Semantic-Web Access to Patent Annotations

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], and Nick Lynch[9]

[1] European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom
agaulton@ebi.ac.uk
[2] SciBite Limited, CB1 Business Centre, 20 Station Road, Cambridge, CB1 2JD, United Kingdom
[3] Stratified Medical, 40 Churchway, London, NW1 1LW, United Kingdom (current)
[4] Department of Computer Science, VU University of Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[5] University of Vienna, Department of Pharmaceutical Chemistry, Althanstraße 14, 1090 Vienna, Austria
[6] School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom
[7] Royal Society of Chemistry, Thomas Graham House (290), Science Park, Milton Road, Cambridge, CB4 0WF, United Kingdom
[8] GlaxoSmithKline, Stevenage, Hertfordshire, SG1 2NY, United Kingdom
[9] Open PHACTS Foundation, c/o Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, United Kingdom

Abstract. SureChEMBL (`https://www.surechembl.org`) is a patent chemistry resource, originally a commercial product developed by SureChem/Digital Science, and recently made freely available at EMBL-EBI [1]. SureChEMBL uses a live and fully automated cloud-based pipeline that combines text-mining and chemistry tools to extract compounds named or depicted in patent documents and make them readily structure searchable by users. Over 50,000 new patent documents and 80,000 new compounds are entered into the system per month and new chemical annotations are usually available in the SureChEMBL interface within 1-7 days of the patent being released by the patent office. While the current SureChEMBL system addresses several chemistry use-cases, such as the identification of novel scaffolds and chemistry, there is an enormous amount of additional knowledge captured within the patent corpus. Much of this information will never be published elsewhere and may be of great value to the drug-discovery and broader life-science community. 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 [2, 3]. 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.

We have therefore developed a pipeline 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 gene or disease within a particular patent document, allowing users to restrict results to only highly relevant entities if they wish.

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/`). A series of API calls have been developed to allow users of the platform to query the data and to integrate it with the extensive range of other data resources included in the platform (*e.g.*, protein, pathway, bioactivity and disease information). In addition, KNIME and Pipeline Pilot nodes have also been created to facilitate the construction of workflows using patent data, for example, identifying all of the compounds from patents that mention a particular target or disease with high relevance. This represents the first large-scale, semantically-annotated life-science patent knowledgebase, freely available to both industrial and academic researchers.

References

1. Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S.A., Pettersson, J., Goncharoff, N., Hersey, A., Overington, J.P.: SureChEMBL: A large-scale, chemically annotated patent document database. Submitted
2. 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)
3. 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)

Acknowledgments. 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 thanks members of the Open PHACTS consortium and the former SureChem/Digital Science team.