

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

Samuel E Adams⁸, Jonathan Alvarsson¹, Igor V Filippov², Rajarshi Guha², Robert Hanson³, Marcus D Hanwell⁴, Geoffrey R Hutchison⁵, Karol M Langner⁶, Daniel M Lowe⁸, Peter Murray-Rust⁸, Noel M O'Boyle^{*9}, Dmitry Pavlov¹⁰, Ola Spjuth¹, Christoph Steinbeck¹¹, Adam L Tenderholt¹², Kevin J Theisen¹³, Egon L Willighagen¹⁴

¹Department of Pharmaceutical Biosciences, Uppsala University, Box 591, 751 24 Uppsala, Sweden

²NIH Center for Translational Therapeutics, 9800 Medical Center Drive, Rockville, MD 20878

³...

⁴Kitware, Inc., 28 Corporate Drive, Clifton Park, NY 12065, USA

⁵Department of Chemistry, University of Pittsburgh, 219 Parkman Avenue, Pittsburgh, PA 15260, USA

⁶Gorlaeus Laboratories, Leiden University, 2333 CC Leiden, The Netherlands

⁷Cheminformatics and Metabolism Team, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

⁸Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, UK

⁹Analytical and Biological Chemistry Research Facility, Cavanagh Pharmacy Building, University College Cork, College Road, Cork, Co. Cork, Ireland

¹⁰GGA Software Services LLC, 41 Nab. Chernoi rechki 194342, Saint Petersburg, Russia

¹¹...

¹²Department of Chemistry, University of Washington, Seattle, WA 98195

¹³iChemLabs, 200 Centennial Ave., Suite 200, Piscataway, NJ 08854, USA

¹⁴Division of Molecular Toxicology, Institute of Environmental Medicine, Nobels vaeg 13, Karolinska Institutet, 171 77 Stockholm, Sweden

Email: Samuel E Adams - sea36@cam.ac.uk; 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; Marcus D Hanwell - marcus.hanwell@kitware.com; Geoffrey R Hutchison - geoffh@pitt.edu; Karol M Langner - langnerkm@chem.leidenuniv.nl; 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; Adam L Tenderholt - adamlt82@u.washington.edu; Kevin J Theisen - kevin@ichemlabs.com; 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 Data, Open Standards, and Open Source 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 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.
- **Open Source in Chemistry.** One can use other people's code without further permis-

sion, including changing it for one's own use and distributing it again.

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

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.

Scope

The Blue Obelisk covers many areas of chemistry and chemical resources used by neighbouring disciplines (*e.g.* biochemistry, materials science). Many of the efforts relate to cheminformatics (the scope of this journal) and we believe that many of the publications in *J. Cheminf.* could be completely carried out using Blue Obelisk resources and other OS chemical tools. The importance of this is that for the first time it would allow reviewers, editors and readers to validate assertions in the journal and also to re-run and re-analyse parts of the calculation.

However, Blue Obelisk software and data is used outside cheminformatics and certainly in the five main areas that, for example, chemical markup language (CML) supports:

- **1.Molecules:** This is probably the largest area for Blue Obelisk software and data, and is reflected by many programs that visualise, transform, convert formats and calculate properties. It is almost certain that any file format

in use can be processed by Blue Obelisk software and properties can be calculated for most (organic compounds).

- **2.Reactions:** Blue Obelisk software can describe the semantics of reactions and provide atom-atom matching and analyse stoichiometric balance in reactions.
- **3.Computational chemistry:** Blue Obelisk software can interpret many of the current output files from calculations and create input for jobs. The Quixote project (see later and elsewhere in this issue) shows that Open Source approaches based on Blue Obelisk resources and principles are increasing the availability and re-usability of computational chemistry.
- **4.Spectra:** 1-D spectra (NMR, IR, UV etc.) are fully supported in Blue Obelisk offerings for conversion and display. There is a limited amount of spectral analysis but the software gives a platform on which it should be straightforward to develop spectral annotation and manipulation. Currently the Blue Obelisk does not support much in multi-dimensional NMR and multi-equipment spectra (*e.g.* GC-MS).
- **5.Crystallography:** The Blue Obelisk software supports the bi-directional processing of crystal structure files (CIF) and also solid-state calculations such as plane-waves with periodic boundary conditions. There is considerable support for the visualisation of both periodic and aperiodic condensed objects.

Many of the current operations in installing and running chemical computations and using the data are integration and customisation rather than fundamental algorithms. It is very difficult to create universal platforms that can be distributed and run by a wide range of different users, and in general, the Blue Obelisk deliberately does not address these. Our approach is to produce components that can be embedded in chemists' environments, including databases and workflows. We believe that a chemical laboratory with reasonable access to common software engineering should be able to build customised applications using Blue Obelisk components and standard infrastructure such as workflows (KNIME or Taverna) and databases (MySQL, CouchDB etc.). Where the Blue Obelisk itself produces data

resources (NMRShiftDB, CrystalEye, Quixote etc.) they are normally done with Open components so that the community can, if necessary, replicate them but not necessarily without additional effort.

Much of the impetus behind Blue Obelisk software is to create an environment for chemical computation (including cheminformatics) where all of the components, data, specifications, semantics, ontology and software are Openly visible and discussable. The largest current uses by the general chemical community are in authoring, visualisation and cheminformatics calculations but we anticipate that this will shortly extend into mainstream computational chemistry and solid-state. Although many of the authors are employed as research scientists, there are also several people who contribute in their spare time and we anticipate an increasing value and use of the Blue Obelisk in education at all levels.

Open Source

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.

The CDK project has been under regular development over the last five years. Several features have been implemented ranging from core components such as an extensible SMARTS matching system, new graph (& subgraph) isomorphism method [8] to more application oriented components such as 3D pharmacophore searching and matching, a variety of structural key and hashed fingerprints. In addition, there have been a number of second generation tools developed on top of the CDK such as OrChem [9] and PaDEL [10]. In addition to the use of the CDK in various applications, it has been deployed in the form of web services [11] and has formed the basis of a variety of web applications.

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 [12] 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 [13] 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. Although a wide range of chemistry software has been built using Blue Obelisk components

(see for example, the "Related Software" link on the Open Babel website, [14] listing over 40 projects as of this writing, or "Software using CDK" at the CDK website), in this section we focus on second-generation tools which themselves have been developed by members of the Blue Obelisk.

Bioclipse [15] (v2.4 released in Aug 2010) and Avogadro [16] (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 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 are designed to be extended or scripted by users through the provision of a plugin architecture and scripting support (using Bioclipse Scripting Language [17], 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.

Both products in turn act as extensible platforms for other software. Bioclipse, for example is used by software such as Brunn [18], 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 automatically generate dose response curves and calculate IC_{50} -values. Avogadro is used by Kalzium [19], a periodic table and chemical editor in KDE, and XtalOpt [20] [21], an evolutionary algorithm for crystal structure prediction. XtalOpt provides a graphical interface using Avogadro and submits calculations using a range of solid-state simulation software to predict stable polymorphs.

A final example of 2nd-generation Blue Obelisk software is the AMBIT2 [22] software, a GUI that facilitates registration of chemicals for the REACH EU directive on toxicity, and which is based on the CDK.

Computational chemistry analysis

Another area where the Blue Obelisk has had a significant impact in the past five years is in supporting quantum chemistry calculations and in interpreting their results. Electronic structure calculations have a long tradition in the chemistry community and a

variety of programs exist, mostly proprietary software with the notable exception of NWChem. [23] However, since each program uses different input formats, and the results are output differently even varying between different versions of the software, preparing calculations and automatically extracting the results is problematic.

Avogadro has already been mentioned as a GUI for preparing calculations. It uses Open Babel to read the output of several electronic structure packages. Avogadro generates input files on the fly in response to user input on forms, as well as allowing inline editing of the files before they are saved to disk. It also features intuitive syntax highlighting for GAMESS input files, allowing expert users to easily spot mistakes before saving an input file to disk.

In addition to this, significant development of new parsing routines took place in an Avogadro plugin to read in basis sets, and electronic structure output in order to calculate molecular orbital and electron density grids. This code was written to be parallel, using desktop shared memory parallelism and high level APIs in order to significantly speed up analysis. Most of this code was recently separated from the plugin, and released as a BSD licensed library, that is now used by the latest version of Avogadro, called OpenQube. Jmol (see below) can also depict computational chemistry results including molecular orbitals.

In 2006, the Blue Obelisk project cclib [24] was established with the goal of parsing the output from computational chemistry programs and presenting it in a standard way so that further analyses could be carried out independently of the quantum package used. cclib is a Python library, and the current version (version 1.0.1) supports 8 different computational chemistry codes and extracts over 30 different calculated attributes. Two related BO projects build upon cclib. GaussSum [25], is a GUI that can monitor the progress of SCF and geometry convergences, and can plot predicted UV/Vis absorption and infrared spectra from appropriate logfiles containing energies and oscillator strengths for easy comparison to experimental data. QMForge [26] provides a GUI for various electronic structure analyses such as Frenking’s charge decomposition analysis [27] and Mulliken or C-squared analyses on user-defined molecular fragments. QMForge also provides a rudimentary Cartesian coordinate editor allowing molecular structures to be saved via OpenBabel.

[Few lines on Quixote here]

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. The web presents new opportunities for scientists as it an open and free medium to distribute scientific knowledge, ideas and education. Web applications are software that runs within the browser, typically implemented in Java or JavaScript. Recently, a new version of the HTML specification, HTML5, defines a well-developed framework for creating native web applications in JavaScript and opens up new possibilities for visualising chemical data.

Jmol, the interactive 3D molecular viewer, is one of the most widely used chemistry applets, and indeed has seen widespread use in other fields such as biology and even mathematics (it is used for 3D depiction of mathematical functions in the Sage Mathematics Projects [28]). It is implemented in Java, and has gone from being a “Rasmol/Chime” replacement to a fully fledged molecular visualisation package, including full support for crystallography [29], 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 and SMARTS, with extensions to conformational and biomolecular substructure searching (Jmol BioSMARTS).

In 2009, iChemLabs released the ChemDoodle Web Components library [30] under the GPL v3 license (with a liberal HTML exception). This library is completely implemented in JavaScript and uses HTML5 to 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.

The business end

Open Source provides a unique opportunity for commercial organisations to work with the cheminformatics community. Traditional business models rely on monetisation of source code, causing every company to repeat the development by other companies,

often combined with a free (gratis) model for people working at academic institutes, to increase adoption and allow contributions from academics. This solution defines the IP on the software return on investment, but has the downside of investment losses due to duplication of software and method development, which becomes visible when proprietary companies merge. Some authors have rationalized that in the chemistry field, the data has value, and few contributors are available to volunteer time to improve codes. [31] If true, this would hamper adoption of Open Source and Open Data in chemistry, and greatly slow the growth of projects such as those in the Blue Obelisk.

The Blue Obelisk community, however, takes advantage that much of the investments needed for development are either paid by academic institutes and funding schemes, and by volunteers investing time and effort who get full access to the source code in return. In fact, all partners get full access up front and the Open Source licencing ensures that they will have access any time in the future. As such, it functions as a social contract between everyone to arrange the immediate return on investment. Effectively, this approach shares the burden of the high investment in having to develop cheminformatics software from scratch, allowing researchers and commercial partners alike to focus on their core business, rather than the development of prerequisites. As such, the rich collection of Open Source cheminformatics tools provided by the Blue Obelisk greatly reduces investment up front for new companies in the cheminformatics market. Such advantages have also been noted in the drug discovery field. [32] [33] [34]

The use of Open Standards allows everyone to select those Blue Obelisk components they find most useful, as they can easily replace one component with another providing the same functionality, taking advantage that they use the same standards for, for example, data exchange. This way, licensing issues are becoming a marginal problem, allowing companies to select a license appropriate for their business model. This too, allows a company to create a successful product with significantly reduced cost and effort.

At the time of writing there are many commercial companies development chemistry solutions around Open Source cheminformatics components provided by the Blue Obelisk community. Examples of such companies include iChemLabs, IdeaConsult, Wingu, Silicos, GenettaSoft, hBar, and Inkspot Sci-

ence. Some of these merely use components, but several actively contribute back to the Blue Obelisk project they use, or donate new Open Source cheminformatics projects to the community.

For example, iChemLabs released the ChemDoodle Web Components library under the GPL v3 license, based on the upcoming HTML5 Open Standard. It allows making web and mobile interfaces for chemical content. The project is already being adopted by others, including iBabel by Chris Swain [35], ChemSpotlight by Geoffrey Hutchison [36] and the RSC ChemSpider [37].

Silicos has release several Open Source utilities based on OpenBabel, such as Pharaos, a tool for pharmacophore searching, Sieve for filtering molecular structure by molecular property, Stripper for removing core scaffold structures from a molecule set, and Pyramid for molecular alignment using shape determined by the Gaussian volumes as a descriptor. Additionally, contributions have been made to the OpenBabel project itself.

Other companies use Blue Obelisk components and contribute patches, smaller and larger. For example, IXELES donated the isomorphism code in the CDK, eMolecules donated canonicalisation code to OpenBabel, and AstraZeneca contributed code to the CDK for signatures. This is just a very minor selection, and the author is encouraged to contact the individual Blue Obelisk projects for an elaborate list.

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) [38] 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 cal-

culations, 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.

OPSIN (Open Parser for Systematic IUPAC Nomenclature) [39] 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 [40, 41], 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. It 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.

Chemical Databases

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, [42] for example, is an open source 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 CDK. 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.

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 [43] 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, [44] MyChem, [45] OrChem [42] and more recently Bingo [46]) 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 [47]. 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 [48]. 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.

Another way through which interoperability of Blue Obelisk projects has been promoted and developed is through integration into workflow software such as Taverna [49] and KNIME [50] (both open source). Such software makes it easy to automate recurring tasks, and to combine analyses or data from a variety of different software and web services. A combination of the Chemistry Development Kit and Taverna, for instance, was reported in 2010 [51]. In the case of KNIME, it comes with built-in basic collection of CDK-based and Open Babel-based nodes, while other nodes for the RDKit and Indigo are available from KNIME’s “Community Updates” site.

Remaining challenges

(Say something here about benchmarks to measure accuracy. Clear examples of performance on open datasets are required. Otherwise there is nothing to counter anecdotal evidence or FUD spread by others.)

(Better engagement with industry...make it clear what and how industry members can engage with projects.)

Open Standards

Chemical Markup Language, CML

Chemical Markup Language (CML) is discussed in several places in this issue, and a brief summary here re-iterates that it is designed primarily to create a validatable semantic representation for chemical objects. The five main areas (molecules, reactions, computational chemistry, spectra and solid-state (see above)) have now all been extensively deployed and tested. CML can therefore be used as a reference for input and output for Blue Obelisk software and a means of representing data in Blue Obelisk resources.

CML, being an XML application, can interoperate with other markup languages and in particular XHTML, SVG, MathML, docx and more specialised applications such as UnitsML and GML (geosciences). We believe that it would be possible using these languages to encode large parts of, say, first year chemistry text books in XML. Similarly, it is possible to create compound documents with word processing or spreadsheet software that

has inter-operating text, graphics and chemistry (as in Chem4Word). Being a markup language, CML is designed for re-purposing, including styling, and therefore a mixture of these languages can be used for chemical catalogues, general publications, log-books and many other types of document in the scientific process.

CML describes much of its semantics through conventions and dictionaries, and the emerging ecosystem (especially in computational chemistry) is available as a semantic resource for many of the applications and specifications in this article.

InChI

The IUPAC InChI identifier is a non-proprietary and unique identifier for chemical substances designed to enable linking of diverse data compilations. Prior to the development of the InChI identifier chemical information systems and databases used a wide variety of (generally proprietary) identifiers, greatly limiting their interoperability. Although its development predates the Blue Obelisk, software such as Open Babel has included InChI support since 2005, and support for InChI in Indigo is due in 2011.

Since the official InChI implementation is in C, it is difficult to access from the other widely used language for cheminformatics toolkits, Java. Early attempts to generate InChI identifiers from within Java involved programatically launching the InChI executable and capturing the output, an approach that was found to be fairly unreliable and broke the ‘write once, run anywhere’ philosophy of Java. The Blue Obelisk project JNI-InChI [52] was established in 2006 to solve this problem by using the Java Native Interface framework to provide transparent access to the InChI library from within Java, and other Java Virtual Machine (JVM) based languages, supporting the wider adoption of this standard identifier by the chemistry community.

The Java Native Interface framework provides a mechanism for code running inside the JVM to place calls to libraries written in languages such as C, C++ and Fortran, and compiled into native, machine specific, code. JNI-InChI provides a thin C wrapper, with corresponding Java code, around the IUPAC InChI library, exposing the InChI library’s functionality to the JVM. To overcome the need to have the correct InChI library pre-installed on a system, JNI-InChI comes with a variety of precompiled native binaries and automatically extracts and deploys the

correct one for the detected operating system and architecture. The JNI-InChI library comes with native binaries supporting a range of operating systems and architectures; the current version has binaries for 32- and 64-bit Windows, Linux and Solaris, 64-bit FreeBSD and 64-bit Intel-based Mac OS X - a number of which are not supported by the original IUPAC distribution of InChI.

The JNI-InChI project has matured to support the full range of functionality of the InChI C library: structure-to-InChI, InChI-to-structure, AuxInfo-to-structure, InChIKey generation, and InChI and InChIKey validation. JNI-InChI provides the InChI functionality for a number of Open Source projects, including the Chemistry Development Kit, Bioclipse and CMLXOM/JUMBO, and is also used by commercial applications and internally in a number of companies. Through its widespread use and Open Source development model, a number of issues in earlier versions of the software have been identified and resolved, and JNI-InChI now offers a robust, and defacto standard tool for working with InChIs in the JVM.

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 [53] is an ex-

tension 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.

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 [54]. 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 [55], 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 [56].

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

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 [57]. 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) [58]. 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). [59]

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, [60] and the Open Notebook Science Solubil-

ity data [61]. 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 [62].

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 [63], where the data behind unpublished figures can be hosted, and Dryad [64] 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.

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 [65] (see Figure 3). This is a website in the style of Stack Overflow [66] 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 [67] allow the sale of low-priced books on chemistry software, and books are now available for purchase on Jmol [68], the Chemistry Development Kit [69] and Open Babel [70].

Conclusions

We have shown that the Blue Obelisk has been very successful in bringing together researchers and devel-

opers with common interests in ODOSSOS, 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 the Blue Obelisk website [71], 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** [http://sf.net/projects/joelib].
8. Rahman S, Bashton M, Holliday R GL an Schrader, Thornton J: **Small Molecule Subgraph Detector (SMSD) Toolkit.** *J. Cheminf.* 2009, **1**.
9. Rijnbeek M, Steinbeck C: **OrChem - An Open Source Chemistry Search Engine for Oracle.** *J. Cheminf.* 2009, **1**.
10. Yap C: **PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints.** *J. Comp. Chem.* 2011, **32**:1466–1474.
11. Dong X, Gilbert K, Guha R, Heiland R, Kim J, Pierce M, Fox G, Wild D: **A Web Service Infrastructure for Chemoinformatics.** *J. Chem. Inf. Model.* 2007, **47**(4):1303–1307.
12. **RDKit** [http://rdkit.org].
13. **Indigo** [http://ggasoftware.com/opensource/indigo].
14. **Open Babel - Related Software** [http://openbabel.org/wiki/Related_Projects].
15. 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.
16. **Avogadro: an open-source molecular builder and visualization tool** [http://avogadro.openmolecules.net].
17. 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].
18. Alvarsson J, Andersson C, Spjuth O, Larsson R, Wikberg J: **Brunn: An open source laboratory information system for microplates with a graphical plate layout design process.** *BMC Bioinformatics* 2011, **12**:179, [http://www.biomedcentral.com/1471-2105/12/179].
19. **Kalzium - Periodic Table and Chemistry in KDE** [http://edu.kde.org/applications/science/kalzium/].
20. **XtalOpt - Evolutionary Crystal Structure Prediction** [http://xtalopt.openmolecules.net].
21. Lonie DC, Zurek E: **XtalOpt: An open-source evolutionary algorithm for crystal structure prediction.** *Computer Physics Communications* 2011, **182**(2):372–387.
22. **AMBIT2** [http://ambit.sf.net/].
23. Valiev M, Bylaska EJ, Govind N, Kowalski K, Straatsma TP, van Dam HJJ, Wang D, Nieplocha J, Apra E, Windus TL, de Jong WA: **NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations.** *Comput. Phys. Commun.* 2010, **181**(9):1477–1489.
24. O’Boyle NM, Tenderholt AL, Langner KM: **cclib: A library for package-independent computational chemistry algorithms.** *J. Comp. Chem.* 2008, **29**(5):839–845.
25. **GaussSum** [http://gausssum.sf.net].
26. **QMForge** [http://qmforge.sf.net].

27. Dapprich S, Frenking G: **Investigation of Donor-Acceptor Interactions: A Charge Decomposition Analysis Using Fragment Molecular Orbitals.** *J. Phys. Chem.* 1995, **99**(23):9352–9362.
28. Stein W, et al.: *Sage Mathematics Software*. The Sage Development Team 2011. [<http://www.sagemath.org>].
29. 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>].
30. **ChemDoodle Web Components: HTML5 Chemistry** [<http://web.chemdoodle.com>].
31. Stahl MT: **Open-source software: not quite endsville.** *Drug Discov Today* 2005, **10**(3):219–22.
32. DeLano WL: **The case for open-source software in drug discovery.** *Drug Discov Today* 2005, **10**(3):213–7.
33. Munos B: **Can open-source R&D reinvigorate drug research?** *Nat Rev Drug Discov* 2006, **5**(9):723–9.
34. Geldenhuys WJ, Gaasch KE, Watson M, Allen DD, Van der Schyf CJ: **Optimizing the use of open-source software applications in drug discovery.** *Drug Discov Today* 2006, **11**(3-4):127–32.
35. **iBabel** [<http://homepage.mac.com/swain/Sites/Macinchem/page65/ibabel3.html>].
36. **ChemSpotlight** [<http://chemspotlight.openmolecules.net>].
37. **iChemLabs and RSC ChemSpider announce partnership** [<http://www.chemspider.com/blog/ichemlabs-and-rsc-chemspider-announce-partnership.html>].
38. **OSRA** [<http://cactus.nci.nih.gov/osra/>].
39. 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>].
40. IUPAC: *Nomenclature of Organic Chemistry*. Pergamon Press, Oxford 1979.
41. IUPAC: *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*. Blackwell Scientific publications 1993.
42. **OrChem** [<http://orchem.sf.net>].
43. Rappé AK, Casewit CJ, Colwell KS, III WAG, Skiff WM: **UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations.** *Journal of the American Chemical Society* 1992, **114**(25):10024–10035.
44. **PgChem** [<http://pgfoundry.org/projects/pgchem/>].
45. **MyChem** [<http://mychem.sf.net>].
46. **Bingo** [<http://ggasoftware.com/opensource/bingo>].
47. **MolCore** [<http://molcore.sf.net>].
48. 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>].
49. 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.
50. **KNIME** [<http://www.knime.org>].
51. 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>].
52. **JNI-InChI** [<http://jni-inchi.sf.net/index.html>].
53. 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>].
54. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JE: **Towards interoperable and reproducible QSAR analyses: Exchange of datasets.** *J Cheminform* 2010, **2**:5.
55. **The Blue Obelisk Descriptor Ontology** [<http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml>]. [<http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml>].
56. 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.
57. **Panton Principles - Principles for Open Data in Science** [<http://pantonprinciples.org>].
58. **About CC0 - “No Rights Reserved”** [<http://creativecommons.org/about/cc0>].
59. **Open Licenses - Data** [<http://www.opendefinition.org/licenses/#Data>].
60. **Crystal Eye** [<http://wwwmm.ch.cam.ac.uk/crystaleye/>].
61. 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>].
62. 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>].
63. **FigShare** [<http://figshare.com>].
64. **Dryad** [<http://datadryad.org>].
65. **Blue Obelisk - Shapado** [<http://blueobelisk.shapado.com>].
66. **Stack Overflow** [<http://stackoverflow.com>].
67. **Lulu** [<http://lulu.com>].
68. Herráez A: *How to use Jmol to study and present molecular structures, Volume 1*. Lulu Enterprises, Morrisville, NC, US 2007.

69. Willighagen E: *Groovy Cheminformatics with the Chemistry Development Kit*. Lulu Enterprises, Morrisville, NC, US 2011.
70. Hutchison GR, Morley C, O'Boyle NM, James C, Swain C, Winter HD, Vandermeersch T: *Open Babel - Official User Guide*. Lulu Enterprises, Morrisville, NC, US 2011.
71. **BlueObelisk** [<http://blueobelisk.org>].
72. NE D, P MR, SM T: **CIFXML: a schema and toolkit for managing CIFs in XML**. *J. Appl. Cryst.* 2011, **44**(3):628–634.
73. 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 XXXXXXXXXXXXXXXX. Add citations to project names.)

Name	Website
CML Tools	
JUMBO & CMLXOM	http://sourceforge.net/projects/cml/ https://bitbucket.org/wwmm/
Cheminformatics Tools	
Chemistry Development Kit (CDK) [4, 5]	http://cdk.sf.net
Cinfony	http://cinfony.googlecode.com
Indigo	http://ggasoftware.com/opensource/indigo
Open Babel	http://openbabel.org
RDKit	http://rdkit.org
Second Generation	
Avogadro	http://avogadro.openmolecules.net
Web Applications	
ChemDoodle Web Components	http://web.chemdoodle.com
Jmol	http://jmol.org
Integration	
CDK-Taverna [51]	http://cdk-taverna.blah.XXX
Lensfield2	https://bitbucket.org/sea36/lensfield2/
A3	..
Interconversion	
OSRA	http://osra.sf.net
OPSIN	http://opsin.ch.cam.ac.uk
JUMBO-Converters	https://bitbucket.org/wwmm/jumbo-converters/
CIFXOM	https://bitbucket.org/wwmm/cifxom/
A3	..
Structure Databases	
Bingo	http://XXX.XXX.XXX
MyChem	http://XXX.sf.net
OrChem [73]	http://orchem.sf.net
PGChem	http://XXX.sf.net
Chempound (Chem#)	https://bitbucket.org/chempound
Text mining	
OSCAR4	https://bitbucket.org/wwmm/oscar4/
ChemicalTagger	http://chemicaltagger.ch.cam.ac.uk/
Other	
cclib	http://cclib.sf.net

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.
NMRShiftDB	?	Database of NMR spectra of organic compounds.

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.