Applying the Semantic Web to Computational Chemistry

Neil S. Ostlund, Mirek Sopek
Chemical Semantics Inc. , Gainesville, Florida, USA
{ostlund, sopek}@chemicalsemantics.com

Abstract. Chemical Semantics is a new start-up devoted to bringing the Semantic Web to computational chemistry, with a future goal to cover chemical results of other kinds, including experimental. We have created a prototype test bed that enables computational chemists to publish the results of their computations (for example: structure, energies and wave functions of *Ab Initio* calculations) on servers (powered by semantic triple stores) holding RDF data. In addition, scientists will be able to search across results produced by themselves and other researchers using SPARQL queries and faceted search built on top of it. The semantics of computational chemistry oriented RDF data is defined by an ontology identified as "GC" (Gainesville Core). Gainesville Core defines concepts such as: theoretical models used in computations, molecular systems (as collections of atoms and molecules), molecular arrangements which describe system evolution in time or in phase space and so on. The poster demonstrates the project during its early stage, including a description and demonstration of the prototype portal created by Chemical Semantics. This paper is the concise description of the poster content.

Keywords: computational chemistry, semantic web, linked data, Gainesville Core, ontology

Chemical Semantics portal main targets

The Chemical Semantic portal, now in its prototype version, has the following long term goals:

1. Interoperable PUBLISHING of Computational Chemistry calculations using Linked Data principles
2. Enhanced search across all published results with basic reasoning enabled by the new ontology
3. FEDERATION of published data with existing web-based chemical datasets
4. Cloud-like ARCHIVING of Computational Chemistry calculations results, input/output files etc.

System architecture

The portal is composed of the following main modules/functional blocks:

1. **Data conversion module** – its role is to convert input files which are the results of Computational Chemistry experiments (computations) from their native format to the RDF format using objects and properties defined by the new ontology. The module operates as a web service end point which, in the current version of the software can accept the CSX file format (see the chapter "Chemical Publisher" below) or CML (Chemical Markup Language) format.
2. **Triple Store** – the standard implementation of a triple store for storing of RDF data sets. In the prototype version of the portal, the community edition of Virtuoso Triple Store is used.
3. **URI generator** for molecular systems, based on PURLs URIs.

4. **Linked Data conformant content negotiation module**. The module delivers RDF datasets representing a molecular system in RDF/XML (or Turtle) serialization when queried by Linked Data/ Semantic agent or a rich web page with different "chemically aware" visualizations of the molecular system.
5. **"Chemically aware" visualization system** – which displays basic information about a molecular system, interactive visualization of the molecular system's structure, electronic orbital energy level graph, federated data – data from other chemical portals (ChemSpider, Chebi, NIST database), visualization of RDF graph, tags (for building folksonomy) etc.
6. **SPARQL End Point** – Virtuoso implementation of the standard SPARQL end point.
7. **Publication Manager** – the module which enables searching and browsing across publications in the portal.
8. **User and access control manager** – enforces access rules both to the portal and to individual publications identified by specific URIs.

The portal has been implemented using Microsoft .NET technology and dotNetRDF semantic library. The triple store technology is OpenLink Software "Virtuoso".

Gainesville Core ontology

For the purpose of definition of shared type system across all published files on the portal we have created a new web ontology named Gainesville Core. The Gainesville Core ontology is identified by IRI: `http://purl.org/gc`

Gainesville Core defines the following types:

1. theoretical models (or experiment foundations) used in computations
2. molecular systems (as collections of atoms and molecules)
3. molecular arrangements which can describe molecular system subparts (like monomers) or system evolutions in time or in phase space (different configurations of the system)
4. types related to computations and specific features of the theoretical model of the system (such as computation method and technology)
5. types related to results of computations (like energies, orbitals etc.)

The Gainesville Core ontology is currently undergoing thorough changes. There is a new update that will soon replace the current content of `http://purl.org/gc`.

"Chemical Publisher"

Within the scope of the Chemical Semantics portal's operations, "Chemical Publisher" is a piece of software (independent or integrated) running along standard Computational Chemistry modeling software that enables "push button" publishing. The purpose of the "Chemical Publisher" is to prepare a consistent data set containing all relevant data that result from Computational Chemistry experiments, and to call the Chemical Semantics portal (using REST based web services) for the publication of these data.
The relevant data will contain, as a minimum, a specific representation of a molecular system (atoms and molecules) but in standard cases, it will also contain all information about the computations being made on the molecular system.

In the current prototype version of the Chemical Semantics Portal, the Chemical Publisher has been realized as an additional module to the popular molecular modeling package, HyperChem. The format used by HyperChem's Chemical Publisher is an XML format named "CSX" which is a modification of the popular CML (Chemical Markup Language). Standard CML format can also be used with the portal in the "manual" upload process.

URI convention assumed by the portal

Chemical Semantics portal uses PURL based URIs.
For example, a typical URI of a specific "publication" on the Chemical Semantics portal may be built as in the following example:

`http://purl.org/chem/pub/2013-08-05-betacyanin`, where:

`http://purl.org` part is owned and controlled by PURL maintainers

`/chem` is owned by PURL maintainers but controlled by Chemical Semantics

`/2013-08-05-betacyanin` is generated by Chemical Semantics for the user of Chemical Publisher. In our model it is "owned" by the user

The URI conventions we have assumed for other components of the published system are:

For Molecular Calculation part of the publication:

`http://purl.org/chem/pub/2013-08-05-betacyanin/mol-calc`

For Molecular System described in the publication:

`http://purl.org/chem/pub/2013-08-05-betacyanin/molSys`

a molecule of the system will have URI: **`http://purl.org/chem/pub/2013-08-05-betacyanin/molSys/m1`**

and a bond in it: **`http://purl.org/chem/pub/2013-08-05-betacyanin/molSys/m1/a1a12`**

"Chemical" rendering of RDF data sets

While the standard "semantic" rendering of the elements of Chemical Graphs stored in the Chemical Portal into RDF/XML or Turtle representation is well defined, it is less obvious what should be returned to the caller when the respective content negotiation recognizes a human user using a browser.

For such cases, we have defined the following "facets" of the view available as standard "tabs" in web page designs. The facets that are available in the prototype are:

"Basic" – containing basic publication data such as its title, abstract, name and author (including his affiliation and contact information)

"Results" – containing summary information about a molecular system, such as its atomic masses distribution, information about calculation method and technology, and basic results of computations (system energies and computed spectra)

“Molecules” – containing three-dimensional, interactive visualization of the molecular structure.

“Wave function” – if the published computations were done with one of the electronic structure methods, this facet will contain a diagram of orbital energies.

“Graph” – containing the interactive rendering of the RDF graph underlying the publication stored in the system

“Data Sets” – the facet contains pointers (URLs) to all files attached to a publication. By definition it will contain the generated RDF file (in Turtle format) and the original file used by Chemical Publisher (CSX described above in the case of the current prototype)

“Tags” – this facet contains all tags assigned by the user which describe a system. The approach to tag usage is typical of “folksonomy” tags present in popular use.

“Data Federation” – the facet that contains references to other sources of molecular data discovered by the Chemical Portal. In the prototype version we have used: ChemSpider, Chebi and NIST web resources to illustrate data federation. The discoveries are made with the help of Inchi identifiers.

SPARQL queries

The Chemical Semantic Portal implements a Virtuoso SPARQL Endpoint as an important element of its functionality. It enables the portal’s users to explore its triple store using standard SPARQL queries.

For example, a query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX gc: <http://purl.org/gc/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?graph WHERE {
    {GRAPH ?graph { ?mol gc:hasAtom ?atom}}
    MINUS
    {GRAPH ?graph { ?a gc:isElement "C" }}
    MINUS
    {GRAPH ?graph { ?b gc:isElement "O" }}
    MINUS
    {GRAPH ?graph { ?b gc:isElement "N" }}
    MINUS
    {GRAPH ?graph { ?b gc:isElement "H" }}
}
```

Will return the URIs of all inorganic molecules stored in the triple store:

graph
http://purl.org/chem/pub/2013-08-06-si_li-test
http://purl.org/chem/pub/2013-08-06-copper-chloride
http://purl.org/chem/pub/2013-08-06-copper-sulfide
http://purl.org/chem/pub/2013-08-06-nobium-pentachloride

References

1. Uschold, M., Grüninger,M. : Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 1996, (http://starlab.vub.ac.be/teaching/uschold.pdf)
2. Hepp,M.: Ontologies: State of the Art, Business Potential, and Grand Challenges. in Hepp, M.; De Leenheer, P.; de Moor, A.; Sure,Y. (Eds.): Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, ISBN 978-0-387-69899-1, Springer, 2007, pp. 3-22. (http://www.heppnetz.de/files/hepp-ontologies-state-of-the%20art.pdf)
3. Szabo,A., Ostlund, N.S.: Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory. Dover Publications, New Edition 1996, ISBN: 0486691861
4. Jensen, F.: Introduction to Computational Chemistry. Wiley; 2 edition (2007), ISBN: 0470011874
5. Vesse,R., Zettlemoyer, R.M., Ahmed, K., Moore, G., Pluskiewicz,T. DotNetRDF - Semantic Web/RDF Library for C#/.Net. http://www.dotnetrdf.org
6. OpenLink Software.: “Virtuoso Universal Server”. http://virtuoso.openlinksw.com/
7. HyperChem(TM) Professional 8.0, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA
8. Murray-Rust, P.; Rzepa, H. S.: Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. J. Chem. Inf. Comput. Sci. 39 (6): 928–942

Contact Us

If you have any further questions regarding the submitted poster, do not hesitate to get in touch with us.

For all questions related to our portal and its implementation details, your contact person is:

Dr. Mirek Sopek, e-mail: sopek@chemicalsemantics.com

For overall scientific questions concerning the science of Computational Chemistry and its use in our work, please contact:

Dr. Neil S. Ostlund, e-mail: ostlund@chemicalsemantics.com