

Linking Open Drug Data to Cheminformatics and Proteochemometrics

Egon L. Willighagen, Jarl E.S. Wikberg

Uppsala University, Department of Pharmaceutical Biosciences,
Box 591, SE-751 24 Uppsala, Sweden
{egon.willighagen, jarl.wikberg}@farmbio.uu.se
<http://www.farmbio.uu.se/>

Abstract. Semantic Web technologies have made great steps forward in data exchange in health care and life sciences in the past years. The work presented here focuses to a some extent on making drug discovery related data available as RDF, and even more so on the integration of RDF approaches with data analysis of molecular information in drug discovery fields like cheminformatics and proteochemometrics. We here show how the chem- and bioinformatics workbench Bioclipse and the Chemistry Development Kit can be used to this purpose.

Key words: Bioclipse, cheminformatics, Resource Description Framework, Chemistry Development Kit, proteochemometrics

1 Introduction

Molecular chemometrics is the field that finds patterns in molecular information and combines methods from statistics and machine learning, cheminformatics, and also includes semantic technologies for lossless exchange of data [1]. While past research in this field focused mostly on the former two, the latter is at least as important: the success of the first two depends very much on the ability to link created models to independent information for validation purposes, and the ability to make assumptions on the (chemical) validness of ones training data and models, numerically [2] as well as visually [3].

Semantic technologies thus play an important role, and the Chemical Markup Language (CML) has met this need in chemistry recently [4, 5]. The use of ontologies and reasoning has, however, been studied earlier than that; For example, Gordon used ontologies and reasoning for *chemical inference* [6].

Bridging semantic data exchange with computation is a current research area, and is acknowledged as important components to improve cheminformatics. The lack of Open Data (training and test data), Open Source (open box software), and Open Standards (understanding what the data means) are useful solutions here, and is promoted by, for example, the Blue Obelisk movement [7].

Clearly, Resource Description Framework (RDF) and derived technologies, including the Web Ontology Language (OWL) and the SPARQL query language, are extremely useful Open Standards. Additionally, the amount of Open Source

software that can use these standards have greatly risen over the past 10 years; these tools now provide the crucial building blocks to handle chemical data expressed in RDF and include Jena [8] and Virtuoso [9].

1.1 The Chemistry behind Drugs

A tremendous effort has been ongoing in recent years to make drug-related data available as RDF using Open Data licenses or by placing it in the *public domain*, allowing modification and redistribution of the data [10]. The Linking Open Drug Data Project [11] and Bio2RDF [12] are two such projects.

Understanding patterns in drug data is important in drug design, where molecular data is linked to chemical and physical properties and biochemical knowledge including binding affinities and ADMETox properties. The statistical modeling of molecular properties requires an efficient representation of the molecular structures, and this is the point where cheminformatics meets the semantic web technologies.

Proteochemometrics is the application of statistics for modeling ligand-protein interactions. It models the binding interaction as function of both the molecular structure of the ligands and the protein sequences. It has been used for many biochemical activities now, including HIV proteases [13], P450 enzymes [14], and G-protein coupled receptors [15]. This modeling approach pulls in a wide variety of data, where semantic technologies help us verify assumption (by explicit facts and strong metadata), validate models (by pulling in validation data) and allows mapping of models onto related data.

Research is therefore ongoing to further integrate semantic web technologies with cheminformatics and chemometrics, and this paper shows the integration of RDF technologies with several Open Source bio- and cheminformatics platforms: the Chemistry Development Kit (CDK) [16, 17], Bioclipse [18], Jmol [19] and JChemPaint [20]. These four libraries provide an extensive set of cheminformatics functionality. The CDK is a chem- and bioinformatics library providing both a universal data model as well as many low level cheminformatics algorithms, including calculation of unique identifiers, calculation of similarity between molecules, substructure search, structure diagram generation for making 2D diagrams, a 3D geometry generator, and (molecular) descriptor calculations. Jmol is a well-known 3D visualization tool, while JChemPaint allows drawing and editing of 2D diagrams.

Bioclipse integrates these tools (and other libraries) into a graphical workbench, and was recently extended by scripting support [21]. The latter makes it possible to share scripts, for example, via social web sites such as MyExperiment.org [22]. Scripts uploaded to this social website as *workflow* can be downloaded directly into Bioclipse again using the Bioclipse MyExperiment plugin. We believe this improves reproducibility of studies. This paper takes advantage of that functionality and gives a few example scripts that use the here presented new RDF functionality. Currently, Bioclipse supports a JavaScript environment into which additional functionality is injected, allowing the functionality in Bioclipse itself to be used directly from the JavaScript environment.

OpenMolecules RDF

About <http://rdf.openmolecules.net/?InChI=1/CH4/h1H4>

Identifier [info:inchi/InChI=1/CH4/h1H4](http://info.inchi.org/InChI=1/CH4/h1H4)

InChI InChI=1/CH4/h1H4

Source Chemical blogspace

Blog Discussion <http://chem-bla-ics.blogspot.com/2008/09/ubiquity-fun-entering-semantic-mark-up.html>

Blog Discussion <http://chem-bla-ics.blogspot.com/2007/09/taqainq-molecules-mashup-of-connotea.html>

PubChem CID [297](http://pubchem.ncbi.nlm.nih.gov/compound/297)

Name methane

Source ChEBI

ChEBI ID [CHEBI:16183](http://www.ebi.ac.uk/chebi/CHEBI:16183)

owl:sameAs <http://bio2rdf.org/chebi:16183>

Source DBPedia

owl:sameAs <http://dbpedia.org/resource/Methane>

Source NMRShiftDB

owl:sameAs <http://pele.farmbio.uu.se/nmrshiftdb/?moleculeId=20029286>

NMRShiftDB mol ID [20029286](http://pele.farmbio.uu.se/nmrshiftdb/?moleculeId=20029286)

[Bookmark: this on del.icio.us](http://del.icio.us)



Fig. 1. Screenshot of the <http://rdf.openmolecules.net/> website for methane, showing an RDF/XML document visualized by the browser with the associated XSLT stylesheet. Links are made to various resources, showing how the website can serve as hub for linking molecular data using the InChI.

2 Applications

The applications in this paper give a few examples of the integration of RDF with cheminformatics. The first example shows the use of the IUPAC International Chemical Identifier (InChI) as unique identifier for molecular structures. This is followed by two examples where data from the Linked Open (Drug) Data network is interpreted and visualized in Bioclipse. The last example shows how RDF is being integrated into the CDK as a new, semantic IO format for its data model, as well as the output of molecular descriptor calculations.

2.1 Molecular Identity and Similarity

Chemical structures can be represented in various ways, but the chemical graph is the most popular in cheminformatics, being a fair trade-off between complexity and information content. However, comparing chemical graphs is computationally expensive, which is why identifiers are more commonly used instead. The SMILES [23] is a popular identifier used in the LODD network, but its identifiers are not unique. The InChI, however, is unique and increasingly used [24]. Even though it is not applicable to all chemical compound classes, it covers the major part of drugs on the market.

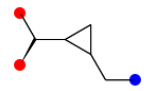
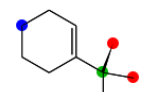
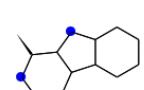
	2D-structure	DBPedia
1	 Generated	http://dbpedia.org/resource/%28%2B%29-cis-2-Aminomethylcyclopropane_carboxylic_acid
2	 Generated	http://dbpedia.org/resource/%28%2C%2C%2C%2C6-Tetrahydropyridin-4-yl%29methylphosphinic_acid
3	 Generated	http://dbpedia.org/resource/%28%2D%29-1-Methyl-2%2C3%2C4%2C9-tetrahydro-1H-pyrido-3%2C4-l

Fig. 2. Screenshot of DBPedia entries with SMILES retrieved with SPARQL and shown in a molecules table, created by a Bioclipse script.

The InChI has a format that includes a *InChI=* prefix, but it is still not in the URI format. To aid the adoption of the InChI in RDF data sets, we have set up a website that provides a one-to-one link between the InChI and a URI, moreover a URI that is dereferenceable, making it suitable for LinkedData networks. For example, Figure 1 shows the URI-based identifier for methane, <http://rdf.openmolecules.net/?InChI=1/CH4/h1H4>. The website does not primarily provide new data, but looks up information from other resources and links to those. The website provides autogenerated RDF content for any InChI. The existence of this website makes it possible for any data set to use *owl:sameAs* triples pointing to these URIs to mark the chemical identity of molecules. Currently, the website acts as a hub in the Linked Data network: links are provided to ChEBI, NMRShiftDB, and DBPedia.

2.2 Visualization: 2D diagrams

An obvious integration of the RDF network with cheminformatics toolkits is the visualization of 2D diagrams of the involved drugs. Bioclipse integrates the CDK and JChemPaint for these purposes and allows data from the LODD network to be read and visualized.

The following Bioclipse script shows this use case, and uses the SPARQL endpoint of DBPedia [25] as starting point. The script queries all entries which have a SMILES, because those are far more abundant than InChIs in Wikipedia, and uses the CDK to create an MDL SD file, while storing the DBPedia resource

URI as property. Clearly, any chemical property can be calculated on the fly, or looked up via additional RDF sources. The results are then opened in a JChemPaint-based molecule table functionality in Bioclipse 2.2 [21], as shown in Figure 2.

The full Bioclipse script for this application is given below, which is also available from MyExperiment.org at <http://www.myexperiment.org/workflows/927>. It shows how Bioclipse integrates RDF resources via the new *rdf* manager. The script first queries a remote SPARQL end point using the *rdf.sparqlRemote(sparql)* call, after which it iterates of all returned hits, extracts the *?compound* and *?smiles* fields for each hit as identified in the SPARQL. For each SMILES, the CDK is used to translate the SMILES into a chemical graph and stored in a list. The list is finally saved as MDL SD file:

```
var sparql = "\
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> \
PREFIX dbpedia: <http://dbpedia.org/ontology/> \
PREFIX dbprop: <http://dbpedia.org/property/> \
\
SELECT DISTINCT ?compound ?smiles WHERE { \
    ?compound dbprop:section ?section . \
    ?section dbprop:smiles ?smiles . \
} ORDER BY ?compound LIMIT 10 OFFSET 0 \
";

var hits = rdf.sparqlRemote("http://dbpedia.org/sparql", sparql);
var compounds = cdk.createMoleculeList()
for (var i=0; i<hits.size(); i++) {
    var hit = hits.get(i);
    var smiles = hit.get(0);
    smiles = smiles.replaceAll("\\s", "");
    if (smiles.endsWith("@en")) {
        smiles = smiles.substring(0, smiles.lastIndexOf('@'));
    }
    var resource = hit.get(1);
    var mol = cdk.fromSMILES(smiles);
    mol.setProperty("DBPedia", resource);
    compounds.add(mol);
}
cdk.saveSDFFile("/Virtual/dbpediaHits.sdf", compounds)
ui.open("/Virtual/dbpediaHits.sdf")
```

2.3 Visualization: 3D geometries

Likewise, Bioclipse can visualize 3D geometries too, using the plugin for Jmol [19]. The following script uses a SPARQL end point for the Bio2RDF data [12], and

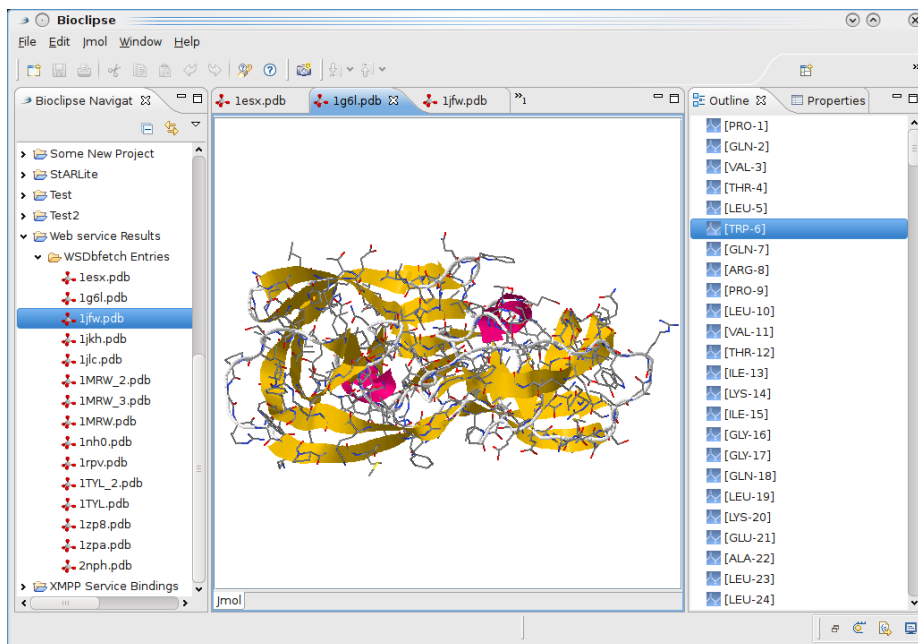


Fig. 3. Screenshot of a Jmol editor in Bioclipse showing a hit for the query against the Bio2RDF SPARQL endpoint for proteins with the string *HIV* in the title.

looks up protein structures which have a title containing *HIV*. The PDB identifier is extracted and used for a webservice call against the PDB database, and opened in the 3D editor.

The script is available at <http://www.myexperiment.org/workflows/928>, and queries a remote SPARQL end point finds all PDB identifiers which have the string *HIV* in its title. The found identifiers are then used to download the entry from the PDB database and opened in a Jmol viewer with a *ui.open()* call:

```
var sparql = "\
select distinct ?i where { \
?s <http://purl.org/dc/elements/1.1/title> ?o . \
?s <http://purl.org/dc/elements/1.1/identifier> ?i . \
FILTER regex(?o, \"HIV\") . \
FILTER regex(?i, \"pdb\") .} \
LIMIT 10";

var hits = rdf.sparqlRemote("http://quebec.bio2rdf.org/sparql", sparql);
for (var i=0; i<hits.size(); i++) {
  var hit = hits.get(i);
  var pdbID = hit.get(0);
  pdbID = pdbID.substring(pdbID.indexOf(":")+1);
```

```

    protein = webservices.downloadPDBAsFile(pdbID)
    ui.open(protein)
}

```

2.4 Molecular Properties and Descriptors

The two previous scripts serve as a starting point for proteochemometrics, and provides links between disease, protein sequences and drugs. The next integration step is to express data created with cheminformatics as RDF too, and in particular the expression of calculated molecular descriptors. For this purpose the data models used by the CDK as well as the Blue Obelisk Descriptor Ontology (BODO) are being expressed as OWL ontologies. This makes it possible to make calculation results part of the Linked Data network.

The following example shows protonated methanol serialized as Notation3 using the OWL-based CDK data model. It defines a molecule with two atoms, one of which is positively charged. Hydrogens are implicitly defined, as commonly done in SMILES too. The bond links to the bound atoms and is of a certain predefined bond order class. The resources in the RDF representation match Java Objects in the CDK library. These objects are typically not identified by URIs, causing the use of *example.com* in the below example. The current code that creates the RDF, however, allows to use an arbitrary domain, and we anticipate that pure URI fields for all Objects in the CDK will become available when the RDF representation becomes more popular. The Dublin Core namespace is reused for the name of the molecule, and an *owl:sameAs* triple linking to the aforementioned OpenMolecule RDF website is given:

```

<http://example.com/model1/atom1>
  a      cdk:Atom ;
  cdk:hasFormalCharge "1" ;
  cdk:symbol "O" .

<http://example.com/model1/atom2>
  a      cdk:Atom ;
  cdk:symbol "C" .

<http://example.com/model1/mol1>
  a      cdk:Molecule ;
  dc:title "Methanol" ;
  owl:sameAs <http://rdf.openmolecules.net/?InChI=1/CH4O/c1-2/h2H,1H3> ;
  cdk:hasAtom <http://example.com/model1/atom2> ,
              <http://example.com/model1/atom1> ;
  cdk:hasBond <http://example.com/model1/bond1> .

<http://example.com/model1/bond1>
  a      cdk:Bond ;
  cdk:bindsAtom <http://example.com/model1/atom1> ,

```

```
    <http://example.com/model1/atom2> ;
    cdk:hasOrder cdk:SingleBond .
```

The OWL-based CDK data model resembles the actually used data model, with its less common hierarchy. Nevertheless, the CDK is used for many different type of cheminformatics studies, showing that it does cover the domain quite sufficiently. The below subset of classes and properties shows the basic components for the representation of a chemical graph:

```
cdk:Atom
    a owl:Class ;
    rdfs:subClassOf cdk:AtomType .
cdk:Bond
    a owl:Class ;
    rdfs:subClassOf cdk:ElectronContainer .

cdk:AtomContainer
    a owl:Class ;
    rdfs:subClassOf cdk:ChemObject .

cdk:hasSymbol
    a owl:DatatypeProperty ;
    rdfs:domain cdk:Element ;
    rdfs:range <xsd:string> .
cdk:hasAtom
    a owl:ObjectProperty ;
    rdfs:domain cdk:AtomContainer ;
    rdfs:range cdk:Atom .
cdk:binds
    a owl:ObjectProperty ;
    rdfs:domain cdk:Bond ;
    rdfs:range cdk:Atom .
```

The here presented CDK data model ontology introduces yet another ontology, whereas the reuse of existing ontologies should be encouraged. However, an exact match of the internal CDK model has the advantage that one knows that documents using the ontology exactly express what is in the CDK data model. At the same time, it leaves the opportunity to use external ontological mappings to express class correspondence for use with reasoning engines. While the above example does not show this, 2D and 3D coordinates can easily be embedded in the RDF document too, thus demonstrating a simple example of how cheminformatics calculation results, 2D diagrams or 3D models in this case, can be represented as RDF triples.

Additionally, being expressed in RDF, it enables full chemical graphs to be embedded in HTML using XHTML+RDFa [26], for example to be used by user scripts [27]. It should also be noted that Bioclipse is capable of extracting RDF from XHTML+RDFa documents, making it a fully supported chemical format.

Calculated molecular descriptors can also be added to such RDF documents, and an extension has been written to the above RDF input/output library for the CDK to serialize those descriptors. Such semantic serialization of descriptors has been proposed earlier to use the Chemical Markup Language [17], and this approach is now extended to directly link to the Blue Obelisk Descriptor Ontology (which is expressed in OWL too), as well as to support listing what algorithm parameter values have been used in the descriptor calculation:

```
<http://example.com/model1/mol1>
  a      cdk:Molecule ;
  bodo:hasDescriptorValue
    [ a      bodo:DescriptorValue ;
      bodo:hasPart
        [ a      bodo:DescriptorValuePoint ;
          bodo:hasValue "0.0"^^xsd:double ;
          bodo:valuePointFor
            [ a      bodo:Descriptor ;
              rdfs:label "TopoPSA"^^xsd:string
            ]
          ] ;
      bodo:isCalculatedBy
        [ a      bodo:DescriptorImplementation ;
          <http://purl.org/dc/elements/1.1/identifier>
            "$Id$"^^xsd:string ;
          <http://purl.org/dc/elements/1.1/title>
            "org.openscience.cdk.qsar.descriptors. ...
            molecular.TPSADescriptor"^^xsd:string ;
          bodo:hasVendor "The Chemistry Development Kit"^^xsd:string ;
          bodo:instanceOf <bodo:tpsa>;
        ] ;
      bodo:isCalculatedWithParameter
        [ a      bodo:ParameterValue ;
          bodo:hasValue "false"^^xsd:boolean ;
          bodo:valueFor
            [ a      bodo:Parameter ;
              rdfs:label "checkAromaticity"
            ]
          ]
        ]
    ] .
```

This example shows the result of calculating the TPSA descriptor using the Chemistry Development Kit, in the RDF representation used by the Blue Obelisk Descriptor Ontology. The algorithm has one parameter which indicates if aromaticity should be detected before the descriptor is calculated, which was set to false. The triple graph also links to an external dictionary of descriptors that also uses Blue Obelisk Descriptor Ontology; in particular, it refers to the entry

describing the TPSA algorithm (*bodo:instanceOf bodo:tpsa*), allowing interoperability as described in the Blue Obelisk paper [7]. The descriptor listing and the underlying ontology are currently found in a single OWL document [28]. Bioclipse accepts further documents defining additional descriptors. Research is ongoing to how use this RDF in webservices using XMPP [29] and SADI [30].

3 Conclusion

This paper shows how RDF data can be integrated with cheminformatics and proteochemometrics using the Chemistry Development Kit and Bioclipse. While ontologies are not new in chemistry in itself, many current cheminformatics libraries do not yet have an RDF interface even though it addresses the important area of interoperability in the field. Our examples show how to go back and forth between a few common cheminformatics representations, including the SMILES, InChI and chemical graphs. The applications show how this link can be used to visualize chemical graphs present online in a line notation in the Linked Open Drug Data. The last example highlighted how cheminformatics calculation results can be represented in RDF. This opens up a world of possibilities for integrating cheminformatical computation into the RDF world, such as proposed by the SADI framework.

What this work does not address is the lack of Open Standard ontologies in chemistry; instead it introduces a new ontology. It is not anticipated that the here used ontologies are final, except perhaps for the CDK data model ontology, which is fixed to the CDK library design. Instead, we explore synchronization with other ontologies, such as ChemAxiom [31]. Additionally, ongoing research is exploring how this work can be linked to MetWare [32], an SKOS-based ontology for metabolomics experiments, focusing on metabolite identification.

Acknowledgments. This research was funded by a KoF grant from Uppsala University (KoF 07) and the Swedish VR-M (04X-05957).

References

1. Willighagen, E., Wehrens, R., Buydens, L.: Molecular Chemometrics. *Crit. Rev. Anal. Chem.* **36** (2006) 189–198
2. Willighagen, E., Denissen, H., Wehrens, R., Buydens, L.: On the use of ^1H and ^{13}C NMR spectra as QSPR descriptors. *Journal of Chemical Information and Modelling* **46**(2) (2006) 487–494
3. Willighagen, E.L., Wehrens, R., Melssen, W., de Gelder, R., Buydens, L.M.C.: Supervised self-organizing maps in crystal property and structure prediction. *Crystal Growth & Design* **7**(9) (September 2007) 1738–1745
4. Willighagen, E.L.: Processing CML conventions in Java. *Internet Journal of Chemistry* **4** (2001) 4+
5. Murray-Rust, P., Rzepa, H.S., Williamson, M.J., Willighagen, E.L.: Chemical markup, xml, and the world wide web. 5. applications of chemical metadata in rss aggregators. *J Chem Inf Comput Sci* **44**(2) (2004) 462–469

6. Gordon, J.E.: Chemical inference. 3. formalization of the language of relational chemistry: ontology and algebra. *Journal of Chemical Information and Computer Sciences* **28**(2) (May 1988) 100–115
7. Guha, R., Howard, M., Hutchison, G., Murray-Rust, P., Rzepa, R., Steinbeck, S., Wegner, J., Willighagen, E.: The Blue Obelisk - interoperability in chemical informatics. *Journal of Chemical Information and Modelling* **46** (2006) 991–998
8. McBride, B.: Jena: A semantic web toolkit. *IEEE Internet Computing* **6**(6) (2002) 55–59
9. Software, O.: Virtuoso Open-Source. <http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF>
10. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the semantic web. *BMC Bioinformatics* **8**(Suppl 3) (2007) S2+
11. Bizer, C., Jentzsch, A.: The Linking Open Drug Data Project. <http://sourceforge.net/projects/lodproject/>
12. Belleau, F., Nolin, M.A.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics* **41**(5) (October 2008) 706–716
13. Lapins, M., Eklund, M., Spjuth, O., Prusis, P., Wikberg, J.: Proteochemometric modeling of hiv protease susceptibility. *BMC Bioinformatics* **9**(1) (April 2008) 181+
14. Kontijevskis, A., Komorowski, J., Wikberg, J.E.: Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors. *Journal of chemical information and modeling* **48**(9) (September 2008) 1840–1850
15. Lapinsh, M., Prusis, P., Lundstedt, T., Wikberg, J.E.: Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Molecular pharmacology* **61**(6) (June 2002) 1465–1475
16. Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E.: The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **43**(2) (2003) 493–500
17. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., Willighagen, E.L.: Recent developments of the Chemistry Development Kit (CDK) - an open-source java library for chemo- and bioinformatics. *Current pharmaceutical design* **12**(17) (2006) 2111–2120
18. Spjuth, O., Helmus, T., Willighagen, E., Kuhn, S., Eklund, M., Wagener, J., Rust, P.M., Steinbeck, C., Wikberg, J.: Bioclipse: An open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* **8**(1) (2007)
19. Willighagen, E.L., Howard, M.: Fast and Scriptable Molecular Graphics in Web Browsers without Java3D. *Nature Precedings* (2007)
20. Krause, S., Willighagen, E., Steinbeck, C.: JChemPaint - using the collaborative forces of the internet to develop a free editor for 2D chemical structures. *Molecules* **5** (2000) 93–98
21. Spjuth, O., Alvarsson, J., Berg, A., Eklund, M., Kuhn, S., Mäsak, C., Torrance, G., Wagener, J., Willighagen, E., Steinbeck, C., Wikberg, J.: Bioclipse 2: A scriptable integration platform for the life sciences. Submitted (2009)
22. De Roure, D., Goble, C., Stevens, R.: The design and realisation of the my-experiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems* **25**(5) (May 2009) 561–567

23. Weininger, D.: SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **28** (1988) 31–36
24. Stein, S., Heller, S., Tchekhovski, D.: An open standard for chemical structure representation - The IUPAC Chemical Identifier. In: *Nimes International Chemical Information Conference Proceedings*. (2003) 131–143
25. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. (2008) 722–735
26. Adida, B., Birbeck, M., McCarron, S., Pemberton, S.: Rdfa in xhtml: Syntax and processing. <http://www.w3.org/TR/rdfa-syntax/>
27. Willighagen, E., O’Boyle, N., Gopalakrishnan, H., Jiao, D., Guha, R., Steinbeck, C., Wild, D.: Userscripts for the Life Sciences. *BMC Bioinformatics* **8**(1) (2007)
28. Willighagen, E.: descriptor-algorithms.owl. <http://qsar.svn.sf.net/viewvc/qsar/trunk/qsar-dicts/descriptor-algorithms.owl?revision=HEAD&view=markup>
29. Wagener, J., Spjuth, O., Willighagen, E.L., Wikberg, J.E.S.: XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous Web services. *BMC Bioinformatics* **10** (September 2009) 279+
30. Wilkinson, M., Vandervalk, B., McCarthy, L.: Sadi - semantic automated discovery and integration. <http://sadiframe.org/>
31. Adams, N., Cannon, E., Murray-Rust, P.: Chemaxiom an ontological framework for chemistry in science. *Nature Precedings* (2009)
32. Willighagen, E., Neumann, S., Van Ham, R.: Metware. <http://www.metware.org/>