Open PHACTS: semantic interoperability for drug discovery

Open PHACTS: semantic interoperability for drug discovery

Antony J. Williams[1], Lee Harland[2], Paul Groth[3], Stephen Pettifer[4], Christine Chichester[5,6], Egon L. Willighagen[7], Chris T. Evelo[6,7], Niklas Blomberg[8], Gerhard Ecker[9], Carole Goble[4] and Barend Mons[6]

[1] Royal Society of Chemistry, ChemSpider, US Office, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[2] Connected Discovery Ltd., 27 Old Gloucester Street, London, WC1N 3AX, UK
[3] VU University Amsterdam, Room T-365, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[4] School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
[5] Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, 1211 Geneva 4, Switzerland
[6] Netherlands Bioinformatics Center, P.O. Box 9101, 6500 HB Nijmegen, and Leiden University Medical Center, The Netherlands
[7] Department of Bioinformatics – BiGCaT, Maastricht University, The Netherlands
[8] Respiratory & Inflammation iMed, AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden
[9] University of Vienna, Department of Medicinal Chemistry, Althanstraße 14, 1090 Wien, Austria

Open PHACTS is a public–private partnership between academia, publishers, small and medium sized enterprises and pharmaceutical companies. The goal of the project is to deliver and sustain an 'open pharmacological space' using and enhancing state-of-the-art semantic web standards and technologies. It is focused on practical and robust applications to solve specific questions in drug discovery research. OPS is intended to facilitate improvements in drug discovery in academia and industry and to support open innovation and in-house non-public drug discovery research. This paper lays out the challenges and how the Open PHACTS project is hoping to address these challenges technically and socially.

Introduction

Open PHACTS (Open Pharmacological Concept Triple Store) is a project funded under a European grant from the Innovative Medicines Initiative (IMI; http://www.imi.europa.eu) [1,2] and was born as a public–private partnership (PPP) between academia, publishers, small and medium-sized enterprises (SMEs) and pharmaceutical companies. The ultimate goal of the project is to deliver and sustain an 'open pharmacological space' (OPS). Open PHACTS will use and enhance state-of-the-art semantic web standards [3] and technologies. It is focused on practical and robust applications to solve specific questions in drug discovery research. The project will deliver a core infrastructure of high-quality, semantically interoperable data accessed by user-friendly software

Antony J. Williams graduated with a PhD in chemistry as an NMR spectroscopist. Antony Williams is currently VP, Strategic development for ChemSpider at the Royal Society of Chemistry. He has written chapters for many books and authored >140 peer reviewed papers and book chapters on NMR, predictive ADME methods, internet-based tools, crowdsourcing and database curation. He is an active blogger and participant in the Internet chemistry network as @ChemConnector.

Lee Harland is the Founder & Chief Technical Officer of ConnectedDiscovery, a company established to promote and manage precompetitive collaboration within the life science industry. Lee Harland received his BSc (Biochemistry) from the University of Manchester, UK, and PhD (Epigenetics & Gene Therapy) from the University of London, UK. He has over 13 years experience leading knowledge management and information integration activities within major pharmaceutical companies. He is also the founder of SciBite.com, an open drug discovery intelligence and alerting service.

Barend Mons is a molecular biologist by training and received his PhD on genetic differentiation of malaria parasites from Leiden University and performed over a decade of research on malaria genetics and vaccines. Barend is currently an associate professor in Bio-Semantics at the Department of Human Genetics at the Leiden University Medical Centre with an honorary appointment in the same discipline at the Department of Medical Informatics, Erasmus Medical Centre, University of Rotterdam, both in The Netherlands. He is also a Scientific Director of the Netherlands Bioinformatics Centre (NBIC). His research is focused on nanopublications as a substrate for *in silico* knowledge discovery.

Corresponding author:. Williams, A.J. (tony27587@gmail.com), (williamsa@rsc.org)

1359-6446 http://dx.doi.org/10.1016/j.drudis.2012.05.016

interfaces that will enhance and accelerate the research process for its users. When fully operational, OPS will facilitate improvements in drug discovery in academia and industry. The OPS platform will support open innovation and in-house non-public drug discovery research and will collaborate on semantic interoperability with other IMI projects regarding modeling languages, standards and data, including supporting European infrastructure projects including ELIXIR (http://www.elixir-europe.org/). This paper lays out the challenges and how the Open PHACTS project is hoping to address these challenges technically and socially.

The challenge

Research and discovery in the life sciences is amazingly complex. Our attempts to understand the complexity of life, and the processes associated with disease and its treatment, are reflected in the complexity of the modern scientific process at many levels. Present technologies now enable us to generate enormous quantities of data. The typical scientific process produces vast amounts of information that is dispersed and hidden in various data sources (e.g. literature and curated databases). Data-driven life science research, including drug discovery, will increasingly rely on a community of collaborating partners to extract knowledge from these sources to solve complex questions.

The development and application of innovative methods and tools for data generation are paralleled by the needs for innovation in data storage, curation, integration, analysis and 'data publication'. Developments in these aspects of 'data stewardship' occur in the public and private sectors and, in general, are still relatively unguided, but offer the opportunity for important contributions to drug discovery research (Fig. 1). The result is a diverse landscape of data sources, with an inherent distribution of data quality, formats, standards, copyright and licensing [4,5]. This makes data sharing, integration, re-use and further exploitation cumbersome, thus hindering knowledge discovery. Therefore, many companies have already expended considerable energy, and continue to invest in mining clinical data, literature, patents and open- and free-access databases. This represents an enormous duplication of effort. Working collectively on precompetitive data integration will reduce costs. Open PHACTS offers a mechanism for project participants and associated collaborators to engage collectively with state-of-the-art cutting edge semantic technology to demonstrate the value of resource description framework (RDF) [6] and Semantic Web technologies [3,7,8], applied directly to state-of-the-art challenges.

The majority of current scientific advances are the result of collective and international collaborative efforts. One of the most important requirements to enable such efforts is that the data and information generated are preserved in a stable, unambiguous, trustworthy and computer readable state; Semantic Web technologies provide open standards that simplify this process, and can significantly contribute to data and information interoperability.

From a methodological perspective the scientific process can be seen as a workflow. Historically the analysis of only individual observations was the principal starting point of hypothesis formation. Presently computerization leads to the availability of thousands and millions of observations. We still consider data as individual objects but at the same time consider patterns linking large amounts of data into common rules. This requires novel data analysis methods to link the results of automation and high-throughput screening approaches to the individual observations. Advances in technologies have led to the generation of so much data that humans cannot capture and synthesize the explicit [9], let alone the implicit, information – this is the so-called 'data deluge' [10]. Most scientific investigations in life science now

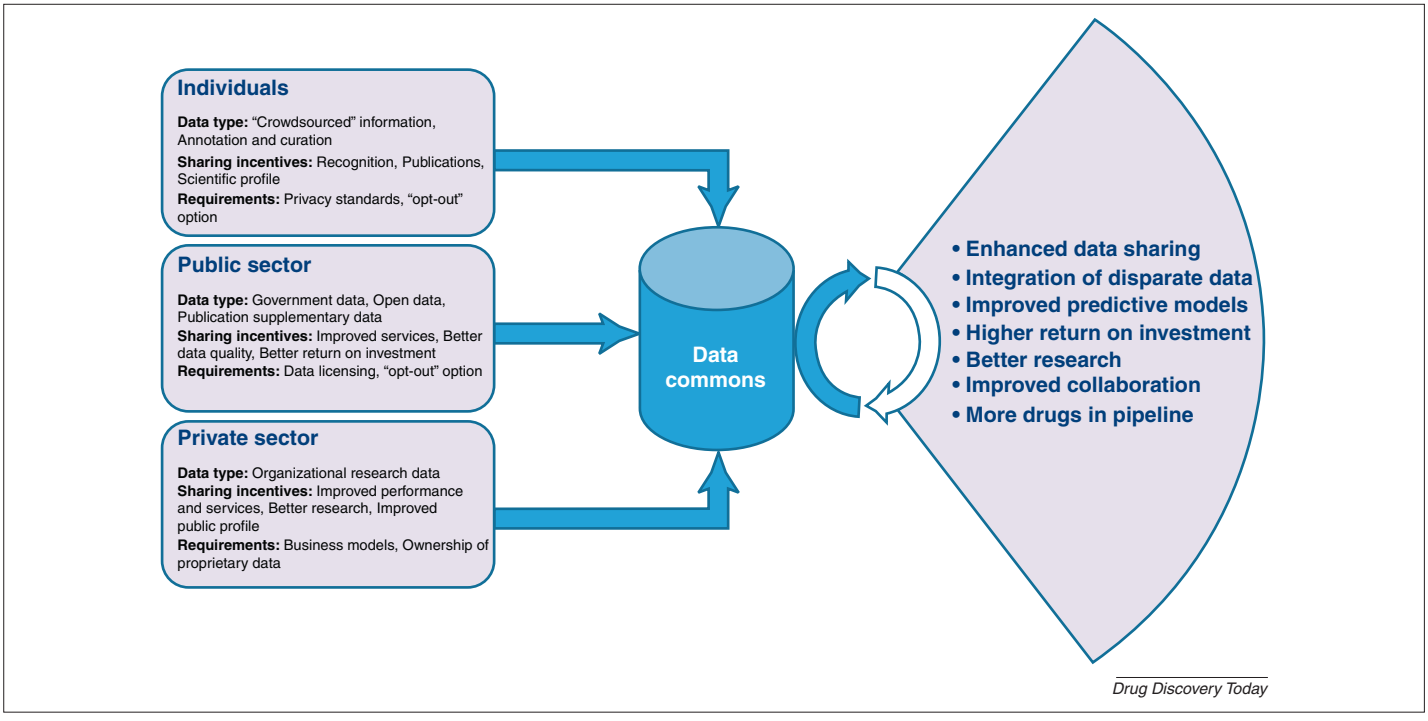

FIGURE 1

The general value of life science data across all sectors and the overall contributions to drug discovery research (inspired by a figure from Ref. [50]).

involve data from DNA and RNA sequencing, proteomics, metabolomics, screening, biomedical imaging, analysis of data in existing databases, data in the narrative literature and medical records. Despite this explosion in available data and information, the volume of data does not fundamentally change the classical biological cycle of observational research driving hypothesis formation and experimental design. The primary challenge in modern biology regards the complexity of the aggregated data which reveals ever increasing layers of complexity in the biology itself. The construction of standardized, re-usable, stable, up-to-date and easy to use workflow elements has emerged as the primary manner by which to work with these complex data. A large interoperable open data space is needed to feed these workflows.

Industry drivers

Addressing the challenges of 21st century therapeutics requires a deep understanding of the complexities associated with the biology of human disease. Increasingly, the industry accepts that this knowledge can never be held within one company alone. These challenges are leading to a change in the way pharmaceutical companies will need to do business.

In synchrony with cost pressures, the result is a new wave of business models in which the pharmaceutical industry is becoming more open [11], leading to increased collaboration with academia and other non-profit groups and to precompetitive collaboration between pharmaceutical companies. Furthermore, there is now an increasing movement within the public sector to apply resources to the discovery of new drugs. The synergy between these developments has led to major partnership activities such as the IMI.

A comprehensive and open repository of pharmacological entities and related chemical, biological and clinical data is crucial to drug discovery. Yet, researchers face myriad resources, databases [12,13], websites, tools and algorithms, each covering a small section of the space and using many different standards. The lack of universally accepted standards significantly inhibits the ability to integrate private, free and commercial pharmacology data to gain a more or less comprehensive picture of the known pharmacological landscape. The result is that experimentalists in industry, who are not computational experts, find it difficult to interact meaningfully with all relevant data. To this end, many companies spend considerable time and energy creating their own integration frameworks to combine such data. Ultimately, these systems mostly achieve the same results, highlighting the considerable duplication of effort in what is arguably a precompetitive activity. As a result there is a great need to develop a resource that addresses this problem.

The role of the Semantic Web

Various projects have been initiated in the past decade to make data relevant to pharmaceutical research available using a Semantic Web context. These include Bio2RDF [14], Chem2Bio2RDF [15], LinkedLifeData (http://linkedlifedata.com/), OpenTox [16], the World Wide Web Consortium Semantic Web Health Care and Life Sciences Interest Group (http://www.w3.org/blog/hcls/) and others [17,18]. This resulted in many tools, including ToxPredict (http://www.toxpredict.org/), ChESS [19] and WENDI [20].

However, these have not been developed from a pharmaceutical industry perspective. In the Open PHACTS consortium exploration of user needs resulted in the following priorities:

i. Providing a sustainable pharmacological information platform that would enable industry and public organizations to share data, analyses and understanding. This should be driven by community-endorsed open data, technical and social standards. Such a platform could serve as a catalyst for increased engagement and data sharing.
ii. Providing accessible, high-quality powerful tools to enable scientists to interact and explore a unified, fully interoperable pharmacological data space.
iii. Provide a layer of quality assessment over the data, to aid non-experts in judging which data to incorporate in their analyses.

The Open PHACTS consortium is committed to addressing these issues and to creating an OPS.

A driving contribution by the industrial partners in Open PHACTS is a list of prioritized research questions often addressed in small-molecule drug discovery and outlined below. The second contribution is data. Companies are now depositing significant datasets as open or free data [21] and are being encouraged to expand their efforts to the benefit of the community [22]. One of the barriers to increased sharing of industry data is the lack of industrial adoption of infrastructures as well as strong incentives to do so [23]. Various frameworks have been developed in recent years, for general purpose and for specific data types, including Data Dryad [24], FigShare (http://www.figshare.com), IsaTab [4] and OpenTox [16]. These will all be evaluated and connected to OPS where appropriate. In addition, Open PHACTS collaborates with other IMI projects, such as eTOX (http://www.etoxproject.eu/), EHR4CR (http://www.ehr4cr.eu/) and DDMoRE (http://www.ddmore.eu/).

The intrinsic value of data

Data contributions need to be made in a consistent manner, ensuring reusability of the data and via a mechanism that will be sustained for the long term. What is equally important is that the data in the system need to be maintained, refreshed and kept up to date. Besides the technological aspects discussed earlier there is also an important social and legal aspect to consider: data copyright and licensing (discussed in detail later). This is a very important element of OPS for the industry: the structured approach to ensure a sustainable infrastructure such that the current seed investments can be built upon ultimately to create an environment that becomes integral to pharmaceutical science. Thus, lowering the barriers for data sharing in industry and academia settings is one of the top priorities of the community-building aspects of Open PHACTS.

The major questions

The OPS should enable research on a wide range of questions that arise in applied pharmaceutical research, broadly covering aspects of chemistry and biology. A multitude of potential uses of the OPS to answer these questions can be envisaged, covering target identification and validation, the interaction profiles of compounds and targets, exploring potential toxic interactions, applications such as the repositioning of existing drugs to new therapeutic areas and many others.

To stay focused and deliver immediate benefit, the platform underlying OPS (i.e. the OPS platform) is being developed in an agile and stepwise, user-driven fashion, focusing on the most stringent issues for the industrial partners. Thus, the aforementioned research questions are central to driving its development and the associated user tools. In this context a research question is a specific query (or a series of queries) typical of the needs of researchers involved in pharmaceutical research, in industry and in academia.

Representatives from the European Federation of Pharmaceutical Industries and Associations (EFPIA) pharmaceutical companies participating in the project generated an initial list of research questions as part of the preparation leading to the formal OPS call from IMI. The list was subsequently extended and refined within the wider Open PHACTS project consortium, including the academic researchers in the pharmacological domain. True to its precompetitive nature, the project has opted to not focus on a single application domain or disease area but rather address the broad interoperability and actual integration of fundamental, quantitative, drug discovery data. The research questions serve several purposes within the project, including:

i. encouraging adoption of agile software development techniques;
ii. forming a basis for prioritizing potential data sources for inclusion;
iii. providing quantitative success measures;
iv. guiding the development of end-user tools and visualizations; and
v. creating a common focus for the project and a user community.

By adopting these principles of user-driven agile development, Open PHACTS addresses the challenges outlined in the introduction. By placing them in the context of the industry-provided use cases, it also ensures practicality.

Scientific quality of the underlying architecture is ensured, because the Open PHACTS project team covers a wide diversity of expertise including fields such as systems biology, computer science, chemo- and bioinformatics, semantics, publishing, pharmaceutical research, assay development and screening; the research questions have provided an excellent vehicle to converge on a common language and to align the diverse communities represented.

The list of potential questions (see Box 1 for examples) is essentially infinite and the list of selected questions will grow as the project evolves and the user community expands. It is also difficult to define at the early stage of the project how many of these questions OPS will be able to address efficiently during the limited project period. Although the intention is to create a generic architecture that can handle a wide range of different business questions, the limited time frame will mean that only a proportion are implemented, prioritized by factors such as business value and availability of high-quality data sources. As a first milestone, the consortium agreed to focus on use cases dealing with compound–target–pathway concepts. Of course, many cases can be answered now by industry researchers or by academic researchers with their internal systems and via bio- and chemo-informaticians. However, as stated above, a shared, fully interoperable and routinely updated scientist-friendly platform for performing these analyses in a non-manual way does not currently exist.

BOX 1

The simplest type of research questions are related directly to compound and/or biological target.

This subset of questions includes examples such as:

- For a given compound, summarize all 'similar compounds' and their activities.
- A lead molecule is characterized by a substructure S. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure S.
- For a specific target, which active compounds have been reported in the literature? What is also known about upstream and downstream targets?
- Find homologous targets to my target of interest and identify active compounds against these homologous targets.
- For a given compound, which targets have been patented in the context of Alzheimer's disease?

Other questions are driven more by genetics, pathways or disease. Examples here include:

- How does the gene variant affect patient survival for this disease?
- Who is working on the most relevant genes concerning Alzheimer's disease?
- For a given disease or indication provide all targets in the pathway and all active compounds hitting them.
- What are the adverse effects of all drugs used in a given disease?
- For a particular disease what is the human expression and distribution in healthy and diseased states?

As the project progresses the questions for each domain are sure to expand in depth and complexity. A key challenge in the project is to provide adequate guidance for inclusion and/or exclusion of different available data sources and data types as well as resolving queries using different vocabularies and mappings, for example the relevant National Center for Biomedical *Ontology* (NCBO) set of vocabularies and ontologies (http://bioportal.bioontology.org/ontologies), including MeSH [25] (http://www.nlm.nih.gov/pubs/factsheets/mesh.html) and SNOMED [26,27] or ICD11 (International Statistical Classification of Diseases and Related Health Problems) for diseases. The latter is an important factor for the combination of queries across domains to address challenging data-mining tasks in high-content biology, target discovery and validation.

The complexities of the challenge

The complexity of data access and licensing

Aligning the ingestion of public data and information sources with internal proprietary data, or data obtained via acquisition, is a challenge for the pharmaceutical industry. The academic medicinal chemistry research community simultaneously suffers from lack of access to large datasets, especially those including unambiguously described chemical structures annotated with curated bioactivity data. In contrast to data from general biology, where whole organism genomes, protein sequences and protein structures are mostly available to everyone (although standardization and curation issues still apply), chemoinformatics information has

traditionally been, and still is, closed and proprietary. Datasets without any license can seem usable at first glance but in fact they pose a particular problem as integration into the OPS can lead to legal issues later. A further aspect is license incompatibilities when data are mixed, as is anticipated in the OPS platform.

International law is not uniform regarding the copyrighting of data and this can lead to many practical problems. This implies that copyright and licensing terms cannot be assumed if they are not explicitly provided by the data providers. Such provenance information is unfortunately often missing. For example, it is common that organizations hosting data do not specify the copyright details associated with data. On a more formal legal side, the 'public domain' concept is ill-defined under international law and this has resulted in the formalization of the idea in copyright waivers, where the data source formally waives any rights it may or may not have. This approach is, for example, taken by the Creative Commons Zero license (http://creativecommons.org/choose/zero/). Otherwise, license incompatibilities are omnipresent in open source software, and possibly apply to data too. An explicit copyright and license statement is therefore crucial to enable the sharing and repurposing of data, which in itself is required for anyone to maintain, correct, mix and redistribute a dataset. Therefore, solving legal and practical issues around data access, sharing and licensing will be a focus of Open PHACTS.

The multiplicity and quality of legacy data sources

Well in excess of 100 individual resources (http://www.oxfordjournals.org/nar/database/c/) in the general field of molecular biology are relevant for medicinal chemistry research (based on informal polling within the Open PHACTS project). However, there is still the urgent need for cleaning, improving and specifically for functionally connecting these data to the public domain bioinformatics data, especially with respect to target validation, safety, efficacy and bioavailability.

The challenges of computational processing

Ideally, to support drug discovery, 'all' information in biology and chemistry contained in the classical narrative literature, datasets such as co-expression data, GWAS data (https://www.gwascentral.org/), curated databases and other research objects should also be made available in a manner that facilitates computational processing. However, it is an unfortunate reality that only a frighteningly small part of the available information is made accessible in such a manner. Providing a limited corpus of pharmacological data that are attuned for computational reasoning has been done before, but Open PHACTS considers the challenge in finding the limits in what is possible with state-of-the-art technologies.

The complexity of the name space

It is well established that the name space of biomedicine is messy and ambiguous. Many synonyms circulate on the web for crucial concept categories (semantic types) such as proteins, genes, drugs and diseases, but also for institutes, authors and, for instance, units of measurement. This historically developed disastrous situation is tackled by various international consortia as they try collectively to optimize the retrospective recovery of entities (concepts) and their relationships from narrative text, databases, among others [28]. It is an unfortunate reality that even in the 'curated' space several synonymous identifiers are being used for the same concept. Chemistry is, in many ways, even more challenged by ambiguity and degeneracy in its identifier systems. A single compound can be represented using systematic names generated according to several conventions [29] including IUPAC, CAS and Beilstein standards. A compound can have tens if not hundreds of trivial and trade names and multiple numeric identifiers that only have value when they can be used as look-ups against databases. The most common numeric identifier in structure space is the proprietary chemical abstracts service (CAS) registry number that can be used as a lookup in a pay-wall protected database or ideally for querying open databases. Unfortunately the uncontrolled proliferation of CAS numbers across public databases has resulted in a disastrous and confused mixing of data to a state of chaos and these numbers should not be trusted to be correct in the public domain – only CAS retains the trusted, and expensive to access, provenance relationship. Chemical compounds can also be represented by numerous electronic data formats that encode their connectivities and atomic makeup and some of these, specifically SMILES [30] and InChIs [31], are treated as identifiers in the online databases. However, these are defined explicitly by the atom–atom connectivities, standardization of the chemical, the chosen canonicalization algorithm and the various algorithmic settings used to generate these identifiers. Some progress has been made with the adoption of a standard InChIKey as a structure identifier and it is increasingly being used to facilitate integration and connectivity between databases. Nevertheless, the challenges of using such confused and ambiguous labels and tokens populated by so much variation has resulted in an enormous challenge in unraveling databases established by mapping together these various identifiers [32,33]. For the computer reasoning and Semantic Web interoperability the identity mapping to unique resolvable resource identifiers is an additional challenge of the Open PHACTS project.

The need for community annotation

The complex and massive quantities of data available in the public domain is of such value, but also of such magnitude and complexity, that curation with any level of comprehensiveness cannot be achieved without solid and continued involvement of the data generating and the data consuming community – human experts in their field [34]. Capturing data and scientific claims at the source (from the authors and investigators) in an interoperable format would be ideal. However, this is far from being common practice and is susceptible to changes in understanding and perception. In addition to the authors and creators of datasets, other contributors and consumers of scientific information can be mobilized as curators, ranging from students to medical practitioners to patients. Such a 'crowdsourced' approach is, of course, a valuable approach to the creation of knowledge as demonstrated by the success of Wikipedia (http://www.wikipedia.org). In the life sciences domain it has been carried over to, for instance, WikiGenes [35], Gene Wiki [36], WikiPathways [37] and the ConceptWiki (http://www.conceptwiki.org/index.php/Main%20Page), and has been applied to the analysis of NIH chemical probes [38].

The same approach was adopted for the curation of chemical space in the ChemSpider database [39]. Unfortunately, for greater success in this area, more-efficient tools and greater participation

are required for it to have the necessary impact. This can be supported by comparing data from different databases, making detection of inconsistencies easier and by creating proper incentives. As stated in one of the early Open PHACTS related publications [40]: 'Data citation and the derivation of semantic constructs directly from datasets have now both found their place in scientific communication. The social challenge facing us is to maintain the value of traditional narrative publications and their relationship to the datasets they report upon while at the same time developing appropriate metrics for citation of data and data constructs'. Therefore, solutions to the central challenge of making 'all' biological and chemical data available in computable format to everyone need to encompass the pure technical interoperability aspects as well as the data and information quality issues and the social aspects enabling modification and sharing, which again are inseparable today from the intimate involvement of the expert community worldwide and, thus, form proper incentives for community curation.

Further complexity is introduced in that human language derives much of its meaning from context. Names of entities are not consistently used to describe the same concept. The drug name sometimes refers to the administered medication, whereas in other occasions is refers to the active, chemical ingredient.

The same is true for mixed references to genes in different organisms and genes and their protein products being referred to by the same and frequently ambiguous terms. Any community annotation system must support addressing this aspect to support the curation process further. These five above mentioned categories reveal the complexity of the daunting challenge facing Open PHACTS.

The Open PHACTS solution

The Open PHACTS platform is, at its core, a data integration platform. However, because of the complexity of the social and technical challenges involved, the consortium decided to take a different approach to classical data warehousing. Instead of imposing a top-down view of data in data warehouses, a more bottom-up view will be taken where information from multiple providers is exposed by adaptive integration of the information. This bottom-up approach will be facilitated by the adoption of open web-based data standards. We characterize this approach as 'semantic data interoperability'. In essence data from all relevant sources will be 'published' in a semantically interoperable web enabled format, extended by community-adopted ontologies, so that an increasing number of tools and services, including those provided by Open PHACTS, can take advantage of the published data layer, with full provenance allowing access to the underlying rough datasets.

Importantly, this semantic approach to data integration has been pioneered by other research efforts in the biomedical domain, as outlined earlier. These research efforts have shown that semantic data integration is feasible for life sciences in principle. However, the addition of rich provenance and context to individual assertions in the Semantic Web has appeared to be crucial. Open PHACTS takes the Semantic Web approach a step further by focusing on creating a robust, up-to-date platform, where each and every association or assertion can be traced back to its origin and can be placed in context, by computers and by the human mind, and is thus designed for direct use by scientists.

The data, tools, workflows and infrastructure used to develop OPS will be open data and open source (http://github.com/openphacts/). All partners are contributing background intellectual property (IP) on data, tools and software to the OPS program in the expectation that all foreground IP generated by the OPS program is contributed with an open license. This is a fundamental principle that will enable the core OPS system to function as a global drug discovery knowledge hub.

There is ample room for the development of proprietary exemplar services taking data feeds from the OPS infrastructure. Academic partners and SMEs in Open PHACTS will also develop innovative exemplar services that further enhance the public drug discovery toolset. Prospective partners that have either services [not requiring changes beyond the Application Programming Interface (API)] or data (published according to Open PHACTS guidelines but not hosted in OPS) can request to become Open PHACTS affiliated projects or partners via a straightforward procedure (essentially taking one week). Partners wishing to influence gradually the core architecture and/or become core data providers can follow the procedures for deeper levels of partnership while being associated with the project already.

This scheme is essential for the sustainability approach of OPS. Clearly, the widespread adoption of the OPS platform and the associated standards defines the success of the project, and the implementation of third-party services on top of the OPS infrastructure will be a key success measure. This includes commercial offerings and we envisage several innovative business models to emerge. For example, a publishing house could offer high-quality assertions as a value-added service within the OPS framework, high-quality reasoning, patent analysis, expert finders, visualization and query-builders interacting with OPS could be developed by start-ups or SMEs and the OPS framework could be used to enrich publisher content semantically.

An initial view of the OPS platform architecture is presented in Fig. 2 but it is clear, given the agile nature of the project, that the approach will be updated and modified over time, based on our experiences.

Data scope

To offer maximum benefit OPS will be set up with the full realization of the need to cover multi-omics and many other data sources. The Open PHACTS consortium develops a software platform with multiple installations designed to host a mixture of in-house data from many major data 'owners', existing public, free and open data, and data from further data sources that wish to expose their content in a safe and trusted way to the OPS. Data creators and owners will be actively approached and invited to join as associated partners. Open PHACTS has an important role to navigate and set policy together with the data providers, enabling users and hosting providers to know their rights and potential legal liability. A series of dedicated workshops is planned during the lifetime of the project to address these crucial issues and the first of these has already been held, in 2011 in The Netherlands (http://www.openphacts.org/ops_workshops.htm).

The enormous and rapidly growing list of potential data sources to be 'semantically published' in the OPS will not be addressed simultaneously. As stated earlier, the prioritized research questions will drive a stepwise inclusion needed to answer an increasing

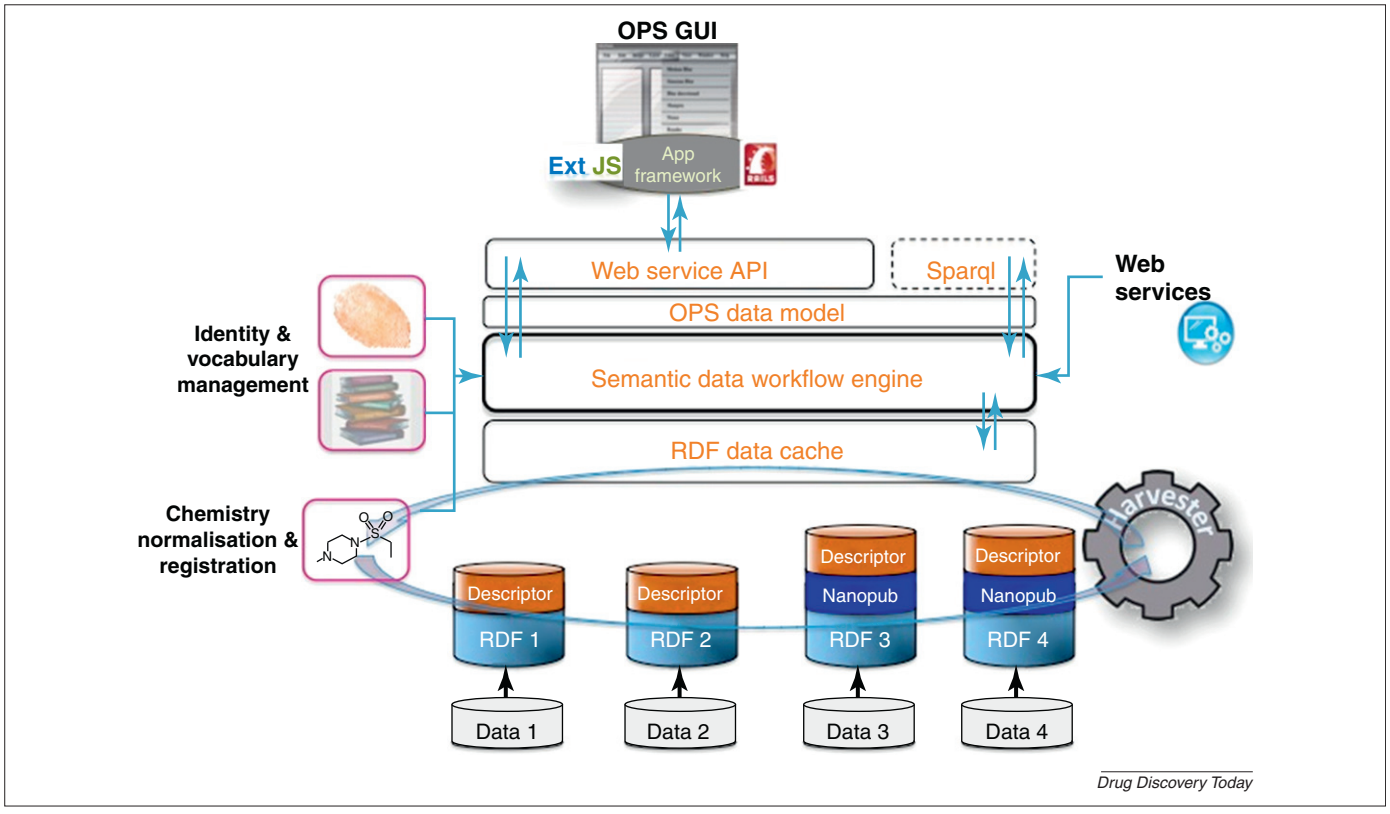

FIGURE 2

The OPS platform architecture highlighting the data feeds, the integration to a chemistry compound registration system, the semantic workflow underpinning the system and the exposure of an API to enable the development of exemplar applications accessing the core system.

number of questions. However, once again the architectural choices made in Open PHACTS will enable a scalable increase of data sources to be covered without unacceptable performance loss.

Data acquisition

The Open PHACTS strategy centers on data publishing – in that as many as possible of the major databases it wishes to connect should publish their data as rich RDF, a standard for web-data interoperability. This enables the OPS 'data harvester' system to detect the presence of data (and subsequent) updates and download or refer to these data, analogous to the way a search engine crawler works. There are already several referring standards such as VoID (http://semanticweb.org/wiki/VoiD) and nanopublications [40], which OPS is considering. Although Open PHACTS needs to appreciate that not every provider will be able to produce perfect RDF, Open PHACTS has developed a set of guidelines for nanopublication (see below) and will assist key data sources to publish (subsets of) their data in compliant formats. Already, Open PHACTS is supporting the development of standards for provenance at the World Wide Web Consortium (http://www.w3.org/2011/prov/). To encourage the publication of data, we are supporting approaches that enable fine-grained credit and attribution. In particular, the project has significantly developed the original notion of nanopublications as developed by the Concept Web Alliance, of which several founder members are key partners in Open PHACTS. A nanopublication is the smallest unit of publishable information: an assertion about anything that can be uniquely identified, attributed to its author and de-referenced. In the Open PHACTS project, nanopublications will increasingly be used to expose existing experimental data as well as assertions retrieved from legacy data and information sources such as scientific articles, for instance MedLine and other databases in a standardized representation, addressing the interoperability challenge. The project has released its first set of guidelines for producing nanopublications that correspond with open standards.

Identity

The complexity of the name space is approached through a system by which the individual concepts constituting the biological concept space (genes, proteins, drugs/chemicals, diseases, among others) as well as the social concept space (authors, articles, datasets, among others) are first treated completely at the individual concept level. Rather than trying to enforce all players globally to refer to these individual concepts in a standard way, the OPS will develop and use on-the-fly identity mapping to combine different terms and internationalized resource identifiers (IRIs) dynamically for the same physical entity. The advantage of this approach is that different rules can be applied at query time to fit the current query best (for instance deciding whether to treat genes and proteins, genes and gene probes or different tautomers or stereoisomers of the same compound as 'the same physical entity' for the purposes of a query). This system will be based upon an identity resolver and mapper and use cross-referencing data from a range of providers, combined with profiles that capture those query rules. Although this means Open PHACTS does not need to mandate the use of any one specific vocabulary for a

certain semantic type (gene, protein, drug, among others), we will still provide recommendations as to how best to represent data within the system. Specifically, Open PHACTS will recommend the use of open public vocabularies and identifier schemes, advocating use of resources such as the NCBO's BioPortal (http://bioportal.bioontology.org/ontologies) and EBI's Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/), approved vocabularies and public identifier mapping services such as BridgeDB [41] (http://www.bridgedb.org/) and identifiers.org (http://identifiers.org/). Wherever possible we aim to re-use existing ontologies and IRIs, and contribute additions and/or corrections back to the original source. Where reliable crossmappings of ontological terms and identifiers do not exist the ConceptWiki (http://www.conceptwiki.org) is available to allow the community as well as authoritative organizations to invoke crowdsourcing of mapping between existing, and the extension of, controlled vocabularies.

Linked data cache

The OPS linked data cache interacts with the harvesting component to gather, update and integrate data. It provides a semantically integrated view of the data by exploiting the identity resolution and identity mapping component to construct appropriate responses based on the contextual aspects of each query (i.e. what sort of user is asking for a query, what is needed to be displayed). It also contributes to performance by caching the semantic representations of frequently used datasets. An important aspect of the cache is the ability to integrate datasets and services; for example, the integration of chemical similarity search or more-advanced reasoning components. To achieve this, Open PHACTS will implement the cache using a semantic workflow engine, the Large Knowledge Collider (LarKC) [42], that is specifically designed for manipulating RDF data at a large scale. By adopting this approach, the OPS should be able to integrate new services, such as large scale reasoning, flexibly into the platform. Additionally, LarKC provides an abstraction over a variety of storage options for RDF data, thus providing for options to scale as the data grows. Other reasoning environments such as associative reasoning [43–45] have already revealed novel concept connections in a complementary manner to LarKC and, for instance, WolframAlpha® (http://www.wolframalpha.com/) will increasingly be able to use the rich semantic resource the OPS will provide to increase output.

Domain-specific services (handling chemistry)

The complexities of handling chemical compound data will be managed by using the proven ChemSpider platform [39]. The resource is a free-access database populated with over 26 million unique chemical compounds linked out to over 400 data sources. The platform includes structure searching, substructure searching and an array of flexible search capabilities via the web-based user interface. ChemSpider also provides access to the underlying data and many of the underlying search routines via web services and is the method by which the structure level data will be accessed programmatically by the OPS system. Each chemical entity in the Open PHACTS system is provided with a unique registration number, a ChemSpider ID, and these IDs are populated into the cache to facilitate linking of the data, and are complemented with the associated SMILES and InChI Keys. Using a combination of structure identifiers and web services the OPS architecture can integrate structure-based queries and the data mappings within the cache to relate chemistry to biology.

One of the largest challenges with handling chemistry is the appropriate representation of chemical compounds in a standardized format, especially one that is acceptable to the pharmaceutical industry. In this regard structure standardization routines will be implemented according to those recommended by the FDA in their substance registry system (SRS) documentation (http://www.fda.gov/forindustry/datastandards/substanceregistrationsystem-uniqueingredientidentifierunii/default.htm). The standardization recommended by the FDA will be adjusted to match the agreed upon needs of the Open PHACTS project defined according to a chemistry committee of EFPIA members and chemoinformatics representatives from the project.

Application programming interface

A key part of the OPS platform is a rich API. This interface provides REST-style based interfaces for common queries. These queries are defined in agile cycles working with application and user interface developers. By driving API development from the user perspective, Open PHACTS will also drive down the complexity of data integration by focusing on those core semantic types and properties that are necessary for the end-users in the project.

The API is managed by LarKC enabling the flexible integration of identifier mapping. Currently, this is implemented by generating SPARQL queries that are run over the triple store managed by LarKC. A key principle in designing the API is to provide an optimized robust experience for user interface and application developers. A key difficulty in many linked data solutions today has been the threat to performance caused by difficult or challenging queries and it is not yet clear in many current prototypic applications how to deal with complexities in the representation of facts in RDF. By adopting a well-defined API, we can optimize the SPARQL queries enabling us to ensure that such performance difficulties do not occur. Finally, the API provides rich information and access to the provenance of the results it returns. This includes the data sources from which those results are provided including the ability to track licensing information.

User interface

Different use cases have varying needs in terms of their interaction with the platform. The architecture of the OPS will be flexible by design and support a variety of end-user applications with the user interfaces optimized to address specific tasks. Indeed, it is important for the project to demonstrate that such a wide range of user interfaces, focusing on different user communities, is a viable output of the system. The OPS platform architecture provides a balanced approach toward enabling community maintained data sources while providing interoperability and performance. A key novelty of the architecture is its emphasis on adaptable data integration driven by end-user needs and customized by multiple user interfaces and analysis workflows stemming directly from research questions. The concept of data publication by local data providers and a harvesting into a cache for reasoning and querying will enable a growing number of data sources to connect to the OPS with minimal disturbance of local procedures or architecture.

Using the platform

Standards are of no value without a working implementation. To make those standards useful, the selection of data sources for the first generation of OPS services is a crucial issue for the success of the Open PHACTS project. Within the project there is an active user community representing experimental academic science as well as industrial drug discovery. These scientists are collaborating to build independent applications, so-called 'exemplars', on top of the core semantic services. As exemplar applications mature, we would hope that each of these be the subject of its own publication. The general approach to select and drive integration will be guided by the defined set of frequently occurring scientific questions and drug discovery scenarios. Exemplar services are being developed simultaneously with the design of the OPS core architecture to enable early demonstration of actual support for ongoing drug discovery programs of the project partners. The experience and the encountered challenges will be used to steer the final stages of development of the wider OPS semantic infrastructure. This is an area where the EFPIA consortium expects to contribute significantly with expertise, curation resources and limited data depositions.

Data quality and annotation

Open PHACTS is going to bring together several data sources from the public domain as well as integrate and mesh data made available by EFPIA members and commercial sources. It is prominently acknowledged by the Open PHACTS project partners that data sources (free and commercial) can be rather poorly curated [32,46], except in those cases where significant funding is provided to the hosting organization to police data quality and ensure that data are kept and validated with ongoing and updated assertions. For instance, in the chemistry space we plan to provide some specific assessments of the quality of small-molecule representations in different databases and also annotate metrics around chemical representation quality. At present a chemistry validation and standardization system is in development to check for valid chemistry representations (e.g. provide warnings when checking for hypervalency, charge imbalance, absent stereochemistry, among others) as well as standardization of the structure representations across various databases to a derivative of the SRS standard agreed upon by Open PHACTS members. These approaches will then be applied to the data sources ingested into Open PHACTS, data quality will be identified, detected errors will be annotated and feedback will be provided to the data suppliers.

Several data sources have already been identified as those bringing value to the Open PHACTS project. Experienced members of the team have judged that some of these require additional levels of curation and validation. These include data from well-known databases such as PubChem [47], DrugBank [48,49] and ChemSpider. Immediate challenges arise with the validation of the chemical entities themselves because a high proportion of the structural representations for the chemicals/drugs/ligands in question do not accurately represent the intended entities. Compounding this issue is the fact that many of the associated chemical names, synonyms and identifiers used for the purpose of text mining, integration and cross-referencing between resources are incorrectly associated. Add to this incorrect mapping to biological targets and the risk of retrieving incorrect data increases significantly. In the biological domain as well, it is well established that even curated databases do not nearly cover all data available in the literature and in dispersed datasets. It is therefore probable that current systems frequently return incomplete (when highly curated) or misleading (when broadly recovered) results.

Quality control, rating and validation of data sources before and after publishing into Open PHACTS will be essential and the development of rigorous processes to define initial quality checking, ongoing community validation and reporting back to the original source providers will be necessary. Agreed upon criteria for acceptance and rejection of data, or integrating using quality flags for inclusion or exclusion in searches will be established across the team. The development of high-quality validated data dictionaries will be essential to many of the projected returns for the system and will depend on a sincere commitment, from all parties, to police data quality throughout the project. Importantly, the OPS system will provide provenance of all the data it makes available enabling users to check themselves whether the results are from what they consider to be quality data sources.

Although data are essential to the platform under development the annotation of these data is equally important. Annotation will result from the ongoing process of layering on additional assertions, relationships and mappings as more information flows from the sources outlined above. However, human annotation will be crucial to developing further the 'assertion base' and the contributions from participants in Open PHACTS. Crowdsourcing capabilities are already delivered via three of the components of the project: ChemSpider, ConceptWiki and WikiPathways, and it is probable that an increasing number of easy interfaces designed to facilitate annotation and curation will be made available. Via the provision of selected interfaces for the purpose of adding annotations to the data the growth of the underlying data will be made more valuable via human intervention for additional knowledge gathering.

Community crowdsourcing across industry and non-profit organizations is a model that will be consistently explored in Open PHACTS to provide the necessary curation resources. Growing collaborations should enable the mitigation of the major perceived bottlenecks for community annotation, including the lack of a globally trusted party, the need to work with sometimes cumbersome structured data entry interfaces and, most importantly, the lack of professional recognition of curation and annotation work. Open PHACTS will actively engage in building a community that will increasingly annotate individual assertions and the combination of these as a routine part of their daily knowledge discovery process. In that process, Open PHACTS will optimally involve the pioneering wiki-type crowdsourcing approaches as well as the professional, dedicated curation communities, including the International Society for Biocuration (http://biocurator.org/).

The international perspective

Although the IMI funding model inherently limits Open PHACTS initially to a mostly European initiative, the issues and solutions to data interoperability are global. Most of the industrial partners in this project have significant research activities in North America and Asia. Many of the important data sources (NCBI, NCBO, PubChem, DrugBank and ChEMBL, UniProt, GO) and knowledge

representation tools are being developed by our scientific colleagues within and outside Europe. Building on the principles of openness and active re-use of established resources the Open PHACTS project has developed an approach of 'associate partners' to enable additional institutions and research groups to participate.

Close synergy with the development of ELIXIR and other developing pan-European infrastructures for life science informatics is essential to ensure that the OPS will become a cornerstone of the global public data infrastructure. Active collaboration will be sought with funding agencies and other groups outside Europe to make sure that the OPS will develop into a global platform. The scientific activities around the globe, notably including rapidly emerging economies in Asia and Latin America, will be engaged as soon as feasible. Especially in community annotation the scientific communities in these continents (including Africa) will be crucial to the long-term success of the OPS. Therefore multilingualism and language independence (both enabled by the Open PHACTS conceptual approach) will be part and parcel of the policy of Open PHACTS.

Sustainability

The development of the OPS as a major open data infrastructure has to have a plan for sustainability, service provision and maintenance in the public domain once IMI project funding ceases. For pharmaceutical industry partners to continue to invest in the development of the OPS there must be a clear route to future sustainability ensuring resource, as well as for investment expectations. The pharmaceutical companies will not invest in the OPS without a sustainability and future plan to deliver on the value proposition. Development of the OPS infrastructure necessitates a robust service layer including change control, 24:7 support, backups, disaster planning and service, which is an order of magnitude more complicated and costly then academic best effort type application hosting.

Open PHACTS therefore is not just another research project with a defined ending. It is supposed to develop and deliver a crucial, intensively used service. It is thus imperative that the OPS as a public–private partnership (PPP) service, once established, will be stable, high performance, user-friendly and sustainable. In addition to the EFPIA partners in the OPS using the system for their core research and business, it is desired and encouraged that other organizations will also use the OPS platform. Actually it will be a significant measure of success when other PPPs and businesses build useful services and applications using the end result of the Open PHACTS project.

It is obvious that making the OPS sustainable beyond the duration of the Open PHACTS and essentially all other interdependent IMI projects is crucial for all parties to enable long-term use of the system. To ensure that the benefits are maximized for the community, it is crucial to engage a wide community of researchers, information providers and service companies in Europe and worldwide at an early stage of the project. More specifically, establishing dedicated human resources and sufficient funding will have to be a mutual responsibility of the OPS partners and the associate partner and user community to be established. With regard to effective sustainability, Open PHACTS adopts the principles of open data, open standards, open source, open infrastructure and open community and free use.

Conclusions

Open PHACTS clearly has the potential to initiate a paradigm shift in the pharmacological space and far beyond. Interest has already been shown by members of the translational research communities, the medical informatics communities, the 'omics communities and even research fields outside the life sciences within the first year of the project. As a result Open PHACTS has recently established a 'waiting room' and a 'gatekeeper' function to manage the participation of interested parties. By raising such interest and expectations across the life science community, Open PHACTS, and consequently IMI, has assumed a daunting responsibility. It would be unacceptable if the OPS appeared to be technically feasible and yet failed to deliver on expectations in terms of accelerating drug discovery and scientific progress far beyond the narrow 'small-molecule space'. It would also be unacceptable if the research performed here appeared to be yet another project that gracefully disappeared when its funding expires. It is therefore of importance that IMI, preferably in close coordination with European Strategy Forum on Research Infrastructures (ESFRI) and comparable initiatives in the USA (i.e. Sage Bionetworks, NCBO, among others), as well as with other continents, develops a comprehensive and widely adopted data, information and knowledge management strategy.

The OPS system, together with the exemplars that will be delivered during the timescale of the originally funded project, is expected to deliver research insights. These will include developing deeper understandings of the application of Semantic Web technologies to the life sciences, the integration and mapping of disparate data types and sources, the influence of data quality, crowdsourcing and curation on decision making as well as the processes and approaches necessary to manage a diverse team to deliver a groundbreaking technology platform. The question as to whether the OPS platform might ultimately replace existing in-house systems will become evident as the system is delivered to the community as an open source software system with its associated open data. Organizations can choose to adopt only certain components of the overall solution or certain data slices appropriate to address their needs. It is probable that integration between existing in-house systems and the OPS platform, using the available programming interfaces, will provide the appropriate solution for certain organizations in some cases. However, in many companies, a major reason for being involved in this project is to replace in-house systems with something that they do not have to solely maintain, freeing up internal resources for higher-priority activities. Ultimately the success of the project will be measured by whether or not new knowledge and wisdom can be extracted. It would of course be beneficial if the work also resulted in new drugs being contributed to the pipeline.

Acknowledgements

The Open PHACTS project consortium consists of leading academics in semantics, pharmacology and informatics from 23 partner organizations including eight pharmaceutical companies and three biotechnology companies. This project will be delivered as a result of the efforts, diligence and collaboration between the consortium members and additional partners. We acknowledge all of the collaborators for their contributions to Open PHACTS.

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution.

References

1 Hunter, A.J. (2008) The Innovative Medicines Initiative: a pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov. Today* 13, 371–373

2 Kamel, N. *et al.* (2008) The Innovative Medicines Initiative (IMI): a new opportunity for scientific collaboration between academia and industry at the European level. *Eur. Respir. J.* 31, 924–926

3 Antoniou, G. and van Harmelen, F., eds) (2004) *A Semantic Web Primer*, The MIT Press, Cambridge, MA/London, England

4 Sansone, S.A. *et al.* (2012) Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126

5 Samwald, M. *et al.* (2011) Linked open drug data for pharmaceutical research and development. *J. Cheminform.* 3, 19

6 Carroll. J.J. and Klyne, G. (2004) Resource description framework (RDF): concepts and abstract syntax. *Tech. rep. W3C*

7 Baker, C.J.O. and Cheung, K.-H., eds) (2010) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, Springer

8 Neumann, E. and Prusak, L. (2007) Knowledge networks in the age of the Semantic Web. *Brief Bioinform.* 8, 141–149

9 Anderson, C. (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired*

10 Lanfear, J. (2002) Dealing with the data deluge. *Nat. Rev. Drug Discov.* 1, 479

11 O'Boyle, N.M. *et al.* (2011) Open data, open source and open standards in chemistry: the blue obelisk five years on. *J. Cheminform.* 3, 37

12 Galperin, M.Y. and Cochrane, G.R. (2011) The 2011 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.* 39 (Database issue), 1–6

13 Williams, A.J. (2008) Public chemical compound databases. *Curr. Opin Drug Discov. Dev.* 11, 393–404

14 Belleau, F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706–716

15 Chen, B. *et al.* (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255

16 Hardy, B. *et al.* (2010) Collaborative development of predictive toxicology applications. *J. Cheminform.* 2, 7

17 Hassanzadeh, O. *et al.* (2009) *LinkedCT: A linked data space for clinical trials*. arXiv:0908.0567v1

18 Chepelev, L.L. and Dumontier, M. (2011) Semantic web integration of cheminformatics resources with the SADI framework. *J. Cheminform.* 3, 16

19 Chepelev, L.L. and Dumontier, M. (2011) Chemical entity semantic specification: knowledge representation for efficient semantic cheminformatics and facile data integration. *J. Cheminform.* 3, 20

20 Zhu, Q. *et al.* (2010) WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminform.* 2, 6

21 Ekins, S. and Williams, A.J. (2010) Meta-analysis of molecular property patterns and filtering of public datasets of antimalarial "hits" and drugs. *MedChemComm* 1, 325–330

22 Ekins, S. and Williams, A.J. (2010) Precompetitive preclinical ADME/Tox data: set it free on the web to facilitate computational model building to assist drug development. *Lab Chip* 10, 13–22

23 Harland, L. *et al.* (2011) Empowering industrial research with shared biomedical vocabularies. *Drug Discov. Today* 16, 940–947

24 Greenberg, J. (2009) Theoretical considerations of lifecycle modeling: an analysis of the dryad repository demonstrating automatic metadata propagation. Inheritance, and value system adoption. *Catalog. Classif. Quart.* 47, 380–402

25 Rogers, F.B. (1963) Medical subject headings. *Bull. Med. Libr. Assoc.* 51, 114–116

26 Elkin, P.L. *et al.* (2009) BioProspecting: novel marker discovery obtained by mining the bibleome. *BMC Bioinformatics* 10 (Suppl. 2), 9

27 Shah, N.H. *et al.* (2009) Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics* 10 (Suppl. 2), 1

28 Altman, R.B. *et al.* (2008) Text mining for biology – the way forward: opinions from leading scientists. *Genome Biol* 9 (Suppl. 2), 7

29 Yerin, A. and Williams, A.J. (1999) The need for systematic naming software tools for exchange of chemical information. *Molecules* 4, 255–263

30 Weininger, D. (1988) SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31

31 The IUPAC International Chemical Identifier (InChI). (2011) September. Available at: http://www.iupac.org/inchi/

32 Williams, A.J. and Ekins, S. (2011) A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today* 16, 747–750

33 Williams, A.J. *et al.* (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov. Today* [Epub ahead of print]

34 Mons, B. *et al.* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* 9, 89

35 Hoffmann, R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.* 40, 1047–1051

36 Good, B.M. *et al.* (2012) The Gene Wiki in 2011: community intelligence applied to human gene annotation. *Nucleic Acids Res.* 40 (Database issue), 1255–1261

37 Kelder, T. *et al.* (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 40 (Database issue), 1301–1307

38 Oprea, T.I. *et al.* (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol.* 5, 441–447

39 Pence, H. and Williams, A.J. (2010) ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87, 1123–1124

40 Mons, B. *et al.* (2011) The value of data. *Nat. Genet.* 43, 281–283

41 van Iersel, M.P. *et al.* (2010) The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11, 5

42 Fensel, D. *et al.* (2008) *Towards LarKC: A Platform for Web-Scale Reasoning.*

43 van Haagen, H.H. *et al.* (2009) Novel protein–protein interactions inferred from literature context. *PLoS ONE* 4, e7894

44 van Haagen, H.H. *et al.* (2011) *In silico* discovery and experimental validation of new protein–protein interactions. *Proteomics* 11, 843–853

45 van Haagen, H. and Mons, B. (2011) *In silico* knowledge and content tracking. *Methods Mol. Biol.* 760, 129–140

46 Williams, A.J. (2008) A perspective of publicly accessible/open-access chemistry databases. *Drug Discov. Today* 13, 495–501

47 Wang, Y. *et al.* (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37 (Web Server issue), 623–633

48 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 34 (Database issue), 668–672

49 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36 (Database issue), 901–906

50 Consulting, V.W. (2012) *Big Data, Big Impact: New Possibilities for International Development.*