graphs and Logic Programming. In *ESWC* (2012), Springer, pp. 330–344.