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. 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.

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/tfidf/termite.html). 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

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

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


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 (FP7/2007-2013) and in-kind contribution of European Federation of Pharmaceutical Industries and Associations (EFPIA) companies. A Strategic Award from the Wellcome Trust (104042/Z/14/R), and the member states of EMBL. We would like to thank members of the Open PHACTS consortium, and the former SureChem/Digital Science team.

Chemical annotations are available under a CC BY-SA licence, biological annotations under a CC BY-NC-SA licence.