The Pharmacology Workspace: A Platform for Drug Discovery

Alasdair J. G. Gray [1], Sune Askjaer [2], Christian Brenninkmeijer [1], Kees Burger [3], Christine Chichester [3], James Eales [1], Chris T. Evelo [4], Carole Goble [1], Paul Groth [5], Lee Harland [6], Antonis Loizou [5], Steve Pettifer [1], Rishi Ramgolam [7], Mark Thompson [3], Andra Waagmeester [4] and Antony J. Williams [8]

[1]University of Manchester [2]H. Lundbeck A/S [3]Netherlands Bioinformatics Center
[4]Maastricht University [5]VU University Amsterdam [6]Connected Discovery
[7]Academic Concept Knowledge Limited [8]Royal Society of Chemistry

ABSTRACT

We present the Open PHACTS linked data platform that is being developed to address a set of example drug discovery research questions and which supports several drug discovery applications. The platform retrieves data from many complementary, but overlapping, data sources to present an integrated view of the data. The platform exploits two entity resolution services: respectively for transforming text and chemical structures to a concept. The single concept URI provided by the resolution service is then expanded to a set of equivalent URIs used by the data sources.

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

EXTENDED ABSTRACT

The investigation and development of new drugs requires that scientists involved in the process deal with multiple information sources. These range from online databases of proteins (e.g. UniProt and Enzyme) and chemicals (e.g. ChEMBL, ChemSpider, and DrugBank), to models of biological pathways (e.g. Reactome, WikiPathways, and KEGG) and 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 space of interest, but overlap in the data they provide and also record different (or even inconsistent) representations of the same data.

A significant challenge to scientists is the labour intensive integration of datasets. The entities of interest must be identified and mapped to each other to allow complementary information from many 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, often varying in the charged state of the compound, e.g. *"Simvastatin"* on ChemSpider[1] and DrugBank[2]. Thus, for successful data integration one must devise strategies that address inconsistencies within the existing data.

The linked data platform being developed in the Open PHACTS project[3] aims to overcome these data integration challenges. There are two key entry points into the system, both of which perform resolution from user input to an identifier for a data concept.

The first is through keyword search, as shown in Figure 1. 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"* does the user mean the chemical menthol, or the menthol receptor protein. The user interface supports this disambiguation by providing different entry points, e.g. compound by name or target by name (shown in Figure 1). The Identifier Resolution Service (IRS) translates user-entered entity names (in free text form), together with the context information, into known entities within the system (i.e. that have a defined URI). The IRS uses several dictionaries including a custom dictionary of chemical names and synonyms from ChemSpider, as well as MeSH, GO, and SwissProt. The IRS provides data for the auto-complete text box including the preferred name for the entity and a link to its definition. This supports the user in disambiguating the entity that they mean. The identified entity URI can then be used to retrieve further information from the linked data platform.

The second entry point is through chemical structure search that uses a tool for drawing chemical structures which are then converted to a standardised chemical structure representation. This is then processed by the ChemSpider structure search service to return a ChemSpider URI for the chemical entity drawn. The service can also be used for substructure and similarity searches.

The linked data platform leverages the comprehensive work already performed by the community in creating RDF-based datasets, which are relevant for the Open PHACTS project. The current platform uses the ChEMBL and ChEBI datasets provided by the Chem2Bio2RDF project (Chen *et al.*, 2010), the conversion of DrugBank provided by the LODD project (Samwald *et al.*, 2011), and the conversion of the Enzyme database sourced from UniProt (Jain *et al.*, 2009). A significant challenge is ensuring that the RDF versions of the datasets are kept up-to-date with the originals from which they are derived. For example, the Chem2Bio2RDF version of ChEMBL is version 8 whereas the original dataset is now at version 13.

The data sources are integrated using parameterized SPARQL queries that are called through an API exposed by the linked

[1] `http://www.chemspider.com/Chemical-Structure.49179.html` accessed May 2012.

[2] `http://www.drugbank.ca/drugs/DB00641` accessed May 2012.

[3] `http://www.openphacts.org/` accessed May 2012.

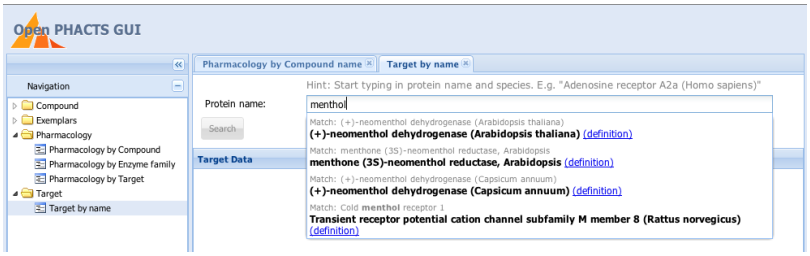

Fig. 1. Screenshot showing a search with the identifier resolution service for the term *"menthol"*.

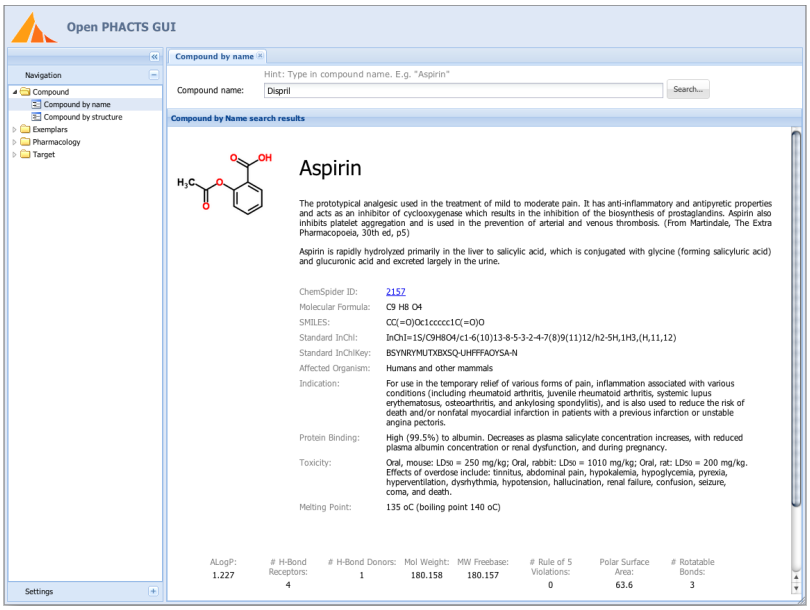

Fig. 2. Screenshot showing the integrated information returned for Aspirin.

data platform. The API call generates a query containing the URI returned by the IRS. The query is then expanded at execution time using an identity mapping service that equates the data entity URIs from the various data sources. To provide adequate interaction speeds, we have cached the datasets in the linked data platform.

The result for doing a compound lookup with the search term *"Aspirin"* is shown in Figure 2. Information about the chemcial structure is sourced from ChemSpider, details of its bioactivity are obtained from ChEMBL, and information about the drugs in which the compound is active are obtained from DrugBank. Currently, the provenance of the data points is not shown in the user interface, although this is planned for the public release.

The linked data platform is being developed to answer a set of pharmacology research questions that require data to be integrated from a variety of data sources (Williams *et al.*, 2012). The platform hides the complexities of interacting with the linked data and concepts by exposing an API that provides the core functionality to support a wide variety of drug discovery applications being developed within the Open PHACTS project, although only one has been shown in this demonstration paper.

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.

REFERENCES

Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., and Wild, D. (2010). Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, **11**(1), 255.

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

Samwald, M., Jentzsch, A., Bouton, C., Kallesoe, C., Willighagen, E., Hajagos, J., Marshall, M., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., and Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, **3**(1), 19+.

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., Evelo, C. T., Blomberg, N., Ecker, G., Goble, C., and Mons, B. (2012). Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today*. To appear.