Semantic Web 5 (2014) 101–113
DOI 10.3233/SW-2012-0088
IOS Press

Applying linked data approaches to pharmacology: Architectural decisions and implementation

Editor(s): Mikel Egaña Aranguren, Technical University of Madrid (UPM), Spain; Michel Dumontier, Carleton University, Ottawa, Canada; Jesualdo Tomás Fernández Breis, University of Murcia (UM), Spain
Solicited review(s): Boris Villazón-Terrazas, Technical University of Madrid (UPM), Spain; Rafael Valencia-Garcia, University of Murcia (UM), Spain; Jesualdo Tomás Fernández Breis, University of Murcia (UM), Spain

Alasdair J.G. Gray [a,*], Paul Groth [b], Antonis Loizou [b], Sune Askjaer [c], Christian Brenninkmeijer [a], Kees Burger [d], Christine Chichester [d], Chris T. Evelo [e], Carole Goble [a], Lee Harland [f], Steve Pettifer [a], Mark Thompson [d], Andra Waagmeester [e] and Antony J. Williams [g]

[a] *University of Manchester, UK*
[b] *VU University Amsterdam, The Netherlands*
[c] *H. Lundbeck A/S, Denmark*
[d] *Netherlands Bioinformatics Center, The Netherlands*
[e] *Maastricht University, The Netherlands*
[f] *Connected Discovery, UK*
[g] *Royal Society of Chemistry, UK*

Abstract. The discovery of new medicines requires pharmacologists to interact with a number of information sources ranging from tabular data to scientific papers, and other specialized formats. In this application report, we describe a linked data platform for integrating multiple pharmacology datasets that form the basis for several drug discovery applications. The functionality offered by the platform has been drawn from a collection of prioritised drug discovery business questions created as part of the Open PHACTS project, a collaboration of research institutions and major pharmaceutical companies. We describe the architecture of the platform focusing on seven design decisions that drove its development with the aim of informing others developing similar software in this or other domains. The utility of the platform is demonstrated by the variety of drug discovery applications being built to access the integrated data.

An alpha version of the OPS platform is currently available to the Open PHACTS consortium and a first public release will be made in late 2012, see http://www.openphacts.org/ for details.

Keywords: pharmacology, linked data, data integration

1. Introduction

The exploration and development of new drugs requires scientists to draw knowledge from multiple sources of information. These range from online databases of proteins (e.g. UniProt [26] and Enzyme [3]) and chemicals (e.g. ChEMBL [19], ChemSpider [37], and DrugBank [30]), to models of biological pathways (e.g. Reactome [33], WikiPathways [29], and KEGG [28]) as well as the scientific literature.

These information sources are often held in different formats and sourced from a wide variety of organizations. Together they cover a wide area of the scientific

*Corresponding author. E-mail: a.gray@cs.man.ac.uk.

space of interest, but at the same time overlap scope, and thus frequently record different (or even inconsistent) representations of the same data. In recent years, several key datasets for drug discovery have been published on the Semantic Web including those provided by Chem2Bio2RDF [11] and Linking Open Drug Data (LODD) [40]. In the latter case, the data is linked following linked data principles [8,24].

The integration of knowledge from these disparate sources presents a significant problem to scientists, with the intellectual and scientific challenges often being overshadowed by the need to repeatedly perform error-prone and tedious mechanical tasks. The first challenge is to identify the entities of interest in the many data sources, and to relate and map these to one another. This allows complementary information from the data sources to be collated in a single record. For example, ChemSpider contains data about chemical compounds and where they can be sourced, while ChEMBL complements this with data about the bioactivity of drug-like molecules and DrugBank provides information on the clinical use of drugs which contain the molecules. These data sources can be linked based on the chemical structure of the compounds. However, differences in scientific or technical approaches to molecular structure representation mean that different data sources will not always be in agreement. For example, searches for the chemical *"Fluvastatin"* on ChemSpider[1] and DrugBank[2] return different compounds: although their basic chemical structure matches, the compounds differ in their stereochemistry[3].

A further challenge is the lack of semantics associated with links in traditional database entries. For example, the entry in UniProt for the protein *"kinase C alpha type homo sapien"*[4] contains a link to the Enzyme database record[5], which has complementary data about the same protein and thus the identifiers can be considered as being equivalent. One approach to resolve this, proposed by Identifiers.org, is to provide a URI for the concept which contains links to the database records about the concept [27]. However, the UniProt entry also contains a link to the DrugBank compound *"Phosphatidylserine"*[6]. Clearly, these concepts are not identical as one is a protein and the other a chemical compound. The link in this case is representative of some interaction between the compound and the protein, but this is left to a human to interpret. Thus, for successful data integration one must devise strategies that address such inconsistencies within the existing data.

In this application paper, we present the architectural decisions made in implementing the Open Pharmacology Space (OPS) platform within the Open PHACTS project[7]. The key goal of the project is to support a variety of common tasks in drug discovery through a technology platform that will integrate pharmacological and other biomedical research data using open standards such as RDF. The OPS platform is being developed by a public-private partnership to address the problem of public domain data integration for both academia and major pharmaceutical companies [44], and will make its first public release in late 2012. A key driver of the project is that integrated data should address concrete pharmacological research questions and integrate with the workflows and applications already used within the drug discovery pipeline.

This paper focuses on two key contributions:

- A set of seven architectural decisions for a data integration platform driven by pharmacological business questions. These decisions should inform others developing similar or related software.
- A discussion of the use of this platform by three separate drug discovery applications.

The rest of this paper is organized as follows. Related work is discussed in Section 2. A summary of the concrete pharmacological research questions that the functionality provided by the OPS linked data platform seeks to address are presented in Section 3. Section 4 discusses the architectural decisions made to provide an integration framework that is capable of integrating public data sources in order to answer the top priority research questions. Section 5 gives details of the implementation of a linked data platform for pharmacology that enables enriched functionality on three separate, existing, drug discovery applications. Finally conclusions are drawn in Section 6.

[1] http://www.chemspider.com/Chemical-Structure.393587.html accessed 17 Sept 2012.

[2] http://www.drugbank.ca/drugs/DB01095 accessed 17 Sept 2012.

[3] For details see http://www.chemconnector.com/2012/07/29/ accessed 17 Sept 2012.

[4] http://www.uniprot.org/uniprot/P17252 accessed 17 Sept 2012.

[5] http://enzyme.expasy.org/EC/2.7.11.13 accessed 17 Sept 2012.

[6] http://www.drugbank.ca/drugs/DB00144 accessed 17 Sept 2012.

[7] http://www.openphacts.org accessed 17 Sept 2012.

2. Related work

The OPS platform deliberately builds upon a large amount of prior work delivered by the Semantic Web community in two key areas: infrastructure and data sources.

In terms of infrastructure, there have been several semantic data integration platforms proposed and developed. ODEMapster [5] is a global-as-view data integration approach for exposing relational data sources through an existing ontology. MASTRO [10] provides support for more expressive queries to be evaluated over a subset of *DL-Lite* ontologies. Fox et al. [18] discuss the requirements for domain ontologies and relationships to upper ontologies when deploying a semantic integration system. FedX [41] provides a framework for federating queries over a set of distributed SPARQL endpoints. All of these systems present a single integrated view of the data to the users and build upon the large body of work on data integration within the database community [22,31]. Central to these data integration systems are the mappings between the data sources (which in many cases are relational databases) and the intended semantics of the integrated result (represented as ontologies in semantic integration systems). Such mappings need to be supplied by domain experts with a good understanding of the data sources and the (ontological) view of the integrated data to be created. Dataspaces [21] is an approach to lower the effort required by the domain expert to generate and maintain mappings. The dataspace system starts with a minimal set of mappings which is incrementally enriched based on user feedback [6].

Another important area of infrastructure work are scientific workflow systems such as Taverna [34], Galaxy [20], and Kepler [32]. These systems are designed to enable users to compose a series of operations into a *workflow* that will extract data from sources, and execute data manipulation and computational services to analyse it. An important aspect of these systems is capturing the provenance of each operation performed by the workflow. A key idea behind workflows is enabling their reuse so that results can be reproduced. Community portals such as MyExperiment [15] support the sharing and reuse of workflows. The OPS infrastructure builds on the Large Knowledge Collider (LarKC) system that provides a pluggable workflow environment for developing scalable Semantic Web applications that integrate large datasets [13].

The Open PHACTS project is, of course, not the first to realize that the use of Semantic Web standards, along with common ontologies, can ease the integration burden in the pharmacology domain [36]. Samwald et al. [39] present the findings of the W3C Health Care and Life Sciences interest group with regards to the availability of data sources, their potential for linking, and present recommendations for best practices in publishing pharmacology data as Linked Data. The Bio2RDF [7], Neurocommons [38], Linking Open Drug Data (LODD) [40], Linked Life Data [35] and Chem2Bio2RDF [11] projects have all made significant sets of biology and chemistry data available in RDF. We built upon the comprehensive work of the community in creating RDF-based data sources to power the OPS platform.

3. Motivation: Pharmacology research questions

A key driver of the OPS system is that the integrated data should address concrete 'real world' pharmacology research questions and integrate with the workflows and applications already used within the drug discovery pipeline. A collection of 83 questions has been created and prioritised by the scientific and pharmaceutical partners in the Open PHACTS project. Full details of the research questions, the generation and prioritisation process, and analysis of the results can be found in [2,25]. We provide a brief summary so as to motivate the functionality provided by the OPS system, and the data sets that have been integrated to provide an information space.

The business questions span a range of areas, however, the focus in developing the OPS platform is on those concerning pharmacology data, specifically the interactions between compounds and molecular targets. Many of the questions can be answered now by industry researchers or by academic researchers with their internal systems and via bio- and cheminformaticians. However, a shared, fully interoperable and routinely updated, scientist-friendly platform for performing these analyses in a non-manual way does not currently exist.

In the simplest case, the questions would be of the form:

"Retrieve all information available about aspirin."

This requires discovering all of the entries for aspirin in the available data sources, e.g. ChemSpider, ChEMBL, and DrugBank, and combining the data. Such a task is complex in itself due to the different focus of the data sources and the fact that each uses its

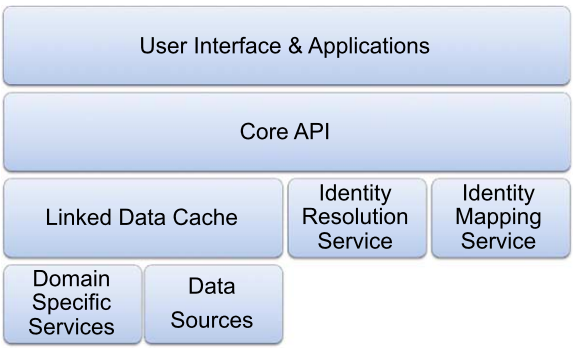

Fig. 1. OPS platform architecture overview.

own set of identifiers, not to mention the inconsistencies in the overlapping data that must be resolved or displayed in an appropriate form to the user.

The more complex research questions involve filtering compounds based on their properties and experimental results, e.g.

> *"Give all oxidoreductase inhibitors with IC50 activity value < 100nM in both human and mouse."*

To answer such questions requires the filtering and comparison of values from different data sources. This is complicated by the fact that the data sources use a variety of units to express this value. For example, even within ChEMBL, a single curated source, the IC50 value is represented in various bases of molars ranging from nano- to milli- and even in other units such as microgram per millilitre ($\mu g.ml^{-1}$).

Another group of the questions are driven by genetics, pathways, or diseases. An example from this group would be:

> *"For a given disease related molecular pathway, give me all targets and known active compounds that modulate them."*

These require a larger set of data sources, e.g. UniProt and WikiPathways, to be associated to compounds, targets, and the interactions between them. Similar issues arise but at a larger scale.

4. Architectural decisions

The OPS platform provides a semantic platform for pharmacology that integrates data from a set of distributed open data sources.

Figure 1 provides an overview of the OPS platform architecture as it is currently implemented. It consists of seven components:

1. Data Sources
2. Linked Data Cache (LDC)
3. Identity Resolution Service (IRS)
4. Identity Mapping Service (IMS)
5. Domain specific services
6. Core API
7. User Interfaces/Applications

These components arise out of corresponding architectural decisions made variously for pragmatic, technical and social reasons. We now discuss each of these seven decisions. Components resulting from these decisions are in italics.

4.1. Rely on existing RDF-based datasets

As a result of the Linked Data movement [8,24], the Semantic Web community have made a large number of existing datasets available as RDF. Particularly relevant datasets to drug discovery are the ChEMBL RDF conversion[8] [45], the ChEBI ontology sourced from the EBI[9] [16], the conversion of DrugBank provided by the LODD project[10] [40], the conversion of the Enzyme database sourced from UniProt[11], and the UniProtKB/Swissprot dataset itself[12] [4]. These *Data Sources* are the foundation of the OPS platform.

The decision to rely on existing data sources was both pragmatic and social. Pragmatically, relying on existing data allowed the project to quickly develop a working system. Socially, it encourages both originating data providers (e.g. UniProt) and third parties to continue to provide RDF-data for pharmacology sources. Such RDF conversions are crucial to the success of Open PHACTS. This decision however has a cost, as data sources will use different schemes instead of one common global schema. Additionally, OPS is reliant on data providers continually updating RDF versions of important datasets. For example, an early version of the OPS platform used the Chem2Bio2RDF conversion of ChEMBL which has not been kept up-to-date with the underlying ChEMBL database from

[8] http://semantics.bigcat.unimaas.nl/chembl/v13_ops/ accessed 17 Sept 2012.

[9] ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi.owl accessed 17 Sept 2012.

[10] http://www4.wiwiss.fu-berlin.de/drugbank/drugbank_dump.nt accessed 17 Sept 2012.

[11] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/enzyme.rdf.gz accessed 17 Sept 2012.

[12] http://www.uniprot.org/uniprot/?query=reviewed%3ayes&force=yes&format=rdf accessed 17 Sept 2012

the EBI[13]. This severely limited the data available to the OPS platform and caused a great deal of confusion to our internal users. We have now moved to a more recent conversion and are working with the EBI to incorporate the RDF conversion process into their data release pipeline.

4.2. Centralize the data

A classic trade-off in the design of data integration systems is whether to federate queries or warehouse data. In this case, we chose to warehouse the data for reasons of reliability and performance. In terms of performance, interactive query speeds are necessary, thus, there may be significant latency overheads when relying on remote services (e.g. SPARQL endpoints). Furthermore, third party services (i.e. external to the consortium) cannot and should not be relied upon to provide consistent access. Essentially, it would be impolite for us to make use of others resources for answering SPARQL queries especially when they may be large or frequent. To address these concerns, the aforementioned datasets are centralized into a *Linked Data Cache (LDC)*. The term cache is used to emphasize that the platform does not maintain data on its own and is just a temporary store to enable data to be queried quickly at interactive speeds.

An issue with caching the data is ensuring that the datasets reflect the latest versions of the underlying datasets. However, it is common for scientific datasets to follow release cycles and these can be used to reload the data in the cache. See for example UniProt [14].

4.3. Separate keyword search and structured queries

A key entry point into the platform is through keyword search. In the pharmacology domain, this is more than just text matching as keywords can often match to multiple often very distinct concepts. For example, when typing *"menthol"* the user could quite reasonably mean the chemical menthol, or the menthol receptor protein. This sort of conceptual keyword search is different from queries over structured databases, thus, a new architectural component was introduced, the *Identity Resolution Service (IRS)*. The role of the IRS is to translate user-entered entity names (in free text form) into known entities within the system (i.e. that have a defined URI). These known entities can then be used in structured queries. The IRS also supports the disambiguation of concepts through interaction with the user interface.

[13]https://www.ebi.ac.uk/chembl/ accessed 17 Sept 2012.

4.4. Equality is context dependent

Datasets often provide links to equivalent concepts in other datasets. These result in a profusion of "equivalent" identifiers for a concept. Identifiers.org provide a single identifier that links to all the underlying equivalent dataset records for a concept. However, this constrains the system to a single view of the data, albeit an important one.

A novel approach to instance level links between the datasets is used in the OPS platform. Scientists care about the types of links between entities: different scientists will accept concepts being linked in different ways and for different tasks they are willing to accept different forms of relationships. For example, when trying to find the targets that a particular compound interacts with, some data sources may have created mappings to gene rather than protein identifiers: in such instances it may be acceptable to users to treat gene and protein IDs as being in some sense equivalent. However, in other situations this may not be acceptable and the OPS platform needs to allow for this dynamic equivalence within a scientific context. As a consequence, rather than hard coding the links into the datasets, the OPS platform defers the instance level links to be resolved during query execution by the *Identity Mapping Service (IMS)*. Thus, by changing the set of dataset links used to execute the query, different interpretations over the data can be provided.

4.5. Leverage domain-specific services

There are a variety of important pharmacological operations that are specific to a domain and have reliable and performant implementations. Thus, instead of creating new implementations, the OPS platform relies on these existing *domain-specific services*. For example, of critical importance for drug discovery is that identical small molecule compounds from across different data sources are integrated accurately, based on structure rather than name or identifier mapping which is less accurate. Instead of reimplementing this feature, we rely on an existing chemistry registration and normalisation service: ChemSpider [37]. This takes each compound from all of the data sources and maps it using structure-based methods to a unique ChemSpider identifier. The mappings between each chemical in each data source and ChemSpider are then loaded into

the system to provide an accurate set of chemical mappings between different databases.

Going forward, we aim to make use of the SADI Framework for accessing additional domain specific services [12].

4.6. Provide a simple API

The initial prototype of the OPS platform only provided a SPARQL endpoint through which the integrated data could be queried. This required each of the drug discovery applications to have an intimate knowledge of the data exposed and the ability to write the required SPARQL queries to retrieve the data desired. This approach also left the OPS platform exposed to poorly formed queries or simply queries that are too open ended (e.g. a `select * where {?s ?p ?o}`). This both impacted developer productivity and hurt the reliability of the service.

To address this problem, an additional component, the *Core API* was introduced into the architecture. The Core API provides a set of common methods that applications can call. This benefits application developers as they no longer need to formulate their own SPARQL queries.

The OPS Core API does more than to simply provide an abstraction layer for the application developer. Such interfaces provide assurances that the data layer will only be exposed to well designed queries, and allows the API developer to make extensive use of optimizations present in RDF stores. We believe this to be key to the successful deployment of Semantic Web infrastructures which consume large scale datasets.

4.7. Leverage the user interface

A key driver of the OPS platform is to enable user-focused drug discovery applications. Thus, the entirety of the OPS platform architecture is designed around enabling these sorts of applications. This also means that we can take advantage of the user interface, for example, by encapsulating queries in an API, thus ensuring that only well-defined queries are run, or using the user interface to ensure that correct URIs are given to the system when performing a query. We also believe that this decision will allow for scalability in the future. However, there is a trade-off as this means that the OPS platform in its current form is not designed for other tasks such as large offline analytics.

5. Open pharmacology space implementation

This section provides details on the implementation of the OPS prototype platform as shaped by the architectural decisions given in Section 4 and the needs of the drug discovery scientists in the Open PHACTS consortium. To clarify the platform architecture, consider the simplest of the three example pharmacology research questions provided in Section 3:

"Retrieve all information available about aspirin."

This can be achieved through the use of two OPS core API methods: `compoundLookup(text)` and `compoundPharmacology(URI)`.

The IRS is invoked by the core API through the `compoundLookup("aspirin")` method and provides a number of alternatives, including *"Aspirin"*, *"FIORINAL"* (which is a compound that contains aspirin), and *"Hypyrin"* (which is an aluminium complex ion of aspirin). The returned items are each annotated with additional metadata (e.g. descriptions and synonyms) to aid the users in their decision.

Once the user selects a particular concept, e.g. Aspirin (`cw:aspirin`[14]), the user interface will make the call `compoundPharmacology(cw:aspirin)`. The workflow to evaluate this method call is depicted in Fig. 2 and is executed as follows[15]. First the compound URI is inserted into the query which is then expanded with equivalent URIs provided by the IMS. The final query is evaluated against the Linked Data Cache. An example result of this API call is shown in Fig. 3, which shows the concept identified by the URI retrieved from the IRS and its links to concepts in ChemSpider, ChEMBL, and DrugBank. Note that it is possible for there to be multiple mappings within a single dataset. Figure provides example properties retrieved from each dataset, e.g. SMILES and InChI from ChemSpider, and activity values from ChEMBL.

5.1. The large knowledge collider

The platform is implemented as a connection of a number of separate software systems. These software systems are coordinated using the Large Knowledge

[14] We use the notation `cw:aspirin` to represent the URI http://www.conceptwiki.org/concept/dd758846-1dac-4f0d-a329-06af9a7fa413/ accessed 17 Sept 2012.

[15] Note that the IRS plugin was used to support the `compoundLookup(text)` call. The Chemical Structure Search plugin will be discussed in Section 5.5.

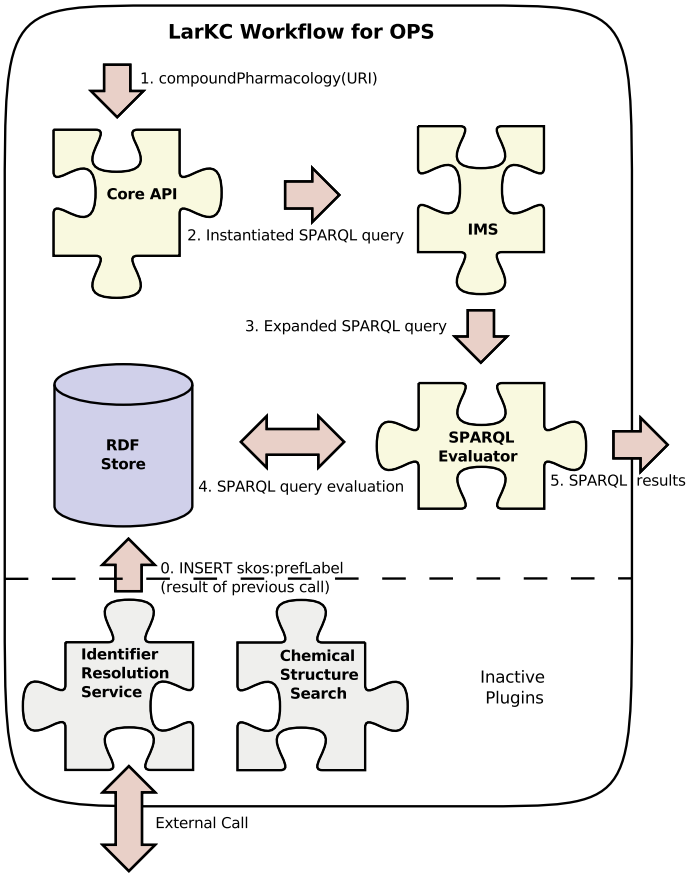

Fig. 2. LarKC plugins and workflow activation for a `compoundPharmacology(URI)` API call.

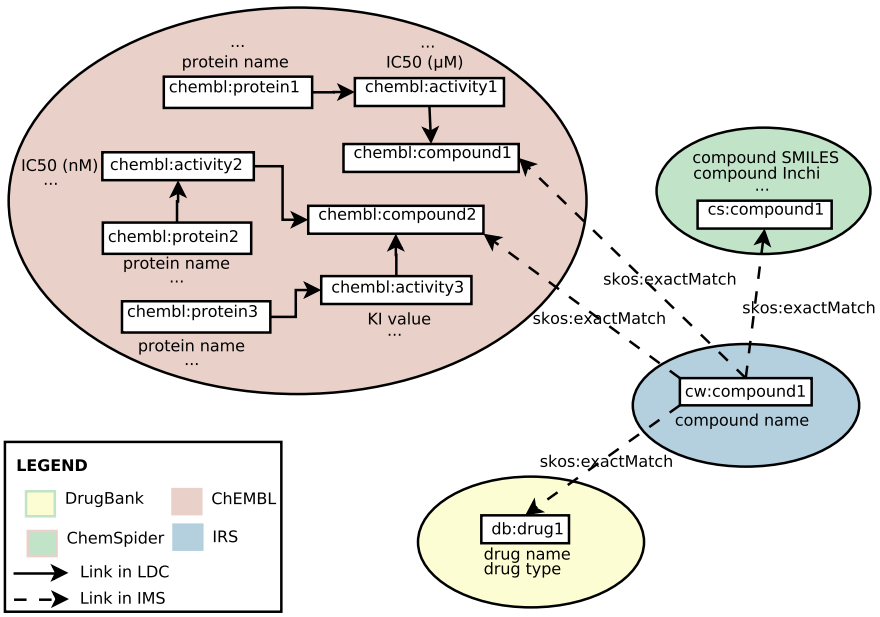

Fig. 3. Sample results of a `compoundPharmacology(URI)` API call.

Collider (LarKC), exploiting its pluggable architecture to call external services [13].

Workflows in LarKC are defined in terms of their input, output and the connections between the plugins that participate in them. In addition, LarKC endpoints are a special type of plugin which expose an interface for receiving input and providing output, accessible through HTTP. A valid workflow must contain a single endpoint and a single plugin connected to its output. When the execution of all participating plugins is completed, the endpoint collects any output from the workflow and responds to the original HTTP request.

Figure 2 presents the LarKC plugins that participate in the main OPS workflow, illustrated as jigsaw puzzle pieces. The figure gives a trace of the information flow between the various plugins as a result of a `compoundPharmacology(URI)` call to the OPS API.

5.2. Linked data cache

Additionally, LarKC acts as the Linked Data Cache (LDC) component within the OPS platform. During the OPS platform implementation, an alternative data layer was developed for LarKC, which implements the OpenRDF Sesame SAIL interface [9]. SAIL stands for "Storage and Inference Layer" and is a connection oriented interface that is implemented by a number of RDF stores to enable their interoperability with the Sesame platform. The SAIL implementation has been contributed to the LarKC open source codebase, enabling the deployment of LarKC with a wide range of alternatives to the default BigOWLIM (which is now deprecated and replaced by Ontotext with OWLIM-SE). More importantly it allows OPS users and developers to set up local OPS instances with an RDF store that is specifically selected to suit their particular data needs and hardware specifications. For example an offline demonstration instance to be installed on a mid-range laptop using a small sample dataset can avoid the installation and configuration overhead of a scalable RDF store by using an in-memory alternative.

We have successfully tested the integration of LarKC with four additional RDF stores: the default Sesame in-memory store [9], bigdata [42], OpenLink Virtuoso [17], and Garlik's 4store [23]. Currently we use the Sesame in-memory store as a number of instability issues were discovered with the remaining alternatives. The in-memory store is deployed on a server with 392 GB of RAM, and an overview of the data currently loaded in the LDC is given in Table 1. As we are fast approaching the limitations of this solution, we are now in the process of moving to a hosted solution provided by OpenLink[16], who have recently joined the Open PHACTS consortium. The Virtuoso triplestore will be enhanced to meet the challenges of scale and complexity faced in deploying the OPS platform.

Table 1
Characteristic properties of RDF data currently present in the LDC.

Data property	Value
Distinct resources	93 118 400
Distinct predicates	293
Named graphs	11
Total quads	460 037 674
Size on disk	63 GB
Size in RAM	292 GB

In addition to the main OPS workflow, the current prototype makes use of an auxiliary workflow for which the IMS, IRS and external services are not used. The purpose of the workflow is to identify the URIs for which data is available in OPS from a large list of candidates. As RDF stores will provide the worst-case performance when queried for URIs that do not appear in their indices (as they do not exist), such queries benefit from circumventing the IMS.

5.3. Identity resolution service

As discussed in Section 4.3 the mapping between user provided text and URIs is a non trivial task that has to take into account the type of entity the user is interested in. This functionality is thus decoupled from the LDC and is provided by the Identity Resolution Service (IRS) plugin.

This plugin is able to make calls to an API provided by ConceptWiki[17] to retrieve a list of resources of the required type with canonical labels curated by the community. The response is then dynamically imported in the underlying RDF store. The IRS plugin responds to explicitly defined predicates in the SPARQL query[18] which are also used to relate the resulting RDF to existing data in the LDC. Updates take place immediately, making the result of the external call directly available to the SPARQL query that triggered it, but also to all future queries.

If such predicates are not present in the query the plugin remains inactive, as is the case in the example

[16] http://www.openlinksw.com/ accessed 17 Sept 2012.
[17] http://ops.conceptwiki.org accessed 17 Sept 2012.
[18] `PREFIX:` <http://wiki.openphacts.org/index.php/ext_function#>

of Fig. 2. However, the example call will retrieve information dynamically asserted by the IRS plugin during a previous `compoundLookup` call – namely the object of the `skos:prefLabel` predicate (arrow #0 in the figure).

5.4. Identity mapping service

The IMS plugin responds to all queries provided by the API endpoint and invokes the IMS service for any instance URIs that appear in the query. Once an appropriate list of mappings for each URI has been retrieved from the IMS service, the plugin will expand the original SPARQL query to include equivalent URIs.

The expanded query is then forwarded to the SPARQL Evaluator plugin which will query the RDF store and write to the output of the workflow. The need to store equivalent URIs in the LDC is thus eliminated, and the platform is able to provide appropriate results to identical queries with different entity equivalence requirements.

5.5. Domain specific services

So far only a single domain-specific service has been integrated with the OPS platform – namely the Chemical Structure Search service provided by ChemSpider. This service enables the retrieval of small molecule compound URIs that match, contain, or are similar to a user provided chemical structure.

As such functionality requires the use of highly specialised and computationally expensive methods and algorithms, we developed a LarKC plugin that invokes the ChemSpider API, dynamically converts search results to RDF and updates the LDC. The inserted data is persistent and reused to answer queries with identical parameters.

Similar to the IRS plugin, the Chemical Structure Search will scan each SPARQL query it receives for a set of predefined predicates which correspond to the types of search supported by ChemSpider.

5.6. Core API

As a key focal point of the OPS platform is to enable the rapid development of user-focused applications in the drug discovery domain, we have developed a simple REST-based API to facilitate the retrieval of data from the LDC. Applications can thus be developed without a requirement for extensive knowledge of the schemas used in each underlying dataset. While the application developers are limited to the methods of the core API, these have been collaboratively defined to meet their functionality needs.

The core API is implemented as an endpoint in LarKC that will receive the input to each OPS workflow, and extract the method name and corresponding parameters from the HTTP POST request. Each method name corresponds to a parameterized SPARQL query which is instantiated using the parameters provided, e.g. a URI corresponding to the compound to lookup. The resulting query is provided as input to all plugins directly connected to the endpoint in the workflow (i.e. the IMS, IRS and Chemical Structure Search plugins).

The parameterized SPARQL queries are used to join the data across the datasets at a schema level, i.e. the specific columns to retrieve from each of the datasets is determined by the SPARQL query issued. For each data retrieval task, e.g. `compoundLookup`, `targetLookup`, etc., the set of columns to retrieve from each of the datasets has been determined in consultation with the scientific users of the platform. The addition of new datasets and the evolution of existing schemas are facilitated since the changes to SPARQL queries are made at a single point. Additionally, the parameterized SPARQL queries have been optimized for performance. This is particularly important due to the potential for large volumes of data and the complexity of performing query time mapping of identifiers.

5.7. User interfaces

Within the Open PHACTS project there are several exemplar user interface applications being developed each with a different drug discovery task as a focus. The OPS platform provides a common semantic web platform which can be accessed by these drug discovery applications through the core API.

For a general browsing interface over the linked data, a version of an internal system used in the pharmaceutical company Lundbeck, named LSP4All, was extended and connected to the platform. This interface, now branded the OPS Explorer, allows for searching, retrieving and browsing of data. As a result some of the typical user interfaces experienced by a pharmaceutical company researcher have been converted to rely on Semantic Web data. Figure 4 shows the OPS Explorer returning the integrated information about the compound aspirin.

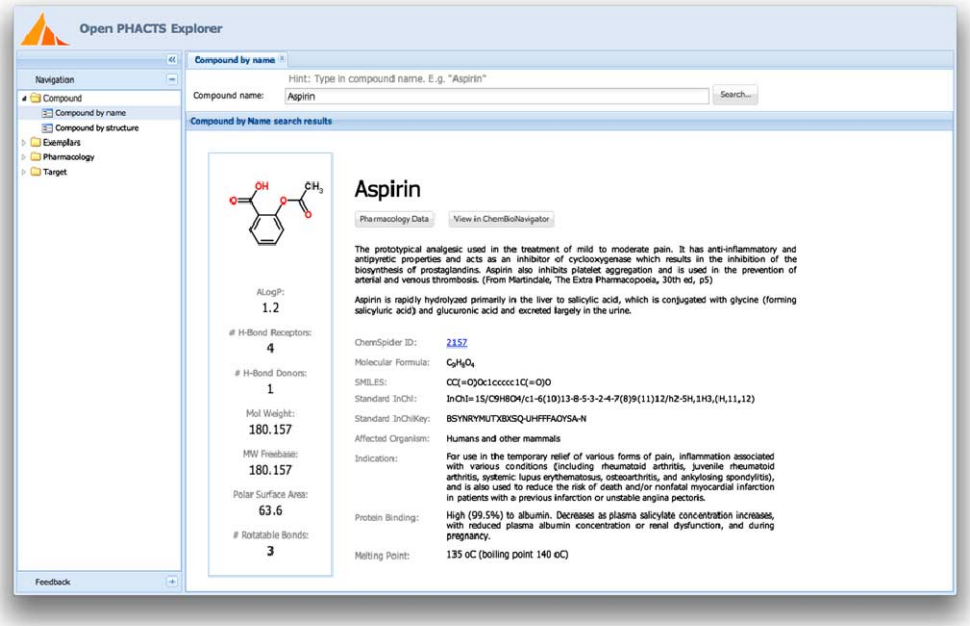

Fig. 4. Compound by name look up through the OPS Explorer interface.

In addition to browsing and searching for data, researchers utilize scientific literature in their work. To support this process the Utopia Documents [1] software was connected to the platform to allow users to view chemical compounds associated with a PDF article, see Fig. 5.

Finally, a more specialized interface for curating biological pathways, PathVisio [43], was connected to the platform, see Fig. 6. This connection allowed PathVisio to provide additional pathway information to the user as they review a pathway.

Thus, based on the same common platform, we have created a prototype for an enriched pharmacology workspace software platform. Additional exemplar applications using the platform are being developed by the Open PHACTS consortium.

6. Conclusions

We presented the architectural decisions for implementing a linked data platform to support drug discovery that integrates the pre-competitive openly available data from many semantic web sources. The motivation for the functionality provided by the OPS linked data platform was driven by real business questions for drug discovery in pharmacology gathered from the Open PHACTS partners. The platform supports different views over the data through stand-off mappings which are applied at query time, i.e. it is possible to have the system relate genes and proteins in one case but not another. The versatility of the platform has been demonstrated by the wide variety of drug discovery applications that can be connected to the integrated data.

Future work is to expand the functionality of the system to a wider set of the business questions. This will also require linking a larger number of data sources, and thus increasing the size and complexity of the linked data cache.

An alpha version of the OPS platform is currently available to the Open PHACTS consortium. A first public release will be made in late 2012, see http://www.openphacts.org/ for details.

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 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.

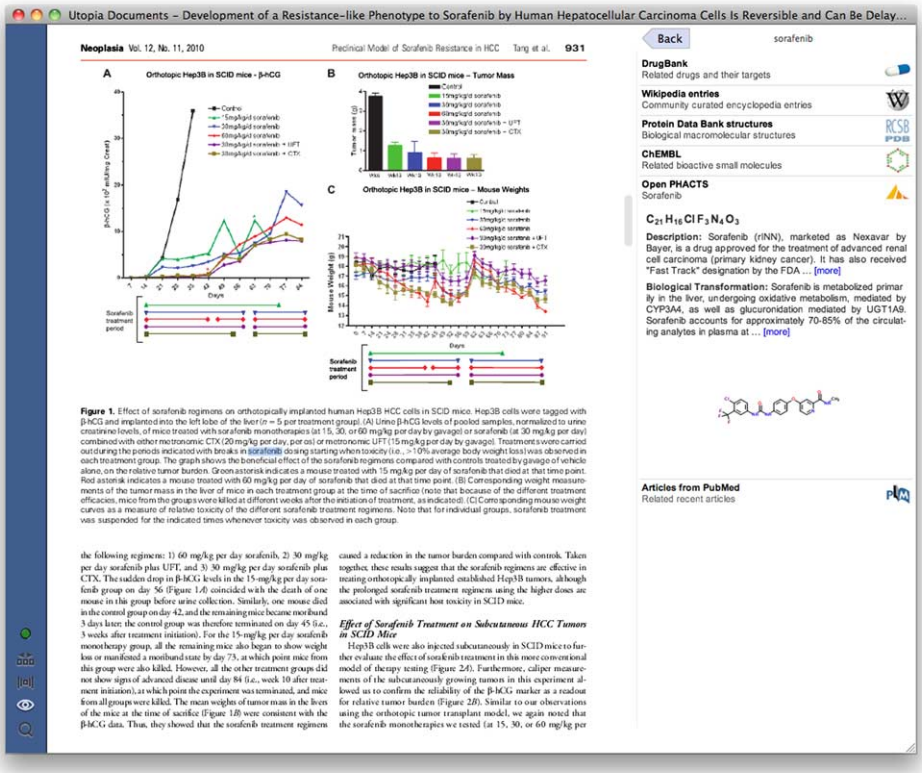

Fig. 5. Article enrichment in utopia documents using the OPS platform.

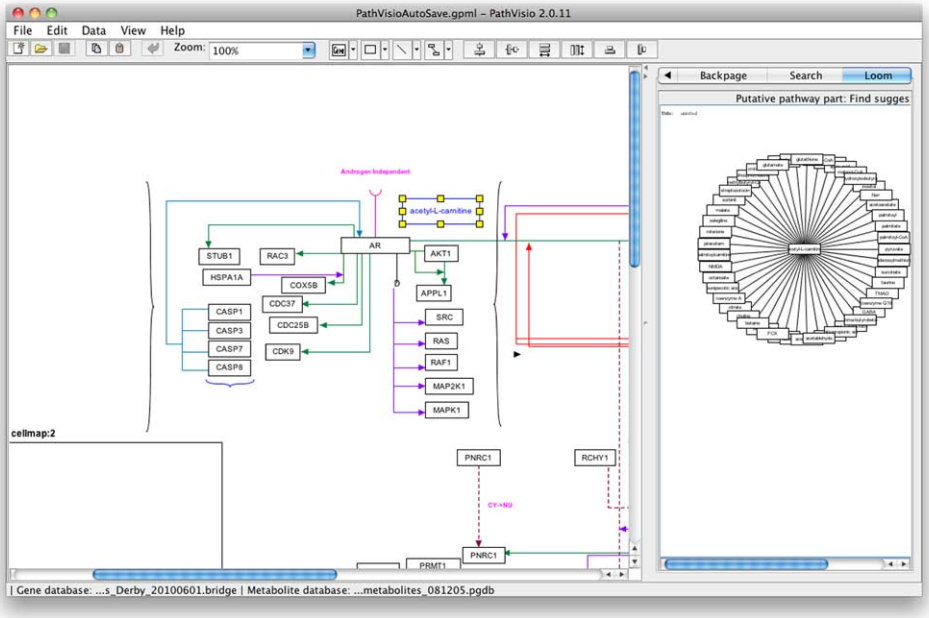

Fig. 6. PathVisio using the OPS platform to suggest relationships with new genes or metabolites (right side window) for a gene selected in a pathway (left side window).

References

[1] T.K. Attwood, D.B. Kell, P. McDermott, J. Marsh, S.R. Pettifer, and D. Thorne, Utopia documents: Linking scholarly literature with research data, *Bioinformatics* (Oxford, England) **26**(18) (2010), i568–i574. doi:10.1093/bioinformatics/btq383.

[2] K. Azzaoui, E. Jacoby, S. Senger, E. Cuadrado Rodríguez, M. Loza, B. Zdrazil, M. Pinto, A.J. Williams, V. de la Torre, J. Mestres, O. Taboureau, M. Rarey, and G.F. Ecker, Analysis of the scientific competency questions followed by the IMI Open PHACTS consortium for the development of the semantic web-based molecular information system OPS, 2012. To be published.

[3] A. Bairoch, The ENZYME database in 2000, *Nucleic Acids Research* **28** (2000), 304–305. doi:10.1093/nar/28.1.304.

[4] A. Bairoch, B. Boeckmann, S. Ferro, and E. Gasteiger, Swiss-prot: Juggling between evolution and stability, *Briefings in Bioinformatics* **5**(1) (2004), 39–58. doi:10.1093/bib/5.1.39.

[5] J. Barrasa Rodríguez and A. Gómez-Pérez, Upgrading relational legacy data to the semantic web, in: *Proc. of the 15th International Conference on World Wide Web (WWW 2006)*, ACM, New York, NY, USA, 2006, pp. 1069–1070. doi:10.1145/1135777.1136019.

[6] K. Belhajjame, N.W. Paton, A.A.A. Fernandes, C. Hedeler, and S.M. Embury, User feedback as a first class citizen in information integration systems, in: *Fifth Biennial Conference on Innovative Data Systems Research (CIDR 2011)*, 2011, pp. 175–183, http://www.cidrdb.org, http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper21.pdf.

[7] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, Bio2RDF: Towards a mashup to build bioinformatics knowledge systems, *Journal of Biomedical Informatics* **41**(5) (2008), 706–716. doi:10.1016/j.jbi.2008.03.004.

[8] T. Berners-Lee, Linked data. Technical report, W3C, 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[9] J. Broekstra, A. Kampman, and F. van Harmelen, Sesame: A generic architecture for storing and querying rdf and rdf schema, in: *First International Semantic Web Conference (ISWC 2002)*, Springer, Berlin/Heidelberg, 2002, pp. 54–68. doi:10.1007/3-540-48005-6_7.

[10] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi, M. Rodriguez-Muro, R. Rosati, M. Ruzzi, and D. Fabio Savo, The mastro system for ontology-based data access, *Semantic Web* **2**(1) (2011), 43–53. doi:10.3233/SW-2011-0029.

[11] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D.J. Wild, Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data, *BMC Bioinformatics* **11**(1) (2010), 255. doi:10.1186/1471-2105-11-255.

[12] L.L. Chepelev and M. Dumontier, Semantic web integration of cheminformatics resources with the SADI framework, *Journal of Cheminformatics* **3**(1) (2011), 16. http://www.jcheminf.com/content/3/1/16.

[13] A. Cheptsov, M. Assel, G. Gallizo, I. Celino, D. DellAglio, L. Bradesko, M. Witbrock, and E. Della Valle, Large Knowledge Collider. A service-oriented platform for large-scale semantic reasoning. in: *Proc. of the International Conference on Web Intelligence, Mining and Semantics (WIMS2011)*, 2011.

[14] The UniProt Consortium, Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Research* **40**(D1) (2012), D71–D75. doi:10.1093/nar/gkr981.

[15] D. De Roure, C. Goble, and R. Stevens, The design and realisation of the myExperiment virtual research environment for social sharing of workflows, *Future Generation Computer Systems* **25** (2009), 561–567. doi:10.1016/j.future.2008.06.010.

[16] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, Chebi: A database and ontology for chemical entities of biological interest, *Nucleic Acids Research* **36**(suppl 1) (2008), D344–D350. http://dx.doi.org/10.1093/nar/gkm791.

[17] O. Erling and I. Mikhailov, *Virtuoso: RDF Support in a Native RDBMS*, 2010, pp. 501. doi:10.1007/978-3-642-04329-1_21.

[18] P. Fox, D. McGuinness, R. Raskin, and K. Sinha, A volcano erupts: Semantically mediated integration of heterogeneous volcanic and atmospheric data, in: *Proc. of the ACM First Workshop on CyberInfrastructure: Information Management in eScience*, ACM, New York, NY, USA, 2007, pp. 1–6. doi:10.1145/1317353.1317355.

[19] A. Gaulton, L. Bellis, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, R. Akhtar, F. Atkinson, A.P. Bento, B. Al-Lazikani, D. Michalovich, and J.P. Overington, ChEMBL: A large-scale bioactivity database for chemical biology and drug discovery, *Nucleic Acids Research. Database Issue.* **40**(D1) (2012), D1100–D1107. doi:10.1093/nar/gkr777.

[20] J. Goecks, A. Nekrutenko, J. Taylor, and The Galaxy Team, Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biology* **11**(8) (2010), R86, 8. doi:10.1186/gb-2010-11-8-r86.

[21] A.Y. Halevy, M.J. Franklin, and D. Maier, Principles of dataspace systems, in: *Proc. of the Twenty-Fifth Symposium on Principles of Database Systems (PODS 2006)*, ACM, New York, NY, USA, 2006, pp. 1–9. doi:10.1145/1142351.1142352.

[22] A.Y. Halevy, A. Rajaraman, and J.J. Ordille, Data integration: The teenage years, in: *Proc. of 32nd International Conference on Very Large Data Bases (VLDB)*, ACM, New York, NY, USA, 2006, pp. 9–16. http://www.vldb.org/conf/2006/p9-halevy.pdf.

[23] S. Harris, N. Lamb, and N. Shadbolt, 4store: The design and implementation of a clustered rdf store. *Time* **42**(8) (2009), 81–96. http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-517/ssws09-paper7.pdf.

[24] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, 1st ed. Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 1, Morgan & Claypool, 2011. doi:10.2200/S00334ED1V01Y201102WBE001.

[25] E. Jacoby, K. Azzaoui, S. Senger, E. Cuadrado Rodríguez, M. Loza, B. Zdrazil, M. Pinto, A.J. Williams, V. de la Torre, J. Mestres, O. Taboureau, M. Rarey, and G.F. Ecker, Scientific requirements for the next generation semantic web-based chemogenomics and systems chemical biology molecular information system OPS, *Computational Chemogenomics*, 2012. To appear.

[26] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. Suzek, M. Martin, P. McGarvey, and E. Gasteiger, Infrastructure for the life sciences: design and implementation of the UniProt website, *BMC Bioinformatics* **10**(1) (2009), 136+. doi:10.1186/1471-2105-10-136.

[27] N. Juty, N. Le Novere, and C. Laibe, Identifiers.org and MIRIAM registry: Community resources to provide persistent identification, *Nucleic Acids Research* **40**(D1) (2012), D580–D586. doi:10.1093/nar/gkr1097.

[28] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, Kegg for integration and interpretation of large-scale molecular datasets, *Nucleic Acids Research* **40** (2012), D109–114. doi:10.1093/nar/gkr988.

[29] T. Kelder, M.P. van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C. Evelo, and A.R. Pico, WikiPathways: Building research communities on biological pathways, *Nucleic Acids Research* **40**(D1) (2012), D1301–D1307. doi:10.1093/nar/gkr1074.

[30] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A.C. Guo, and D.S. Wishart, DrugBank 3.0: A comprehensive resource for 'omics' research on drugs, *Nucleic Acids Research* **39** (2011), D1035–1041. doi:10.1093/nar/gkq1126.

[31] M. Lenzerini, Data integration: A theoretical perspective, in: *Proc. of 21st ACM Symposium on Principles of Database Systems (PODS 2002)*, ACM, New York, NY, USA, 2002, pp. 233–246. doi:10.1145/543613.543644.

[32] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao, Scientific workflow management and the kepler system, *Special Issue: Workflow in Grid Systems. Concurrency and Computation: Practice & Experience* **18**(10) (2006), 1039–1065. doi:10.1002/cpe.v18:10.

[33] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, Reactome knowledgebase of human biological pathways and processes, *Nucleic Acids Research* **37**(Database issue) (2009), D619–622. doi:10.1093/nar/gkn863.

[34] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble, Taverna, reloaded, in: *Proc. of the 22nd International Conference on Scientific and Statistical Database (SSDBM 2010)*, Springer-Verlag, Heidelberg, Germany, 2010, pp. 471–481. doi:10.1007/978-3-642-13818-8_33.

[35] V. Momtchev, Linked life data: Knowledge extraction and semantic data integration in the pharmaceutical domain. LarKC Pharma Workshop at High Performance Computing Center Stuttgart (HLRS), Stuttgart, Germany, 2011.

[36] E.K. Neumann, E. Miller, and J. Wilbanks, What the semantic web could do for the life sciences, *Drug Discovery Today: BIOSILICO* **2**(6) (2004), 228–236. doi:10.1016/S1741-8364(04)02420-5.

[37] H.E. Pence and A. Williams, ChemSpider: An online chemical information resource, *Journal of Chemical Education* **87**(11) (2010), 1123–1124. doi:10.1021/ed100697w.

[38] A. Ruttenberg, J.A. Rees, M. Samwald, and M.S. Marshall, Life sciences on the Semantic Web: The Neurocommons and beyond, *Briefings in Bioinformatics* **10**(2) (2009), 193–204. doi:10.1093/bib/bbp004.

[39] M. Samwald, A. Coulet, I. Huerga, R.L. Powers, J.S. Luciano, R.R. Freimuth, F. Whipple, E. Pichler, E. Prud'hommeaux, M. Dumontier, and M.S. Marshall, Semantically enabling pharmacogenomic data for the realization of personalized medicine, *Pharmacogenomics* **13**(2) (2012), 201–212. doi:10.2217/PGS.11.179.

[40] M. Samwald, A. Jentzsch, C. Bouton, C. Kallesoe, E. Willighagen, J. Hajagos, M. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens, Linked open drug data for pharmaceutical research and development, *Journal of Cheminformatics* **3**(1) (2011), 19. doi:10.1186/1758-2946-3-19.

[41] A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt, Fedx: Optimization techniques for federated query processing on linked data, in: *International Semantic Web Conference (1)*, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 601–616. doi:10.1007/978-3-642-25073-6_38.

[42] SYSTAP, LLC, Bigdata: Approaching web scale for the semantic web, Whitepaper, SYSTAP, LLC, 2009. http://www.bigdata.com/whitepapers/bigdata_whitepaper_07-08-2009.pdf.

[43] M. van Iersel, T. Kelder, A. Pico, K. Hanspers, S. Coort, B. Conklin, and C. Evelo, Presenting and exploring biological pathways with PathVisio, *BMC Bioinformatics* **9**(1) (2008), 399. doi:10.1186/1471-2105-9-399.

[44] A.J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E.L. Willighagen, C.T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons, Open PHACTS: Semantic interoperability for drug discovery, *Drug Discovery Today* **17**(21–22) (2012), 1188–1198.

[45] E. Willighagen, J. Alvarsson, A. Andersson, M. Eklund, S. Lampa, M. Lapins, O. Spjuth, and J. Wikberg, Linking the resource description framework to cheminformatics and proteochemometrics, *Journal of Biomedical Semantics* **2**(Suppl 1) (2011) S6. doi:10.1186/2041-1480-2-S1-S6.