Open PHACTS: A semantic knowledge infrastructure for public and commercial drug discovery research

The Open PHACTS Consortium
Corresponding Author: Lee Harland
ConnectedDiscovery Ltd, London, UK
pmu@openphacts.org

1 Introduction

Technology advances in the last decade have led to a "digital revolution" in biomedical research. Much greater volumes of data can be generated in much less time, transforming the way researchers work [1]. Yet, for those seeking to develop new drugs to treat human disease, the task of assembling a coherent picture of existing knowledge from molecular biology to clinical investigation, can be daunting and frustrating. Individual electronic resources remain mostly disconnected, making it difficult to follow information between them. Those that contain similar types of data can describe them very differently, compounding the confusion. It can also be difficult to understand exactly where specific facts or data points originated or how to judge their quality or reliability. Finally, scientists routinely wish to ask questions that the system does not allow, or ask questions that span multiple different resources. Often the result of this is to simply abandon the enquiry, significantly diminishing the value to be gained from existing knowledge. Within pharmaceutical companies, such concerns have led to major programmes in data integration; downloading, parsing, mapping, transforming and presenting public, commercial and private data. Much of this work is redundant between companies and significant resources could be saved by collaboration [2]. In an industry facing major economic pressures [3], the idea of combining forces to "get more for less" is very attractive and is arguably the only feasible route to dealing with the exponentially growing information landscape.

The development of a scalable, semantic system holding critical, interoperable decision-making data could provide many benefits regarding the issues outlined above. Approaches based in semantic web technologies are an attractive option in part due to their technical capabilities, but critically, they also provide an open standard at the core of such a shared infrastructure. Pharmaceutical companies, academia and non-profit drug-discovery organisations increasingly work in a network of partnerships with other organisations. This makes common, open standards critical to the success of any integration effort for modern drug discovery [4]. In addition, a vendor-neutral infrastructure based on an open platform would ensure the widest possible potential for adoption across commercial and non-profit sectors alike. It is with this goal in mind that the Open PHACTS (Open **Pha**rmacological **C**oncept **T**riple **S**tore) project [5] was conceived. Open PHACTS is a major public-private partnership involving organisations from major pharmaceutical companies, academia and small-medium enterprises (SMEs). The project is funded by the European Federation of Pharmaceutical Industries and Associations (EFPIA) and the European Union through the Innovative Medicines Initiative [6] and scheduled to complete in early 2014.

2 Strategic Aims

Semantic technologies have already gained much traction within biomedicine with initiatives such as the World Wide Web Consortium Semantic Web Health Care and Life Sciences Interest Group [7] Bio2RDF [8], Chem2Bio2RDF [9] and many others. Open PHACTS is complementary to such efforts, focusing on the creation of a task focused, production-grade and sustainable public-private infrastructure. This latter point is critical, as the output from the project should form the foundation for future pre-competitive efforts. As the scientific domain is large, the initial phase of the project focused on defining scope by generation and ranking of the key scientific questions that scientists within drug discovery would like to be able to ask of such a system [10]. An excerpt from these questions is shown in Box 1 and provides clear direction for both the data required and software functionality, as well as strong criteria for measuring success.

- 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 disease/indication provide all targets in the pathway and all active compounds hitting them

Box 1. Example scientific questions the Open PHACTS system will address

3 Data Standards & Quality

High quality data is at the heart of this endeavour. Wherever possible, Open PHACTS will re-use and augment existing efforts in the life-sciences domain by promoting the publishing of pharmacological data in RDF using W3C standards and publically available ontologies. By actively engaging with data providers, we hope to further develop the standards and tooling require for full semantic representation of required content. For instance, we have collaborated with the ChEMBL [11] group, to implement the Quantities, Units, Dimensions and Types ontology (QUDT, [12]) to enable quantitative data queries and interconversion between different classes of measurement. Quality issues are also being addressed, particularly the representation of chemicals on the semantic web. This is still far from optimal and often due to inaccuracies in the electronic representation of these molecules in their source databases through error or subtly different algorithms for creating them. When these are published as linked data, incorrect or missing links have a significantly detrimental effect on data analysis. To address this, the project has developed very detailed guidelines for structure processing and normalisation (based on guidelines from the US Food and Drug Administration [13]) that will deliver more consistency between different databases. In addition, RDF describing novel chemical structure quality metrics and multiple physiochemical properties is being generated for each chemical using software from ChemSpider [14] and ACD/Labs [15] respectively, contributing novel and important data to the community.

Prominence is also given to the provenance of the data within the system, to benefit both consumers (who know where a particular "fact" came from) and producers (crediting the original source). Each Open PHACTS data set is accompanied by a VoID [16] based specification, enhanced with provenance information encoded using the Provenance Authoring and Versioning ontology [17]. We are also actively contributing to the W3C Provenance task [18] to help define, and ensure alignment with this emerging standard. Finally, the project is using nanopublications [19] to record information for individual assertions, both from databases and those generated by individuals through annotation tools. By enabling users to understand where the answer to their query actually came from, we hope to promote data citation [20] and provide credit to those producing important scientific assertions.

4 Technical Architecture

The overall architecture of the Open PHACTS core platform is shown schematically in Fig. 1. The approach is to create a modular system, based on the reuse of existing software and protocols rather than developing these from scratch. While this still requires development to make these components robust enough for a "production-grade" system, it leverages and supports existing efforts by the life science informatics community.

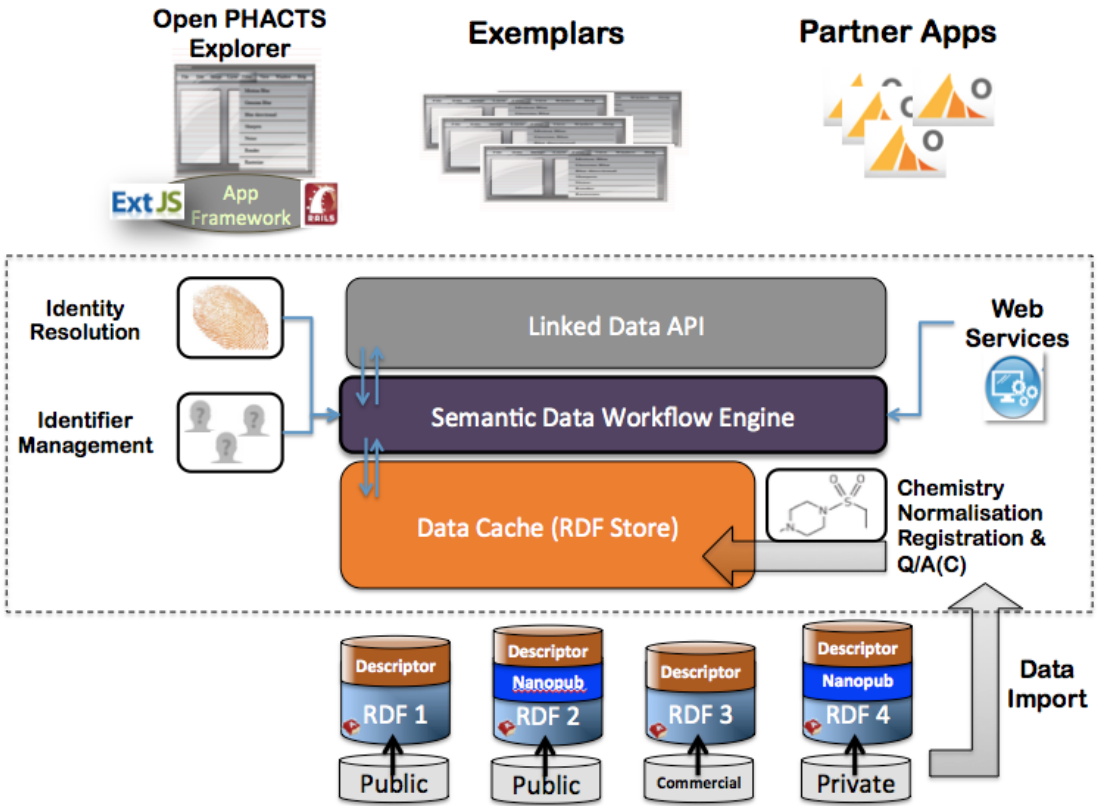

Fig. 1. The Open PHACTs Core Architecture

As outlined above, data are sourced as RDF and VoID descriptors are created if not already present. These data are loaded into a triple store and are expected to be in the range of 50-80 billion triples by

the end of the project. As humans think in terms of natural language and not RDF, the Identity Resolution Service (provided by ConceptWiki [21]) recognises the objects within user-entered text strings and output corresponding URIs for downstream queries. Unfortunately, concepts in life science suffer from a proliferation of identifiers and URIs [22] often making cross-linking of data difficult. However, prescribing rigid rules for URIs that data must use (e.g. "Open PHACTS only accepts "purl"-based Uniprot identifiers for proteins) would significantly limit the content that can be consumed. Even worse, conversion and over-writing of identifiers in the data source would mean that the system no longer faithfully represents the original data and make it more difficult to track down the cause of unexpected results. Therefore, Open PHACTS takes a liberal approach and accepts any identifier namespace and syntax for a given entity class, as long as mappings to URIs that are already known to the system are provided. These mappings are resolved at query time using the Identifier Mapping Service (IMS), which incorporates a semantic version of the BridgeDB system [23]. An additional benefit of this is that the definition of what is "the same" can be adjusted to suit the query. These definitions will be governed by the use of profiles that instruct the IMS which URI classes can be combined. For instance, under certain circumstances it may be reasonable to join data attached to URIs for a protein and its equivalent gene or between a chemical compound and a drug of which it is a part; but not in others.

The IRS and IMS are distinct modules, exposing functionality via web-services. The Large Knowledge Collider (LarKC, [24]) provides the necessary middleware required to combine invocation of these services with data in the triple store within a single SPARQL query. This allows for further federation, for example, incorporating the results of live chemical structure similarity searches (provided by the ChemSpider API) into the current query. Ultimately, it is at this point that the SPARQL queries integrate data by joining across multiple different data sets and address the specific questions that cannot be otherwise easily answered by scientists. These queries are packaged into the Open PHACTS API, implementing the Linked Data API specification [25] and providing a highly accessible mechanism for non-semantic web developers to access the system via the RESTful/JSON paradigm. The API is used by the "Open PHACTS Explorer" which is a web-based user interface to the system, built on the Lundbeck Life Science Platform [26]. The Explorer will allow end-users to browse data within the platform, perform simple queries and aggregations and export results for use in downstream tools such as chemistry analysis software, pathway analyses and Excel.

5 Sustainability

While the Open PHACTS Explorer will address the immediate need to deliver a tool to end-users, it is not intended to be the only portal to the system. Rather, the major goal of the project is to create an "application ecosystem" in which non-profit and commercial organisations consume data via the Open PHACTS API for specific scientific applications. We hope that the availability of a professionally hosted, high-quality, integrated pharmacology data platform, developed using standards endorsed by many major pharmaceutical companies should present an attractive resource that commercial organisations can licence and use to enhance their own offerings. Combined with other revenue

streams, this will be crucial in tackling perhaps the biggest challenge within the project, namely sustaining the development of this unique infrastructure after the current funding expires. To that end, the project includes a number of exemplar applications, designed to demonstrate this principle and "kick-start" the application offerings. In addition, Open PHACTS has a flourishing partnership programme, by which interested parties can learn more and identify opportunities for collaboration or API deployment.

6 Conclusion

Open PHACTS is a unique initiative, bringing major industry and non-profit groups together to develop a shared platform for integration and knowledge discovery. The project aims to deliver on multiple fronts, enhancing the quality of relevant RDF data, addressing key scientific bottlenecks, developing and promoting open standards and creating a lasting infrastructure for cross-sector collaboration. A first version of the software is due for release in late 2012 and will be announced via the project website [5]. We hope that the project will demonstrate that semantic technologies are ready for "prime-time" use as a dependable, core infrastructure for future initiatives in drug discovery.

Acknowledgements

The Open PHACTS project consortium consists of leading experts in semantics, pharmacology and informatics including academic institutions, pharmaceutical companies and scientific software companies. The work described here is the result of contribution by all collaborators in the project. The financial support provided by the IMI-JU project Open PHACTS, grant agreement n◦ 115191 is gratefully acknowledged. Finally, I would like to thank Paul Groth and Antony Williams for valuable feedback on the abstract.

References

1. Kell DB, Oliver SG: Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays. 26, 99-105 (2004)
2. Barnes MR, Harland L, Ford SM, Hall MD, Dix I, Thomas S, Williams-Jones BI, et al.: Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. Nat. Rev. Drug Discov. 8, 701-8 (2009)
3. Scannell JW, Blanckley A, Boldon H, Warrington B: Diagnosing the decline in pharmaceutical R&D efficiency. Nat. Rev. Drug Discov. 11, 191-200 (2012)
4. Harland L, Larminie C, Sansone SA, Popa S, Marshall MS, Braxenthaler M, Cantor M, et al.: Empowering industrial research with shared biomedical vocabularies, Drug Discov. Today. 21, 940-7 (2011)
5. http://openphacts.org
6. http://imi.europa.eu
7. http://www.w3.org/blog/hcls

8. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette JJ: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. 41, 706-16 (2008)
9. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics. 12, 256 (2010)
10. Azzaoui K, Jacoby E, Senger S, Rodríguez EC, Loza M, Zdrazil B: Scientific questions to a next generation semantically enriched biomolecular internet resource. *In Preparation*
11. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, et al.: ChEMBL: a large-scale bioactivity database for drug discovery. 40, D1100-7 (2011)
12. http://www.qudt.org
13. http://1.usa.gov/snNNdn
14. http://www.chemspider.com
15. http://www.acdlabs.com
16. http://vocab.deri.ie/void
17. http://swan.mindinformatics.org/spec/1.2/pav.html
18. http://www.w3.org/TR/prov-primer
19. Groth P, Gibson A, Velterop, J: The anatomy of a nanopublication Paul Groth, Andrew Gibson, Jan Velterop Information Services and Use. 30, 51-56 (2010)
20. Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, van Mulligen E, et al.: The value of data. Nat. Genet. 43, 281–283 (2011)
21. http://conceptwiki.org
22. http://www.slideshare.net/dullhunk/the-seven-deadly-sins-of-bioinformatics
23. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, et al.: The BridgeDB framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11, 5 (2010)
24. Fensel D, van Harmelen F, Andersson B, Brennan P, Cunningham H, Emanuele DV, Fischer F, et al.: Towards LarKC: a Platform for Web-scale Reasoning Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2008) 8 (2008)
25. http://code.google.com/p/linked-data-api/
26. Kallesøe, CS: Building scalable solutions with open source tools. In: Harland L and Forster F. (eds.) Open source software in life science research (Woodhead Publishing Series in Biomedicine), London 2012