Accessing the Open PHACTS Linked Data API with KNIME

James A. Lumley
Research IT, Eli Lilly
April 2017
The KNIME Analytics Platform

Open source platform for data analytics. Over 1000 modules (or nodes) to connect to all major data sources; support for many data types inc. XML/JSON/Images/Docs/Chemical Formats; Math and Stats functions, Predictive modelling and machine learning; Tool blending for Python/R/Weka/SQL/Java; Interactive data views and reporting. “a toolbox for any data scientist”.  
https://www.knime.org/knime-analytics-platform
OPS-KNIME Nodes

♦ 2016 (VU Amsterdam)*
  • Original Nodes and workflows by Ronald Siebes, VU Amsterdam
  • OPS_Swagger and OPS_JSON nodes used to create and execute the parameterized API calls, as well as transforming the output to a tabular form

♦ Q2 2017 (Eli Lilly)
  • Update of Erl Wood KNIME Nodes will add new OPS node developed internally at Eli Lilly with input from OPS
    – KNIME Node: Luke Bullard
    – Team input: James Lumley / Derek Marren (Lilly); Daniella Digles / Nick Lynch (OPS); Randy Kerber (d2discovery)
    – Workflows: James Lumley
  • Single Node allows user to select the call of interest and return both JSON and Tabular results
  • Focus of development: Updating to new API, improving usability
  • Further iterations possible once feedback received

OPS & Erl Wood Community Nodes

- View based on *internal Beta* of Lilly opensource Erl Wood nodes *due for release Q2 2017*

- Community → Erlwood Nodes → Open PHACTS

- Open PHACTS sub-folder contains single OPS Linked Data API node that will allow a configured call/return
Preferences panel allows client/workflow level control of API URL Endpoint and API Id/Key, avoiding the need to configure these in the node.
Using the OPS Linked Data API node

App Id and App Key fields are automatically populated if they are set in the preferences.

Drop down ‘Select Method Type’ allows selection of API call.
Using the OPS Linked Data API node

Input port is optional. Toggle on input field allows user string input or selection of input table column.

First output port returns formatted data table (corresponding to API param "_format=tsv").
Using the OPS Linked Data API node

Drop down ‘Select Method Type’ allows selection of API call

Logically grouped methods match developer API docs (swagger) at https://dev.openphacts.org/docs/2.1
Allows formatted results table or full JSON/XML return for debug/analysis

First output port returns formatted data table (corresponding to API param “_format=tsv”)

Second output port is optional and if requested, will return JSON or XML response (via second API call without _format param)
User input and example return

User input
User input and example return

Raw Tabular Return:

Pivoted to show Column Names and Values:
User input and example return

Optional JSON Output as raw JSON Object
User input and example return

Rather than parsing the JSON to understand the raw output, the node also has an attached ‘View’ with a hierarchically formatted tree view of the JSON output:
User input and example return

Generic JSON Extraction to flat table shows additional data returned from API, deeper JSON processing can be done using KNIME JSON nodes.
Extensive native support for JSON or XML parsing with KNIME 3.3 allows complete/custom parsing of the return JSON object for full debugging.
Chemistry Support on input SMI

Input columns of differing chemical types are automatically converted to SMILES via Marvin if the API param is SMILES based.
API Timeouts and URL changes

Advanced developers can change the API timeout value or edit the API URL on a single node using the Web Service panel.
1. A new KNIME 3.3 compatible “OpenPHACTS Linked Data API” node will be released in Q2 2017

2. Designed for users, it provides easy configuration of API settings and parameters with easy to user tabular data return (via API _format parameter)

3. Designed for developers it allows additional full JSON/XML response that can be viewed/parsed by the expert user to see raw response

4. Further example workflows will be release once the node is available