Open Data, Open Source and Open Standards in chemistry - The Blue Obelisk five years on

Charles A Darwin*[1,2], Egon L Willighagen[3] , Rajarshi Guha[4] , Christoph Steinbeck[6] , Noel M O'Boyle[5] , Peter Murray-Rust[7] , Jane E Doe*[2] , Dmitry Pavlov[8]

[1]Life Sciences Department, Kings College London, Cornwall House,Waterloo Road, London, UK
[2]Department of Zoology, Cambridge, Waterloo Road, London, UK
[3]Division of Molecular Toxicology, Institute of Environmental Medicine, Nobels vaeg 13, Karolinska Institutet, 171 77 Stockholm, Sweden
[5]Analytical and Biological Chemistry Research Facility, Cavanagh Pharmacy Building, University College Cork, College Road, Cork, Co. Cork, Ireland
[6]Cheminformatics and Metabolism Team, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK [7]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, UK [8]GGA Software Services LLC, 41 Nab. Chernoi rechki 194342, Saint Petersburg, Russia

Email: Charles A Darwin*- charles@londonzoo.co.uk; Egon L Willighagen - egon.willighagen@ki.se; Rajarshi Guha - guhar@mail.nih.gov; Christoph Steinbeck - steinbeck@ebi.ac.uk; Noel M O'Boyle - n.oboyle@ucc.ie; Peter Murray-Rust - pm286@cam.ac.uk; Jane E Doe*- jane.e.doe@cambridge.co.uk; Dmitry Pavlov - dpavlov@ggasoftware.com;

*Corresponding author

Abstract

Background: The Blue Obelisk movement was established in 2005 as a response to the lack of open data, open standards and open source (ODOSOS) in chemistry. While other scientific disciplines such as physics, biology and astronomy (to name a few) were embracing new ways of doing science and reaping the benefits of community efforts, there was little if any innovation in the field of chemistry and scientific progress was actively hampered by the lack of access to data and tools.

Results: This contribution looks back on the past 5 years and surveys progress and remaining challenges in the areas of Open Source, Open Data and Open Standards in chemistry.

Conclusions: Here we show that the Blue Obelisk has been very successful in bringing together researchers and developers with common interests in ODOSOS, leading to development of many useful resources freely available to the chemistry community. But how best to engage with the wider chemistry community outside of the Blue Obelisk remains an open question.

Background

The Blue Obelisk movement was established in 2005 at the 229th National Meeting of the American Chemistry Society as a response to the lack of open data, open standards and open source (ODOSOS) in chemistry. While other scientific disciplines such as physics, biology and astronomy (to name a few) were embracing new ways of doing science and reaping the benefits of community efforts, there was little if any innovation in the field of chemistry and scientific progress was actively hampered by the lack of access to data and tools. Since 2005 it has become evident that a good amount of development in open chemical information and information is driven by the demands of neighbouring scientific fields. In many areas in biology, for example, the importance of small molecules and their interactions and reactions in biological systems has been realised. In fact, one of the first free and open databases and ontologies of small molecules was created as a resource about chemical structure and nomenclature by biologists [1]. Open workflow systems such as Taverna [2] and KNIME [3] have opened the door for an easy and seamless integration of FOS components into executable workflows or data pipelines. A combination of the Chemistry Development Kit and Taverna, for instance, was reported in 2010 [4].

KNIME currently has a built-in collection of CDK-based and OpenBabel-based nodes [5] for implementing chemistry workflows. Another collection of chemistry nodes—based on the RDKit and its Java wrapper—appeared in 2010 and is available to KNIME users from the KNIME's official "Community Updates" site. Indigo nodes for KNIME are in preparation and will appear nearby the RDKit nodes by May 2011.

The formation of the Blue Obelisk group is somewhat unusual in that it is not a funded network, nor does it follow the industry consortium model. Rather it is a grassroots organisation, catalysed by an initial core of interested scientists, but with membership open to all who share one or more of the goals of the group:

- Open Source. One can use other people's code without further permission, including changing it for one's own use and distributing it again.
- Open Standards. One can find visible community mechanisms for protocols and communicating information. The mechanisms for creating and maintaining these standards cover a wide spectrum of

human organisations, including various degrees of consent.

- Open Data. One can obtain all data in the public domain when wanted and reuse it for whatever purpose.

The Blue Obelisk was first described in the CDK News [**?**] and later as a formal paper by Guha et al. [6] in 2006. This contribution looks back on the past 5 years and surveys progress and remaining challenges in the areas of Open Source, Open Data and Open Standards in chemistry.

Open Source

Progress

Cheminformatics toolkits

Open Source toolkits for cheminformatics have now existed for nearly ten years. During this period, some toolkits were developed from scratch in academia, whereas others were made Open Source by releasing in-house codebases under liberal licenses. When the Blue Obelisk was established five years ago, the primary toolkits under active development were the Chemistry Development Kit (CDK) [7,8], Open Babel [9], and JOELib [10]. Of these, both the CDK and Open Babel continue to be actively developed. [Insert main focus of work on CDK in last 5 years] Since 2006, major new features of Open Babel include 3D structure generation and 2D structure-diagram generation, UFF and MMFF94 forcefields, and significantly expanded support for computational chemistry calculations. In addition, a major focus of Open Babel development has been to fix problems in the areas of stereochemistry, kekulisation and canonicalisation.

Two new Open Source cheminformatics toolkits have appeared since the original paper. In 2006 Rational Discovery, a cheminformatics service company (since closed down), released RDKit [11] under the BSD License. RDKit continues to be actively developed and includes code donated by Novartis.

More recently, GGA Software Services (a contract programming company) released the Indigo toolkit [12] and associated software in 2009 under the GPL. Indigo is written from scratch in C++ and has high-level wrappers for using it in C, Java, Python, or .NET environment. Like other toolkits, Indigo provides support for tetrahedral ans cis-trans stereochemistry, 2D coordinate generation, exact/substructure/SMARTS matching, fingerprint generation (all these both for molecules and reactions), and canonical SMILES computation. It also provides some less common functionality, like matching tautomers ans resonance substructures, enumeration of subgraps, finding maximum common substructure

of N input structures, and enumerating reaction products. Indigo also has its rendering plugin, which is used in a lot of commercial and open-sourse projects as the main 2D chemical rendering engine.
One important aspect of the above toolkits is that they originated for the support of desktop and server-side applications. More recently, a number of workers have developed light-weight toolkits that are targeted towards web applications. One example of this form is the ChemDoodle library from iChemLabs. The focus of these toolkits is to support client-side cheminformatics functionality, allowing for rich, chemically-aware browser-based applications.

Second-generation tools

Although feature-rich and robust cheminformatics toolkits are useful in and of themselves, they can also be seen as providing a base layer on which additional tools and applications can be built. This is one of the reasons that cheminformatics toolkits are so important to the open source 'ecosystem'; their availability lowers the barrier for the development of a 'second generation' of chemistry software that no longer needs to concern itself with the low-level details of manipulating chemical structures, and can focus on providing additional functionality and ease-of-use.
Bioclipse [13] (v1.0 released in Aug 2006) and Avogadro [14] (v1.0 in Oct 2009) are two examples of such software, based on the CDK and Open Babel, respectively. Bioclipse is an award-winning 'molecular workbench' for life sciences that ..., while Avogadro is a 3D molecular editor and viewer aimed at preparing and analysing computational chemistry calculations. An interesting aspect of both of these projects is that they share some developers with the underlying toolkit and this has driven the development of new features in the CDK and Open Babel.
CDK-Taverna [4] as an open-source cheminformatics workflow solution. Commonly used in this context are workflow engines for cheminformatics, where numerous recurring tasks can be automated, including tasks for chemical data filtering, transformation, curation and migration workflows, chemical documentation and information retrieval related workflows (structures, reactions, pharma- cophores, object relational data etc.) or data analysis workflows (statistics and clustering/ machine learning for QSAR, diversity analysis etc. CDK-Taverna has matured to become a freely available and increasingly powerful tool for the biosciences. It was recently ported to Taverna version 2.x. The combination of CDK-Taverna workers with the multitude of workers already published on myexperiment.org by the large and active user community of Taverna enables scientists to quickly build workflows to process a diverse set of data as typically found in today's systems biology scenarios.

Some other recent projects that build on Blue Obelisk software include AMBIT (a GUI that facilitates registration of chemicals for the REACH EU directive on toxicity, based on the CDK), etc.

The business end

Although the software developed by the Blue Obelisk is Open Source on principle, this does not rule out the establishment of a business based on open source chemistry software. Rather than supporting itself through sales, such a business generally relies on providing support, customisation and services around the software. The interaction between an open source project and a company is important, and works best where there is some form of 'giving back' to the project (such as developer time, code contributions, or financial assistance with server costs).

An example of such an interaction was the donation by eMolecules, Inc. to Open Babel of code for the canonicalisation of molecules and fragments (Nov 2006). eMolecules is an online vendor of chemicals that uses Open Babel under-the-hood to manage its compound collections. Another more recent example occurred in July 2010 when Silicos, a Belgian company that provides services in the area of cheminformatics, released several command line applications based on Open Babel as well as donating code to the project. For example, the Pharao tool released by them is a comprehensive solution for pharmacophore searching that provides extensive support for a variety of pharmacophore searches. Other tools released by them include Sieve for filtering by molecular property and XXX.

Another example of Open Source tools originating from commercial groups is the ChemCraft tool from Molecular Networks GmBH, that does XXX.

Rather than releasing new tools, an alternative approach to provide an interface to existing tools. hBar Solutions has developed an online portal, hBar Lab, for managing and performing computational chemistry calculations in the cloud. To do so it leverages two Blue Obelisk projects, Jmol and Open Babel, as well as the open source quantum mechanics packages MPQC.

Collaboration and interoperability

One of the effects of the Blue Obelisk has been to bring developers together from different Open Source chemistry projects so that they look for opportunities to collaborate rather than compete, and to leverage work done by other projects to avoid duplication of effort. As an example of this, when in March 2008 the Jmol development team were looking to add support for energy minimisation, rather than implement a forcefield from scratch they ported the UFF forcefield implementation from Open Babel to Jmol. This

code has allowed Jmol to support 2D to 3D conversion of structures (through energy minimisation). Similarly, efficient Jmol code for atom-atom rebonding has been ported to the CDK.

Another collaborative initiative between Blue Obelisk projects was the establishment in May 2008 of the ChemiSQL project. This brought together the developers of several open source chemistry database cartridges (PgChem, MyChem, OrChem and more recently Bingo) with a view to making their database APIs more similar and collaborating on benchmark datasets for assessing performance. For two of these projects, PgChem and MyChem, which are both based on Open Babel, there is the additional possibility of working together on a shared codebase.

In the area of cheminformatics tookits, two of the existing toolkits Open Babel and RDKit are planning to work together on a common underlying framework called MolCore. This project is still in the planning stage, but if it is a success it will mean that the the two libraries will be interoperable (while retaining their existing focus) but also that the cost of maintaining the code will be shared among more developers, freeing time for the development of new features.

One of the goals of the Blue Obelisk is to promote interoperability in chemical informatics. When barriers exist to moving chemical data between different software, the community becomes fragmented and there is the danger of vendor lock-in (where users are constrained to using a particular software, a situation which puts them at a disadvantage). This applies as much to Open Source software as to proprietary software. Cinfony is a project (first release in May 2008) whose goal is to tackle this problem in the area of cheminformatics toolkits [15]. It is a Python library that enables Open Babel, the CDK, and RDKit to be used using the same API; this makes it easy, for example, to read a molecule using Open Babel, calculate descriptors using the CDK and create a depiction using RDKit.

Remaining challenges

Accuracy. Very often software work at the 95% to 98% level. The variety of chemical structures is such that with a large enough dataset, 'unusual' structures are always found which may be mishandled by software. Given that much of the development of open source software is unfunded, and relies heavily on developer motivation, it is understandable that working on the final N% of problem structures (which may require substantial work) is not the most exciting of tasks. Putting a bounty system in place may be useful for these situations where a researcher needs a problem to be fixed because it affects a particular dataset of interest.

Performance. There is a famous quote by Knuth that "Premature optimization is root of all evil (or

something)". Given that compute time is cheap, ...

Chemical Structure Registration Systems, Databases

Databases? (XXX CS: A couple of database-related paragraphs could go here XXX)

XXX the paragraph below on OrChem needs some work Registration, indexing and searching of chemical structures in relational databases is one of the core areas of cheminformatics. A number of structure registration systems have been published in the last five years, exploiting the fact that free cheminformatics toolkits such as OpenBabel and the CDK were available. OrChem, for example, is an extension for the Oracle 11G database that adds registration and indexing of chemical structures to support fast substructure and similarity searching. The cheminformatics functionality is provided by the Chemistry Development Kit. OrChem provides similarity searching with response times in the order of seconds for databases with millions of compounds, depending on a given similarity cut-off. For substructure searching, it can make use of multiple processor cores on today's powerful database servers to provide fast response times in equally large data sets. OrChem is free software and can be redistributed and/or modified under the terms of the GNU Lesser General Public License as published by the Free Software Foundation. All software is available via http://orchem.sourceforge.net.

Open Standards

Progress

The IUPAC InChI identifier is a non-proprietary and unique identifier for chemical substances designed to enable linking of diverse data compilations. Although its development predates the Blue Obelisk, software such as Open Babel has included InChI support since 2005. Since the official InChI implementation is in C, it is difficult to access from the other widely used language for cheminformatics toolkits, Java. The Blue Obelisk project JNI-InChI has been set up to solve this problem by using the Java Native Interface to link the InChI binary to Java. In this way, it promotes the wider adoption of this standard identifier by the chemistry community.

Also, Indigo is going to provide its own plugin with an interface to the InChI implementation in May 2011. This will enable developers to calculate InChI and InChIKey-s within Python and .NET projects (not to mention Java), eliminating the trouble of compiling and using binary modules directly.

OpenSMILES

One of the most widely used ways to store chemical structures is the SMILES format (or SMILES string). This is a linear notation depicted by Daylight Information Systems that describes the connection table of a molecule and may optionally encode chirality. Its popularity stems from the fact that it is a compact representation of the chemical structure that is human readable and writable, and is convenient to manipulate (e.g. to include in spreadsheets, or copy from a Wikipedia article).

Despite its widespread use, a formal definition of the language did not exist beyond Daylight's SMILES Theory Manual and tutorials. This caused some confusion in the implementation and interpretation of corner cases, for example the handling of cis/trans bond symbols at ring closures. In 2007, Craig James (eMolecules) initiated work on the OpenSMILES specification, a complete specification of the SMILES language as an Open Standard developed through a community process. The specification is largely complete and contains guidelines on reading SMILES, a formal grammar, recommendations on standard forms when writing SMILES, as well as proposed extensions.

Recently proposed CurlySMILES [16] is an extension of the SMILES notation, which allows to define crystals structures, polymers, electron delocalization charges, molecule interations, and many other features absent in the initial SMILES specification. There has been discussions about including parts of the CurlySMILES notation into OpenSMILES, especially polymers.

CML?

Pistoia Alliance (and others?) should be mentioned at this point

Remaining challenges

A core requirement for chemical structure databases and chemical registration systems in general is the notion of structure standardization. That is, for a given input structure, multiple representations should be converted to one canonical form. Structure canonicalization routines partially address this aspect, converting multiple alternative topologies to a single canonical form. However, the problem of standardization is broader than just topological canonicalization. Features that must be considered include

- topological canonicalization
- handling of charges
- tautomer enumeration and canonicalization
- normalization of functional groups

Currently, most of the individual components of a 'standardization pipeline' can be implemented using BO tools. The larger problem is that there is no agreed upon list of steps for a standardization process. While some specifications have been published (e.g., Pubchem) and some standardization services and tools are available (Pubchem provides an online service to standardise molecules and the NCGC provides a stand alone tool) each group has their own set of rules. A common reference specification for standardization would be of immense value in interoperability between structure repositories as well as between toolkits (though the latter is still confounded by differences in lower level cheminformatic features such as aromaticity models).

We have already discussed the development of an Open SMILES standard. While much progress has been made towards a complete specification, more remains to be done before this can be considered finished. After that point, the next logical step would be to start work on a standard for the SMARTS language, the extension to SMILES that specifies patterns that match chemical substructures.

Open Data

Progress

A considerable stumbling block in advocating the release of scientific data as Open Data has been how exactly to define Open. A major step forward was the launch in 2010 of the Panton Principles for Open Data in Science [17]. This formalises the idea that Open Data maximises the possibility of reuse and repurposing, the fundamental basis of how science works. These principles recommend that published data be licensed explicitly, and preferably under CC0 (Creative Commons 'No Rights Reserved') [18]. This license allows others to use the data for any purpose whatsoever without any barriers.

Remaining challenges

Other areas of activity

Web services? (NIH, CDK and Chemspider)

Books (Angel, Egon, OB)

Conclusions

We have shown that the Blue Obelisk has been very successful in bringing together researchers and developers with common interests in ODOSOS, leading to development of many useful resources freely available to the chemistry community. But how best to engage with the wider chemistry community

outside of the Blue Obelisk remains an open question. If the Blue Obelisk is truly to make an impact, then an attempt must be made to reach beyond the subscribers to the BO mailing list and blogs of members. One recent development that attempts to address this issue was the establishment in Apr 2010 of a question-and-answer website related to Blue Obelisk projects and themes at http://blueobelisk.shapado.com. This is a website in the style of Stack Overflow that encourages high quality answers (and questions) through the use of a voting system. In the year since it was established, over 200 users have registered, many of whom had no previous involvement with the Blue Obelisk.
We hope to see this involvement between the Blue Obelisk and the wider community grow in future. To this end, we encourage the reader to visit http://blueobelisk.org, send a message to our mailing list, investigate related projects or read our blogs.

Authors contributions

Charles Darwin did all the work. The others stole the glory.

Acknowledgements

Thanks to everyone.

References

1. Matos PD, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical Entities of Biological Interest: an update**. *Nucleic Acids Research* 2009, :gkp886v1, [http://nar.oxfordjournals.org/cgi/content/full/gkp886v1].
2. Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T: **Taverna: a tool for building and running workflows of services**. *Nucleic Acids Research* 2006, **34**(Web Server):W729–W732.
3. **KNIME** [http://www.knime.org].
4. Kuhn T, Willighagen E, Zielesny A, Steinbeck C: **CDK-Taverna: an open workflow environment for cheminformatics**. *BMC Bioinformatics* 2010, **11**:159, [http://www.biomedcentral.com/1471-2105/11/159].
5. **KNIME** [http://www.knime.org/downloads/extensions].
6. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL: **The Blue Obelisk - Interoperability in Chemical Informatics**. *Journal of Chemical Information and Modeling* 2006, **46**(3):991–998, [http://dx.doi.org/10.1021/ci050400b].
7. 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**. *Journal of Chemical Information and Computer Sciences* 2003, **43**(2):493–500, [http://dx.doi.org/10.1021/ci025584y].
8. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL: **Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics.** *Current pharmaceutical design* 2006, **12**(17):2111–2120, [http://view.ncbi.nlm.nih.gov/pubmed/16796559].
9. **Open Babel** [http://openbabel.org].
10. **JOELib** [https://sourceforge.net/projects/joelib/].
11. **RDKit** [http://rdkit.org].

12. **Indigo** [http://ggasoftware.com/opensource/indigo].

13. Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C, Torrance G, Wagener J, Willighagen E, Steinbeck C, Wikberg J: **Bioclipse 2: A scriptable integration platform for the life sciences**. *BMC Bioinformatics* 2009, **10**:397+, [http://dx.doi.org/10.1186/1471-2105-10-397].

14. **Avogadro: an open-source molecular builder and visualization tool** [http://avogadro.openmolecules.net].

15. O'Boyle NM, Hutchison GR: **Cinfony–combining Open Source cheminformatics toolkits behind a common interface.** *Chemistry Central journal* 2008, **2**:24+, [http://dx.doi.org/10.1186/1752-153X-2-24].

16. Drefahl A: **CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures**. *Journal of Cheminformatics* 2011, **3**, [http://dx.doi.org/10.1186/1758-2946-3-1].

17. **Panton Principles - Principles for Open Data in Science** [http://pantonprinciples.org].

18. **About CC0 - "No Rights Reserved"** [http://creativecommons.org/about/cc0].

19. Rijnbeek M, Steinbeck C: **OrChem - An open source chemistry search engine for Oracle(R)**. *Journal of Cheminformatics* 2009, **1**:17+, [http://dx.doi.org/10.1186/1758-2946-1-17].

Figures

Figure 1 - Blue Obelisk logo

Figure 2 - Sample figure title

Figure legend text.

Tables

Table 1 - Blue Obelisk Open Source software projects

(Description if necessary XXXXXXXXXXXXXXX. Add citations to project names.)

Name	Website	Description
Cheminformatics toolkits		
Chemistry Development Kit (CDK) [7, 8]	http://cdk.sf.net	
Cinfony	http://cinfony.googlecode.com	Noel O'Boyle, P
Indigo	http://ggasoftware.com/opensource/indigo	G
Open Babel	http://openbabel.org	Geoffre
RDKit	http://rdkit.org	G
Integration		
CDK-Taverna [4]	http://cdk-taverna.blah.XXX	Christoph
A3	..	
Interconversion		
OSRA	http://osra.sf.net	Igor Filippo
A3	..	
Structure Databases		
OrChem [19]	http://orchem.sourceforge.net	Christoph Steinbeck, Ora
MyChem	http://XXX.sourceforge.net	Jerome Pansanel, MySQ
PGChem	http://XXX.sourceforge.net	Ernst-Georg Schmid, Postgr
Bingo	http://XXX.XXX.XXX	Dmitry Pavlov, Oracle-

Table 2 - Sample table title

Large tables are attached as separate files but should still be described here.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.