Journal of Cheminformatics

EDITORIAL Open Access

Resource description framework technologies in chemistry

Egon L Willighagen[1*] and Martin P Brändle[2]

Editorial

The Resource Description Framework (RDF) is providing the life sciences with new standards around data and knowledge management. The uptake in the life sciences is significantly higher than the uptake of the eXtensible Markup Language (XML) and even relational databases, as was recently shown by Splendiani et al. [1] Chemistry is adopting these methods too. For example, Murray-Rust and co-workers used RDF already in 2004 to distribute news items where chemical structures were embedded using RDF Site Summary 1.0 [2]. Frey implemented a system which would now be referred to as an electronic lab notebook (ELN) [3]. The use of the SPARQL query language goes back to 2007 where it was used in a system to annotate crystal structures [4].

The American Chemical Society (ACS) Division of Chemical Information (CINF) invited scientists from around the world to present their use of RDF technologies in chemistry on 22nd-23rd August 2010 at the 240th ACS National Meeting in Boston, USA. During three half-day sessions, the speakers demonstrated a mix of smaller and larger initiatives where RDF and related technologies are used in cheminformatics and bioinformatics as Open Standards for data exchange, common languages (ontologies), and problem solving. The fifteen presentations were grouped in the themes computation, ontologies, and chemical applications. Figures 1, 2 and 3 display the most important keywords reflecting the abstracts of the talks in each session as word clouds [5].

The goal of the meeting was to make more chemists aware of what the RDF Open Standard has to offer to chemistry. We are delighted to continue this effort with this Thematic Series, for which the speakers (and others) were invited to present their work in more detail to a wider chemistry community. The choice of an Open Access journal follows this goal. At this place, we would like to thank Pfizer, Inc., who had partially funded the article processing charges for this Thematic Series. Pfizer, Inc. has had no input into the content of the publication or the articles themselves. All articles in the series were independently prepared by the authors and were subjected to the journal's standard peer review process.

In the remainder of this editorial, we will briefly outline the various RDF technologies and how they have been used in chemistry so far.

1 Concepts

The core RDF specification was introduced by the World Wide Web Consortium (W3C) in 1999 [6] and defines the foundation of the RDF technologies. It has evolved into a set of recommendations by the W3C published in 2004 (See Table 1). RDF specifies a very simple data structure linking a subject to an object or a value (literal) using a predicate. Cheminformaticians will recognize this data structure as an edge from graph theory. This structure allows us to represent facts like "vanillin dissolves in methyl alcohol" [7]. RDF uses Uniform Resource Identifiers (URIs) to identify things. Therefore, the RDF equivalent of the solution statement could be like this so-called triple:

http://dbpedia.org/resource/Vanillinhttp://example.com/dissolvesInhttp://dbpedia.org/resource/Methanol.

Since URIs may be used to reference resources on any server worldwide, RDF triples allow to span a global graph data structure. This is not surprising, since RDF is the core technology behind the proposed Semantic Web [8]. In fact, the Web nature is clear here, as one can follow both the URIs for vanillin and methanol to obtain further information on those two chemicals. These molecules' URIs are said to be dereferencable, allowing agents to spider the Web for information following the hyperlinks, quite like how you follow hyperlinks on websites. Hence, the term Semantic Web.

Recent projects such as Bio2RDF [9], Chem2Bio2RDF [10], and OpenTox [11] have brought genomic, chemical

* Correspondence: egon.willighagen@gmail.com
[1]Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, SE-17177 Stockholm, Sweden
Full list of author information is available at the end of the article

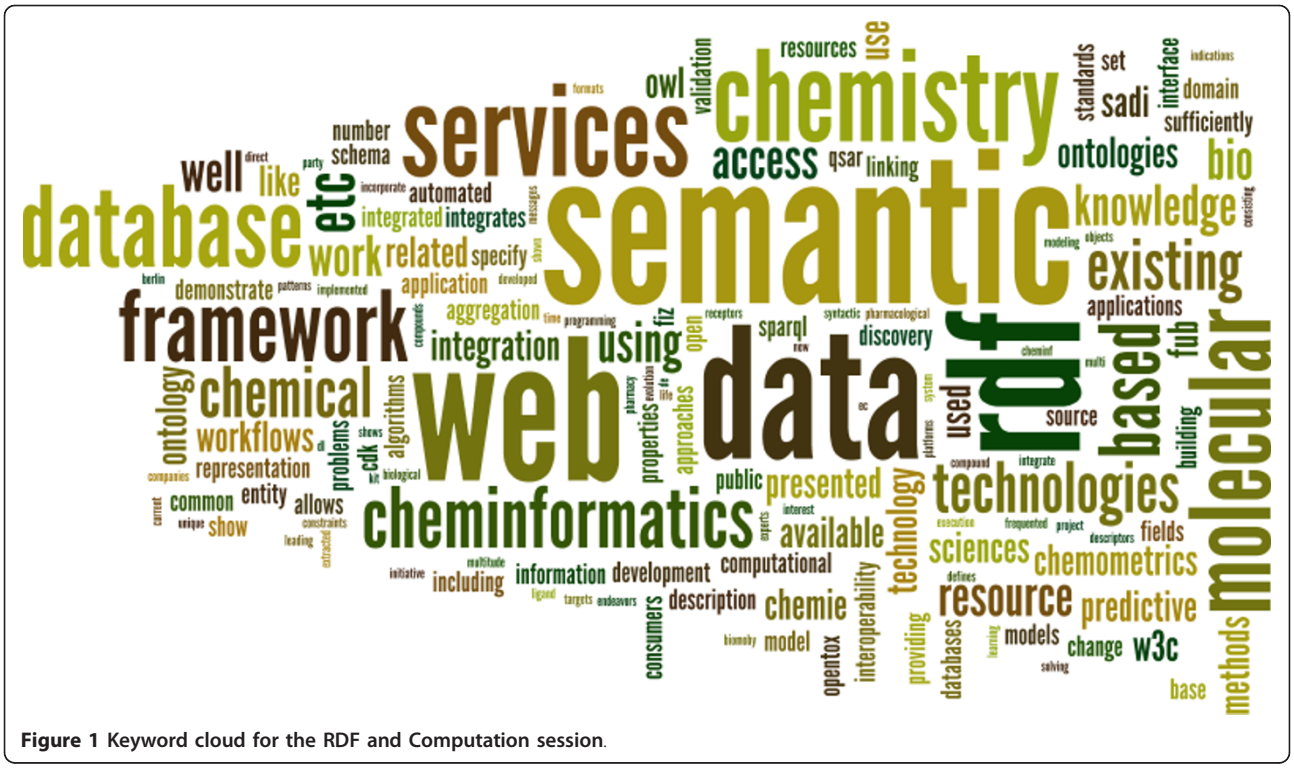

Figure 1 Keyword cloud for the RDF and Computation session.

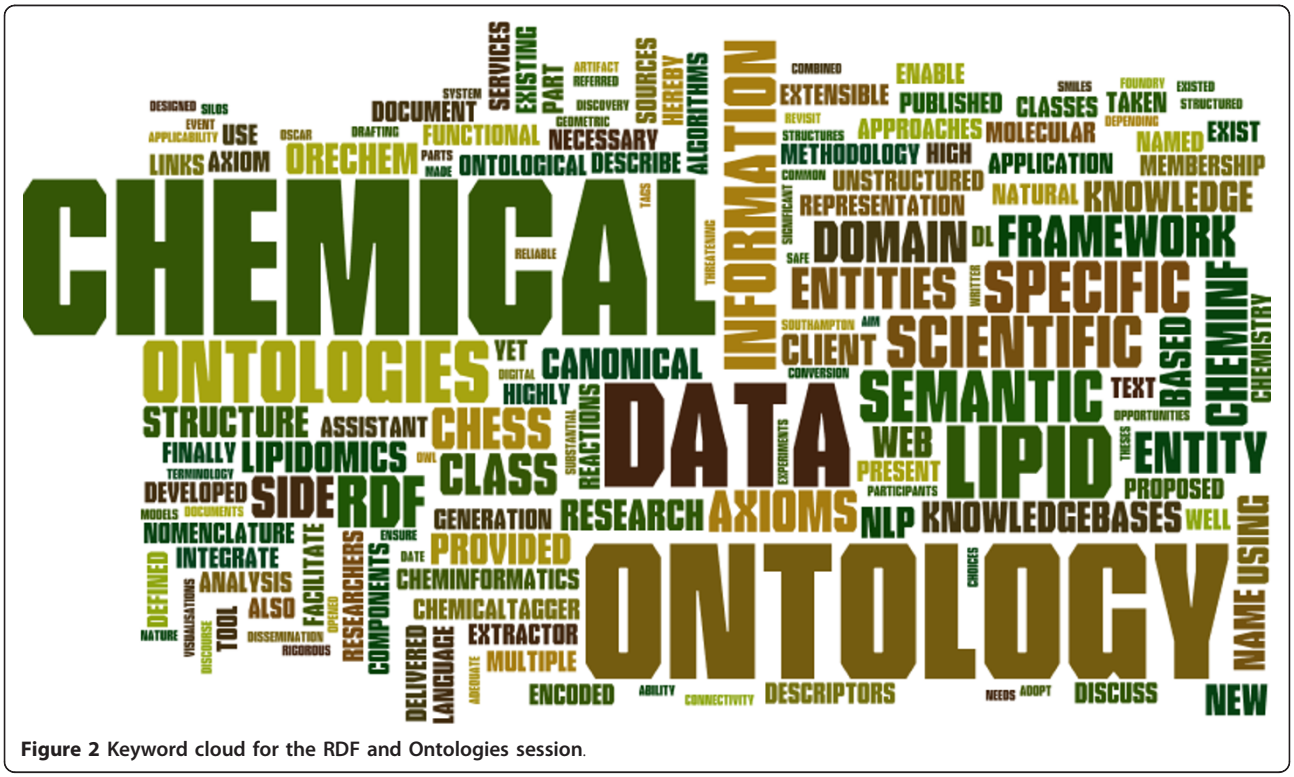

Figure 2 Keyword cloud for the RDF and Ontologies session.

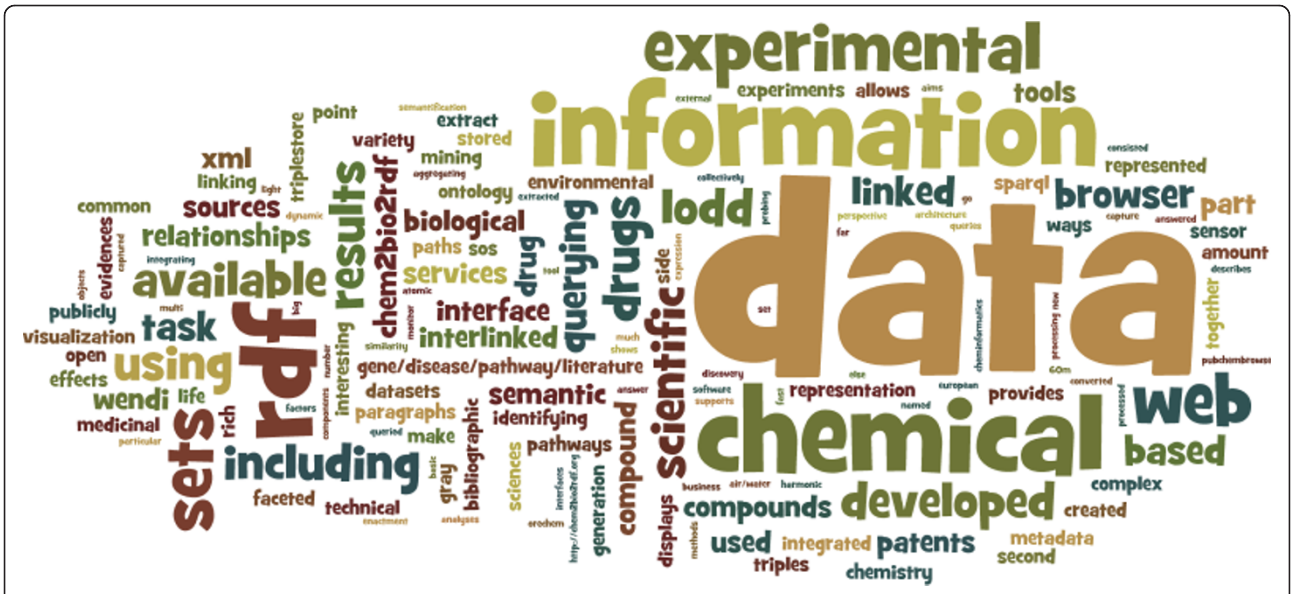

Figure 3 Keyword cloud for the RDF and Chemical Applications session.

and pharmaceutical knowledge to the Semantic Web by expressing it in RDF. These three projects aim at making databases with chemical knowledge available from a central access point, interlinking the individual data sets. Smaller data sets are also becoming available as RDF, such as the Open Notebook Science Solubility data [12].

2 Formats

The actual use of RDF depends on various further standards. For example, standards were required that describe how RDF statements are exchanged. Several standards serve this purpose: RDF/XML is an XML-based serialization [13], while simpler formats exist with N-Triples [14] and Notation3 [15]. For integration with current web practices, RDFa has been defined to allow RDF triples to be embedded in HTML pages [16]. Additionally, a proposal has been written that describes how RDF can be serialized as Javascript Object Notation (JSON) [17], and while this is not a formal specification yet, a new RDF working group will formalize this into a new standard [18]. Several of these serialization standards are used in the papers in this Series.

Using these serializations, RDF can be downloaded directly from pure RDF documents (RDF/XML, Notation3), or extracted from RDFa-based web pages using online RDF extraction web services, like http://www.w3.org/2007/08/pyRdfa/. These approaches make it simple to aggregate chemical data from web pages.

Table 1 Key W3C Specifications

Year	Technology	Description
1999	RDF	Resource Description Framework (RDF) Model and Syntax Specification [6]
2004	RDF/XML	RDF/XML Syntax Specification (Revised) [13]
	RDF	Resource Description Framework (RDF): Concepts and Abstract Syntax [32]
	OWL	OWL Web Ontology Language Overview [33]
2007	OWL2	OWL 2 Web Ontology Language Document Overview [33]
2008	RDFa	RDFa in XHTML: Syntax and Processing [16]
	SPARQL	SPARQL Query Language for RDF [19]

Several of the key specifications and when they were recommended by the World Wide Web Consortium.

3 Querying the World Wide Web

The most promising technology in the RDF family is the SPARQL Protocol and RDF Query Language (SPARQL) [19], which has been applied by Chen et al. in three chemogenomics use cases [10]. One of the use cases shows how SPARQL queries are used to find compounds that are active in bioassay for genes related to proteins to which the chemical dexamethasone binds, using information from PubChem, Uniprot, and DrugBank, all made available as RDF in the Chem2Bio2RDF database. The other use cases in this paper use the same approach by aggregating data sources before querying them. As such, it is similar to querying data stored in a relational database. However, an important difference between SPARQL and SQL query engines is the underlying data they act on: a graph of triples for RDF data, and rectangular tables in relational databases. This difference implies that RDF resources must have at least some common elements, whereas a relational DBMS assumes an identical data structure for all records of a table.

```
PREFIX cc: <http://github.com/egonw/cheminformatics.classics/1/#>
SELECT *
FROM
<http://www.w3.org/2007/08/pyRdfa/extract?
uri=http%3A%2F%2Fegonw.github.com%2Fcheminformatics.classics%2Fclassic1.html&forma
t=pretty-xml&warnings=false&parser=lax&space-preserve=true>
{
    ?mol cc:p0 ?p ;
         cc:t0 ?t .
}
```

Plot Results

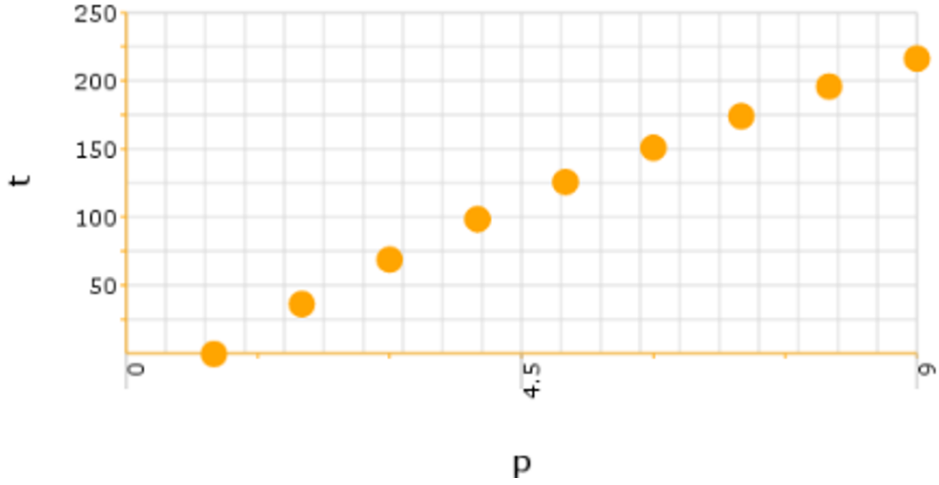

Figure 4 Web page using SPARQL to visualize alkane boiling points extracted from another web page. Web page with JavaScript by Jankowski visualizing the boiling point of a series of alkanes from Wiener [31] extracted with SPARQL from a second, XHTML+RDFa web page at http://egonw.github.com/cheminformatics.classics/classic1.html

For example, Jankowski used a public SPARQL service to extract boiling points of a series of alkanes from an XHTML webpage with the data made machine readable with RDFa, and visualized that using Javascript in another web page dynamically [20] (see Figure 4). A second important difference is that SPARQL queries can be *federated* [21]. Federated SPARQL allows one to query various RDF providers in one query. This has been used recently in the Receptor Explorer tool to help translational research by connecting basic neuroscience research with clinical trials [22]. Being able to query resources in this manner, brings us a step closer to systems biology approaches.

4 Ontologies

With RDF we have a data structure to link resources and provide details about those resources, and SPARQL provides us with the tools to query and aggregate that data. The next standard we will discuss now is the Web Ontology Language (OWL) which brought the RDF technology to the ontology community [23]. Ontologies are most certainly not new to chemistry [24] nor biology or life sciences, but the OWL standard makes it much easier to use ontologies, partly because they are formulated in RDF themselves. Ontologies, like controlled vocabularies and thesauri, describe what things mean, by linking terms to a human-readable definition. As such, ontologies are used for sharing knowledge in a common language, as well as to organize that knowledge. While linking resources is not new either, expressing the content of resources in explicit terms allows humans and software to reason formally on the content and to find possible sources of error. For example,

Konyk et al. have used OWL to link PubChem, DrugBank, and DBPedia, noting that it offers new ways to discover knowledge [25].

There are currently not many ontologies in chemistry, but many OBO Foundry-based ontologies can be reused using an OBO to OWL mapping [26]. This makes available chemical ontologies like the CO ontology [27], the ontology of Chemical Entities of Biological Interest (ChEBI) [28,29], and the Chemical Information Ontology http://code.google.com/p/semanticchemistry/, but also other ontologies in the life sciences, such as the Gene Ontology [30]. This way, OWL provides a universal standard to link data sources in life sciences, transcending traditional boundaries between the various domains.

The current state is that different RDF resources are using different ontologies. This does not necessarily have to be a problem, because the ontologies can be explicitly mapped to each other. This way, equivalent terms from two ontologies can be formally defined as equivalent, using the OWL predicates owl:equivalentClass and owl:equivalentProperty for classes, and owl:sameAs for instance. Making the equivalence explicit this way helps to illustrate the provenance of data integration efforts.

5 Discussion

This Thematic Series shows the current state of the use of RDF in chemistry, as presented at the ACS RDF 2010 meeting in Boston, and provides an insight into the progress of these methods. Much of the research is currently explorative, rather than formative, though standards are being proposed. It may very well turn out that some aspects of chemistry will never be expressed in RDF, and some computation will be done without ontology-based reasoning. It is important to realize here where RDF is positioned, namely for linking resources.

However, the use of RDF for already well-defined data structures in chemistry is not obvious. Data types like connection tables and various matrices are possible, but the use of URIs makes such structures needlessly verbose. Moreover, there is no need to format already well-formalized data structures into RDF, such as the various uses of matrices in computational chemistry as RDF triples. In fact, several papers in this series outline how to combine knowledge expressed with RDF with computational services. This shows that RDF is not an isolated framework, but one that can be integrated into existing cheminformatics workflows.

What RDF does not solve, are the following issues that remain in cheminformatics. RDF is about knowledge representation, and while ontologies take care of meaning and provide requirements to verify formal data consistency, it does not enforce any data quality, data structure, or data availability. This is, in fact, similar to other ways of providing data. For example, a data set with boiling points may or may not include information about experimental error. Metabolomics data may name the molecules for which concentration profiles have been measured, or the original accurate masses from which the identity was deduced.

It must be clear, therefore, that the RDF technologies are not the solution to everything. Their use does not guarantee an impressive scientific scenario. Instead, it can help simplify data analysis and particularly data integration, making it easier to handle large volumes of data accurately, or at least, with an explicitly defined accuracy.

As such, the use of explicit, semantic formats can be considered a gold standard of scientific practise. It is about adding as much detail to your lab notebook as you need. But, it does not inhibit you from writing nonsense in your notebook.

6 Outlook

The future of the use of RDF technologies as open standards in chemistry looks bright, and fills the needs in chemistry for semantically linking chemical data to other data sources. RDF technologies provide a domain-independent way for representing knowledge and their open nature assures many alternative approaches for making data available as RDF. This Thematic Series shows a few novel and creative applications of these RDF technologies, and we hope they may serve as seminal work in cheminformatics for future years.

Author details
[1]Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, SE-17177 Stockholm, Sweden. [2]Chemistry Biology Pharmacy Information Center, ETH Zürich, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland.

Received: 19 April 2011 Accepted: 13 May 2011 Published: 13 May 2011

References

1. Splendiani A, Burger A, Paschke A, Romano P, Marshall M: **Biomedical semantics in the Semantic Web.** *J Biomed Semantics* 2011, **2(Suppl 1)**:S1.
2. Murray-Rust P, Rzepa HS, Williamson MJ, Willighagen EL: **Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators.** *J Chem Inf Comput Sci* 2004, **44(2)**:462-469.
3. Frey JG: **Dark Lab or Smart Lab: The Challenges for 21st Century Laboratory Software.** *Org Process Res Dev* 2004, **8(6)**:1024-1035.
4. Hunter J, Henderson M, Khan I: **Collaborative Annotation of 3D Crystallographic Models.** *J Chem Inf Model* 2007, **47(6)**:2475-2484.
5. **Wordle.** [http://www.wordle.net/].
6. Lassila O, Swick RR: **Resource Description Framework(RDF) Model and Syntax Specification.** 1999 [http://www.w3.org/TR/1999/REC-rdf-syntax-19990222].
7. Bradley JC, Neylon C, Guha R, Williams A, Hooker B, Lang ASID, Friesen B, Bohinski T, Bulger D, Federici M, Hale J, Mancinelli J, Mirza KB, Moritz MJ, Rein D, Tchakounte C, Truong HT: **Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents.** *Nature Precedings* 2010 [http://dx.doi.org/10.1038/npre.2010.4243.2].

8. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web.** *Sci Am* 2001, **284(5)**:34-43.
9. Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J: **Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.** *J Biomed Inf* 2008, **41(5)**:706-716.
10. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild D: **Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data.** *BMC Bioinf* 2010, **11**:255.
11. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, Girschick T, Buchwald F, Wicker J, Karwath A, Gutlein M, Maunz A, Sarimveis H, Melagraki G, Afantitis A, Sopasakis P, Gallagher D, Poroikov V, Filimonov D, Zakharov A, Lagunin A, Gloriozova T, Novikov S, Skvortsova N, Druzhilovsky D, Chawla S, Ghosh I, Ray S, Patel H, Escher S: **Collaborative development of predictive toxicology applications.** *J Cheminf* 2010, **2**:7.
12. Bradley JC, Guha R, Lang A, Lindenbaum P, Neylon C, Williams A, Willighagen EL: *Beautifying Data in the Real World, Sebastopol, US: O'Reilly Media, Inc* 2009, **chap 16.**
13. Beckett D: **RDF/XML Syntax Specification (Revised).** 2004 [http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/].
14. Beckett D, Grant J: **RDF Test Cases.** 2004 [http://www.w3.org/TR/2004/REC-rdf-testcases-20040210/].
15. Berners-Lee T: **Notation 3 - An readable language for data on the Web.** 2006 [http://www.w3.org/DesignIssues/Notation3.html].
16. Adida B, Birbeck M, McCarron S, Pemberton S: **RDFa in XHTML: Syntax and Processing.** 2008 [http://www.w3.org/TR/2008/REC-rdfa-syntax-20081014/], W3C.
17. Alexander K: **RDF in JSON.** *Proceedings of the 4th Workshop on Scripting for the Semantic Web, Tenerife, Spain, June 02, 2008, Volume 368 of CEUR Workshop Proceedings* 2008.
18. Herman I: **New RDF Working Group, RDF/JSON, RDF API...** 2010 [http://www.w3.org/QA/2010/12/new_rdf_working_group_rdfjson.html].
19. Prud'hommeaux E, Seaborne A: **SPARQL Query Language for RDF.** 2008 [http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/].
20. Jankowski O: **SPARQL to Chart.** 2010 [http://www.pharmash.com/posts/2010-09-27-sparql-to-chart.html].
21. Prud'hommeaux E: **Federated SPARQL.** 2007 [http://www.w3.org/2007/05/SPARQLfed/].
22. Cheung KH, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, Paschke A: **A journey to Semantic Web query federation in the life sciences.** *BMC Bioinf* 2009, **10(Suppl 10)**:S10.
23. McGuinness DL, Van Harmelen F: **OWL Web Ontology Language Overview.** 2004 [http://www.w3.org/TR/2004/REC-owl-features-20040210/].
24. Gordon JE: **Chemical inference. 3. Formalization of the language of relational chemistry: ontology and algebra.** *J Chem Inf Comput Sci* 1988, **28(2)**:100-115.
25. Konyk M, De Leon A, Dumontier M: **Chemical Knowledge for the Semantic Web.** *J Chem Inf Comput Sci* 2008, 169-176.
26. Moreira DA, Musen MA: **OBO to OWL: a protege OWL tab to read/save OBO ontologies.** *Bioinformatics (Oxford, England)* 2007, **23(14)**:1868-1870.
27. Feldman HJ, Dumontier M, Ling S, Haider N, Hogue CWV: **CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules.** *FEBS Lett* 2005, **579(21)**:4685-4691.
28. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest.** *Nucleic Acids Res* 2008, **36(suppl 1)**:D344-D350.
29. Hull D: **GO faster ChEBI with Reasonable Biochemistry.** *Nature Precedings* 2008 [http://dx.doi.org/10.1038/npre.2008.2435.1].
30. Aranguren M, Bechhofer S, Lord P, Sattler U, Stevens R: **Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL.** *BMC Bioinf* 2007, **8**:57.
31. Wiener H: **Structural Determination of Paraffin Boiling Points.** *J Am Chem Soc* 1947, **69**:17-20.
32. Carroll JJ, Klyne G: **Resource Description Framework (RDF): Concepts and Abstract Syntax.** 2004 [http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/].
33. W3C OWL Working Group: **OWL 2 Web Ontology Language Document Overview.** 2009 [http://www.w3.org/TR/2009/REC-owl2-overview-20091027/].

doi:10.1186/1758-2946-3-15
Cite this article as: Willighagen and Brändle: **Resource description framework technologies in chemistry.** *Journal of Cheminformatics* 2011 **3**:15.