Anna Gaulton[1], Lee Harland[2], Mark Davies[1,3], George Papadatos[1], Antonis Loizou[4], Nathan Dedman[1], Daniela Digles[5], Stian Soiland-Reyes[6], Valery Tkachenko[7], Stefan Senger[8], John P Overington[1,3], Nick Lynch[9]

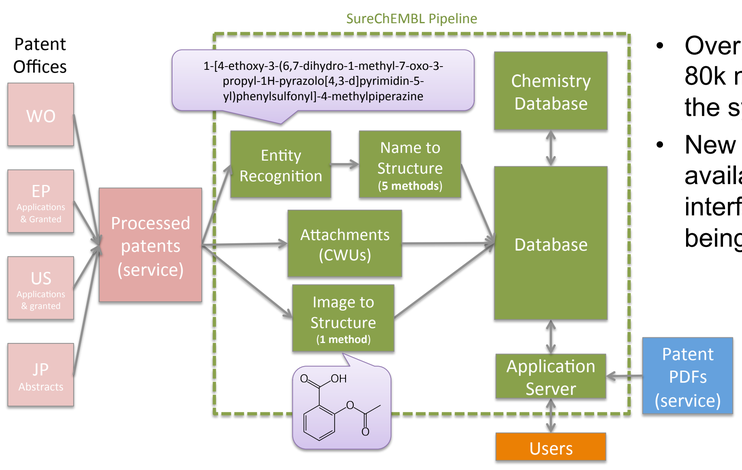
[1] European Molecular Biology Laboratory - European Bioinformatics Institute, [2] SciBite Limited, [3] Stratified Medical (current address), [4] VU University of Amsterdam, [5] University of Vienna, [6] University of Manchester, [7] Royal Society of Chemistry, [8] GlaxoSmithKline, [9] Open PHACTS Foundation

Email: agaulton@ebi.ac.uk

Semantic-Web Access to Patent Annotations

SureChEMBL

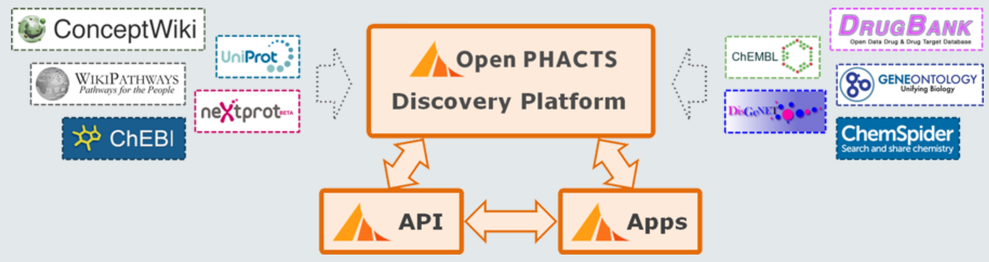
SureChEMBL (https://www.surechembl.org) is a patent chemistry resource, originally a commercial product developed by SureChem/Digital Science, and recently donated to EMBL-EBI[1,2]. SureChEMBL uses a live and fully automated pipeline that combines text-mining and chemistry tools to extract compounds named or depicted in patent documents and make them structure searchable by users.



- Over 50k new patent documents and 80k new compounds are entered into the system per month.
- New chemical annotations are usually available in the SureChEMBL interface within 1-7 days of the patent being released by the patent office.

Open PHACTS

The Open PHACTS Discovery Platform is a semantic-web data integration platform, developed for the purpose of providing both the pharmaceutical industry and academic researchers with open access to interoperable drug discovery information[3,4].



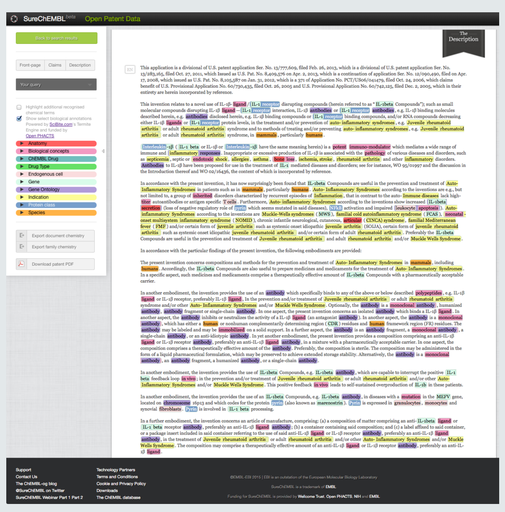
The platform currently includes data from a wide variety of public databases and provides API access to the integrated information. However, the further addition of biological and chemical patent information to the platform was considered to be of great potential utility. Much of this information will never be published elsewhere and may be of great value to the drug-discovery and broader life-science community.

Biological Annotation & Relevance

A pipeline was developed to identify and annotate additional entities (namely genes and diseases) within the SureChEMBL patent corpus using the Termite text-mining tool (https://scibite.com/content/termite.html). Since patent documents are often designed to obfuscate the key subject matter, it was essential to also develop an algorithm to assess the relevance of each entity within a particular patent document, allowing users to restrict results to only highly relevant entities if they wish.

Gene/Disease relevance:

Various features such as term frequency, position and distribution were used to create a biological relevance score for each entity, indicating the importance of that entity in the patent.



Compound relevance:

A set of chemical and frequency filters were applied to remove likely 'irrelevant' compounds (e.g., buffers, ions, fragments). A method has also been developed to identify the likely 'claimed' compounds within a patent based on scaffold and similarity analysis. This will be applied to all SureChEMBL patents/compounds and included in API results.

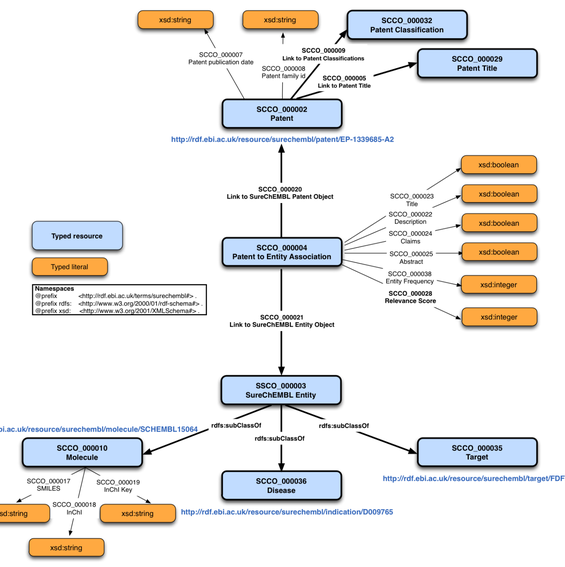
Patent relevance:

Patents were also filtered according to International Patent Classification codes to remove non life-science documents.

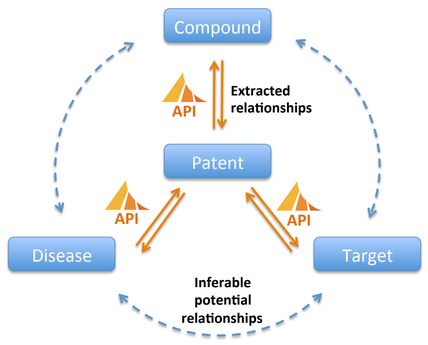
Patent Data Integration

An RDF model has been developed to capture the relationships between patent documents and annotated compounds, genes and diseases, and annotations for more than 6 million life-science patents have been made available in this format via the Open PHACTS platform:

https://dev.openphacts.org/docs/develop



A series of API calls have been developed to allow users of the platform to query the data. Interoperability with other data sets is provided by the Open PHACTS Identifier Mapping Service and Chemistry Registry Service, and users to integrate patent data with the extensive range of other resources included in the platform (e.g., protein, pathway, bioactivity and disease information).



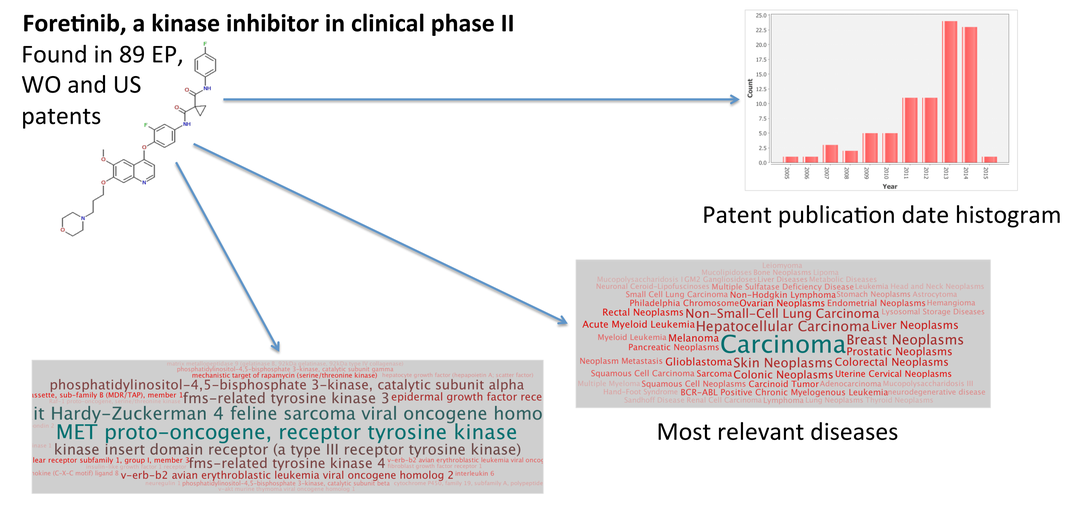
KNIME Workflows

Open PHACTS provides KNIME nodes and Pipeline Pilot components to facilitate the development of complex workflows using the Open PHACTS API (see https://dev.openphacts.org/resources for more information):

- KNIME nodes: https://github.com/openphacts/OPS-Knime
- PP components: https://exchange.sciencecloud.com/exchange/browse#details:216,239

Example KNIME workflows have also been constructed to demonstrate the use of the patent data API calls: for example, identifying the most relevant targets or diseases for a compound from the patent corpus. These workflows will be made available alongside other Open PHACTS example workflows:

- Open PHACTS KNIME workflows: http://www.myexperiment.org/groups/1125.html



References and Acknowledgements

1) Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S.A.,Pettersson, J., Goncharoff, N., Hersey, A. and Overington, J.P. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Research DOI:10.1093/nar/gkv1253 (2015).

2) Senger, S., Bartek, L., Papadatos, G. and Gaulton, A. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. Journal of Cheminformatics 7(1), 49 (2015).

3) Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B. Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today, 17(21-22), 1188-1198 (2012).

4) Azzaoui, K., Jacoby, E., Senger, S., Rodriguez, E.C., Loza, M., Zdrazil, B., Pinto, M., Williams, A.J., de la Torre, V., Mestres, J., Pastor, M., Taboureau, O., Rarey, M., Chichester, C., Pettifer, S., Blomberg, N., Harland, L., Williams-Jones, B., Ecker, G.F.: Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discovery Today, 18, 843-852 (2013).

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which comprise financial contribution from the European Union's Seventh Framework Programme (FP7/20072013) and in-kind contribution of European Federation of Pharmaceutical Industries and Associations (EFPIA) companies; a Strategic Award from the Wellcome Trust[104104/Z/14/Z]; and the member states of EMBL. We would like to thank members of the Open PHACTS consortium, and the former SureChem/Digital Science team.

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK

Tel. +44 (0) 1223 494 444
comms@ebi.ac.uk
www.ebi.ac.uk