Niklas Blomberg
DECS Computational Chemistry,AstraZeneca R&D Mölndal, Sweden
Gerhard F. Ecker
Dept. Medicinal Chemistry, Univ. Vienna, Austria
Richard Kidd
Royal Society of Chemistry, Cambridge, UK
Barend Mons
Netherlands Bioinformatics Center
and Leiden University Medical Center, Leiden, The Netherlands.
Bryn Williams-Jones
Pfizer Worldwide R&D, Sandwich Research, UK

Knowledge Driven Drug Discovery goes Semantic

While the availability of freely accessible information sources relevant to medicinal chemistry and drug discovery has grown over the past few years, the knowledge management challenges of this data have also grown enormously: how to get consistent answers, how to manage different interfaces, how to judge data quality, and how to combine and overlay the data to generate new knowledge. Open PHACTS (Open Pharmaological Concept Triple Store), a consortium of 22 partners, is poised to address this knowledge management challenge with semantic web technology to accelerate drug discovery. Here we describe the brief rationale, history and approach of the OpenPHACTS consortium with a final aim to create an Open Pharmacological Space (OPS).

Modern drug discovery research is increasingly dependent on the availability, processing and mining of high quality data. Analysis and hypothesis generation for drug-discovery projects requires careful assembly, overlay and comparison of data from many sources. For example, expression profiles and data from genome-wide association studies (GWAS)[1] need to be overlaid with gene and pathway identifiers and reports on compounds *in vitro* and *in vivo* pharmacologyUtility of data-driven research goes from virtual screening, HTS analysis, via target fishing and secondary pharmacology to biomarker identification.

Over the last 15 years industry has spend significant resources to integrate public data and information sources and align this with internal, proprietary data while the academic Medicinal Chemistry research community suffered from lack of access to large data sets, especially those including curated bioactivity data. In contrast to data from the bioinformatics world, where whole organism genomes, protein sequences, and protein structures are available to everyone, the chemoinformatics community traditionally is closed and proprietary. Access to commercial databases of high quality crystal structures and chemical information requires licenses, as do most of the software packages needed. Medium size and large sets of bioactivity data per se are rare, as large scale screening efforts have almost exclusively been performed in industrial laboratories. This has disconnected industrial and academic drug discovery efforts and directed academia more towards method development. Furthermore, *in silico* models developed in academia have largely been restricted to the small and scattered publicly available chemical space.

This setting changed drastically with the NIH roadmap, which led to the creation of PubChem, a public available depository of screening data. PubChem currently comprises 31 million compounds, 73 million substances and 490.000 bioassay results. Others like DrugBank, ChemBank, IUPHAR, and ChEMBLdb followed soon and today there is a panel of databases available which can be searched for compounds and associated biological data. The current release of the ChEMBLdb contains more than 2,4 million activities of approx. 623.000 compounds measured against almost 7.200 targets.[2] The latest issue of the Nucleic acid research database summary lists almost 140 individual resources in the general field of molecular biology. However, there is still the urgent need for cleaning, improving and connecting these data to the public domain bioinformatics data, especially with respect to target validation, safety, efficacy and bioavailability. As an example, ChemSpider was set up as a free access platform for the aggregation, deposition and curation of community chemistry data, which has collected almost 25 million unique chemical structures linked out to over 400 data sources. ChemSpider has addressed the issues around data quality of chemical compound information available across the public data sources by running automated cleaning algorithms and providing wiki-like manual tools for commenting on and editing the aggregated compound information. Another milestone was the publication of the GSK 'Tres Cantos Antimalarial Compound Set', a set of 13 533 annotated compound structures shown to inhibit Plasmodium growth.[3] This for the first time allows academia to get access to a large data set derived under industry settings and standards and is likely to stimulate anti-malarial research through chemical biology and medicinal chemistry.

Public access to large amounts of information, by means of the open access policy for databases and papers, now will assist academia to contribute in a meaningful manner to drug discovery. Medicinal Chemists are thus expected to familiarize themselves with a multiplicity of highly variable sources and data formats, and their focus is shifting from data acquisition, to problem-solving skills, knowledge management and data integration.[4]

Nevertheless, there is a real danger that the high capacity to generate more data will not be in sync with our ability to manage the data well enough and, more importantly, to transform these data into biological and biochemical knowledge. Data integration over multiple sources is not sufficient to understand biology, but it is a prerequisite to even start to understand the complexity of any process in living organisms. As traditional medicinal chemistry research is embracing chemical biology and make increased use of phenotypic screening and high content biology for SAR, the requirements on data-analysis and integration will increase.[5] Unfortunately, this cumbersome and costly process is repeated across companies, institutes and academic laboratories. This represents a significant waste and an opportunity cost and effectively slows down scientific progress and consequently biomedical intervention.[6]

The emerging semantic web technologies and approaches are one way to address this major bottleneck in contemporary high throughput science. Simply put, "semantic web approaches" aims to establish unique identifiers for the concepts and entities within a given domain to allow effective connections to be made between data sources. More ambitiously unique identifiers are not only assigned to "things" (e.g. a compound) but also assigned to concepts such as "hydrolyses", "is a", etc (e.g "NaCl is a salt") to allow more advanced search and reasoning over large data-sets. In the chemistry domain the CAS-number is an example of such a unique identifier; another one is the IUPAC InChI which rapidly gains popularity and support. As data volumes increase we would want to address increasingly complex search questions and rapid answers to questions such as "*provide all compounds which have been associated with liver toxicity and list their interaction profiles with the transporters expressed in the liver*" are becoming crucial for the success of drug discovery research programs. Currently, answering such a broad question requires cumbersome parsings, reading and integration efforts. The ideal situation would be an immediate answer to this question, with the full possibility to 'drill down' in the underlying information resources for deeper investigation.

The Concept Web Alliance (CWA), formed in 2009 and comprising over 90 participants from academia and the private sector worldwide, is a collaborative community seeking to apply semantic concepts to deal with the massive amounts of information flooding the biological sciences and (later) other scientific disciplines. Many participants are also participants or members of other related networks and alliances, such as the W3C, the Pistoia Alliance and SageBionetworks, to name just a few. Collaboration of these like-minded alliances is a stated aim of CWA, which is built on the principles of Open Source, Open Access and Open Data. The rationale for the CWA semantic web approach is that classical data warehousing methods are no longer scalable to the size, spread and complexity of life science datasets, information resources and data analysis needs. These aims fit very well with the challenges already identified in harnessing public data for drug discovery.

A first step towards better global data integration and innovative ways to manage these data to produce meaningful information and finally knowledge is obviously the interoperability of the various data and information sets. Although standards are indispensible to this process, the *discussions about* standards can be lengthy and in fact may have an inherent potential side effect of blocking the very process it aims to accelerate. The consequences of not agreeing to common standards are evident when looking on public bioactivity data bases such as PubChem, BindingDB, and DrugBank. All of them host a broad range of pharmacological activity data, mostly manually curated, which are accessible through various web-based tools. However, the lack of common standards for representation of these data makes it excruciatingly time consuming to exploit the information present. Nevertheless, in a fully connected world where many different teams are providing valuable, but specialist, data-sources top down approaches as e.g. previously enforced by IUPAC for chemical nomenclature, will most likely fail in the biomedical domain. As an example, the CWA has adopted the approach of 'bottom up standard setting by best practices'. Recognizing that the power of standards lies in their widespread adoption the CWA firmly believe that the only long-term sustainable model for a scientific system to support global computational biology approaches as needed is full openness around the core semantic components, vocabularies and interfaces. However, this does not preclude the exposure and the delivery of proprietary content, nor does it preclude value-added closed-source or commercial services delivered on top of this system.

In order to succeed and gain widespread community adoption, this approach will need to go way beyond the current state of the art of tools in the life sciences and semantic web domain. The transition from classical, hypothesis driven research towards systems approaches requires rigorous new methodology. A further key challenge is that current data sources are largely incompatible with massive computational approaches and the vast majority of drug-discovery sources cannot easily interoperate. Recently developed data and text mining approaches, improved data capture standards, and leveraging semantic web technology open a first-time-opportunity to achieve interoperability through the semantic harmonization of data in key data sources. A priori data interoperability 'at the source' would therefore be a desired long-term effect of this distributed approach. Regarding questions around long-term sustainability of available resources, recent studies in the scope of the ELIXIR project have shown that out of 531 databases surveyed, 63 were either not online anymore or had not been updated since 2005 and, for a further 78, the update status was unclear. More importantly, less than 10% of the biomolecular resources surveyed indicated that they had multi-annual funding secured. The data resource landscape is therefore very fragile and a large and influential consortium involving academic as well as industrial drug-discovery partners can play an important role in capturing the most important 'assertional content' globally in a stable, interoperable and sustainable format.

The semantic approach, is based on the extraction and encoding of free-text, table, image, molecular sequence and structured information in Resource Description Framework (RDF) assertions that together with provenance data to form the basic building block of interoperability. The concept of RDF "triple"s (extracted simple assertions, also called nanopublications by CWA) has already been adopted by a wide and rapidly expanding community and is being implemented both for bioinformatics and chemogenomics. For instance, the Bio2RDF.org project aims to transform different sources of bioinformatics data into a distributed platform for biological knowledge discovery.[7] Initially, the authors focused on building a public database of open-linked data with web-resolvable identifiers that provides information about named entities. This involved the conversion of open data represented in a various formats to RDF-based linked data with normalized names. Bio2RDF entities also make reference to other open linked data networks thus facilitating traversal across information spaces. Bio2RDF is currently indexing around 5 billion triples, and is built with the open source Virtuoso database. However, currently the redundancy problem is not yet handled in Bio2RDF and in the linked open "data-cloud" in general. One step further is the Chem2Bio2RDF initiative, which comprises a repository aggregating data from multiple chemogenomic data sources that is cross-linked to Bio2RDF. Chem2Bio2RDF also includes a tool to facilitate query generation as well as a set of extended functions to address specific chemical/biological search needs. Potential usefulness in specific examples of polypharmacology, multiple pathway inhibition and adverse drug reaction–pathway mapping has been demonstrated.[8] Another very valuable source recently launched is ChemProt (www.cbs.dtu.dk/services/ChemProt/), a disease chemical biology database, which is based on a compilation of multiple chemical–protein annotation resources, as well as disease-associated protein–protein interactions (PPIs).[9] ChemProt comprises more than 700.000 unique chemicals with biological annotation for 30.578 proteins, leading to more than 2 million chemical–protein interactions, which were integrated in a quality scored human PPI network of 428.429 interactions. ChemProt can assist in the *in silico* evaluation of environmental chemicals, natural products and approved drugs, as well as the selection of new compounds based on their activity profile against most known biological targets, including those related to adverse drug events. Nevertheless, the increasing availability of linked data sources also requires innovative browsing and navigation tools, such as iPHACE (cgl.imim.es/iphace/).[10] iPHACE represents an integrative web-based tool to navigate in the pharmacological space defined by small molecule drugs contained in the IUPHAR-DB, with additional interactions present in PDSP. Extending beyond traditional querying and filtering tools, iPHACE offers a means to extract knowledge from the target profile of drugs as well as from the drug profile of protein targets.

Open Pharmacological Space

In light of all these developments and in order to foster public-private partnership the Innovative Medicines Initiative launched a call for development of an Open Pharmacological Space to establish a set of practical standards for the major public drug discovery resources and to implement these standards in a public infrastructure to the benefit of both pharma and academic drug-discovery communities. Open PHACTS, the winning consortium, will concentrate on a semantic web approach to develop an open source, open standards and open access innovation platform (OPS). The Open PHACTS project will be one of the first international attempts to create a reliable and scalable system, a common product beyond collective prototyping. OPS aims to deliver a sustainable, reliable web based environment through proven agile software engineering models. OPS will comprise data, vocabularies and infrastructure needed to accelerate drug-oriented research. This semantic integration hub will address key bottlenecks in small molecule drug discovery: disparate information sources, lack of standards and shared concept identifiers, guided by well defined research questions assembled from participating drug discovery teams. Workflows for data capture, processing, interoperability, visualization, and chemogenomics will be developed to create a comprehensive Systems Chemical Biology Analysis Network. Security issues around proprietary data, shared via the CWA nanopublication system and accessible for safe querying and reasoning will be properly addressed with expert trusted parties. The core Open PHACTS consortium comprises 14 European core academic and SME partners as well as 8 EFPIA members, with leading experts in the fields of data mining, annotation, small molecule data storage and manipulation, target related bioinformatics, RDF-type information handling, massive *in silico* reasoning and chemical biology. Noteworthy, Open PHACTS is not only open in terms of data but also in terms of the consortium itself. With the alignment of other Knowledge Management projects in IMI and a wider community of like-minded partners it is likely that already in 2011 the approach taken by Open PHACTS will be actively followed, co-developed and implemented by close to 100 partners world-wide. Any partner with an interest and an ability to contribute data, software or expertise is principally considered as an 'associated partner'.

Dissemination and community engagement will also utilize the manifold channels EFMC and RSC can provide. A large and influential consortium like this, involving academic groups, learned societies, as well as industrial drug-discovery partners collaborating in the context of OPS is likely to increasingly drive researchers around the globe to capture and distribute data and information in a semantically interoperable and computer readable format, as their data will 'connect' and 'mean' more from the onset. We therefore emphasize that - if successful and sustainable – the OPS project is likely to significantly contribute to more successful and cost-effective development of drugs and vaccines in human and animal health, as well as in nutrition and personal genomics. Turning data into knowledge is the cornerstone of successful drug discovery, but is the core business of science in general. A future driven by the open sharing of data, tools, services and workflows benefits the whole scientific community.

References (Endnotes)

1 Johnson AD, O'Donnell CJ (2009). An Open Access Database of Genome-wide Association Results. BMC Medical Genetics 10:6

2 Bender A (2010). Compound bioactivities go public. Nature Chem Biol 6, 309 (2010)

3 Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF (2010). Thousands of chemical starting points for antimalarial lead identification. Nature 465, 305-310.

4 Broccatelli F, Carosati E, Cruciani G, Oprea TI (2010). Tranporter-mediated efflux influences CNS side effects: ABCB1, from antitargets to target. Mol Inf 29, 16-26.

5 Hoffmann T, Bishop C (2010). The future of discovery chemistry: quo vadis? Academic to industrial – the maturation of medicinal chemistry to chemical biology. Drug Discovery Today 15, 260-264

6 Barnes MR, Harland L, Foord SM, Hall MD, Dix I, Thomas S, Williams-Jones BI, Brouwer C (2009). Lowering industry firewalls_pre-competitive informatics initiatives in drug discovery. Nature Reviews Drug Discovery 8, 701-708

7 Belleau, F., Nolin, M., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems Journal of Biomedical Informatics, 41, 706-716.

8 Chen, B., Dong. X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 2010, 11, 255

9 Tabourou O, Nielsen SK, Audouze K, Weinhold N, Edsgard D, Roque FS, Kouskoumvekaki I, Bora A, Curpan R, Jensen TS, Brunak S, Oprea TI (2010). Nucleic Acid Res, published online Oct 8.

10 Garcia-Serna R, Ursu O, Oprea TI, Mestres J (2010). iPHACE: integrative navigation in pharmacological space. Bioinformatics 26, 985-986.