Describing chemical functional groups in OWL-DL for the classification of chemical compounds

Natalia Villanueva-Rosales[1], Michel Dumontier[1,2,3]

[1] School of Computer Science, [2]Department of Biology, [3]Institute of Biochemistry,
Carleton University, 1125 Colonel By Drive,
K1S 5B6, Ottawa, Canada
nvillanu@scs.carleton.ca, michel_dumontier@carleton.ca

Abstract. Functional groups describe the semantics of chemical reactivity in terms of atoms and their connectivity, which exhibit characteristic chemical behavior when present in a compound. In this paper, we take a first step towards designing an OWL-DL ontology of functional groups for the classification of chemical compounds. We highlight the capabilities and limitations OWL 1.0 and the proposed OWL 1.1 in terms of our domain requirements. We also illustrate how cyclic structures may be identified from SWRL rules and suggest extensions for reasoners to achieve this objective. This work represents a preliminary step towards describing, reasoning and querying about structure and function of molecules.

Keywords: Chemistry, functional groups, compound, classification, knowledge discovery, OWL 1.1, SWRL, ontology.

1 Introduction

Biochemistry is the study of how the interactions and transformations of molecular compounds are part of biological processes that define living organisms. These chemical transformations are made possible due to the chemical properties of molecules, defined in part by 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. Therefore, compounds may be classified based on the presence of functional groups [1]. Unfortunately, chemical records often lack functional group annotation and compound classification is often done manually. Importantly, knowledge of the presence or absence of functional groups is an important component in chemical synthesis, pharmaceutical design and lead optimization.

A biochemist's interest in chemical compounds extends from structure to function, and includes everything from chemical properties, their bioactivity, the chemical reactions they participate in and the roles they may play in the viability of living systems. Despite the availability of 80 file formats to store chemical information, none provide lossless information storage or have the ability to encode chemical functional groups in terms of atoms and their connectivity. Most file formats like the XML-based Chemical Markup Language (CML) [2] or the RDF-based Comb*e*Chem project [3] have shallow data models to store common features such as atoms, bonds, and stereochemistry, but are based on relational models rather than using more appropriate formal semantics. In CML, the molecule entity has a child element “atomlist” and “bondlist”, when it is more correct to state that a molecule is composed of atoms, and these atoms make bonds with other atoms. In addition, chemical file format converters (openbabel, oechem) also have minimal data models for common features such as atoms, bonds, and stereochemistry, but other chemical properties are not explicit. More expressive formats are required to capture not only basic chemical properties, but must also be extensible so as to be able to associate functional attributes with respect to structure.

While ontologies have been designed to list types of chemical functional groups (CO [1]) or compounds (ChEBI [4]), they are simply used for the manual annotation of chemicals or navigation of search results. Since these ontologies only contain textual descriptions, rather than formal logical descriptions, they cannot be directly interpreted by computer programs. In addition, ChEBI terms may have multiple parents, and it will become gradually more difficult to establish and maintain relationships in a growing ontology, as it was found for classification of terms in the medical domain [5]. Multiple relationships are better handled by formal expressiveness and the reasoning capabilities of an underlying description logic such as OWL, which has motivated the development of a new methodology for widely used ontologies like the Gene Ontology (GO) to increase its formal explicit semantic content [6]. *OWL*, the Web Ontology Language [7], is the recommended knowledge representation language for building semantic web ontologies. OWL-DL, a variant that is based on a family of description logics (DL), facilitates the description of complex concepts from simpler ones with an emphasis on decidability of reasoning tasks [8]. In other words, a feature of DL is that reasoning tasks terminate after a finite amount of time and that the inferences drawn are valid. Reasoning tasks like checking ontology *consistency*, computing *inferences*, and *realization* (classifying real world objects into their most specific category) can be executed by a reasoner (e.g., Pellet [9] and Racer Pro [10]) over DL ontologies [11]. In addition, reasoners support query answering about any concept described in the ontology, thereby providing new ways to query knowledge across various levels of granularity and vastly different domain knowledge. Thus, OWL offers a promising framework for the design of highly expressive chemical ontologies.

In this work, we take a first step in providing a logical description of chemical structure such that it may be used to define functional groups for the purpose of compound classification. We describe the capabilities and limitations of using OWL-DL for the design of ontologies to represent chemical concepts with both, the current 1.0 and the proposed 1.1 specifications. We also describe how cyclic chemical structures may be identified from SWRL rules and suggest extensions for reasoners which may achieve the same objective. This work represents a preliminary step towards describing, reasoning and querying about structure and function of molecules.

2 Structure and Function

2.1 OWL Ontology for the identification of Chemical Functional Groups and Classification of Organic Compounds

The ontology of chemical functional groups and organic compounds with example instances may be obtained at http://ontology.dumontierlab.com/cfg-owled-2007. The most current ontology will be available at http://ontology.dumontierlab.com/cfg. The model for this ontology is illustrated in **Fig. 1** and relates compounds, molecules, atoms and functional groups with a minimal set of properties. In this model, molecules have atoms as proper parts, and atoms are connected to each other by a bond. Chemical bonds are represented using a symmetric property between two atoms. While the most general bond property is *hasBondWith*, several sub-properties are also available to specify bond order i.e. *hasSingleBondWith*, *hasDoubleBondWith*, *hasTripleBondWith*, and *hasAromaticBondWith*. Functional groups consider composition and connectivity to define a specific chemical substructure. Specific organic compounds may be defined by virtue of the presence of specific functional groups.

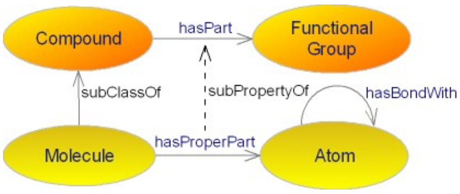

Fig. 1. Ontology entities and relations. Defined class (orange) and primitive class (yellow).

2.1.1 Defining Functional Groups

We define 35 chemical functional groups (**Fig. 3**) in OWL-DL by describing the necessary and sufficient atomic composition and connectivity (referred herein as the chemical substructure). This substructure is attached to the molecule backbone (often referred to as an “R” group) which may consist of carbons in aliphatic (alkyl) or aromatic (aryl) substructures (**Fig. 2A**; CarbonGroup) or even include hydrogen atoms (**Fig. 2A**; OrganicGroup). Thus, the necessary and sufficient conditions to describe a functional group involve the specification of the R group and the chemical substructure, as illustrated for the hydroxyl functional group in **Fig. 2B**.

Expressing these conditions for the hydroxyl functional group using the Manchester OWL syntax is as follows (in cursive font):

HydroxylGroup: *CarbonGroup that (hasSingleBondWith some (OxygenAtom that hasSingleBondWith some HydrogenAtom)*

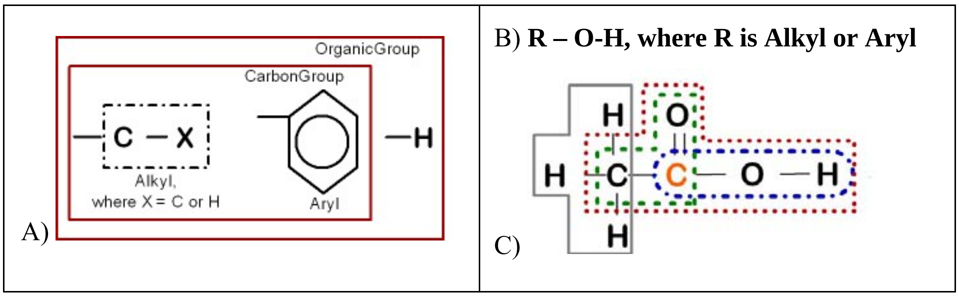

Fig. 2. A) R groups may be combinations of alkyl or aryl groups (CarbonGroup) and/or hydrogen (OrganicGroup). B) The hydroxyl functional group is defined by a CarbonGroup connected to Oxygen bound to Hydrogen; C) Overlapping functional groups in Ethanoic acid: hydroxyl (blue;dot-dash), carbonyl (green; dash), carboxylic acid (red; dot), methyl (gray; solid)

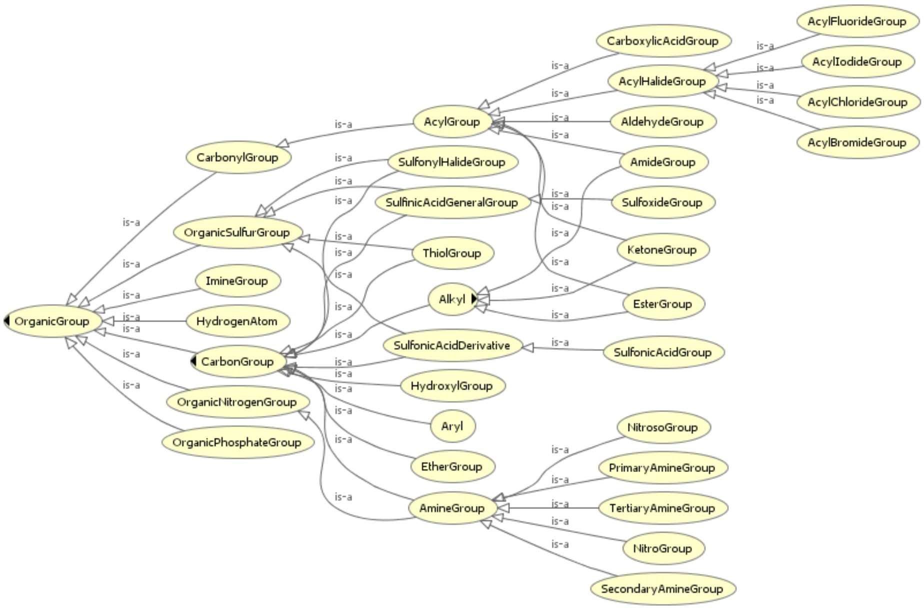

Fig. 3. Fully Inferred Ontology of Functional Groups

2.1.2 Defining Organic Compounds

The organic compounds in this ontology are shown in **Fig. 4**. We define 28 organic compounds by virtue of containing certain functional groups. These compounds include: alcohols, amines, amides, ketones and carboxylic acids. Using the Manchester syntax, the class describing an alcohol contains the following necessary and sufficient conditions (in cursive font):

Alcohol: *OrganicCompound that (hasPart some HydroxylGroup)*

Thus, molecules would be inferred to be an alcohol if they contain an atom that is classified as a hydroxyl group.

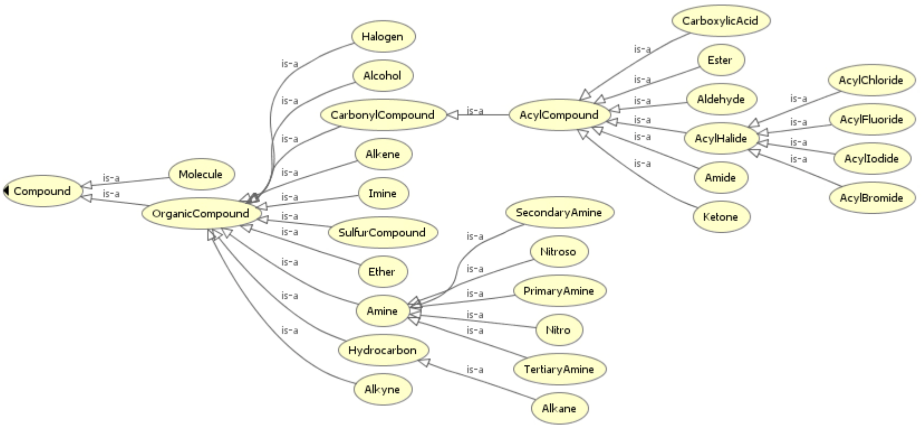

Fig. 4. Fully Inferred Ontology of Organic Compounds

2.2 Classification

We tested the ontology for its ability to identify functional groups and classify compounds using selected chemicals from the PubChem database that covered the defined functional groups and organic compounds. PHP scripts converted the SDF files to OWL files which were reasoned about using Protégé 4.0 (alpha v.29) and Pellet 1.4. We compared classification results against prior work [1], and determined that our results matched equivalent definitions. New identifications were possible with the addition of new classes with respect to previous work, such as the amide group/compound and the hydrocarbon compound. The ontology did not identify ring structures, but methods to do this are described in sections 2.4 and 3.1.3. In addition, R group atoms were classified as instances of functional groups, opening the door to identifying all atoms of the functional group, which was not previously available. Identifying all atoms of the functional group during the classification process might require either the use of rules or descriptions for each atom in the group. A more feasible approach would be to unroll class descriptions at query time.

2.3 OWL-DL Features

The expressivity of the ontology is ALCHOIQ, containing atomic and complex concept negation, concept intersection, existential and universal restrictions, role hierarchy, enumerated classes, and qualified cardinality restrictions. In the following subsections we will provide a brief description of some of the functional groups / compounds to illustrate their feature requirements and how OWL 1.0 and OWL 1.1 support these features (**Table 1**).

2.3.1 Existential and Qualified Cardinality Restrictions

Aside from primitive classes, all class expressions in this ontology include existential restrictions. For instance, the amine group consists of a carbon R group having a bond with a nitrogen atom. Further specialization of amine groups is described by the presence (or absence) of a bond with a hydrogen atom, as illustrated for amines in **Fig. 5**; Primary amines are those where the nitrogen atom has a bond with exactly 2 hydrogen atoms, secondary amines are those where the nitrogen atom has a bond with exactly 1 hydrogen atom, and tertiary amines are those where the nitrogen atom has all the bonds with hydrogen atoms substituted with bonds to other atoms. It is now evident that the definition of these classes relies not only in the quantification of the number of atom bonds (Cardinality Restrictions), but also in the qualification of the atoms that are bonded (Qualified Cardinality Restrictions).

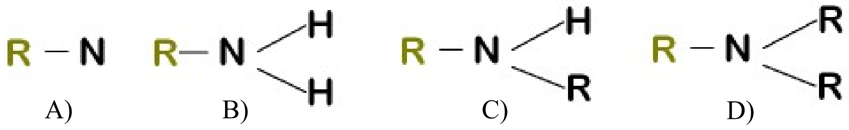

Fig. 5. A) Amine Group, B) Primary Amine Group, C) Secondary Amine Group and D) Tertiary Amine Group.

Table 1. Features required for classifying compounds from functional groups.

Structure	Feature	Details	OWL 1.0	OWL 1.1	Note
Amine Group	Existential Restriction	hasBondWith some Atom	☺	☺	
Hydrocarbon	Universal Restriction	hasProperPart only Carbon or Hydrogen	☺	☺	CWA
1' Amine Group	Qualified Cardinality Restriction	hasBondWith exactly 2 HydrogenAtom	☹	☺	CWA
2' Amine Group	Negation	hasBondWith exactly 1 HydrogenAtom	☹	☺	CWA
--	Disjoint axiom for set	100+ disjoint Atom types	☹	☺	
--	Symmetric role	hasBondWith	☺	☺	
	Complex role inclusion axiom	hasPart ○ isLocatedIn → isLocatedIn	☹	☺	
Cyclic	Local reflexive	isConnectedTo "Self"	☹	☺	
RingAtom	Partial order		☹	☹	

2.3.2 Universal restrictions

Universal restrictions make possible the identification of compounds where the chemical structure composition must be constrained. For instance, a hydrocarbon is a chemical compound that has *only* the presence of carbon and hydrogen atoms.

Universal restrictions may also be useful in defining functional groups, but care must be taken since the chemical structure of functional groups can overlap and such restrictions might exclude identifying all the functional groups present in a chemical structure. For instance, ethanoic acid (**Fig. 2C**) contains four functional groups in which two (hydroxyl group, carbonyl group) are fully contained by a third (carboxylic acid group). The inferred ontology identifies the carbonyl group as a more general concept than the more specific carboxylic acid due to the presence of the R group. Since chemists typically prefer to know the largest group, it will be important to return the most specific concept in a query answering application. While the use of universal restrictions requires the application of the closed world assumption, it will rarely be the case that only a partial set of atoms for a molecule are known. Generally speaking, the atomic composition is either fully known or unknown. Thus, it's unlikely that invalid inferences will be obtained.

2.3.3 Cycles in Ring Structures

Monocyclic and polycyclic ring structures are important parts of molecules that participate in several kinds of chemical reactions. The identification of ring structures and their constituent atoms would be an asset in finding suitable molecules for chemical synthesis. Examples of ring structures can be seen in **Fig. 6**.

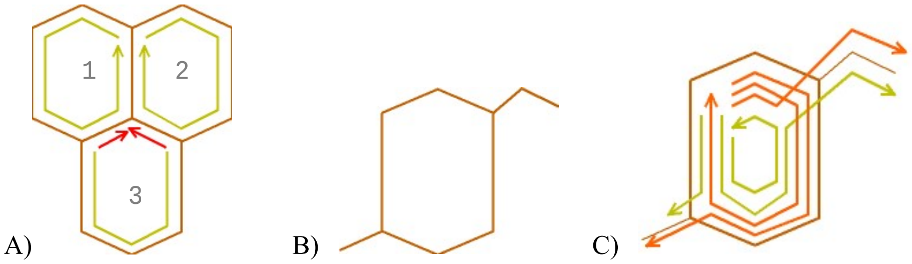

Fig. 6. A) A three ring structure represented with a transitive, anti-symmetric, reflexive property would preclude the identification of a ring atom (red arrow). B) Molecule with 6 carbon ring. C) Directed search over a transitive, symmetric, reflexive property would enable representation and identification of ring members.

Identification of members of a ring structure minimally requires a transitive (OWL1.0), symmetric (OWL1.0) and reflexive property (OWL1.1). However, the use of such a property will result in every atom in the molecule being recognized as a "ring" member. Changing the symmetric attribute to an anti-symmetric attribute (OWL1.1) would help control the order of identification, and lead to classification of some atoms, but not others as anti-symmetry would preclude a path between atoms for certain rings (**Fig. 6A**). In addition, it will be difficult to assert a unique path through a molecule or ring *a priori*, because there is no directionality in bonds.

2.4 SWRL Rules for identifying rings

While proposed for OWL 1.1, we were unable at this time to reason about concepts with local reflexive properties using Protégé 4 [10]. As an alternative, we designed

DL-safe SWRL [12, 13] rules to describe and infer the presence of cyclic structures. Since each variable in DL-safe rules must bind an entity, we designed rules for different sized cycle structures. For example, we designed rules to identify 6 member rings such as benzene (1). A disadvantage of this approach is that a rule is required for each structure, which limits the scalability of this approach.

CarbonAtom(?x) ^ CarbonAtom(?y) ^ CarbonAtom(?z) ^ CarbonAtom(?w) ^
CarbonAtom(?u) ^ CarbonAtom(?v) ^ hasBondWith(?x,?y) ^ **(1)**
hasBondWith(?y,?z) ^ hasBondWith(?z,?w) ^ hasBondWith(?w,?u) ^
hasBondWith(?u,?v) ^ hasBondWith(?v,?x) → RingAtom (?x) .

Some of our rules required as input the inferences returned by the DL reasoner. The Protégé 3.2 SWRL tab plug-in [13, 14] only considers as an input the assertions in the ontology and not the inferences generated by a DL reasoner. We implemented an additional plug-in to integrate the Jess rule engine with DL reasoners to facilitate passing assertions and inferences to the rule engine, and also provide an interface to query either assertions or the full set of inferences [15].

3 Discussion

3.1.1 Significance

This work is significant in that it describes for the first time, to the best of our knowledge, an OWL-DL based ontology comprising of a non-trivial number of chemical functional groups that may be used for the classification of organic compounds. The ontology is suitably expressive to provide precise logic-based descriptions that match well defined chemical substructures, providing evidence that semantic web technologies are sufficient to represent and reason about the chemistry domain. We anticipate that our work will be compatible the description of chemical properties which will facilitate support semantic query answering across structure and function. This will enable sophisticated approaches by which chemists can locate chemicals in ever growing databases. As well, we expect our work will provide new opportunities to describe standard reactions that involve specific functional groups.

3.1.2 General considerations

Although most of the inferences anticipated in our ontology are obtained under open world semantics, the classification of individuals into classes that use closure axioms (e.g., an atom that has bond with *some* carbon atom and *only* with carbon atoms or an atom with *exactly* 3 bonds with a carbon atom) requires closed world reasoning. As an initial effort, we added axioms to enforce closed world assumption and obtain the inferences desired (refer to the published ontology for more details about these axioms). However, we are aware that this is not a very scalable nor is it a long term solution. We have also found it useful for testing to query our ontology with nRQL [10], a query language that implements negation as failure (NAF). Given that our domain assumes that we have all the relevant knowledge at the time of classification, we can safely “close the world”. Future directions in our research include to explore logic programming and related formalisms like the one described in [16] and the use of implementations like KAON2 [17] as an alternative to achieve our goal.

To the best of our knowledge, the QCR proposed in OWL 1.1 are already supported in Pellet [9] and Racer [10], but given that our ontologies were created using Protégé 4 alpha with preliminary support of OWL 1.1, the inferences were not drawn with the qualification on a cardinality restriction. Thus, the class of organic compounds that have exactly 2 oxygen atoms was found to be equivalent to the class of organic compounds that have exactly 2 hydrogen atoms. Clearly, this is not semantically correct and has consequences in our application domain, but we are confident that we will soon have implementations that fully support OWL 1.1.

3.1.3 Simulated Partial Order

We believe that the implementation of an algorithm, or a possible combination of role properties beyond the ones proposed for OWL 1.1, for applying a directed path search simulating partial order over a *symmetric* property with local reflexive attributes would have the desired effect of uniquely identifying ring atoms (**Fig. 6C**). The approach can also be seen as a search along all available paths in a directed manner, without falling back on atoms already explored. This mechanism would infer the presence ring atoms, but not acyclic atoms.

3.1.4 Complex roles for spatial knowledge discovery

While the current ontology is geared towards describing wholly self-connected and self-contained molecules, we aim to investigate spatial relations by considering the spatial regions they occupy. The construction of complex roles proposed in OWL 1.1 will support these goals. For instance, a material continuant is located in another if the spatial region that it occupies is part of the spatial region of the other [18]. This will facilitate the inference that a heme molecule is located in the heme-iron complex.

4 Conclusions

In this paper, we strove to describe how the current and proposed features for OWL may be used for the description of chemical functional groups towards the classification of organic compounds. We highlight the importance of several new OWL 1.1 features, including qualified cardinality restrictions and complex properties, and describe the simulation of partial order over symmetric properties that could be implemented. Finally, we suggest the tighter integration of DL-safe rules with DL reasoners to facilitate more sophisticated reasoning.

OWL-DL is a very natural knowledge representation language for the chemistry domain: primitive concepts (atoms, molecules) form the basis for constructing more complex concepts (functional groups, organic compounds). Together with other OWL ontologies being developed for the life sciences, we expect this will enable querying knowledge at various levels of granularity – from structure and reactivity of chemicals to cellular processes and biological outcomes. This knowledge will play an important role from chemical synthesis to pharmaceutical design.

Acknowledgments: This work was supported with CONACYT scholarship #150581 for NVR and Carleton University startup grant for MD.

References

1. 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. Vol. 579 (2005) 4685-4691
2. 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. Vol. 41 (2001) 1113-1123
3. Taylor, K.R., Gledhill, R.J., Essex, J.R., Frey, J.G., Harris, S.W., De Roure, D.C.: Bringing Chemical Data onto the Semantic Web. J. Chem. Inf & Mod. Vol. 46 (2006) 939-952
4. Brooksbank, C., Cameron, G., Thornton, J.: The European Bioinformatics Institute's data resources: towards systems biology. Nucleic Acids Res. Vol. 33 (2005) D46-53
5. Heja, G., Varga, P., Pallinger, P., Surjan, G.: Restructuring the foundational model of anatomy. Stud Health Technol Inform. Vol. 124 (2006) 755-760
6. Wroe, C.J., Stevens, R., Goble, C.A., Ashburner, M.: A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. Pac Symp Biocomput. Vol. (2003) 624-635
7. W3C: OWL Web Ontology Language Guide. In: Smith, M.K., Welty, C., McGuinness, D.L. (eds.) (2004)
8. Horrocks, I.: Applications of Description Logics: State of the Art and Research Challenges. ICCS2005, Kassel, Germany (2005) 78-90
9. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: a practical owl-dl reasoner. 3rd International Semantic Web Conference (ISWC2004) (2004)
10. Haarslev, V., Möller, R., Wessel, M.: Querying the Semantic Web with Racer + nRQL. KI-04 Workshop on Applications on Description Logics, Ulm (2004)
11. Wolstencroft, K., Lord, P., Tabernero, L., Brass, A., Stevens, R.: Protein classification using ontology classification. Bioinformatics. Vol. 22 (2006) e530-538
12. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. (2004)
13. O'Connor, M.J., Knublauch, H., Tu, S.W., Musen, M.A.: Writing Rules for the Semantic Web Using SWRL and Jess. Protege with Rules Workshop 2005, Madrid, Spain (2005)
14. O'Connor, M.J., Knublauch, H., Tu, S.W., Grossof, B., Dean, M., Grosso, W.E., Musen, M.A.: Supporting Rule System Interoperability on the Semantic Web with SWRL. Fourth International Semantic Web Conference (ISWC2005), Galway, Ireland (2005)
15. Bellinger, C., Villanueva-Rosales, N., Dumontier, M.: Protege 3.2 plugin to query assertions and inferences from the Jess SWRL rule engine and DL reasoners. School of Computer Science, Vol. BCSc Honors Thesis. Carleton University, Ottawa (2006)
16. Motik, B., Horrocks, I., Rosati, R., Sattler, U.: Can OWL and Logic Programming Live Together Happily Ever After? Proc. of the 2006 International Semantic Web Conference (ISWC 2006) (2006)
17. Motik, B.: Practical DL Reasoning over Large ABoxes with KAON2. (2005)
18. 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. Vol. 6 (2005) R46