Open PHACTS

Milestone 4

Core services released to broader community

Prepared by Pfizer, CD, VUA, UNIVIE
Approved by UNIVIE, GSK, VUA

March 2013
Version 1.0

Project title: An open, integrated and sustainable chemistry, biology and pharmacology knowledge resource for drug discovery
Instrument: IMI JU
Contract no: 115191

Start date: 01 March 2011
Duration: 3 years

Nature of the Deliverable	
Report	**x**
Prototype	
Other	
Dissemination level	
Public dissemination level	**x**
For internal use only	

Definitions

- Partners of the Open PHACTS Consortium are referred to herein according to the following codes:

  Pfizer – Pfizer limited – **Coordinator**
  UNIVIE – Universität Wien – **Managing entity of IMI JU funding**
  DTU – Technical University of Denmark – DTU
  UHAM – University of Hamburg, Center for Bioinformatics
  BIT – BioSolveIT GmbH
  PSMAR – Consorci Mar Parc de Salut de Barcelona
  LUMC – Leiden University Medical Centre
  RSC – Royal Society of Chemistry
  VUA – Vrije Universiteit Amsterdam
  CNIO – Spanish National Cancer Research Centre
  UNIMAN – University of Manchester
  UM – University of Maastricht
  ACK – ACKnowledge
  USC – University of Santiago de Compostela
  UBO – Rheinische Friedrich-Wilhelms-Universität Bonn
  AZ – AstraZeneca
  GSK – GlaxoSmithKline
  Esteve – Laboratorios del Dr. Esteve, S.A.
  Novartis – Novartis
  ME – Merck Serono
  HLU – H. Lundbeck A/S
  E.Lilly – Eli Lilly
  NBIC – Stichting Netherlands Bioinformatics Centre
  SIB – Swiss Institute of Bioinformatics
  ConnDisc – Connected Discovery
  EBI – European Bioinformatics Institute
  Janssen – Janssen Pharmaceutica
  OGL – OpenLink Software

- **Grant Agreement**: The agreement signed between the beneficiaries and the IMI JU for the undertaking of the Open PHACTS project.
- **Project**: The sum of all activities carried out in the framework of the Grant Agreement.
- **Work plan**: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried, out as specified in the Grant Agreement.
- **Consortium**: The Open PHACTS Consortium composed of the above-mentioned legal entities.
- **Project Agreement**: Agreement concluded amongst Open PHACTS participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

1 Introduction

This document sets out the delivery of the Open PHACTS Discovery Platform, a set of software and services designed to address knowledge management issues in the pharmacology domain. It serves as a summary of the first major release of the system to the wider community. However, as the project is moving quickly, readers are encouraged to look to the materials at the Open PHACTS website (http://openphacts.org) to obtain the latest guides to the available software, functionality and data.

2 Timeline & Background

The challenge facing the Open PHACTS development team was considerable: develop a pharmacology-based system that matches the levels of functionality found in mature software of this type, in little over one year, using cutting-edge technology and a distributed development and research team. Building a successful and competent team following a distributed agile development process has been a major achievement within the project, and has resulted in the delivery of the first version of the Open PHACTS Discovery Platform. Below we describe the overall vision and the services that external parties will be able to use in their own scientific analyses.

As the project progressed we took the decision to stagger the release of the Open PHACTS core services to the broader community, rather than simply open up the services to all at once. While the system is technically ready, the rationale here is that we need to manage this process - it is critical that those who use the system get a positive impression and become advocates for the Open PHACTS approach. To do this we needed to prepare the necessary tutorials, communication channels (feedback processes), documentation, and have time to engage with the external community. Simply opening up would leave us spread too thinly, unable to reasonably respond to the many different consumer groups. Thus, we made the decision to open the services in waves, each targeted at a specific community.

3 Launch of the Open PHACTS Explorer

In preparation of a wider release, we decided to open up the first view of the beta of the Open PHACTS Explorer to our Associate Partners. On December 20, 2012 we invited them to register for testing the Explorer and to commit to spending some time in January to do this. 17 Associated Partners were keen to do so.

On January 14, 2013 these 17 volunteers received the link to the Open PHACTS Explorer beta release, the first release of the Explorer, providing a search interface to query the data in different ways, then display and download the integrated information. The Open PHACTS Explorer could be reached via the landing page[1], which contains links to more information, details on the data sources, terms of use and licensing details, video and PDF tutorials, and

[1] www.openphacts.org/explorer

known issues. Associated Partners were asked to register before they access the Open PHACTS Explorer for the first time.
V1.1 of the Open PHACTS Explorer was publicly released on March 4 by activating the link to the landing page on our website under “Open PHACTS Explorer”. On March 6 Gerhard Ecker personally invited 40 Med-Chem groups throughout the world where he personally knows the group leader for further testing of the system.

Next steps in our staggered release plan are to personally invite the participants of EuroQSAR 2012 and ISMC 2012 by end of March as well as members of the EFMC by end of April to use and test the system. Furthermore, for potential customers of the API (software vendors), who are such an important user group, we felt it best to tie this to the community workshop in April 2013, where we can provide face-to-face discussions on the system benefits.

4 The Open PHACTS Discovery Platform

The overall vision of the platform is one that:

- Allows multiple databases covering different scientific domains to be integrated, to allow complex questions to be asked by scientists
- Uses semantic technology to significantly increase the fidelity in representing and integrating this data. Essentially connecting pre-clinical data in a robust, reliable and standardised way
- Enhances the reliability of integrated chemistry by applying strict normalisation and quality control metrics to chemicals from source databases
- Focuses on the provenance of data. The who, what, when and how of every data-point is preserved and accessible, allowing scientists to develop trust in the data they are working with through the system
- Reuses existing standards and software as much as possible. Show how existing, mature community efforts can directly contribute to drug discovery systems rather than “reinventing the wheel”
- Creates a robust, task-orientated user interface that hides the technology and allows scientists to query the system in ways they are already familiar with
- Delivers an API that provides integrated, high-quality chemistry and biology data to application developers. Promoting the use of the system and providing an essential route to sustainability.

With this first release of the Open PHACTS Discovery Platform, we provide a demonstration that bleeding-edge semantic technologies can successfully be applied to core pharmacology use-cases provided by EFPIA companies.

5 Data of Version 1.0

The following data sources are integrated within version 1.0 of the Open PHACTS Discovery Platform:

Source	Version	Supplier	Downloaded	Initial Records	Triples	Properties
Chembl	Chembl 13 RDF (11-Jun-2012)	Maastricht	08 Aug 2012	1,149,792 (~ 1,091,462 compounds, 8845targets)	146,079,194	17 (for compounds) 13 (for targets)
DrugBank	Aug 2008	Bio2Rdf (www4.wiwiss.fu-berlin.de)	08 Aug 2012	19,628 (~14,000 targets, 5000 drugs)	517,584	74
SwissProt	2012_07 (July 11, 2012)	SIB	07 Aug 2012	536,789	156,569,764	78
ENZYME	July 11, 2012	SIB	07 Aug 2012	6,187	73,838	2
ChEBI	Release 94	EBI	08 Aug 2012	35,584	905,189	2
ChemSpider ACD Labs		ChemSpider	08 Aug 2012	1,194,437	161,336,857	22ACD4CS
ConceptWiki		NBIC	07 Aug 2012	2,828,966	3,739,884	1

It should be noted that to integrate these data required extensive cross-referencing of the same scientific concept across all of these databases. This currently stands at over 18 million links across 8 distinct databases and 15 different datasets. Where this data was generated by Open PHACTS, it was released to the scientific community under an open licence[2]. Further, the ACD/Labs data represents over 22 million new data points generated as part of the project, again accessible to the broader scientific community[3].

6 The Software

The vast majority of components are available to developers outside of the project under open source licences, giving others the chance to use all, or parts of the technology in their own projects. A system overview is presented below, along with a brief description of each part.

[2] http://www.openphacts.org/terms-and-conditions
[3] http://www.acdlabs.com/company/media/pr/120522_phacts.php

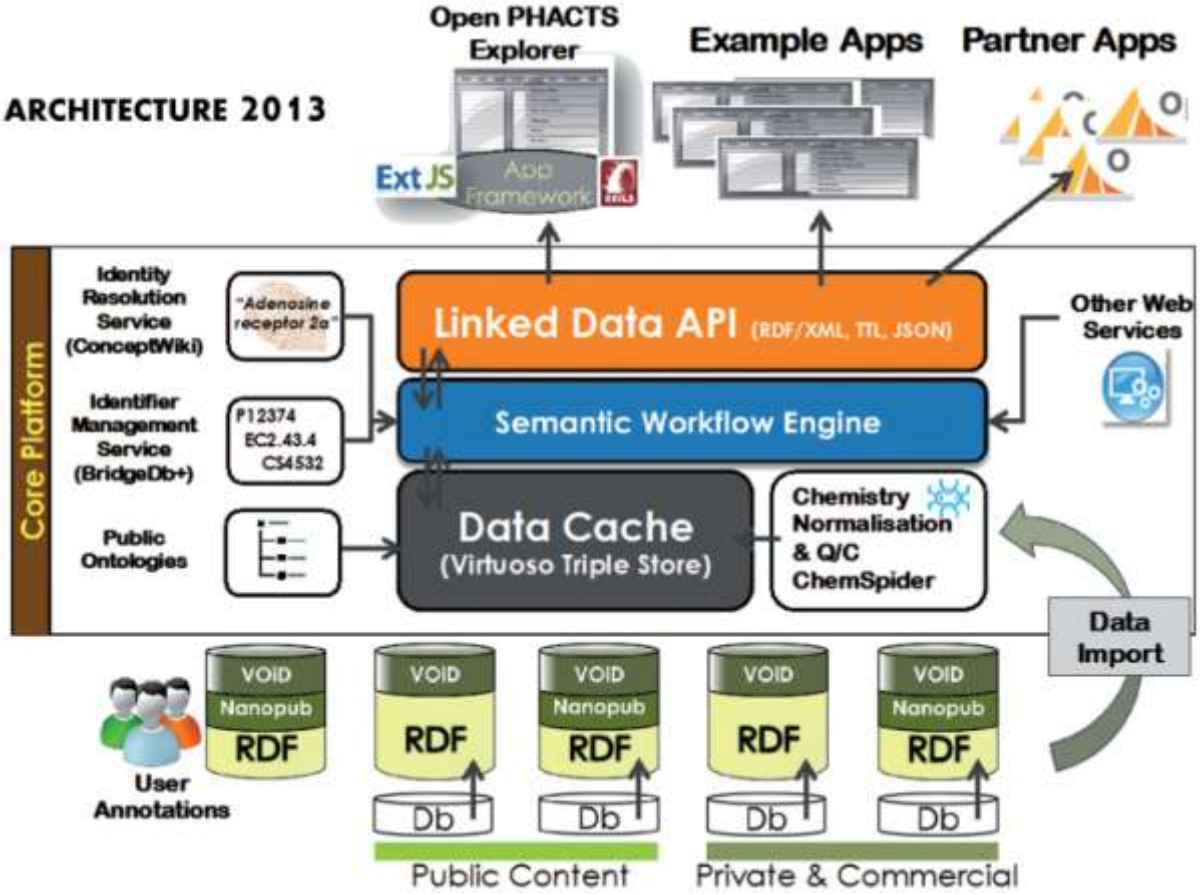

Figure 1: The Open PHACTS System

Component	Role
Data	Data is sourced in RDF format from both public and commercial data sources. Where this does not yet exist, we have developed guidance for data providers[4] and will work with them to convert the data in the correct manner
VoID	The Vocabulary of Integrated Datasets[5] provides a community standard for describing the meta-data regarding each dataset, such as version, licence and who provided the content. Each database is required to provide a VoID description.
Chemistry Normalisation	All chemical molecules loaded into the system are processed through a set of rules that aims to uniquely identify any compound, and provide high-quality mappings between the same compound in different source databases.
Data Cache	This is the underlying database (triple store) that stores the data in the system. Importantly, the cache can be reloaded at any point. Data is not maintained so data providers can trust they remain the stewards of their data.

[4] http://www.openphacts.org/specs/datadesc
[5] http://www.w3.org/TR/void

Semantic Workflow Engine	The engine that allows developers to fold in web-service queries (federated services) into queries over the data cache. Allowing us to expand and adjust queries using the results of web services.
Identifier Management Service	Provides support to connect the large number of different identifiers used for the same concept across different life science databases.
Identity Resolution Service	Provides translation from human entered text to concept identifiers, required for data queries
Linked Data Application Programmable Interface (API)	A community based tool[6] for providing access to semantic data in a developer-friendly manner
Applications	A multitude of different user-facing applications can consume data from the Linked Data API

7 The service

The Open PHACTS Discovery Platform is primarily a *cloud service*. The first version of the platform is hosted by a professional semantic hosting company Open Link at the Amsterdam Internet Exchange (AMS-IX) one of the largest internet exchanges in the world guaranteeing high speed access to the platform. Being a cloud service facilitates developers and organizations by allowing them to use the platform through an API without the need for hosting or maintaining massive amounts of data. Furthermore, it allows us to quickly add new functionality and update data so as the platform improves so do the applications that use it.

8 The Open PHACTS API

The core philosophy within the project is that the API is the main “product” of the effort, representing a software component that multiple applications can be developed from. The vision is akin to the role of Apple’s iOS operating system and the App store, each app using the same core but providing different, tailored functionality. The API provides a set of developer friendly RESTful calls allowing developers easy access to an integrated view over the various databases.

This includes methods for acquiring data that is both filtered and normalized, getting access to data using a variety of query inputs (e.g. URLS, smiles, class definition). The API includes the ability to retrieve hierarchies and search on using class based reasoning. State of the art chemistry structure search is provided by Chemspider the leader in open chemistry. In addition, applications have access to a comprehensive synonym and concept look-up API. Finally, we provide access to the underlying mappings that are used within the Open PHACTS Discovery Platform itself.

[6] http://code.google.com/p/linked-data-api

Our aim is to provide the facilities so that developers can create great pharmacology applications.

The platform API documentation can be found at http://dev.openphacts.org, which will be the home for developer documentation going forward.

9 The Open PHACTS Explorer

A critical element of the project is to provide a core interface to provide user-friendly access to the data in the system and support many of the common questions a scientist might wish to ask. This is provided by the Open PHACTS Explorer. This system was developed using code donated by Lundbeck as its base, and deliberately designed to "hide the technology" and provide users with an easy to use, familiar portal to pharmacology data.

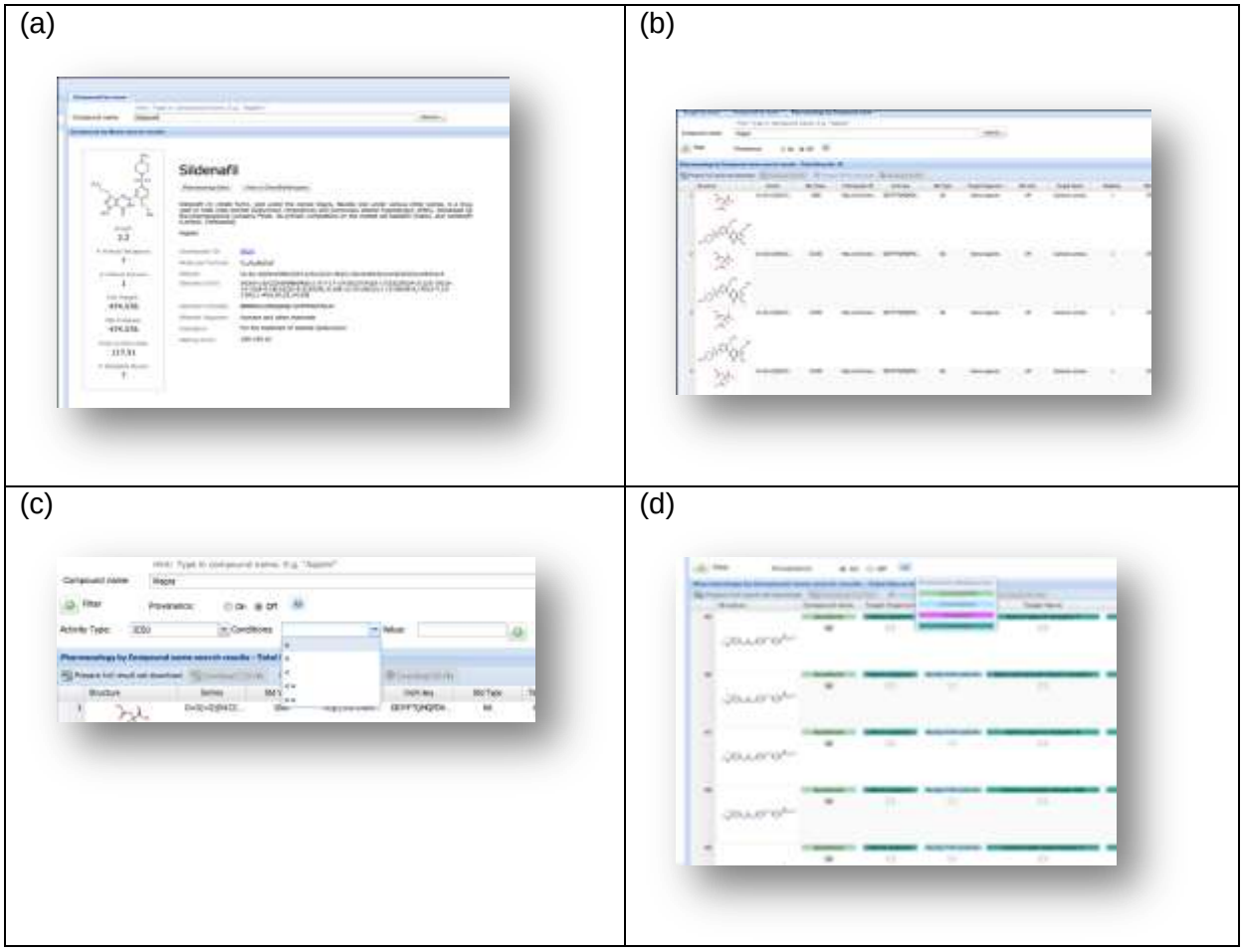

Figure 2: The Open PHACTS Explorer. (a) Viewing a single compound. (b) A single click takes the user to extensive pharmacology data. (c) Data can be filtered on common parameters such as activity value. (d) Switching provenance highlighting "on" shows where every data point originated.

As a result of the approach we have taken, users can also ask more complex questions using workflow tools such as Pipeline Pilot[7] and the popular open platform, KNIME[8]. The Open PHACTS API is fully compatible with KNIME and can be incorporated into workflows as illustrated in Figure 3.

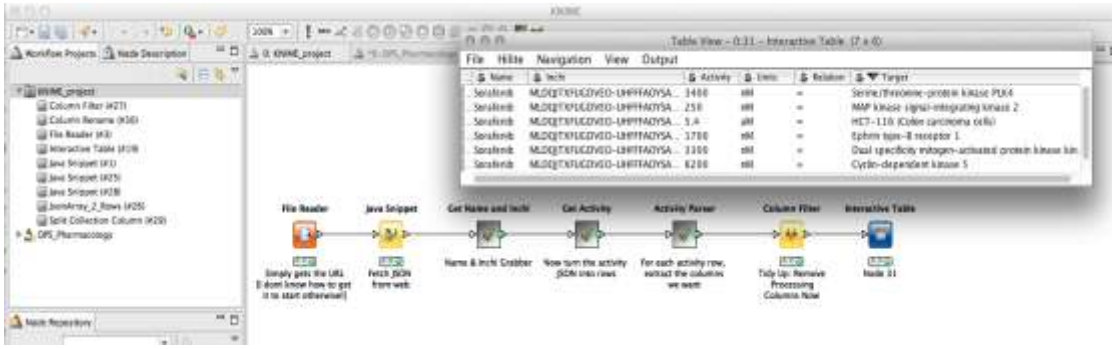

Figure 3: Accessing Open PHACTS through KNIME.

As the project goes on the Open PHACTS consortium will be releasing a set of reference eApps that show how to use the platform in specialized ways to facilitate more specific drug discovery questions.

The Open PHACTS code repository is now available as an open source community project on GitHub[9]. The central GUI API, GUI components code, API configuration files, Open PHACTS Explorer code and much more can be found there.

10 Support

A support network is being established to provide feedback to users of the system. This includes documentation, participation in portals such as twitter and forums, and a robust internal procedure for identifying and resolving user issues. We will also be publishing numerous papers and technical documentation for those interested in understanding the system in more depth. These efforts are aligned to our wider partnership program to promote the system to potential application developers.

Importantly, because our development cycles are agile and the platform is provided as a cloud service we are able to quickly respond to the needs of developers and users.

[7] https://community.accelrys.com/groups/openphacts
[8] http://www.knime.org/
[9] https://github.com/openphacts/

11 The Future

The release of the first version of the Open PHACTS Discovery Platform was just the start of an ambitious program of features to deliver to the scientific community. Over the second half of the project we will be delivering features such as:

- **Dynamic mapping**: Seamlessly connecting data between highly related topics such as compound parents and salt forms, stereo-isomers and genes and proteins. All at the users-control.
- **Chemistry Quality:** Extensive assessments on the quality of structures in the Open PHACTS database by the application of EFPIA-agreed quality metrics
- **Additional use cases, pathways and disease data**: Connecting chemistry and biology
- **Publisher assertions**: Providing access to deep knowledge not accessible elsewhere
- **eApps** – reference applications that show how to leverage the Open PHACTS Discovery Platform

This release of the platform provides a stable base upon which to build a cutting edge system that will address real issues in drug discovery today.

12 Summary

To summarise, the 1.0beta version of the Open PHACTS Discovery Platform delivers:

- Integration across major pre-clinical databases
- A rich user interface to access pharmacology data
- A working API that can be used in workflow tools and by application developers
- A validation of semantic technology , demonstrating it works well in this domain and is a viable long term investment
- A stable platform on which to build novel, critical functionality for drug discovery
- A launch and customer support process is set-up to allow best user-support, stakeholder management and learning for next releases.