Representing Chemicals using OWL, Description Graphs and Rules

Janna Hastings[1,2,3⋆], Michel Dumontier[4], Duncan Hull[1], Matthew Horridge[5], Christoph Steinbeck[1], Ulrike Sattler[5], Robert Stevens[5], Tertia Hörne[2], and Katarina Britz[2,3]

[1] European Bioinformatics Institute, UK
[2] University of South Africa
[3] Meraka Institute, South Africa
[4] Carleton University, Canada
[5] University of Manchester, UK

Abstract. Objects can be said to be structured when their representation also contains their parts. While OWL in general can describe structured objects, description graphs are a recent, decidable extension to OWL which support the description of classes of structured objects whose parts are related in complex ways. Classes of chemical entities such as molecules, ions and groups (parts of molecules) are often characterised by the way in which the constituent atoms of their instances are connected via chemical bonds. For chemoinformatics tools and applications, this internal structure is represented using chemical graphs. We here present a chemical knowledge base based on the standard chemical graph model using description graphs, OWL and rules. We include in our ontology chemical classes, groups, and molecules, together with their structures encoded as description graphs. We show how role-safe rules can be used to determine parthood between groups and molecules based on the graph structures and to determine basic chemical properties. Finally, we investigate the scalability of the technology used through the development of an automatic utility to convert standard chemical graphs into description graphs, and converting a large number of diverse graphs obtained from a publicly available chemical database.

Key words: chemistry, ontology, description graphs, rules

1 Introduction

Objects can be said to be structured when their representation also contains their parts. While OWL in general can describe structured objects, description graphs are a recent, decidable extension to OWL which support the description of classes of structured objects whose parts are related in complex ways [1–3].

Classes of chemical entities such as molecules, ions and groups (parts of molecules) are often characterised by the way in which the constituent atoms of

⋆ To whom correspondence should be addressed, hastings@ebi.ac.uk.

their instances are connected via chemical bonds. For example, a cyclic hydrocarbon such as benzene is characterised as six carbon atoms, each of which is connected to two other carbon atoms in such a way that it forms a single cycle (or ring). For various cheminformatics applications, chemical structures are represented as chemical graphs, comprising of atoms as vertices and bonds as edges. These can be encoded as connection tables [4].

A classic chemoinformatic application is chemical classification by comparing all substructures such general descriptions are subsumed by more complex and refined substructures. The Web Ontology Language (OWL), as it currently stands, is incapable of representing the required complex structures, particularly cycles [7]. The chemical graph formalism has previously been reported as a candidate application for substructure classification using description graphs [5, 6].

In this paper, we present a method for transforming chemical graphs into description graphs, and apply this method to create an OWL knowledge base of chemical entities enhanced with the structures of the chemical entities as description graphs. We will consider to what extent the formalism of description graphs, together with rules for expressing conditionality, supports the type of reasoning which domain experts would expect from a structure-enhanced chemical knowledge base, such as classification based on chemical structures, and determination of chemical properties based on the structures. Finally, we assess the scalability of the technology by evaluating the times taken to reason over knowledge bases of varying sizes.

2 Background

2.1 OWL 2, Description Graphs, and Rules

OWL 2 [7] is the latest release of the Web Ontology Language (OWL) family of languages. While OWL provides an extensive collection of constructs for logic-based ontology development, decidability of reasoning problems—e.g., testing consistency of an ontology, satisfiability of classes or computing its inferred class hierarchy—is obtained by making sure that OWL has a *tree model property* [8]: in a nutshell, that means that every consistent ontology has a model, i.e., a state of affairs that satisfies all axioms in the ontology, whose relational structure looks like a tree. For this reason, OWL has not traditionally been able to describe arbitrarily structured objects, but only those which had structures which could be expressed in the shape of trees. *Description Graphs* are a formalism which has been introduced by Motik et al. [1–3] to address this weakness of OWL in representing structured objects, while still preserving the decidability of reasoning on ontologies containing such structured objects.

A description graph is a directed graph $G = (V, E, \lambda)$ in which each vertex $i \in V$ is labeled with a set of (possibly negated) class names $\lambda\langle i\rangle$; and each edge $\langle i, j\rangle \in E$ is labeled with a set of atomic properties $\lambda\langle i, j\rangle$. Each description graph has a *main class*, which indicates the object whose structure is being

modelled in the graph, and it is this main class that will be used to link to the remainder of the ontology that the description graph is a part of.

In order to preserve the decidability of reasoning, some important constraints must be observed within a graph-enhanced knowledge base [1]. For our purposes, the most significant of these is that the properties which are used in the description graphs (i.e. the graph edges) must not be referred to in the main ontology axioms, which is known as the *strong separation* requirement. The full set of properties in the knowledge base has thus to be separated into *tree properties* and *graph properties*. This provides a limitation in terms of the possibility for reasoning over the information encoded in the graphs, as the graph properties cannot be referred to in OWL axioms, an example of which might be `SubClassOf`(*has_atom* `only` (`CarbonAtom` or `HydrogenAtom`)) `HydrocarbonMol`

Thus chemical classificiation must be expressed with rules. Further, these rules must be role-safe, that is, they must not refer simultaneously to properties used in the graphs and those used in the OWL ontology axioms.

A graph-extended OWL knowledge base is thus a 4-tuple $K = (T, G, P, A)$ where T is a set of OWL class axioms, G is a set of description graphs, P is a set of rules, and A is a set of OWL assertions. T is allowed to refer only to tree properties, G and P are allowed to refer only to the graph properties, and A is allowed to refer to both graph and tree properties [1–3].

2.2 Chemical entities and graphs

At the molecular level, all of matter is composed of *atoms* of different kinds (such as Carbon and Oxygen) joined together through chemical bonds of different strengths. Covalent bonds (the strongest kind of chemical bond) join atoms together into composite units called *molecules*. Chemical entities are usually categorised into chemical classes by virtue of sharing common substructure or activity. An example of a chemical class is 'carboxylic acids', which groups together all molecules that share the important carboxy functional group and therefore hold the disposition to behave similarly in certain chemical reactions involving that group.

The structure of a molecule is nicely represented by a chemical graph, which describes the atomic connectivity within a molecule in terms of labelled nodes for the atoms or groups within the molecule, and labelled edges for the (usually covalent) bonds between the atoms or groups [4]. The chemical graph formalism is widely used in the field of cheminformatics to calculate many properties of chemical entities. Chemical graphs are encoded in a variety of standard formats, prominent among which is the MOLFile connection table-based format [9].

3 Methods

The purpose of our experiment is to evaluate the utility and scalability of description graphs and rules for the representation of, and reasoning over, chemi-

cal structures. The knowledge base[6] consists of i) a simple ontology describing classes pertaining to chemical entities, ii) auto-generated description graphs from structures in the ChEBI database, and iii) rules for structure-based classification. We used the HermiT [11] reasoner[7] for reasoning about the ontology, description graphs and rules [5]. ChEBI [12] was used as a source for chemical structures, which were parsed using the Chemical Development Kit [13].

Our evaluation criteria considers the following three aspects expected by domain experts:

- Can chemical entities be classified based on their substructures?
- Can basic chemical properties be determined from the description graphs?
- How scalable is the resulting knowledge base?

We now describe the structure of the implemented knowledge base.

3.1 Ontology

At the root of the ontology is the node 'chemical entity', beneath which are nodes for the primary division in kind of entity, namely 'group', 'atom', 'molecule', and 'ion'. 'Atom' is further divided into the concrete types of atoms as per the periodic table, such as 'carbon atom' and 'oxygen atom'.

An illustration of the overall structure of the core terms of the ontology is shown in Figure 1.

3.2 Description Graphs

Description graphs were automatically generated[8] from a MOLfile connection table format [9]. The standard MOLfile format consists of an atom table, which provides information about the atoms included in the molecule such as their types, and a bond table, which provides information about the bonds included in the molecule such as which atoms they connect and their order (single, double, etc.).

Each description graph consists of a vertex for the description graph *main class* which is a subclass of 'molecule' in the ontology, a vertex for each atom which is a subclass of the atom type e.g. 'carbon atom' in the ontology, and a vertex for each bond which is a subclass of the bond type in the ontology e.g. 'single'. Each atom vertex is connected to the molecule by the *has_atom* property. Atom vertices are associated with bonds with *has_bond*. Figure 2 shows an illustration of the description graph for cyclobutane.

[6] Ontology availabe in two files, the main ontology at http://www.ebi.ac.uk/~hastings/owled2010/chemistry_dgs_ontology.owl and the graphs at http://www.ebi.ac.uk/~hastings/owled2010/chemistry_dgs_graphs.owl

[7] Version 1.2.2 with slight customisation for input and output of graphs which is currently only partially supported by the HermiT library and the OWL API.

[8] Our software for this experiment is available in source and binary at http://www.ebi.ac.uk/~hastings/owled2010/descgraphs.zip

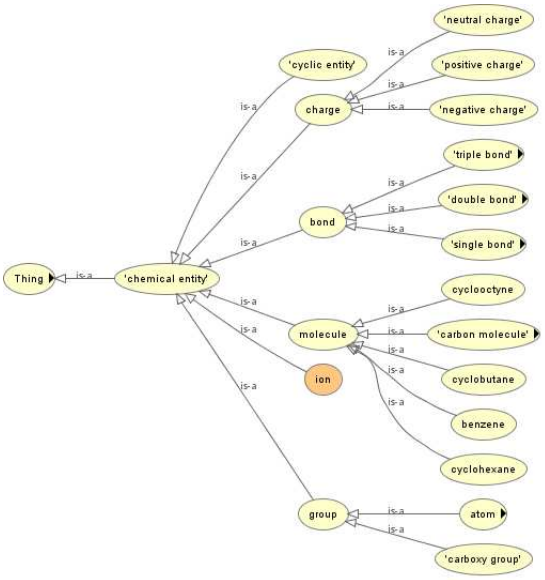

Fig. 1. Core ontology structure

The vertices (but not the properties) of the description graphs are also classified in the main OWL ontology. The main class is classified beneath 'molecule' in the main ontology, the atoms beneath 'atom' and the bonds beneath 'bond'.

3.3 Rules

We implemented rules to classify chemical structures based on their composition and their connectivity. Rules were devised for the classification of *cyclic* compounds, which contained a cycle of connected atoms.

For example, a rule to determine cycles of length three atoms is (slightly simplified for readability, the full generated version also includes `DifferentFrom` statements to ensure non-trivial cycles)

$$\begin{aligned}
&\texttt{Molecule}(M) \wedge \\
\texttt{has_atom}(M, A_1) \wedge \texttt{has_atom}(M, A_2) \wedge \texttt{has_atom}(M, A_3) \wedge &\\
\texttt{Atom}(M, A_1) \wedge \texttt{Atom}(M, A_2) \wedge \texttt{Atom}(M, A_3) \wedge &\\
\texttt{Bond}(M, B_1) \wedge \texttt{Bond}(M, B_2) \wedge \texttt{Bond}(M, B_3) \wedge \wedge &\\
\texttt{has_bond}(A_1, B_1) \wedge \texttt{has_bond}(A_2, B_1) \wedge &\\
\texttt{has_bond}(A_2, B_2) \wedge \texttt{has_bond}(A_3, B_2) \wedge &\\
\texttt{has_bond}(A_1, B_3) \wedge \texttt{has_bond}(A_3, B_3) \wedge &\\
\rightarrow \texttt{instanceOf}(\texttt{M}, \texttt{CyclicMolecule}) \qquad (1)&
\end{aligned}$$

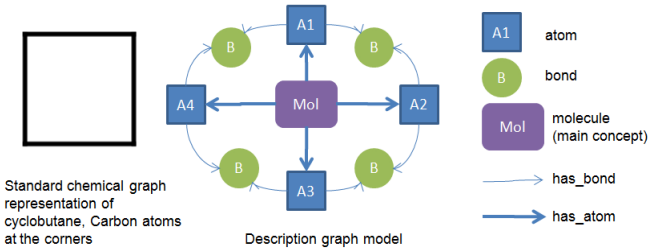

Fig. 2. Illustration of the cyclobutane description graph

Rules were also devised to determine *parthood* between chemical structures. If all atoms of A are atoms of B, and all bonds of A are bonds of B, then A is a *subgraph* of B. The term *group* is commonly used to denote arbitrary chemical parts, while the terms *molecule*, *ion* and so on refer to entire (complete) structures. Rules were devised for each group so as to identify these groups in the molecule. However, a consequence of the strong separation requirement is that a single rule cannot refer to both graph properties and tree properties. For this reason, even if we determine that a given graph is a subgraph of another graph, we cannot assert a relationship such as *has_part* between the two main classes at the ontology level. A workaround for this is to create a class for every group, such that if the group's structure is a subgraph of the molecule's structure, then the molecule can be classified as belonging to that class. Rules for parthood determination are of the form

$$\begin{aligned} &\texttt{Molecule}(M) \wedge \\ &\texttt{has\_atom}(M, A_1) \wedge ... \wedge \texttt{has\_atom}(M, A_n) \wedge \\ &\texttt{Atom}(A_1) \wedge ... \wedge \texttt{Atom}(A_n) \wedge \\ &\texttt{Bond}(B_1) \wedge ... \wedge \texttt{Bond}(B_m) \wedge \\ &\texttt{has\_bond}(A_{i1}, B_{j1}) \wedge ... \wedge \texttt{has\_bond}(A_{in}, B_{jm}) \\ &\rightarrow \texttt{instanceOf}(\texttt{M}, \texttt{Class}) \end{aligned} \tag{2}$$

where `M` is an arbitrary individual of type *molecule*; A_1 – A_n are *group* atoms; bonds exist between the group atoms A_{i1} – A_{in} and A_{j1} – A_{jn}, and `Class` is a class the identity of which depends on the group used to generated the rule, for example 'carboxylic acid' for the 'carboxy group'.

The properties used as the graph edges (has_atom, has_bond) are available for use in the rules, as long as a rule does not mix graph properties with properties used in OWL axioms in the main ontology.

In the next section, we present the results of reasoning over the knowledge base.

4 Results

Reasoning with the rules over the combined knowledge base resulted in classification of description graph-enriched classes as cyclic molecules and as classes containing specific defined groups such as carboxylic acids[9]. We find this result positive in terms of overcoming the previously explicit limitation of OWL knowledge bases in expressing arbitrarily structured objects at the class level and performing classification based on the structure.

However, we acknowledge that the types of conditionality that can be expressed in rules potentially provide the facility for only a limited set of chemical properties relative to those required by chemists. For example, it is difficult to express rules that must apply to *all* atoms from a given molecule's graph without specifically naming those atoms, since there is no `forAll` operator in SWRL.

To evaluate the performance, we executed reasoning over iteratively increasing sizes of the knowledge base, both with and without rules. The results are summarised in Figure 3[10]. We do not attempt to control the size of the graphs which we randomly selected for inclusion into our knowledge base, but note that the average size of a molecule in the ChEBI database is around 30 atoms[11].

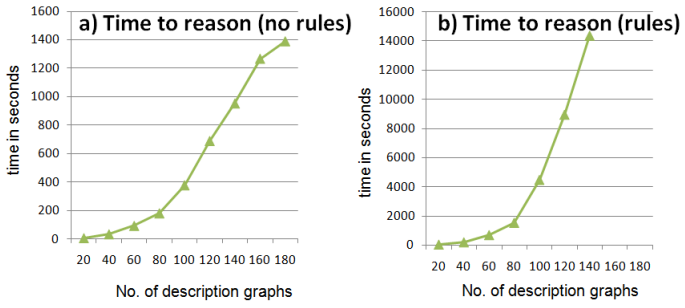

Fig. 3. Performance results of classification

The scalability of the reasoner against the knowledge base enriched with desciption graphs appears workable, with reasoning time growing to a maximum of 23 minutes (1388 seconds) for a knowledge base enriched with 180 graphs. However, including the graphs alone – without rules – does not allow for any classification based on the information encoded in the graphs. The rules are this essential to expose the structure in the graphs to the reasoner. Unfortunately, we find that reasoning over the knowledge base enriched with graphs and rules

[9] The resulting inferred ontology is available at http://www.ebi.ac.uk/~hastings/owled2010/chemistry_dgs_inferred.owl

[10] Tested on a Dell twin core laptop.

[11] Excluding hydrogen atoms, which are commonly implicit, as these can be 'added back' by calculations to determine their predicted positions.

appears to grow very rapidly into unmanageable durations, with the highest duration that we recorded for a knowledge base enriched with 140 graphs taking four hours (14380 seconds) to classify. This scalability is affected dramatically by the number and complexity of the generated rules, therefore this would appear to be a limiting factor in following our approach in a more complete fashion, where many chemical properties and subgraph relations might be reasonably expected to be included in the same knowledge base.

5 Discussion

Our results have shown that it is possible to create a chemical knowledge base using OWL, description graphs and rules. The main strength of our approach is the direct encoding of complex structures at the class level in the ontology, and the encoding of rules for determining properties such as being cyclic, which are not able to be expressed as OWL axioms. We thereby show that this approach allows properties to be calculated by the reasoner rather than requiring these to be pre-computed and added to the asserted hierarchy of the ontology.

The main weaknesses are the limitations of rules for arbitrary property encoding and in particular the lack of quantification operators; and that there seems to be a scalability performance problem with using rules in this fashion. Pragmatically, the performance of the system was not where it would need to be to handle thousands or even millions of chemical graphs as are included in public databases. However, if ontologies are restricted to particular sub-domain areas of limited size, this might not be too much of a limitation.

Other approaches for partially including chemical structural information in knowledge bases have been described in recent years. Armengol and Plaza (2005) [14] describe an ontology-like, formal encoding of chemical structural features using *feature terms*. Key to their approach is the representation of the main structural unit of a chemical entity and then the explicit representation of the additions and modifications to that structural unit. However, their knowledge base is not straightforwardly translatable into OWL and therefore it is not clear to what extent a comparison can be drawn in terms of conclusions that can be drawn with a reasoner.

The ChEBI ontology is a well known ontology for chemical entities, providing a deep classification according to the physical composition and chemical structure of chemical entities. While containing an ever-growing number of chemical entities, ChEBI is maintained entirely by hand, with no automated link between the structure of the chemicals captured in the chemical database and the structural definition of ontology classes. As a result, the ChEBI database has been able to grow at a much faster rate than the ChEBI ontology, with the sizes currently[12] at around 550000 for the database and 22000 for the ontology. Chemical structures are exported into the ontology as *annotations* in the InChI [15] format.

[12] As of Release 69.

In 2007, Dumontier and Villaneuva-Rosales developed an OWL ontology for the classification of chemical compounds based on the presence of specified chemical functional groups [16]. A key aspect of the approach was that tree-like expressions specified the necessary and sufficient conditions for functional groups such that the taxonomy of functional groups would be discovered on reasoning (thus reducing the burden of curating such an ontology). However, they were unable to express *arbitrary* structure at the class level, and they therefore used SWRL rules to classify instances having more sophisticated structures such as cycles.

Taking the desiderata of chemical ontology as the ability to center the knowledge base around an accurate representation of the structures of chemical entities, and to automatically determine the properties of chemical entities from those structures within the knowledge base, we find that the description graphs and rules extensions to OWL are a big step forward on the standard OWL language for this purpose.

6 Conclusion

Our approach uses OWL, description graphs, and rules to implement a structure-enriched knowledge base for chemicals with classification based on the chemical structures and rules. We see this work as a contribution to the evaluation of new OWL-related technology towards the requirements of the chemistry application domain.

Cheminformatics tools and techniques do already exist to detect chemical properties and subgraphs / graph isomorphisms, and the CDK [13] provides a well-developed open source library of such algorithms. These well-developed and optimised graph manipulation algorithms already in widespread use in the field of cheminformatics could provide input into the relatively new development of graph-enriched ontologies.

Next steps will be to investigate whether different representation strategies and/or rule implementations could alleviate the performance overhead in reasoning with the rules; to implement a system to allow visualisation of the chemical description graphs being created; to extend the rules to determine several additional chemical properties; and to investigate the incorporation of a 'chemical datatype' into OWL based on InChI strings.

Acknowledgements

We acknowledge the detailed comments from three anonymous reviewers whose input helped to significantly improve the final result. We further wish to acknowledge invaluable discussions and suggestions from Kirill Degtyarenko, Stefan Schulz, Colin Batchelor, and Birte Glimm. This work has been partially supported by the BBSRC, grant agreement number BB/G022747/1 within the "Bioinformatics and biological resources" fund.

References

1. Motik, B., Cuenca Grau, B., and Sattler, U. (2008) Structured Objects in OWL: Representation and Reasoning. In Proc. of the 17th International World Wide Web Conference (WWW 2008), Beijing, China, 21-25 April 2008. ACM.
2. Motik, B., Cuenca Grau, B., Horrocks, I. and Sattler, U. (2008) Representing Structured Objects using Description Graphs. In Proc. of the 11th Int. Joint Conf. on Principles of Knowledge Representation and Reasoning (KR 2008), AAAI Press, 2008.
3. Motik, B., Cuenca Grau, B., Horrocks, I. and Sattler, U. (2008) Modeling Ontologies using OWL, Description Graphs, and Rules. In Proc. of the 5th OWLED Workshop on OWL: Experiences and Directions, Karlsruhe, Germany, October 26-27, 2008.
4. N. Trinajstic. (1992) Chemical Graph Theory. CRC Press, Florida, USA.
5. Glimm, B., Horridge, M., Parsia, B., Patel-Schneider, P.F. (2009). A Syntax for Rules in OWL 2. In Proc. of OWL Experiences and Directions 2009 (OWLED 2009).
6. Konyk, M., De Leon, A., Dumontier, M. (2008) Chemical Knowledge for the Semantic Web. 2008. Proceedings of Data Integration in the Life Sciences (DILS2008), Lecture Notes in Computer Science. LNBI 5109:169-176, Evry, France.
7. Grau, B. and Horrocks, I. and Motik, B. and Parsia, B. and Patel-Schneider, P. and Sattler, U (2008) OWL 2: The next step for OWL. In Journal of Web Semantics, 4:6 309–322.
8. Vardi, M. Y. (1996) Why Is Modal Logic So Robustly Decidable? In Proc. DIMACS Workshop, volume 31, pages 149184, 1996.
9. http://www.mdl.com/company/about/history.jsp, last accessed April 2010.
10. Horridge, M. and Bechhofer, S. (2009). The OWL API: A Java API for Working with OWL 2 Ontologies. In Proc. of OWL Experiences and Directions 2009 (OWLED 2009), R. Hoekstra and P. F. Patel-Schneider, eds.
11. Shearer, R., Motik, B. and Horrocks, I. (2008) HermiT: A Highly-Efficient OWL Reasoner. In Dolbear, C., Ruttenberg, A. and Sattler, U. (Eds.), Proceedings of the 5th Workshop on OWL: Experiences and Directions, Karlsruhe, Germany, October 2627, 2008.
12. de Matos, P.; Alcntara, R.; Dekker, A.; Ennis, M.; Hastings, J.; Haug, K.; Spiteri, I.; Turner, S.; and Steinbeck, C. (2010) Chemical Entities of Biological Interest: an update. Nucl. Acids Res. 2010 38: D249–D254.
13. Steinbeck C., Hoppe C., Kuhn S., Floris M., Guha R., Willighagen E.L. (2006) Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. Curr. Pharm. Des. 2006; 12(17):2111-2120.
14. Armengol, A. and Plaza, E. (2005) An ontological approach to represent molecular structure information. In J.L. Oliviera et al. (Eds.): ISMBA 2005, LNBI 3745, pp. 294-304, 2005.
15. http://www.iupac.org/inchi/, last accessed April 2010.
16. Villanueva-Rosales, N. and Dumontier, M. (2007) Describing chemical functional groups in OWL-DL for the classification of chemical compounds. OWL: Experiences and Directions (OWLED 2007).