IMI1 Final Project Report Public Summary

Project Acronym: OPEN PHACTS

Project Title: The Open Pharmacological Concepts Triple Store

Grant Agreement: 115191
Project Duration: 01/03/2011 - 29/02/2016

1. Executive summary

1.1 Project rationale and overall objectives of the project

Drug discovery is increasingly relying on big data, and all major pharmaceutical companies maintain extensive in-house instances of public biomedical and chemical data alongside internal data. Analysis and hypothesis generation for drug-discovery projects requires careful assembly, overlay and comparison of data from many sources, requiring shared identifiers and common semantics. Utility of data-driven research goes from virtual screening, high-throughput screening (HTS) analysis, via target fishing and secondary pharmacology to biomarker identification. Alignment and integration of internal and public data and information sources is a significant effort and the process is repeated across companies, institutes and academic laboratories. This represents a significant waste and an opportunity cost. In order to overcome this major barrier, the Open PHACTS project aimed at building an Open Pharmacological Space, freely accessible to both Industry and Academia.

This was achieved by developing and providing the Open PHACTS Discovery Platform, which integrates publicly available pharmacological data from a variety of information resources. The latest version connects 13 data sources and allows queries for compounds, targets, pathways, and diseases. The whole system is built upon a generic semantic data integration infrastructure, the Open Pharmacological Concepts Triple Store. Additionally, an API and appropriate services are provided for complex pharmacological research use cases. In the course of the project, Example Applications have been built which demonstrate its possible use, both via workflows and through commercial applications. In a broader context Open PHACTS also developed community best practices in data provenance and licensing and defined standards for data integration in the life science domain.

The platform, the API and most of the data are available under an open source and open access model. To leverage and sustain the results of the Open PHACTS project, a well-considered sustainability plan was developed, which resulted in the creation of the Open PHACTS Foundation, a not-for-profit membership organisation.

1.2 Overall deliverables of the project

The execution of the project was based on 195 deliverables and 10 milestones, which all have been achieved. A first and major step was the selection of a set of data sources to be interlinked, out of a set of more than 1.400 publicly available databases. This was achieved by creating a set of prioritised use cases at the very beginning (research business questions), which formed the basis for all strategic decisions with respect to selection of data sources[2] (see M1). For the identification of additional data sources relevant for the Open PHACTS Discovery Platform, the concept of Researchathons was developed and implemented. In a Researchathon, scientists are meeting face to face, discuss new challenges, solidify them in concrete research questions, identify the relevant data sources, and put this forward to the technical team for a feasibility study.

[2] Williams A.J, Harland L, Groth P, et al: Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today, Volume 17, Issues 21-22, pp 1188-1198, November 2012, doi:10.1016/j.drudis.2012.05.016

Since one of the major aims of the project was to enable both non-commercial and commercial exploitation, licensing was immediately identified as one of the major risks for implementation. This was solved by selecting and developing a set of standard licenses[3] for the RDF versions of the data provided to the Open PHACTS Triple Store, which are considered as best practice by the community. Concrete implementation has been shown by a set of Example Applications, which have been developed during the course of the project[4]. These include also an application jointly developed with the IMI project eTOX (Collector).

For the technical implementation, the concept of hackathons (computer scientists meet and code first prototypes) and Usathons (computer scientists and medicinal chemists work together on implementation of concrete tasks) was pursued. This allowed the internal release of a first prototype already after 12 months (M3), and the release of a prototype to the broader community after 24 months (M4). Currently, Version 2.0 of the Open PHACTS Discovery Platform is online, and secure access for EFPIA companies is established (M6, M8). During technical implementation, numerous new and innovative solutions for newly identified problems were developed, such as *scientific lenses*[5] to support task specific views of the data, taxonomies for important target classes such as transporters (see D 1.5.1), receptors (see D 1.5.2) and ADME (see D 1.5.3), just to mention a few.

A major achievement of the project constitutes a set of standardised API calls, which have also been implemented in KNIME and Pipeline Pilot. These allow the fast and easy setup of complex queries combined with powerful analysis features. Examples of the exploitation of these workflow tools in daily research are already documented in several publications[6], and reach out to other IMI projects, such as eTOX (M9), as well as to H2020 projects (EU-ToxRisk). Engagement of the scientific community was not only achieved by over 40 publications in peer reviewed scientific journals, and over 160 presentations at scientific conferences, the project also developed and implemented a process for identifying and selecting Associate Partners[7], who sign a memorandum of understanding and are the first to hear about the latest developments and have the opportunity to present their ideas at Open PHACTS Community Workshops[8] and Researchathons (M5).

Finally, sustainability of the Open PHACTS Discovery Platform is secured by the Open PHACTS Foundation (OPF), a not-for-profit membership organisation which was set up as a result of the sustainability plan of the Open PHACTS project. The OPF has currently 4 members and is already partner in two large H2020 projects (BDE and EU-ToxRisk).

---

[3] http://support.openphacts.org/support/solutions/articles/131704-what-data-is-in-the-system-exactly-what-are-the-versions-of-data-sources-

[4] http://www.openphacts.org/2/sci/apps.html

[5] Batchelor C, Brenninkmeijer C.Y.A, Chichester C, et al: Scientific Lenses to Support Multiple Views over Linked Chemistry Data. The Semantic Web-ISWC 2014, Volume 8769 of the series Lecture Notes in Computer Science, 2014, pp 98-113 (http://link.springer.com/chapter/10.1007%2F978-3-319-11964-9_7)

[6] Montanari F, Zdrazil B, Digles D, Ecker G.F: Selectivity profiling of BCRP versus P-gp inhibition: from automated collection of polypharmacology data to multi-label learning. Journal of Cheminformatics, 4 February 2016, doi:10.1186/s13321-016-0121-y, and
Chichester C, Digles D, Siebes R, et al: Drug Discovery FAQs: workflows for answering multidomain drug discovery questions. Drug Discovery Today, Volume 20, Issue 4, April 2015, pp 399-405, doi:10.1016/j.drudis.2014.11.006

[7] http://www.openphactsfoundation.org/people/associated-partners/

[8] http://www.openphactsfoundation.org/category/events/community-workshops/

1.3 Summary of progress versus plan since last period

Besides all the achievements listed under 1.4, the final year of the project was devoted to finalise and complete the following tasks:

Connecting EFPIA in house data to the Open PHACTS Discovery Platform
Although secure access to the Open PHACTS Discovery Platform has been achieved, EFPIA companies are reluctant to store their in house data at an external service provider. Thus, the development and implementation of a virtual machine, which contains the entire application suite and can be deployed within the boundaries of the company, was seen as the technical solution for this task. In a pilot study, a virtual machine based on API version 1.5 was delivered to Lilly and successfully tested in house. The Docker images developed for this are available at https://hub.docker.com/u/openphacts/.

Integration of patent data (SureChEMBL) into the Open PHACTS Discovery Platform
Integration of patent data into the Open PHACTS Discovery Platform is seen as key to engaging EFPIA companies for financial long term commitments to the Open PHACTS Foundation and thus sustainability of the Open PHACTS Discovery Platform. SureChEMBL data have been compiled (including data to the required RDF format, along with Linksets and Chemistry entities registered in the database) and were integrated into the Open PHACTS Discovery Platform. However, in the current version the SMILES codes for the compounds have to be taken from EBI, as RSC could not deliver them in time. This does not affect the functionality of the system, but is only the second best solution since although all the patent compounds have undergone the RSC-based chemistry standardisation process this is currently not reflected in the SMILES returned by the 'Compound information' API call. A testing procedure has been developed to assess the completeness of the data, and pilot studies have been performed by GSK & SureChEMBL team and also as part of a hackathon looking further ahead to future extensions. Furthermore, patents have been annotated with targets and diseases, where appropriate.

Integration of commercial data
Major commercial data sources identified were GVK BIO, Thomson Reuters, and Aureus. Based on the level of interest by pharma companies the decision was taken to collaborate with GVK BIO. However, although there was willingness to participate in a pilot study, it was not feasible to successfully complete a pilot study in the time available due to circumstances outside of our control (e.g. staff turnover at GVK BIO). Thus, we decided to take a data set which is part of an ongoing collaboration involving Lilly, NIH-NCATS, and Data2Discovery. The role of Data2Discovery is informatics: transforming and integrating data to enhance semantic value, development of a Knowledge Network (KN), a publicly shared Open Phenotypic Drug Discovery Resource (OPDDR, aka PD2) which can be used to identify relationships between National Pharmaceutical Collection (NPC) compounds, phenotypic assays, ontological classes of assays, and associated public data on related molecular targets. The data set used for the pilot refers to the Open Phenotypic Drug Discovery Resource, which includes the NIH NCATS phenotypic data (in-house generated by Lilly) and has been published openly later on. The resulting Phenotypic assay metadata have been made accessible via the Open PHACTS Discovery Platform via the loading of the linked data set of phenotypic assays aligned to BioAssay Ontology. They are surfacing through new specific Assay API calls.

Allowing the use of directionality (causality) relationships when evaluating pathways.
This task was directed at allowing better evaluations of different drugs that could target different but related targets that are part of the same biological pathway. For this purpose the

formalisation of interaction directionality of pathways at WikiPathways itself was improved and content from the expert curated Reactome pathway collection was added. All that information was captured in an extended version of the pathway RDF that is publicly available for download and that was loaded into the Open PHACTS Discovery Platform. New API calls were developed that allow the use of that information to find genes and proteins as drug targets upstream or downstream from a selected target.

1.4 Significant achievements since last report

Technical Task Force (TTF) (WP 1-4):

1. Release of three new API versions with both new data and new functionality:
    a. V1.4 in April 2015[9]
    b. V1.5 in May 2015[10] and corresponding Pipeline Pilot[11] and KNIME nodes
    c. V2.0 in January 2016[12] and corresponding Pipeline Pilot[13] nodes (in April 2016)
2. Managed decommission of old API V1.3 (March 2015) and V1.4 (April 2016)
3. Delivery of Container based (Virtual Machine) hosted version of Open PHACTS to support EFPIA needs for adding their own data into the platform in a secure way
4. Security auditing by two Pharma companies & acceptance of Open PHACTS security approaches
5. Extension and updates to the provision of key platform components such as IMS, and IRS early beta, refresh of the chemistry platform data sources including ChEMBL 20 update
6. Preview calls for SureChEMBL to support early dissemination of target and disease annotation of the SureChEMBL data
7. Using the BioJS (http://biojs.net) framework, Explorer based HTML and JavaScript widgets have been developed which can be embedded in any web page to provide visualizations of Open PHACTS data with the minimum of effort.

Scientific Task Force (STF) (WP 5-6):

1. Continuation of ENSO implementation work for SureChEMBL, Pathway and Cell Localisation use cases
2. Release of a new version of the Open PHACTS Explorer V3.0[14], including a completely new user interface, new data sources, mobile compatibility and the ability for users to favourite entities
3. Further usage of support portal[15] for building the science community
4. Documentation on the user support portal continues to be expanded
5. Delivery of an eApp (CBN) hosted at a Pharma partner site

Management Task Force (MTF) (WP 7-9):

1. Sustainability - The Open PHACTS Foundation:
    a. Became a charity in 2015

[9] http://support.openphacts.org/support/solutions/articles/4000020797-1-4-api
[10] http://www.openphactsfoundation.org/open-phacts-api-version-1-5-is-here/
[11] http://www.openphactsfoundation.org/pipeline-pilot-component-collection-1-5-released/
[12] http://www.openphactsfoundation.org/open-phacts-api-2-0-is-here/
[13] http://www.openphactsfoundation.org/open-phacts-pipeline-pilot-component-collection-2-0/
[14] http://www.openphactsfoundation.org/new-version-of-open-phacts-explorer-released/
[15] https://support.openphacts.org/

   b. Became partner in the H2020 project EU-ToxRisk, which aims to achieve a paradigm shift in toxicology towards a more efficient and animal-free chemical safety assessment
   c. End of 2015 Lilly joined as new member (in total the Foundation currently has 3 industrial members)
   d. The University of Vienna joined as the first academic member in October 2015.
   e. Successful knowledge transfer from the project to the Foundation
2. Successful collaborations with other IMI projects and joint use cases in place (eTOX, K4DD, ELF, DDMoRe - see Milestone 9)
3. Active communication & community engagement:
   a. Open PHACTS received the European Linked Data Award[16]
   b. Project website[17] has been modified to make sure that visitors can easily find and access the information and resources they need. In particular, we now have a new sub-section of the site dedicated to meeting researchers' needs. This transitions the site from having a project focus, to its post-project role supporting the Open PHACTS Discovery Platform and Foundation.
   c. Twitter[18] followers increased to 1.042 (up from 812 in 2014; status 15 April 2016)
   d. Open PHACTS YouTube channel[19] continues to demonstrate use of the API and example applications
   e. 4 project newsletters in 2015/16[20] (1.353 subscribers as of 18 April 2016 with 15-20 new sign ups per month)
   f. high conference attendance in 2015/16[21]
   g. 10 publications in 2015/16[22]
4. Conferences & workshops organised by Open PHACTS:
   a. The 2nd Open PHACTS Researchathon was held on "Understanding the knowledge management needs of phenotypic screening" in Santiago de Compostela in February 2015[23]
   b. We organised an Open PHACTS Pipeline Pilot & KNIME Workshop in Amsterdam in May 2015 to introduce users to how the Open PHACTS API can help answer scientific questions[24]
   c. In February 2016 we organised an Open PHACTS closing conference entitled "Linking Life Science Data: Design to Implementation, and Beyond" in Vienna[25]
5. Project Management:
   a. Organisation of two amendments of the Grant Agreement (12 and 13), including the termination of a participant due to bankruptcy and the addition of a new Third Party
   b. Implementation of an improved teleconference structure to better meet the needs to achieve the additional aims defined in the ENSO proposal

[16] http://www.openphactsfoundation.org/open-phacts-wins-the-european-linked-data-contest/
[17] http://www.openphacts.org
[18] https://twitter.com/Open_PHACTS
[19] https://www.youtube.com/user/OpenPHACTS
[20] http://www.openphactsfoundation.org/category/news/newsletters/
[21] http://www.openphactsfoundation.org/category/events/
[22] http://www.openphactsfoundation.org/category/research-outputs/publications/
[23] http://www.openphactsfoundation.org/open-phacts-phenotypic-screening-workshop/
[24] http://www.openphactsfoundation.org/open-phacts-pipeline-pilot-knime-workshop/
[25] http://www.openphactsfoundation.org/completion-of-the-open-phacts-project/

1.5 Scientific and technical results/foregrounds of the project

The Open PHACTS project has been able to develop new technical deliverables that have enabled to establish the overall platform and wider services. The project has also utilised its experience to set core processes and best practices that have supported deeper understanding of the role of Linked Data to support Life Science informatics and data integration. Standards, technical solutions, as well as processes are well documented and fully open to the public domain (publications, GitHub, entries on our web site).
We list some of the main scientific and technical results/foregrounds of the project, although this is not an exhaustive list and we refer you to other publications and other communications especially with respect to the scientific research:

- Enablement of Bioannotations of SureChEMBL and exposure of the SureChEMBL patent corpus through an API linked to the other data sources in Open PHACTS
- Open PHACTS API is now relied on by a range of informatics applications and workflows and has been deployed within host organisations. Focus on delivering an API as a core architecture principle has also helped other projects in this analysis.
- Research questions to drive the overall requirements analysis has enabled others to take this approach in subsequent projects including Elixir.
- Lenses over Linked Data: An approach to support task specific views of the data has enabled users to investigate different views of data using the flexibility of Linked data associations[26].
- The RDF guide provides details about modelling data as RDF. This specification builds on the RDF Guidelines by defining the metadata that should be published to describe the dataset and the links to other datasets. Guidelines for exposing data as RDF in Open PHACTS: http://www.openphacts.org/specs/2013/WD-rdfguide-20131007/
- DataSet descriptions: http://www.openphacts.org/specs/datadesc/ - The dataset description defined in this specification declares the properties that should be included in the description of dataset or its links. The information is exchanged using the Vocabulary of Interlinked Datasets (VoID). This specification provides details of the metadata expected to describe the datasets and the links that relate the instances in those datasets.
- Identity Mapping Service - Based on BridgeDb (Maastricht University and University of Manchester developed framework) has been extended to support some of the key deliverables of the project around identity mapping and underpins much of the Lens work. There has been interest in using the key learnings in subsequent projects, e.g. the eNanoMapper project is actively adopting it[27].
- Explorer: As well as delivering an updated version of Explorer 3.0, the project has created a series of reusable components that can be used beyond the project.
- VM/Docker – the effort in delivering the Container-based version of Open PHACTS has enabled some key analysis of how to publish data such that it can be versioned in the same way as software.

[26] Batchelor C, Brenninkmeijer C.Y.A, Chichester C, et al: Scientific Lenses to Support Multiple Views over Linked Chemistry Data. The Semantic Web-ISWC 2014, Volume 8769 of the series Lecture Notes in Computer Science, 2014, pp 98-113 (http://link.springer.com/chapter/10.1007%2F978-3-319-11964-9_7)
[27] http://www.enanomapper.net/

- Chemistry Validation – Agreement on a standard set of chemistry validation and normalisation rules, implemented on our chemistry processing workflows via CVSP and the Chemistry Registry System (CRS). CVSP and CRS were planned to be transferred to the public domain by end of the project; due to prioritisation for data processing this is currently not the case, but RSC has committed to the open-sourcing of these following the project end.

1.6 Potential impact and main dissemination activities and exploitation of results

Drug discovery is a very data-intensive discipline, and all major pharmaceutical companies maintain extensive in-house instances of public biomedical and chemical data alongside internal data. Analysis and hypothesis generation for drug-discovery projects requires careful assembly, overlay and comparison of data from many sources, requiring shared identifiers and common semantics. Alignment and integration of internal and public data and information sources is a significant effort and the process is repeated across companies, institutes and academic laboratories. This represents a significant waste and an opportunity cost.

To address these challenges, the Open PHACTS project developed an open source, open standards and open access innovation platform, the Open PHACTS Discovery Platform, using a semantic web approach. This semantic integration hub comprises data, vocabularies and infrastructure needed to accelerate drug-oriented research. It addresses key bottlenecks in small molecule drug discovery, such as disparate information sources, lack of standards and shared concept identifiers, guided by well-defined research questions assembled from participating drug discovery teams from Europe. Thus, the development of the Open PHACTS Discovery Platform was driven by pharmaceutical companies and SMEs based in Europe, which definitely increases the competitiveness of Europe and helps to establish Europe as an attractive place for pharmaceutical research and development, especially in the area of data integration and data interoperability. Furthermore, results obtained from the Open PHACTS Discovery Platform will speed up and economise the drug discovery and development process. Successful exploitation of the system for analysis of phenotypic screening data, neglected diseases and safety assessment has been shown and demonstrates that the system has been taken up by the scientific and industrial community. It is also used by other IMI projects such as the European Lead Factory, and constitutes a core technology in the H2020 projects EU-ToxRisk and BigDataEurope. Via its strong community engagement activities, documented in 65 publications, over 160 presentations at international conferences, 7 workshops and 2 Researchathons, Open PHACTS has driven the development of new standards and influenced major European initiatives, such as ELIXIR and CORBEL. Finally, by founding the Open PHACTS Foundation, a not-for-profit membership organisation based in UK, the Open PHACTS project lays the ground for long term sustainability of the Open PHACTS Discovery Platform.

1.7 Lessons learned and further opportunities for research

The main objective of the Open PHACTS project was to overcome the duplicated efforts in pharmaceutical industry for integrating data from the public domain with their in house data warehouses. This was achieved in a collaborative effort of academia and pharmaceutical industry by developing the Open PHACTS Discovery Platform. Working together in a public private partnership (PPP) had mutual benefits to the overall objectives of the project:

- from day one on the platform was designed to fit the needs of pharmaceutical industry, such as the definition of industry driven research questions, a careful

assessment of legal issues such as licenses, development of standards adopted by both academia and industry, and development and implementation of a sustainability plan, just to mention a few

- presence of academic institutions in the consortium ensured broad dissemination of the activities and results, as this is one of the main assessment criteria in academia;
- SMEs drove the early adoption of the project results and developed prototype applications. This also fostered dissemination and sustainability of the platform
- the agile and flexible way of operation in academic groups led to fast adaption to new technologies and solutions, such as the scientific lenses
- finally, having industry and academia working together on a joint scientific problem also influenced teaching in the academic institutions

However, having different domains and different scientific cultures (industry & academia, chemists & computer scientists) working together in a large consortium (31 partners, >100 scientists) also represented quite a challenge to the consortium. This requires a structured execution of tasks guided by a strong project management. One of the main factors for success was the large amount of face to face meetings, which brought the different communities together right from the beginning of the project. Furthermore, the concepts of Usathons and Researchathons facilitated the formation of mixed teams comprising chemists and computer scientists both from academia and industry, thus ensuring cross domain collaboration. In addition, the implementation of Hackathons and the development of a lash-up within the first 6 months formed a strong technical team, led by a CTO. This allowed the user community in the project to influence the “product” in a very early phase of the development.

During the execution of the Open PHACTS project, several potential new research topics to further advance the field were identified and partly picked up in an ENSO application. These include, among others, the implementation in KNIME and Pipeline Pilot workflows, the application of the Open PHACTS Discovery Platform for phenotypic screening results and for safety assessment, as well as its potential for repurposing of drugs. In the near future we expect an automated creation of RDF files and a stronger involvement of the community in providing and curating data.