

# Chemical Knowledge for the Semantic Web

Mykola Konyk<sup>1</sup>, Alexander De Leon<sup>1</sup>, and Michel Dumontier<sup>1,2,3</sup>

<sup>1</sup> School of Computer Science

<sup>2</sup> Department of Biology

<sup>3</sup> Institute of Biochemistry,

Carleton University, 1125 Colonel By Drive,

K1S 5B6, Ottawa, Canada

mkonyk@gmail.com, alexjdl@gmail.com,

michel\_dumontier@carleton.ca

**Abstract.** With over 80 file formats to represent various chemical attributes, the conversion between one format and another is invariably lossy due to informal specifications. In contrast, the use of a formal knowledge representation language such as the Web Ontology Language (OWL) enables precise molecular descriptions that can be reasoned about in a logically valid manner. In this paper, we describe a chemical knowledge representation using OWL. We demonstrate its utility in querying a new drug repository created from PubChem, DrugBank and DBpedia. By leveraging Semantic Web technologies, it becomes possible to integrate chemical information at differing levels of detail and granularity, opening new avenues for life science knowledge discovery.

**Keywords:** semantic web, knowledge representation, knowledge engineering, ontology, life sciences, question answering, OWL, chemistry, molecule, mashup.

## 1 Introduction

While powerful web search engines can sift through enormous amounts of biochemical information online, it is still difficult to find compounds having a set of desirable attributes i.e. can form specific derivatives, or are stable at room temperature and have a non-toxic metabolic profile. Although over 80 file formats exist to represent chemical data, none, including the Chemical Markup Language (CML) [1], are capable of encoding arbitrarily knowledge in such a way that the meaning is wholly preserved. Controlled vocabularies have been designed for chemical functional groups (CO [2]) or compounds (ChEBI [3]), but they are generally used for the annotation of chemicals or in navigation of search results. In contrast, Semantic Web ontologies aim to explicitly describe and relate objects using formal, logic-based representations that a machine can understand and process [4]. This will facilitate knowledge representation, integration and question answering in areas of critical importance to the life sciences.

In this paper, we describe a knowledge representation for chemical information using OWL, the Web Ontology Language [5]. OWL facilitates the description of

complex concepts from simpler ones and can be used for consistency checking and classification [6]. We describe our efforts to integrate DrugBank and PubChem, two popular chemical databases and DBpedia, an RDF version of Wikipedia. Finally, we illustrate the value of using semantic web technologies to seamlessly integrate and query diverse biochemical knowledge in a manner that opens new avenues for knowledge discovery in the life sciences.

## 2 Methods

### 2.1 Chemical Knowledge Representation

Upper level ontologies increase interoperability and semantic coherency of domain ontologies by grounding the basic types of domain entities and imposing restrictions on the relationships that these entities may hold. We use the Basic Formal Ontology (BFO) [7] because it offers a simple framework that distinguishes objects, qualities, processes and spatial regions. Our Basic Relation Ontology<sup>1</sup> (BRO) provides object-process, object-quality, parthood, spatial, temporal relations drawn from foundational work [8]. The New Upper Level Ontology<sup>2</sup> (NULO) maps the domain and range values of BRO properties to BFO concepts, and further constraints on relations are specified in NULO-constraints<sup>3</sup>. Reflexive, irreflexive, asymmetric, disjoint roles and role chains have been added to the BRO-OWL11 ontology<sup>4</sup> so as to maximize reasoning capability [9].

An outline of the chemical knowledge representation is illustrated in Fig 1. Briefly, molecules, atoms and rings are types of objects that bear qualities and may be located in spatial regions.

**Objects:** Molecules, atoms, rings are types of objects that are spatially extended, maximally self-connected and self-contained and bear any number of qualities appropriate to their type.

**Qualities:** A quality is a categorical property that exists in some object. Qualities have been defined for each kind of object. For instance, a molecule might bear the quality of monoisotopic mass whereas the partial charge is an atom quality. Some quality types may be borne by multiple types of objects (i.e. atoms or molecules may bear a chiral quality). We have identified over 50 types of qualities, largely defined from OpenBabel and PubChem descriptors.

**Mereology:** A molecule is composed of at least two or more atoms and has zero or more ring parts. Molecules or Rings are related to Atoms by *hasProperPart*, an asymmetric relation. Molecules and rings are related to each other by *hasPart*, a transitive (if a *hasPart* b and b *hasPart* c, then a *hasPart* c) and reflexive (one can have itself as a part) relation. Thus, rings may also be a molecule (i.e. benzene).

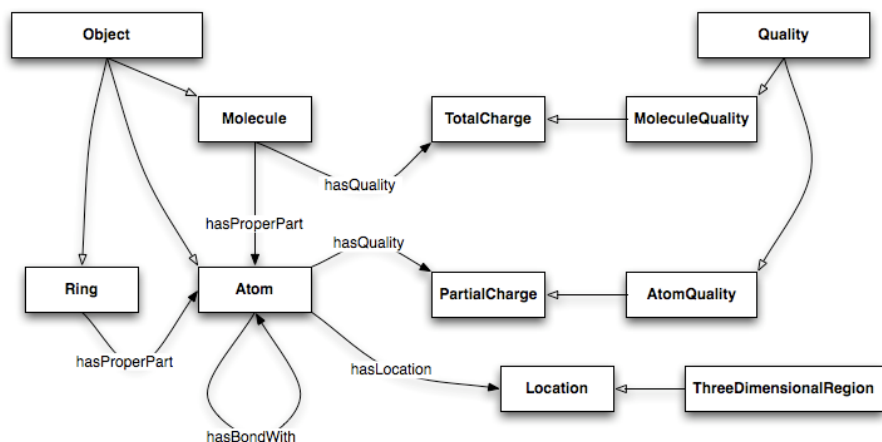
---

<sup>1</sup> <http://ontology.dumontierlab.com/bro>

<sup>2</sup> <http://ontology.dumontierlab.com/nulo>

<sup>3</sup> <http://ontology.dumontierlab.com/nulo-constraints>

<sup>4</sup> <http://ontology.dumontierlab.com/bro-owl11>



**Fig. 1.** Overview of major ontological components and their relationships in the chemical knowledge representation

**Connectivity:** Atoms are connected to each other via symmetric *hasBondWith* object properties. Specification of number of shared electrons is done via sub-properties (e.g. *hasSingleBondWith*, *hasAromaticBondWith*).

**Stereochemistry:** The spatial arrangement of atoms within molecules affects behavior and function. Stereochemical knowledge is reflected at the molecule (the molecule is a *ChiralMolecule*), atom (the atom is a *ChiralAtom*) and bonds (*hasWedgeBondWith* and its inverse *hasHashBondWith*) levels.

**Location:** Physical objects such as molecules or atoms may be spatially located in two or dimensional spatial regions to which specific coordinates may be assigned. The Cartesian coordinates of a three dimensional spatial region are assigned via datatype properties (*coordinateX*, *coordinateY* and *coordinateZ*). Since atoms are parts of molecules, and the region of space that atoms occupy is part of the region of space that molecules occupy, we can say that an atom *isLocatedIn* molecule [8].

## 2.2 Open Babel: Chemical File Conversion to OWL

We implemented a plugin for the widely used and freely available Open Babel software suite to convert any of the 80 chemical file formats into an OWL chemical knowledge model. OB provides an application programming interface (API) for reading and writing chemical file formats, accessing information about molecules, atoms, bonds, rings and for computing chemical attributes. Since each file format is different and contains an arbitrary set of information, we compute missing information using Open Babel built-in routines, where possible.

The plugin architecture is highly flexible and allows one to create mappings from ontology classes and their attributes to main classes of the OB data model. The mappings are defined within 7 major sections of an XML based configuration file.

**Generate:** Specifies how the ontology should be generated. For example, it deals with adding comments or time stamps in the ontology header.

**Base:** Specifies the namespace of the ontology.

**URIs:** Specifies which namespaces will be used in the ontology.

**Import:** Specifies which ontologies should be imported to provide additional information of named entities.

**Classes:** Contains mapping rules for establishing class type and membership. OWL classes allow the grouping data with similar properties by defining the necessary and sufficient conditions for class membership. For example, one can define the HydrogenAtom class as an Atom (to which all of the atoms present in the OB data model get mapped) that have 1 as their atomic number. More complex mappings may be generated through unions and intersections of restrictions (or combinations of both) and nested conditions.

**DataProperties:** Contains mapping rules to specify datatype properties. Datatype properties describe binary relations between OWL individuals and RDF literals or XML schema datatypes. For example, the Atom class would be the domain of hold the *atomicNumber* datatype property whereas Location would be the domain of *coordinateX*, *coordinateY* datatype properties.

**ObjectProperties:** Contains mapping rules to specify object properties. Object properties describe relations between OWL individuals. A domain and range may be specified for each property; hence, we may define the *hasProperPart* object property with *Molecule* as a domain and *Atom* as a range. In this case, *hasProperPart* object property will be created between every single atom individual and a molecule individual (an individual is an instance of an OWL class).

### 2.3 DrugBank

DrugBank is comprehensive drug knowledge base that is freely available on the web [10]. It combines clinical and chemical information about drug molecules and also provides detailed information about their drug targets. DrugBank contains nearly all drugs that have been approved in North America, Europe and Asia. These have been tagged as approved, experimental, biotech, nutraceutical, illicit and withdrawn drugs.

OWL classes are generated from each DrugBank "drugcard" records using Apache Group's open-source implementation of UIMA<sup>5</sup>. UIMA is a framework to analyze large amount of unstructured information using a workflow of annotators. Each annotator uses information from the original input and/or from previous annotators in the workflow and produces new information that is made available to other annotators further in the workflow. We designed an RDF/XML template to allow UIMA annotators to collaborate in converting DrugBank records into an OWL class. This flexible approach decouples the OWL representation from the software.

Drugs are types of objects represented as OWL classes. By importing the ontology into an existing OWL knowledge base, one can automatically classify instances based

---

<sup>5</sup> <http://incubator.apache.org/uima/>

on their characteristics. For example, the drug Leuprolide is equivalent to the class of all things that have *pubchemcompoundid* = 3911. On reasoning, we discover that all individuals asserted as instances of this class will inherit the property of having *pubchemcompoundid* = 3911 and that an individual that contains the value 3911 for the data property *pubchemcompoundid* will be inferred as an instance of the class.

## 2.4 DBpedia Integration

DBpedia makes the encyclopedia-like information from Wikipedia available in RDF. We mapped the Wikipedia link found in some DrugBank records to the corresponding DBpedia entry. The corresponding URI was found by querying DBpedia's SPARQL endpoint for the resource that is the subject of the given Wikipedia page. When adding the DBpedia RDF graph, the record is visible to the ontology as an OWL individual. To strengthen the relationship between the DBpedia instance and the corresponding drug class from the Drugbank ontology, we assert that the class is equivalent to the set containing the DBpedia instance. This is expressed in OWL using *enumerations* (owl:oneOf).

## 3 Results

We created an example OWL knowledge base<sup>6</sup> from some of the i) 4422 UIMA-generated OWL ontologies from DrugBank records with PubChem identifiers, ii) Open Babel plugin generated OWL ontologies from PubChem SDF records and iii) script generated OWL import documents for DBpedia URIs from DrugBank Wikipedia links. We will demonstrate querying this knowledge base using the simple Manchester OWL syntax [11].

### Use Case 1: Querying Substructures, Functional Groups and Compounds

An important aspect of chemical synthesis, pharmaceutical design and lead optimization involves searching chemical databases for compounds having certain kinds of substructures. Our knowledge model provides the means to define and search for substructures. As an example, let us search our knowledge base the -OH substructure.

DLQuery: *OxygenAtom that hasSingleBondWith some HydrogenAtom*

Such queries can be captured in an ontology of functional groups. A functional group describes the semantics of chemical reactivity in terms of atoms and their connectivity, and exhibits characteristic chemical behavior when present in a compound. In our ontology of major functional groups found in organic compounds<sup>7</sup>, the organic alcohol group is defined as R-OH, where R is any alkyl or aryl carbon. Importing the functional group ontology into the chemical knowledge base enables the reasoner to automatically discover which atoms are part of known substructures, and we can query accordingly:

<sup>6</sup> <http://ontology.dumontierlab.com/ckb-dils2008>

<sup>7</sup> <http://ontology.dumontierlab.com/organic-functional-group-complex>

DLQuery: *Molecule that hasPart some AlcoholGroup*

An ontology of organic compounds<sup>8</sup> provides the necessary and sufficient conditions to automatically classify molecules based on the presence of functional groups. Hence, this ontology allows us to refer to the encapsulated concept in future queries:

DLQuery: *Alcohol*

In this way, once a substructure or functional group is defined, it can be captured as an ontology concept and published on the semantic web for sharing and reuse.

### Use Case 2: Simultaneous Querying of Chemical Qualities and Substructures

A chemical knowledge base generated from the Open Babel OWL plugin will have structural information and a wide variety of descriptors, including identifiers. To ask about the set of descriptors for leuprolide using the PubChem identifier:

DLQuery: *isQualityOf some (Molecule and pubchemcompoundid value 3911)*

Answers to this query involve inferences drawn from i) the domain value of *isQualityOf* and ii) the *hasQuality* inverse property. First, PubChem descriptors are inferred to be qualities of an object due to fact that a Quality is the domain of the *isQualityOf*. Second, the knowledge base contains *hasQuality* assertions between molecules, atoms and rings and since the inverse of *isQualityOf* is *hasQuality*, it is possible to answer this query. Qualities including total charge, heat of formation, molecular mass, among others in the example knowledge base.

### Use Case 3: Query over PubChem, DrugBank and DBpedia

Taken together, we can pose a fairly sophisticated query across our expressive ontologies and the three resources to ask about biotech drugs (DrugBank) that have an alcohol moiety (PubChem) and are eliminated within an hour (DBpedia):

DLQuery: *Alcohol and BiotechDrug and eliminationHalfLife value "Hour"*

## 4 Discussion

**Knowledge representation.** Representing chemical knowledge using an expressive formal language like OWL enables new opportunities for data integration and classification that are not possible with XML or RDF (on their own). Here, we take a step forward towards a more *realist* representation with respect to how molecules are composed, the qualities they bear, and the spatial locations they occupy. Having a regular and coherent representation rooted in reality should facilitate the classification of a feature and how it will be added to our knowledge.

While our approach is guided by the Basic Formal Ontology, it is insufficient in several respects. The first is that the BFO would like to define types, and ensure that instantiated types are those that really do exist (and that we can point to). Unfortunately, several problems arise. *First*, OWL is inadequate to specify the full

---

<sup>8</sup> <http://ontology.dumontierlab.com/organic-compound-complex>

molecular structure at the class level. This is because OWL class descriptions may not contain cycles. To overcome this limitation, we define classes in which only a single instance, containing the structural description, is the member. In this way, every instance of that class will inherit the properties of the equivalent instance. However, this approach has a serious consequence: that all instances of that class are equivalent to that single instance and therefore not differ. Thus, it will never be possible to have a collection of instances. So the solution is mostly to integrate information, rather than to have a realistic representation. As such, the solution is unsatisfying. However, recent work [12] to represent structured objects in OWL should prove adequate in this regard, and provide the means by which we can describe full molecular structure at the class level. *Second*, the integration of RDF-data from DBpedia forces an instance-level representation. This is because for a proper class description, triples must be converted into class restrictions, which are syntactically different. This poses a major problem that we overcome in the same manner as was used for molecule structure, but remains restricted to data integration, rather than realistic representation. Moreover, there exists an exceptional challenge in interpreting DBpedia data properties. Some have cryptic single letter names (i.e. “c” or “r”) for which no definition is provided. Longer term goals include creating an OWL mapping to DBpedia types identified with molecule records.

**Data integration:** While data integration is trivially satisfiable when using the same URIs, it is also possible to integrate data at the class level. Using OWL, we have defined class membership based solely on one or more identifiers, and therefore can yield logical equivalence between different data records. Class based representations mean that all instances will inherit the attributes of their type, and hence the challenge is to identify which identifiers in fact are equivalent. For instance, PubChem identifiers are unique, but several structures can map to CAS numbers. Using the logical framework of OWL it is possible to generate an inconsistency when two records are said to be different when they are in fact the same. More work is required in this regard to identify logically consistent identifier mappings.

**Conversion of legacy data:** There exists major challenge in creating ontologically structured knowledge from textual or semi-structured data. While DrugBank is a good resource for finding information about drug classes or drug targets, the meaning of this data is in free text rather than having been selected from controlled vocabularies. Our conversion is largely limited to highly regular fields such as FDA status or links to PubMed papers, and the remainder lies in the form of annotations (rather than class restrictions). Since DrugBank annotates their drugs using a shallow set of drug categories, we expect to further refine these into a nicely structured ontology and/or mapped to existing drug ontologies (i.e. ChEBI). There is a need for investigating new techniques to mine the large amount of textual information embedded as general descriptions, indications, toxicity, mechanism of action, absorption, dosage forms, among others into coherent structured knowledge.

**Chemical conversion:** Our plugin offers enormous flexibility in converting unstructured descriptors from sources other than PubChem. The configuration file can be modified so as to create a minimal knowledge base with only essential information, or can be used to map a wide variety of descriptors to ontological concepts, whether

ours or their own. To maintain compatibility, users can specify a number of relationships (equivalence, subclass, type, sub-property) to concepts defined in our ontologies.

## 5 Conclusion

In this paper, we described a chemical knowledge representation for an OWL-based knowledge model. We integrate and query across PubChem, DrugBank and DBpedia in a way that is not possible using traditional database technologies. Indeed, by leveraging Semantic Web technologies, it becomes possible to integrate chemical information at differing levels of detail and granularity, opening new avenues for life science knowledge discovery.

## References

1. Murray-Rust, P., Rzepa, H.S.: Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* 41, 1113–1123 (2001)
2. Feldman, H.J., Dumontier, M., Ling, S., Haider, N., Hogue, C.W.: CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Lett.* 579, 4685–4691 (2005)
3. Brooksbank, C., Cameron, G., Thornton, J.: The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* 33, 46–53 (2005)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* (2001)
5. W3C: OWL Web Ontology Language Guide. In: Smith, M.K., Welty, C., McGuinness, D.L. (eds.): W3C Recommendation (2004)
6. Horrocks, I.: Applications of Description Logics: State of the Art and Research Challenges. In: Dau, F., Mugnier, M.-L., Stumme, G. (eds.) ICCS 2005. LNCS (LNAI), vol. 3596, pp. 78–90. Springer, Heidelberg (2005)
7. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: applying BFO in the biomedical domain. *Stud. Health Technol. Inform.* 102, 20–38 (2004)
8. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. *Genome Biol.* 6, R46 (2005)
9. Horrocks, I., Patel-Schneider, P., Sattler, U., Parsia, B., Motik, B., Bechhofer, S., Calvanese, D., Giacomo, G.d., Lutz, C.: OWL 1.1 Specification (2006)
10. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, 668–672 (2006)
11. Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., Wang, H.H.: The Manchester OWL Syntax. OWL Experiences and Design, Athens, Georgia (2006)
12. Motik, B., Grau, B.C., Sattler, U.: Structured Objects in OWL: Representation and Reasoning. In: 17th Int. World Wide Web Conference (WWW 2008), pp. 169–182. ACM Press, Beijing, China (2008)