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

Jonathan Alvarsson[1] , Igor V Filippov*[2] , Rajarshi Guha[3] , Robert Hanson[4] , Geoffrey R Hutchison[5] , Daniel M Lowe[6] , Peter Murray-Rust[7] , Noel M O'Boyle[8] , Dmitry Pavlov[9] , Ola Spjuth[1] , Christoph Steinbeck[10] , Egon L Willighagen[11]

[1]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
[9]Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 751 24 Uppsala, Sweden
[1]0Department of Chemistry, University of Pittsburgh, 219 Parkman Avenue, Pittsburgh, PA 15260, USA
[1]1NIH, Somewhere in the US, USA

Email: Jonathan Alvarsson - jonathan.alvarsson@farmbio.uu.se; Igor V Filippov*- igorf@helix.nih.gov; Rajarshi Guha - guhar@mail.nih.gov; Robert Hanson - hansonr@stolaf.edu; Geoffrey R Hutchison - geoffh@pitt.edu; Daniel Lowe - dl387@cam.ac.uk; Peter Murray-Rust - pm286@cam.ac.uk; Noel M O'Boyle - n.oboyle@ucc.ie; Dmitry Pavlov - dpavlov@ggasoftware.com; ola.spjuth@farmbio.uu.se; Christoph Steinbeck - steinbeck@ebi.ac.uk; Egon L Willighagen - egon.willighagen@ki.se;

*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 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]

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 Data in Chemistry. One can obtain all scientific data in the public domain when wanted and reuse it for whatever purpose.
- Open Source in Chemistry. One can use other people's code without further permission, including changing it for one's own use and distributing it again.
- Open Standards in Chemistry. 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.

Note that while some may advocate also for Open Access to publications, the Blue Obelisk goals (ODOSOS) focus more on the availability of code (to reproduce results), standards (to exchange data), and the scientific data itself. All three of these goals stem from the fundamental tenants of the scientific method for data sharing and reproducibilty.

The Blue Obelisk was first described in the CDK News [2] and later as a formal paper by Guha et al. [3] in 2006. This contribution looks back on the past 5 years and surveys progress and remaining challenges in the areas of Open Data, Open Source, 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) [4, 5], Open Babel [6], and JOELib [7]. 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 provide for accurate conversion and representation in areas of stereochemistry, kekulisation, and canonicalisation. The project has also grown, in terms of new contributors, new support from commercial companies, and second-generation tools applying Open Babel to a variety of end-user applications, from molecular editors to chemical database systems.

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 [8] under the BSD License. This is a C++ library with Python and (more recently) Java bindings. RDKit is actively developed and includes code donated by Novartis. Recent developments include the Java bindings, as well as performance improvements for its database cartridge. More recently, GGA Software Services (a contract programming company) released the Indigo toolkit [9] and associated software in 2009 under the GPL. Indigo is a C++ library with high-level wrappers in C, Java, Python, and the .NET environment. Like RDKit and other toolkits, Indigo provides support for tetrahedral and cis-trans stereochemistry, 2D coordinate generation, exact/substructure/SMARTS matching, fingerprint generation, and canonical

SMILES computation. It also provides some less common functionality, like matching tautomers and resonance substructures, enumeration of subgraphs, finding maximum common substructure of N input structures, and enumerating reaction products.

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 [10] (v2.4 released in Aug 2010) and Avogadro [11] (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 which wraps cheminformatics functionality behind user-friendly interfaces and graphical editors while Avogadro is a 3D molecular editor and viewer aimed at preparing and analysing computational chemistry calculations. Both projects have are designed to be extended or scripted by users through a plugin architective and the provision of scripting support (using Bioclipse Scripting Language [12], or Python in the case of Avogadro). An interesting aspect of both Avogadro and Bioclipse is that they share some developers with the underlying toolkits and this has driven the development of new features in the CDK and Open Babel.

Bioclipse in turn acts as base for other software such as Brunn [13], a laboratory information system for microplate based high-throughput screening. Brunn provides a graphical interface for handling different plate layouts and dilution series and can automaticly generate dose response curves and calculate IC_{50}-values.

CDK-Taverna [14] 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, pharmacophores, 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.

Web applications

While desktop software has composed the majority of scientific tools since the computer was introduced, the internet continues to change how applications and content are distributed and presented. As such, there is a growing need to present chemical and scientific information in a web-based medium. Recently, a new version of the HTML specification, HTML5 [?], defines a well-developed framework for creating native web applications in JavaScript. This native framework provides a significant benefit to scientists as the web is an open and free medium to distribute scientific knowledge, ideas and education. It also poses significant new problems, as the cost associated with learning and developing tools for new technologies is not trivial. As such, it is an important goal for the Blue Obelisk to provide such tools and encourage the community to utilise these new technologies to improve products and education. A specific example of this initiative is the ChemDoodle Web Components library [?], produced by iChemLabs. The ChemDoodle Web Components library is open source under the GPL v3 license with a liberal HTML exception. ChemDoodle Web Components allow the scientist to present publication quality 2D and 3D graphics and animations for chemical structures, reactions and spectra. Beyond graphics, this tool provides a framework for user interaction to create dynamic applications through web browsers, desktop platforms and mobile devices such as the iPhone, iPad and Android devices. It is now easier than ever for scientists to create web-based scientific content, without having to master complicated technologies such as Java applets. iChemLabs is dedicated to funding, developing and supporting the library to ensure that the next generation of scientific applications is easily achievable by academia, government and industry. This goal helps to make sure that the cost of education decreases while using the web to further spread science.
Some projects have also been adopted by other fields, including biology and crystallography. A good example here is Jmol, which has, in the past five years, gone through two major revisions as part of an accelerated development focus. Starting with Jmol 10.2, released in April of 2006, Jmol 11.0 added over 100 new features and was released in March of 2007. Three years later, Jmol 12.0 followed with over 700 new features. Based on the outstanding work of Michael (Miguel) Howard and others, this development

path took Jmol from a “Rasmol/Chime replacement” to a fully fledged molecular visualisation package, including full support for crystallography [15], efficient delivery of surface data using fast solvent-excluded surface generation and the highly compressed JVXL format [http://chemapps.stolaf.edu/jmol/docs/misc/JVXL-format.pdf], display of molecular orbitals from standard basis set/coefficient data, the inclusion of dynamic minimisation using the UFF force field, and a full implementation of Daylight SMILES[ref] and SMARTS[ref], with extensions to conformational and biomolecular substructure searching in the form of Jmol BioSMARTS.[http://jmol.svn.sourceforge.net/viewvc/jmol/trunk/Jmol/src/org/jmol/smiles/package.html] Behind the scenes, Jmol has been essentially completely rewritten. A full mathematical expression evaluator was added, and the scripting language was folded into a versatile JavaScript-like command language, including a complete set of program flow commands, several new data types, and user-defined variables and functions. An exciting non-chemical application of Jmol has been its incorporation into the open source Sage Mathematics Project [http://www.sagemath.org/], where Jmol is being used to deliver three-dimensional depiction of mathematical functions.

The business end

Open Source provides a unique opportunity for commercial organisations to work with the community. As opposed to traditional business models which focus on maximising revenue through monetisation of all aspects of software, working with the community allows a business to maximise profits through greater community adoption and popularity. This is an ideal solution for some commercial organisations because only they can afford the significant investment in funding and time that is required to develop a successful open source project. Regardless, there are always challenges to growing a project, and there are specific issues that complicate making a business out of open source software in the scientific industry. A goal of the Blue Obelisk is to reduce these challenges and bring commercial organisations and the community together to benefit science.

Open source products attract interest because they are easily modifiable, and they can usually be used freely under a standard open source license. With proper advertising, a company can grow a user base more quickly than with a proprietary product, which requires significantly more marketing resources. Additionally, an open source product may integrate with other products that are governed by compatible open source licensing, allowing for faster development. Adopters benefit from a quickly established community of users, that not only know how to use that software, but may also have experience developing it. These benefits allow a company to create a successful product with significantly reduced cost and effort. Successful open projects require the dedication of a number of individuals to ensure that the product is continually developed, supported and funded. In many non-commercial cases, an author will produce an open project as a hobby, and stop work on the project before the development reaches a certain quality or community interest reaches a critical mass. This is because the hobbyist eventually needs to focus on income or he/she has a change of priorities. On the other hand, commercial organisations have the means to devote resources to a project to ensure that it achieves an appropriate level of quality and adoption. In return, that commercial organisation gains recognition from that product and is the first choice for commercial support, consulting and custom development. This is the primary goal of commercial open source, as selling licenses for an open product is very unlikely to be feasible.

In the scientific industry, it is incredibly difficult to start successful open projects. This is because the expertise needed to contribute to such a project requires a higher degree of education and the population of end users is much smaller than other industries, such as entertainment. Because of this, there is an increased barrier to starting open projects in the scientific industry. The Blue Obelisk attempts to solve this issue, by uniting those interested in open source and by helping to increase awareness for new open projects. Commercial organisations can take advantage of this goal, as open products will provide a faster return on investment, and the community benefits from open software and new job opportunities.

For instance, iChemLabs distributes several scientific software solutions under various open and proprietary licensing. The license is chosen to best ensure the success of the project. iChemLabs released the ChemDoodle Web Components library under the GPL v3 license with a liberal HTML exception. iChemLabs is dedicated to developing, funding and supporting the library so that the scientific community can immediately utilise the maturing HTML5 technologies and quickly make web and mobile interfaces for scientific content. Other open source developers can benefit from this library freely and expect continued development, while commercial organisations that use the library under an appropriate license (GPL for open projects or purchasing a proprietary license for proprietary projects) will benefit from an already developed market. Of course, no one will need to waste time rewriting functionality that is already provided with quality. Several open projects have already incorporated the ChemDoodle Web Components, such as iBabel by Chris Swain [**?**], ChemSpotlight by Geoffrey Hutchison [**?**] and the RSC ChemSpider [**?**]. 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 Silicos include Sieve for filtering by molecular property, Stripper for removing core scaffold structures from a molecule set, and Piramid for molecular alignment using shape determined by the Gaussian volumes as a descriptor.

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 is 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 package MPQC.

Converting chemical names and images to structures

The majority of chemical information is not stored in machine-readable formats, but rather as chemical names or depictions. The OSRA and OPSIN projects focus on extracting chemical information from these sources. Such software plays a particularly important role for data mining the chemical literature, including patents and theses.

Optical Structure Recognition Application (OSRA) [16] was started in early 2007 with the goal to create the first free and open source tool for extraction and conversion of molecular images into SMILES and SD files. From the very beginning the underlying philosophy was to integrate existing open source libraries and to avoid "reinventing the wheel" wherever possible. OSRA relies on a variety of open source components: Open Babel for chemical format conversion and molecular property calculations, GraphicsMagick for image manipulation, Potrace for vectorization, GOCR and OCRAD for optical character recognition. The growing importance of image recognition technology can be seen in the fact that only a few years ago there was only one widely available software package for chemical structure recognition - CLiDE (commercially developed at Keymodule, Ltd), but today there are as many as seven available programs.

The Blue Obelisk project OPSIN (Open Parser for Systematic IUPAC Nomenclature) [17] focuses instead on interpreting chemical names. The chemical name is the oldest form of communication used to describe chemicals, predating even the knowledge of the atomic structure of compounds. Chemical names are abundant in the scientific literature and encode valuable structural information. Through successive books of recommendations [18, 19], IUPAC has tried to codify and to an extent standardise naming practices. OPSIN aims to make this abundance of chemical names machine readable by translating them to SMILES, CML or InChI. The program is based around the use of a regular grammar to guide tokenisation and parsing of chemical names, followed by step-wise application of nomenclature rules. OPSIN is able to offer fast and precise conversions for the majority of names using IUPAC organic nomenclature, and is available as a web service, Java library and standalone application for maximum interoperability.

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 toolkits, 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 [20]. 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 optimisation is the 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 [21] is an extension of the SMILES notation, which allows to define crystals structures, polymers, electron delocalisation charges, molecule interactions, 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

?

QSAR-ML

The field of QSAR has long been hampered by the lack of open standards, which makes it difficult to share and reproduce descriptor calculations and analyses. QSAR-ML was recently proposed as an open standard for exchanging QSAR datasets [22]. A dataset in QSAR-ML includes the chemical structures (preferably described in CML) with InChI to protect integrity, chemical descriptors by linking to the Blue Obelisk Descriptor Ontology [23], response values, units, and versioned descriptor implementations to allow for integrating several descriptor software in the same calculation. Hence, a dataset described in QSAR-ML is completely reproducible. To allow for easy setup of QSAR-ML compliant datasets, a plugin for Bioclipse was created with graphical interfaces that can be used to set up QSAR datasets and perform calculations. Descriptor implementations were initially available from CDK and JOELib, as well as via remote web services such as XMPP [24].

Others...?

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 standardisation. That is, for a given input structure, multiple representations should be converted to one canonical form. Structure canonicalisation routines partially address this aspect, converting multiple alternative topologies to a single canonical form. However, the problem of standardisation is broader than just topological canonicalisation. Features that must be considered include

- topological canonicalisation
- handling of charges
- tautomer enumeration and canonicalisation
- normalisation of functional groups

Currently, most of the individual components of a ‘standardisation pipeline’ can be implemented using BO tools. The larger problem is that there is no agreed upon list of steps for a standardisation process. While some specifications have been published (e.g., PubChem) and some standardisation 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 standardisation 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 [25]. 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', also known as CCZero) [26]. This license allows others to use the data for any purpose whatsoever without any barriers. Other licenses compatible with the Panton Principles include the Open Data Commons Public Domain Dedication and Licence (PDDL), the Open Data Commons Attribution License, and the Open Data Commons Open Database License (ODbL) (see http://www.opendefinition.org/licenses/#Data).

Despite this positive news, little chemical data has become available from the traditional chemical fields of organic, inorganic, solid state chemistry. Table 2 lists a few notable exceptions: ChemPedia (a now discontinued crowd-sourcing project), CrystalEye (http://wwmm.ch.cam.ac.uk/crystaleye/), and the Open Notebook Science Solubility data [27]. There is also data available using licenses not compatible with the Panton Principles, but where the user is allowed to modify and redistribute the data. A new data set in this category is the data from the ChEMBL database, which is available under the Creative Commons Share-Alike Attribution license [28].

Importantly, publishing data as CC0 is becoming easier now that websites are becoming available to simplify publishing data. Two projects that can be mentioned in this context are FigShare (http://figshare.com/), where the data behind unpublished figures can be hosted, and Dryad (http://datadryad.org/) where data behind publications can be hosted. Initiatives like this make it possible to host small amounts of data, and those combined are expected to become soon a substantial knowledge base.

Remaining challenges

Other areas of activity

While each Blue Obelisk project has its own website and point of contact (typically a mailing list), because of the breadth of BO projects it can be difficult for a newcomer to understand which of them, if any, can best address a particular problem. To address this issue, members of the Blue Obelisk established a Question & Answer website at http://blueobelisk.shapado.com (see Figure 3). 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, showing that the Q&A website complements earlier existing channels of communication.

The rise of self-publishing and print-on-demand services has meant that publishing a book is now as straightforward as uploading to an appropriate website. Unlike the traditional publishing route where books with projected low sales volume would be expensive, websites such as Lulu allow the sale of low-priced books on chemistry software, and books are now available for purchase on Jmol [29], the Chemistry Development Kit [30] and Open Babel [31].

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. Figure 2 shows how the various Blue Obelisk projects collaborate. 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.

We hope to see this involvement between the Blue Obelisk and the wider community grow in the 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. Murray-Rust P: **The Blue Obelisk**. *CDK News* 2005, **2**(2):43–46.
3. 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].
4. 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].
5. 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].
6. **Open Babel** [http://openbabel.org].
7. **JOELib** [https://sourceforge.net/projects/joelib/].
8. **RDKit** [http://rdkit.org].
9. **Indigo** [http://ggasoftware.com/opensource/indigo].
10. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg JES: **Bioclipse: an open source workbench for chemo- and bioinformatics**. *BMC Bioinformatics* 2007, **8**:59.
11. **Avogadro: an open-source molecular builder and visualization tool** [http://avogadro.openmolecules.net].
12. 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].
13. Alvarsson J, Andersson C, Spjuth O, Larsson R, Wikberg JES: **Brunn: An open source laboratory information system for microplates with a graphical plate layout design process**. *BMC Bioinformatics* Accepted.
14. 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].
15. Hanson RM: **Jmol - a paradigm shift in crystallographic visualization**. *Journal of Applied Crystallography* 2010, **43**(5 Part 2):1250–1260, [http://dx.doi.org/10.1107/S0021889810030256].
16. **OSRA** [http://cactus.nci.nih.gov/osra/].
17. Lowe DM, Corbett PT, Murray-Rust P, Glen RC: **Chemical Name to Structure: OPSIN, an Open Source Solution**. *Journal of Chemical Information and Modeling* 2011, **51**(3):739–753, [http://dx.doi.org/10.1021/ci100384d].
18. IUPAC: *Nomenclature of Organic Chemistry*. Pergamon Press, Oxford 1979.

19. IUPAC: *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*. Blackwell Scientific publications 1993.

20. 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].

21. 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].

22. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JE: **Towards interoperable and reproducible QSAR analyses: Exchange of datasets**. *J Cheminform* 2010, **2**:5.

23. **The Blue Obelisk Descriptor Ontology**[http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml]. [http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml].

24. Wagener J, Spjuth O, Willighagen EL, Wikberg JES: **XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services**. *BMC Bioinformatics* 2009, **10**:279.

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

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

27. Bradley JC, Friesen B, Mancinelli J, Bohinski T, Mirza K, Bulger D, Moritz M, Federici M, Rein D, Tchakounte C, Bradley JC, Truong H, Neylon C, Guha R, Williams A, Hooker B, Hale J, Lang A, Bradley JC, Neylon C, Guha R, Williams AJ, 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, (713), [http://dx.doi.org/10.1038/npre.2010.4243.3].

28. Overington J: **ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr.** *Journal of computer-aided molecular design* 2009, **23**(4):195–198, [http://dx.doi.org/10.1007/s10822-009-9260-9].

29. Angel Herráez: *How to use Jmol to study and present molecular structures*, *Volume 1*. Lulu Enterprises, Morrisville, NC, US 2007.

30. Egon Willighagen: *Groovy Cheminformatics with the Chemistry Development Kit*. Lulu Enterprises, Morrisville, NC, US 2011.

31. Geoffrey R Hutchison, Chris Morley, Noel M O'Boyle, Craig James, Chris Swain, Hans De Winter, Tim Vandermeersch: *Open Babel - Official User Guide*. Lulu Enterprises, Morrisville, NC, US 2011.

32. 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 - Dependency diagram of Blue Obelisk projects.

Each block represents a project. Square blocks show Open Data, ovals are Open Source, and diamonds are Open Standards. Colors represent license: LGPL is green, GPL is orange, and BSD is blue.

Figure 3 - Screenshot of the Blue Obelisk eXchange Question and Answer website.

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) [4, 5]	http://cdk.sf.net	
Cinfony	http://cinfony.googlecode.com	Noel O'Boyle, F
Indigo	http://ggasoftware.com/opensource/indigo	G
Open Babel	http://openbabel.org	Geoffre
RDKit	http://rdkit.org	G
ChemDoodle Web Components	http://web.chemdoodle.com	
Integration		
CDK-Taverna [14]	http://cdk-taverna.blah.XXX	Christoph
A3	..	
Interconversion		
OSRA	http://osra.sf.net	Igor Filippo
OPSIN	http://opsin.ch.cam.ac.uk	Daniel Low
A3	..	
Structure Databases		
OrChem [32]	http://orchem.sourceforge.net	Christoph Steinbeck, Orac
MyChem	http://XXX.sourceforge.net	Jerome Pansanel, MySQ
PGChem	http://XXX.sourceforge.net	Ernst-Georg Schmid, Postgre
Bingo	http://XXX.XXX.XXX	Dmitry Pavlov, Oracle-

Table 2 - Open Data in chemistry.

Overview of major open chemical data available under a license or waiver compatible with the Panton Principles.

Name	License/Waiver	Description
ChemPedia	CC0	Crowd-sourced chemical names. Project discontinued.
CrystalEye	PPDL	Crystal structures from primary literature.
ONS Solubility	CC0	Solubility data for various solvents.

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.