

CEVOpen: Can machines curate the Open phytochemistry literature?

Clive Adams¹, Clyde Davies¹, Emanuel Faria², Ambarish Kumar³, Peter Murray-Rust^{4,5,*}, Mike Williams¹, Andy Wright¹, Gitanjali Yadav^{3,4}

¹Chem4Word; ²Verriclear.com ; ³National Institute of Plant Genome Research, New Delhi, IN; ⁴University of Cambridge, UK ; ⁵ContentMine Ltd, Cambridge, UK,

CEVOpen* is a distributed virtual citizen science community which creates Open modular knowledge. We create tools (software, dictionaries) which read the Open published scientific literature and make it semantic and useful. The approach is general; here we apply it to plants and their medicinal oils.

References and Acknowledgements

- CEVOpen at <https://github.com/petermr/cevopen> ContentMine.org software: <http://github.com/contentmine> and : <http://github.com/petermr/ami3/> Wikidata at <https://www.wikidata.org/>
- * Chem4Word <https://www.facebook.com/groups/chem4word> EuropePMC at <http://europepmc.org>
- Images from Wikimedia Commons; Licence CC0 and CC BY-SA. GY, AK, thank NIPGR for financial support. CEV = ContentMine, Chem4Word, EsoilDB and Verriclear

The story: EuropePMC has 16000 Open articles reporting medicinal properties of plant oils.

We Automatically download articles adding sectioning, disambiguation, validation, normalization and semantics, From 100,000 compound references we create a semantic phytochemical dictionary. Everything (chemicals, plants, countries, activities, bibliography) is linked to Wikidata.

SEMANTICS: Wikidata (a Wikimedia project) has nearly 70 million items of Open semantic knowledge. Items have unique identifiers (Q numbers) related by properties (P). Example: "The plant thyme produces thymol which is an antiinfective" is represented by the triples:

Q408883 P2868 Q50377176 (thymol has_role antiinfective)

Q408883 P1582 Q148668 (thymol isNaturalProductOf T. vulgaris)

SEARCH "((essential oil) AND ((antimicrobial) OR (antibacterial) OR (antioxidant) OR (anthelmintic)))" using ContentMine's "getpapers" we got 16006 open access results in 2019-11 (EuPMCVVersion: 6.2)

We use ContentMine dictionaries to support text-mining ...

... and text-mining to populate Dictionaries!

Automatic sectioning of article and template-based table extraction

Heuristic template automatically identifies columns and head/body/footer

compound=(Constituents)	percentage=(%)
Trans-geraniol (Lemonol)	35.38
α-citral (Trans-citral)	20.37
Cis-citral (Cis-citral)	14.76
Cis-geraniol (nerol)	7.38
-----SNIPPED-----	
β-pinene	0.07
Total	98.62
Yield (w/w) %	1.40 %
Number of constituents	27
Hydrocarbon monoterpenoid	2.06
Oxygenated monoterpenoid	88.59
Sesquiterpenoid hydrocarbon	2
Oxygenated sesquiterpenoid	0.17
Others	5.8

Commonest compounds in 1000 articles

limonene x 125
α-pinene x 123
β-pinene x 118
linalool x 107
caryophyllene oxide x 100
β-caryophyllene x 92
camphene x 90
sabinene x 86
γ-terpinene x 69

Synonyms or isomers?

caryophyllene x 25
α-caryophyllene x 9
(E)-caryophyllene x 7
Isocaryophyllene x 6
caryophyllene <E>-x 6
(E)-caryophyllene x 5
trans-caryophyllene x 3
trans-β-caryophyllene x 2
E-caryophyllene x 2
(E)-β-caryophyllene x 2

Other dictionaries

COUNTRY
PLANT
PLANT_PART
INSTRUMENT
APPARATUS
SOLVENT
FUNDER
DISEASE
INSECT

Phytochemical dictionary in Chem4Word

Pinene ($C_{10}H_{16}$) is a bicyclic monoterpene chemical compound. There are two structural isomers of pinene found in nature: α -pinene and β -pinene. As the name suggests, both forms are important constituents of pine resin; they are also found in the resins of many other conifers, as well as in non-coniferous plants such as camphorweed (*Heterotheca*)^[3] and big sagebrush (*Artemisia tridentata*). Both isomers are used by many insects in their chemical communication system. The two isomers of pinene constitute the major component of turpentine.

KNIME workflow: extract annotate cheminformatics

Queries that CEVOpen/Wikidata can support

- How much does the composition of essential oils vary with geography?
- Can chemical fingerprints be correlated with medicinal activity?
- Who funds, or is likely to fund, essential oil research?
- What new phytochemicals have been reported in 2019 preprints?
- Identify plants that produce chemicals similar to known medicinal drugs.

Developments in progress

- Integration with KNIME (fingerprints, searching) OSCAR/OPSIN (NLP)
- Semantic annotation of articles
- Templates for extracting medicinal activity
- Checking tables and content for semantic consistency
- Synonym resolution and fuzzy chemistry
- What would YOU like?
- What can you help with?