Open PHACTS

Deliverable 8.3.2

Run OPS Workshop 2
"Working with OPS"

Prepared by RSC, UNIVIE, GSK
Approved by RSC, UNIVIE, GSK, Pfizer, ConnDisc

May 2012
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

- **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.

Content

1 Introduction

On 16-17 April 2012 the Second Open PHACTS Community Workshop with the topic „Working with OPS" was held at the RSC's Chemistry Centre, Burlington House, London. This workshop provided an understanding of the scientific and technical background to Open PHACTS, featured the current status, and gave the opportunity to discuss particular issues around data integration, licensing and sustainability with new consortium members, other database hosts (public and commercial) and publishers.

The audience included representatives from ChEMBL, ChEBI, IUPHAR-DB, Elsevier, Thomson-Reuters, Springer, ChemistryCentral, NPG, Selventa, Syngenta, Aureus Sciences, Janssen and Harvard Medical School, and their attendance made the workshop a great success!

2 Summary of the Workshop

Find all presentations attached in chapter 5 and on the Open PHACTS website (http://www.openphacts.org/index.php?option=com_content&view=article&id=114&Itemid=128).

2.1 Background and current status of the project

2.1.1 Industry and academic view

First of all, Bryn Williams-Jones and Gerhard Ecker were talking about the background and reason of the project from the industry and academic perspective. Currently, pharmaceutical companies expend significant and duplicated efforts aligning and integrating internal information with public data sources. This process is largely incompatible with massive computational approaches and the vast majority of drug discovery sources cannot easily interoperate. Open PHACTS will create a precompetitive infrastructure to make these approaches available both to industry and also to academia and smaller companies, who have historically not had access to large-scale integrated pharmacological data resources.

The project will deliver an online platform with a set of integrated publicly available pharmacological data. The software and data will be available for download and local installation, under an open source and open access model. Tools and services will be provided to query and visualise this data, and a sustainability plan will be in place, continuing the operation of the OPS platform after the project funding ends.
Throughout the project, a series of recommendations will be developed in conjunction with the community, building on open standards, to ensure wide applicability of the approaches used for integration of data.
The OPS platform is a specific achievable result from this project, but the semantic technology and principles are deliberately being adopted to enable the platform to be adapted for application to other subject domains in the future.

2.1.2   Semantic context

After this, Barend Mons was talking about the semantic context of the project and the emerging role of nanopublications. The Semantic web allows computers to assist in finding associations in datasets that are just too large for any human brain to comprehend. The purpose of nanopublications is:

- Publishing machine-readable assertions
- Bottom up standard setting/multiple standards
- Fine grained attribution to authors
- Rapid dissemination and sharing of data

2.1.3   Project progress after the first year

To conclude, Richard Kidd was talking about the progress the Open PHACTS consortium has already achieved since the beginning. Additionally, he explained how to get involved into the project and showed an alpha demo of the Open PHACTS platform.

Open PHACTS is a 3-year knowledge management project of the Innovative Medicines Initiative (IMI), a unique partnership between the European Community and the European Federation of Pharmaceutical Industries and Associations (EFPIA). The project is due to end in March 2014. The project consortium consists of leading academics in semantics, pharmacology and informatics, driven by solid industry business requirements: 27 partners, including 9 pharmaceutical companies and 3 biotechs. The project is on track, with 5 new partners joining (we started with 22). At the moment, the project is seeking a partnership with a commercial service provider to reliably host the publicly accessible system.

As we near the half-way point of this project we have met all of our planned milestones, including the roll-out of an initial version of the platform for consortium partners to test which was delivered in February 2012. A public beta version of the platform is scheduled for delivery in Autumn 2012. Beyond the technical aspects we are now working on further development of our community and plan to ensure sustainability beyond the initial funded period.

The 3 pillars of Open PHACTS:

- **A Precompetitive Infrastructure**
  - Begin the task of creating an environment that can also power future collaborative efforts (public & industry)
  - Expose Industry Experience: Create drug-discovery focused tools outside of the firewall, influenced by decades of practical experience
- **A Pharmacology Use Case**
  - Showcase one application this technology: a stable, responsive, user-orientated system for Pharmacology Analysis
- **A Data Publishing Methodology**

o Develop standards and methodologies to promote good data sharing and interoperability
o An exemplar project for the use of the Nanopublication concept
o A technical approach that can be repeated in other areas

The development of the OPS infrastructure is being driven by a set of prioritised research questions from the EFPIA partners. Based on the main research questions, prioritised data sources have been identified to be integrated into the system. Three Use cases have been defined to benchmark the OPS system towards current standard workflows in data retrieval and mining. Four Exemplars are being developed to demonstrate the capabilites of the OPS System and to define interfaces and input/output standards.

Exemplar Services:

- **Chem-Bio Navigator:** querying and visualization of sets of pharmacologically annotated small molecules, on basis of chemical substructures, pharmacophores, biological activities
- **Target Dossier:** *in silico* dossiers about targets, incorporating related information on sequences, structures, pathways, diseases and small molecules
- **Polypharmacology Browser:** map coverage of the chemo-biological space, to facilitate the polypharmacological profiling of small molecules
- **Utopia Documents:** link publications to related data

See here the Open PHACTS 6 month lash-up demo, which demonstrates the results of the first 6 months of the project: http://www.youtube.com/OpenPHACTS.

Since it is well suited to describing complex data, open and supported by a growing body of tools and scientists, RDF is our chosen format. The Open PHACTS platform will be built on open source technology. This includes the data harvester, the semantic workflow engine/API code, the Open PHACTS Explorer and associated widgets. The standards for producing RDF/Nanopublications will all be open and available. An open version of the system will be available at www.openphacts.org later, fully functional with public data. Interested parties will also be able to download the core platform and instantiate it on their own servers, having everything they need to run a local system should this be required. The Open PHACTS platform is decoupled from any specific RDF database engine. It should be possible to run the platform on a range of free and commercial platforms that meet certain criteria (to be published).

2.1.4 *How to get involved into the project*

Open systems need an engaged community, to grow, develop and sustain. We actively manage our partners, and the wider community. We term this the Open PHACTS Waiting Room, managed by a Gatekeeper (Bryn Williams-Jones):

- Our relationships with all partners are visible: what we are doing together and why
- Opportunities to engage and develop are open and are based on project needs

We hold regular community workshops and events (like this one):

- Learn more about Open PHACTS and the Open Pharmacological Space
- Participate in new ideas and functions
- Engage in development of new use cases, help us answer new questions
- Contribute to development, and engage in plans for sustaining the Open Pharmacological Space

These are the various levels at which one can engage with the Open PHACTS community:

- **Associated Partners:** We have a Memorandum of Understanding (MoU) ready to sign for mutual support and exchange of ideas, data or technology. As an Associated Partner you will be the first to hear about the latest developments in the Open PHACTS project and you will also have the opportunity to present ideas and use cases to the core Open PHACTS team.
- **Development Partnerships:** Once you are an Associated Partner and want to do some more specific development work together with us (e.g. develop APIs, new data, algorithms etc), you can enter a Development Partnership with the Open PHACTS project. This will give you greater access to the core of the project. Development Partnerships are focused on defined pieces of work of mutual interest and an agreed collaborative annex is added to the MoU.
- **Joining the consortium:** If you would like to become an integral part of the project, it might be worth to consider the option to join the Open PHACTS consortium.

Check our website for more details on how to get involved into the project and work with the community:
http://www.openphacts.org/index.php?option=com_content&view=article&id=102&Itemid=55

2.2 Interaction with public databases

In this session, John Overington was invited to speak about ChEMBL and Anthony Harmar to talk about IUPHAR-DB. Please find their presentations in chapter 5.

2.3 Interaction with open access and commercial publishers

First, Stefan Senger gave an overview on the EFPIA vision in regards to Open PHACTS and commercial data. The aim is to achieve differential access. We are committed to performing a proof-of-concept experiment. What we want to avoid is making decisions during the development of the Open PHACTS platform that limit our options in regards to differential access. Hence, we decided to start the dialogue with commercial data providers early.

Afterwards, Jan Velterop was talking about nanopublications and the integration of full text. The Open PHACTS consortium has put together a set of guidelines on nanopublications. A nanopublication is the smallest unit of publishable information: an assertion about anything

that can be uniquely identified and attributed to its author. The basic building block of the nanopublication is the triple: Subject – Predicate – Object. The Open PHACTS project plans to use nanopublications in its Open Pharmacological Space (OPS) tool; the OPS will use semantic web technology to allow scientists to analyse diverse texts and databases as part of their drug discovery efforts. The guidelines explain the elements of a nanopublication and set out the steps required to create one. The guidelines should help data owners understand the Open PHACTS approach to nanopublication and adopt it in their own projects. This first set of guidelines will guide the next stage of implementation within the Open PHACTS project where nanopublications will be integrated with other forms of data to provide a unique view across pharmacology-related data.

Find the current version of the guidelines on our website (http://www.openphacts.org/documents/publications/OpenPHACTS_Nanopublication_Guidlines_v1.8.1.pdf).

2.4 Data integration and mapping

Antony Williams and Christine Chichester gave an overview on data integration and mapping in Open PHACTS. The data sources we presently have in the Open PHACTS platform were selected by:

- Choosing data sources in order to answer business questions
- Selecting appropriate data sources based on licensing
- Filtering criteria of judged data quality
- Chemistry Data: ChEBI, ChEMBL, Drugbank aggregated through ChemSpider – initial focus on pharmacology
- Biology Data: SwissProt, GO/GOA and ENZYME
- Data underpins public version plus all related exemplars
- Data are intended to be updated quarterly for the near future
- Curation/annotation flow is intended to be continuous

How to integrate the data:

- Chemicals imported to ChemSpider and registered with CSIDs
- Registered compounds passed to ConceptWiki
- ConceptWiki passes data to Identifier Mapping Service
- RDF generated from ChemSpider for deposition to cache
- Linked Data Cache receives RDF from various original data sources – RDF is provided by data source provider, or is sourced from alternate provider, or is generated by Open PHACTS team

As already mentioned RDF is well suited to describe complex data and is open and supported by a growing body of tools and scientists. Producers of RDF can enhance their RDF with information required to create Nanopublications and we will support data providers to generate RDF as necessary. We will publish guidelines for producing RDF. These are community-endorsed best practices. RDF will be used to produce and integrate nanopublications.

2.5 Licensing

Open PHACTS hopes to accommodate both **open** and **license-restricted** data. We will go live in Autumn with data sets that have well-defined licensing. We will develop a licensing strategy around public data to ensure OPS complies with the terms for use and distribution and will support restricted datasets within the platform in the future.

Curation and annotation:

- Develop systems to validate and standardize data: our preference is to be proactive in data quality when possible
- Provide interactive ability to curate and annotate data from data sources
- Existing systems to be integrated and enhanced: ConceptWiki and ChemSpider already provide curation/annotation features
- Provide feedback to data suppliers regarding identified issues and help clean up existing data to contribute back to providers
- Our primary concern is participation, not the systems

There was a discussion around how to encourage users to actively contribute to the curation of the resources that are relevant for Open PHACTS (e.g. the ConceptWiki, Chemspider). There was agreement that we must make it easy in order to get users to participate. The preference would obviously be that users register first when they want to contribute to the curation so that everything is traceable. However, it might be good to also offer a simple feedback mechanism that doesn't require registration.

2.6 Sustainability

Barend Mons concluded the workshop with a presentation on sustainability of the Open PHACTS project.

O-XS4all for the Life Sciences
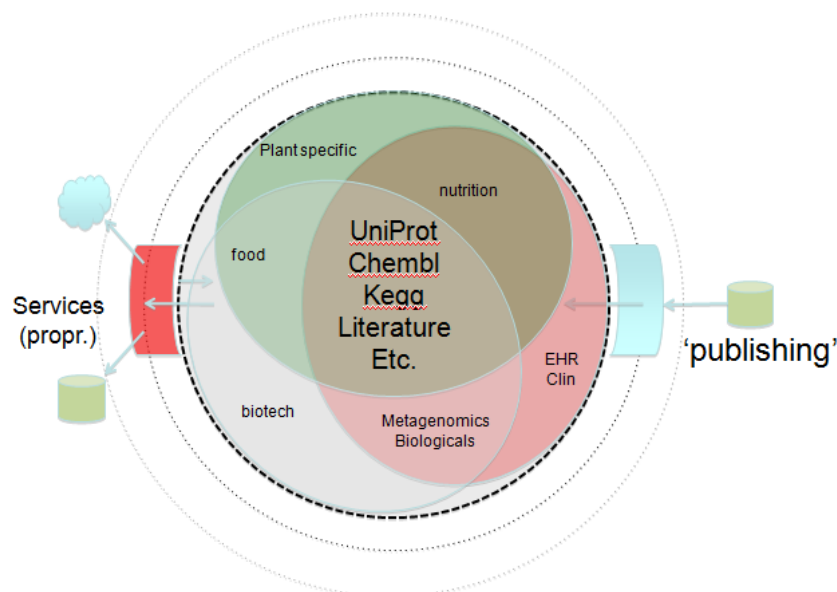
Barend Mons mentioned that the number of services that will be built by third parties on top of the Open PHACTS infrastructure/platform is going to be one of our success measures. This was a timely reminder to prompt us to look at how we are planning to encourage third parties to build services and also how we are planning to support third parties in doing this. Obviously, this has resource implications and needs to be discussed. However, since it takes a significant amount of time to build services we have to start this process early.

As part of the sustainability discussion it was pointed out that it will help to attract future funding if we can show how popular and useful the OPS infrastructure is. In order to be able to do this it will be very helpful to be able to refer to published reports about the benefit of the Open PHACTS platform. Also, the more organization publish such reports the more weight the argument will have.

3 Participants

Jan	Velterop	Academic Concept Knowledge
Jason	Theodosiou	Aureus Sciences
Jan	Kuras	ChemistryCentral
Bryn	Williams-Jones	ConnectedDiscovery
Marcus	Ennis	EBI
John	Overington	EBI
Jabe	Wilson	Elsevier
Michael	Taylor	Elsevier Labs
Stefan	Senger	GSK
Paolo	Ciccarese	Harvard Medical School
Ian	Harrow	Ian Harrow Consulting
Tony	Harmar	IUPHAR-DB
Eveline	Hoekx	Janssen
Jean-Marc	Neefs	Janssen
Dmitrii	Rassokhin	Janssen
Tony	Hammond	Nature Publishing Group
Barend	Mons	NBIC
Christine	Chichester	NBIC/SIB
Maxim	Popov	Novartis
Mark	Schreiber	Novartis
Richard	Kidd	RSC
David	Sharpe	RSC
Antony	Williams	RSC
Allan	Price	Selventa
Ted	Slater	Selventa
Wim	Van der Stelt	Springer
Mark	Forster	Syngenta
Andrew	Garrow	Thomson-Reuters
Mabel	Loza	Universidad de Santiago de Compostela
Emiliano Cuadrado	Rodriguez	Universidad de Santiago de Compostela
Katrin	Stierand	Universität Hamburg
Gerhard	Ecker	University of Vienna
Franz	Gatterer	University of Vienna
Anika	Robl	University of Vienna

4 Agenda

Monday, 16 April 2012

12.00-13.00   Registration and Lunch

13.00-17.30   Workshop

- Background and current status of project

    Bryn Williams-Jones – background and industry view
    Gerhard Ecker – academic view
    Barend Mons – semantic context
    Richard Kidd – project progress, how to get involved and alpha demo

- Interaction with public databases

    John Overington – ChEMBL
    Tony Harmar – IUPHAR-DB

- Interaction with open access and commercial publishers

    Stefan Senger – pilot for commercial data within OPS
    Jan Velterop – nanopublications and the integration of full text

Tuesday, 17 April 2012

09:00-12:00   Workshop

- Data integration and mapping

- Licensing issues (Antony Williams and Christine Chichester)

- Sustainability (Barend Mons)

12.00-13.00   Lunch and depart

5 Presentations