Phase II Final Scientific/Technical Report

Period: January 1, 2015 to June 30, 2017

Report of June 30, 2017

Award DE-SC0011735

Chemical Semantics, Inc.

PI: Neil S. Ostlund

Title: Publication and Retrieval of Computational Chemical-Physics Data via the Semantic Web

This Final Report is for Phase II of our Fast-Track Grant, DE-SC0011735, originally ending Dec 31, 2016. A no-cost extension of the grant was made until June 30, 2017. This final report is for the whole period of Phase II, i.e. Jan 1, 2015 to June 30, 2017.

Unlimited Distribution

Executive Summary

This research showed the feasibility of applying the concepts of the Semantic Web to Computation Chemistry. We have created the first web portal (www.chemsem.com) that allows data created in the calculations of quantum chemistry, and other such chemistry calculations to be placed on the web in a way that makes the data accessible to scientists in a semantic form never before possible. The semantic web nature of the portal allows data to be searched, found, and used as an advance over the usual approach of a relational database. The semantic data on our portal has the nature of a Giant Global Graph (GGG) that can be easily merged with related data and searched globally via a SPARQL Protocol and RDF Query Language (SPARQL) that makes global searches for data easier than with traditional methods.

Our Semantic Web Portal requires that the data be understood by a computer and hence defined by an ontology (vocabulary). This ontology is used by the computer in understanding the data. We have created such an ontology for computational chemistry (purl.org/gc) that encapsulates a broad knowledge of the field of computational chemistry. We refer to this ontology as the Gainesville Core. While it is perhaps the first ontology for computational chemistry and is used by our portal, it is only a start of what must be a long multi-partner effort to define computational chemistry.

In conjunction with the above efforts we have defined a new potential file standard (Common Standard for eXchange – CSX for computational chemistry data). This CSX file is the precursor of data in the Resource Description Framework (RDF) form that the semantic web requires. Our portal translates CSX files (as well as other computational chemistry data files) into RDF files that are part of the graph database that the semantic web employs. We propose a CSX file as a convenient way to encapsulate computational chemistry data.

Accomplishments

Most of the goals of the project have been accomplished. Primarily, what we have done is created a web portal that illustrates the semantic web and what it can accomplish. This was the fundamental goal of the project and it has been accomplished such that we have a usable public portal. Along the way various related aspects such as ontologies, file standards, and web structures have been explored and used or abandoned. The portal (www.chemsem.com) is publically available and usable by computational chemists. If we have one regret, it is that we haven’t had the time and resources, nor made the effort to make the portal more widely known and used. This is primarily a Phase III objective, i.e. finding customers for our technology and performing a serious marketing effort. We would have liked to be further along in bringing in other members of the computational chemistry community. We have had to focus our efforts on making the portal robust and with this in mind its use has been primarily internal to the project.

Our goal of creating a “Gainesville Core” ontology is accomplished although it is limited to those concepts currently used in our portal. The most recent version of this ontology is publically available at purl.org/gc. We would hope that others would use this current version and extend it in conjunction with ourselves. No ontology as ambitious as one for computational chemistry can be the creation of a single group.

The goal of creating a semantic web portal for computational chemistry begins with a scientist performing calculations on a molecular system. The results of these calculation usually reside in an output file. These files can be uploaded to our portal where they are translated into semantic data of the RDF form appropriate for using with the semantic web. Alternatively, with certain software packages CSX data can be created immediately and then uploaded to the portal. With either path, the portal then contains computational data which is displayed in a variety of ways at the portal. One can then search the portal for such data using SPARQL queries. Our portal thus accomplishes the task of being a usable semantic web portal for computational chemists.

Project Activities

1. Generate Ontology. Significant time and energy were spent with software developed by Stanford called Protégé. This software enables the creation of ontologies and their display. It has become the standard for manipulating ontologies and considerable effort was spent in learning and using it to create our Gainsville Core ontology.
2. Define CSX. Our precursor to semantic data is an XLM file we refer to as CSX. It defines most of the data of computational chemistry and large amounts of effort were spent defining it and getting it “correct” for our purposes. An earlier standard, CML, was used as well but found to be lacking compared to CSX. A CSX schema exits and was created to define the CSX standard. Various versions of CSX evolved as we added more and more computational results to the schema.
3. Portal Functionality. Our portal was required to ingest data, display it, convert it to RDF format, store it in a triple store, and allow SPARQL queries of the data. Each of these required implementing the appropriate functionality. The portal has a REST capability so data can be sent to it over the network from a client machine. Alternatively, data can just be uploaded manually to the portal. Data that is in the format of the original software package (output file) must be parsed into a structured form such as CSX. We

developed our own parsers as well as worked with 3-rd party parsers such as those developed by CCLIB. We contributed additions to the CCLIB group's capabilities and adopted their parsers. Once data was in a structured form it has to be converted to RDF and placed onto a triple store (semantic web graphic database). We used Virtuoso as our triple store but experimented with others. All these chemistry functionalities had to be invented and developed in the context of our portal.

4. Portal GUI. Our portal displays data from computational chemistry packages such as dipole moments, orbitals, electron densities, vibrational spectra, optical spectra frequencies, etc. and code had to be developed to display these various quantities. We used JMOL for much of the rendering but developed our own renderings code as well. In addition we had to organize the data as a publication with authors, title, abstract, etc. A GUI had to be developed for these capabilities. In addition, we created searching for publications based on tags and other aspects of a specific publication. Finally, the initial organization of publications was flat and we developed a folder structure for hierarchical display of our data publications.

Products Developed

a. Publications.

1. **Computational Chemistry Data Management Platform Based on the Semantic Web:**

Bing Wang[†], Paul A. Dobosh[†], Stuart Chalk[‡], Mirek Sopek[†], and Neil S. Ostlund[*†]

[†] Chemical Semantics Inc., 2772 NW 43rd Street, Suite B1, Gainesville, Florida 32606, United States

[‡] Department of Chemistry, University of North Florida, Jacksonville, Florida 32224, United States

J. Phys. Chem. A, **2017**, *121* (1), pp 298–307.

DOI: **10.1021/acs.jpca.6b10489**

2. **A Portal for Quantum Chemistry Data Based on the Semantic Web:**

Bing Wang,[1] Paul A. Dobosh,[1] Stuart Chalk,[2] Keigo Ito,[1] Mirek Sopek[1] and Neil S. Ostlund[1,*]

[1.] Chemical Semantics Inc., 2135 NW 15th Ave, Gainesville, FL 32605

[2.] Department of Chemistry, University of North Florida, Jacksonville, FL 32224

Springer, Proceedings of the "21st International Workshop on Quantum Systems in Chemistry, Physics, and Biology" (QSCP-XXI), in press.

b. Web sites.

1. www.chemical semantics.com

This is our main web site describing the company and its mandate

2. www.chemsem.com

This is our main portal where customer publications occur.

3. Purl.org/gc

This is our Gainesville Core ontology for computational chemistry.

4. www.staging.chemsem.com

This is the development site for our portal.

c. Collaborations

We have attempted to establish various collaborations to stimulate customers using our portal but these are very preliminary unfortunately. We have established collaborations with scientists that are developing computational chemistry software packages with data that we can incorporate into our portal. These principally include groups around the following software packages:

- PSI4
- Gamess
- NWChem
- Gaussian

d. Technologies

We believe we are the first ones to develop the technology of a true semantic web portal for computational chemistry. This involved developing xml technologies and a schema for incorporating structured computational data into a portal. Then there was the new technology for translating this data into an RDF form appropriate for a triple store database using an ontology to drive the translation. In the process of creating our portal there were many proprietary pieces of software developed.

e. Inventions/Patents/Licensing

No specific technology was patented or licensed.

f. Other products

Our portal is essentially a product. We have yet, however, to commercialize it.

Computer Modeling

a. Model Description

The computer model we used is RDF, SPARQL, and other standards of the semantic web as defined by the world wide web consortium (W3C). In addition we use XML files defined by their schema and also JSON and JSON-LD files as ways to structure data. These were incorporated in to a web site that was our portal, www.chemsem.com

b. Performance

No specific and elaborate performance measurements were made. Obvious performance issues arrived as we develop our code and were dealt with in real time.

c. Test Results

No specific test results are available. The popularity of the portal will be determined by its users.

d. Theory

The fundamental theory involved in his research is that of the semantic web. Its basic properties and description are given by the W3C.

e. Mathematics

No specific new mathematics was involved in this research.

f. Peer Review

The fundamentals of the generic semantic web are well-known and described by the W3C.

g. Hardware Requirements

Hosting of web sites was obviously required for this research. We sites such as www.staging.chemsem.com and www.chemsem.com were hosted initially at Chemical Se\mantics, Inc. and later at Microsoft's azure.

h. Documentation

Documentation of our portal and other facets of our efforts are on the relevant web sites. No hard copy documentation was created or needed.