CEVOpen: Can machines curate the Open phytochemistry literature?

Clive Adams [1], Clyde Davies[1], Emanuel Faria[2], Ambarish Kumar[3], Peter Murray-Rust[4,5,*], Mike Williams[1], Andy Wright[1], Gitanjali Yadav[3,4]

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

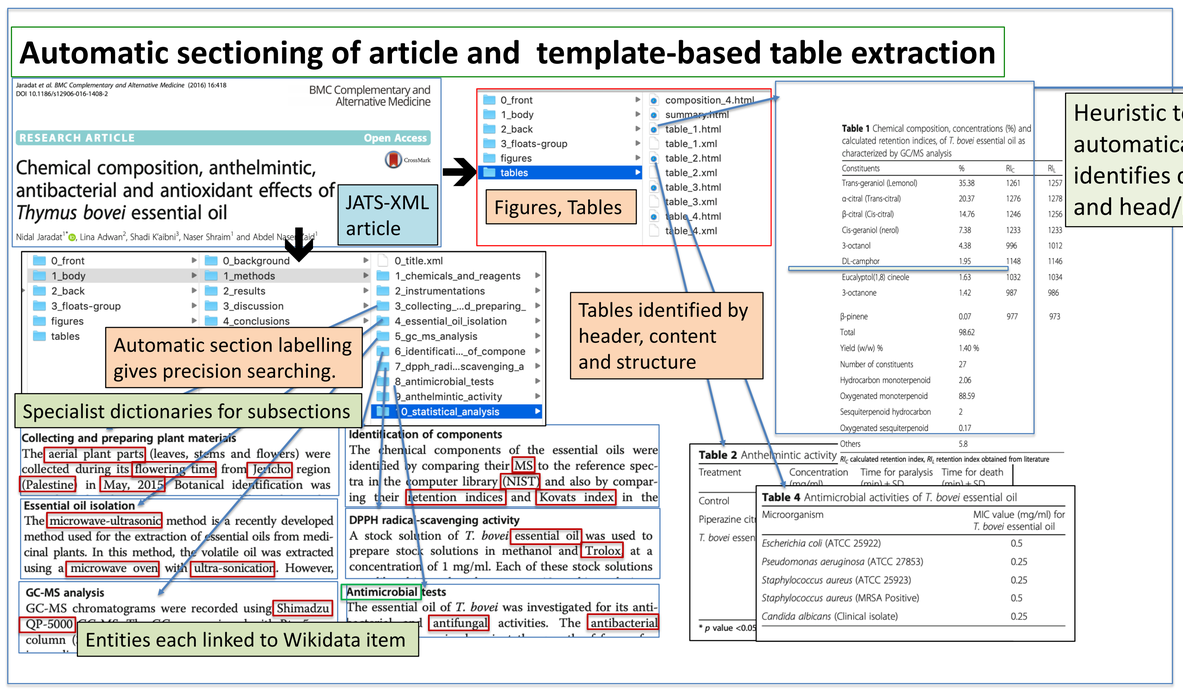
SEMANTICS: Wikidata (a Wikimedia project) has nearly 70 million **items** of Open semantic knowledge. **Items** have unique identifiers (**Q numbers**) related by **properties (P). Example: "*The plant thyme produces thymol which is an antiinfective*"** is represented by the triples:

`Q408883 P2868 Q50377176` (thymol **has_role** antiiinfective)

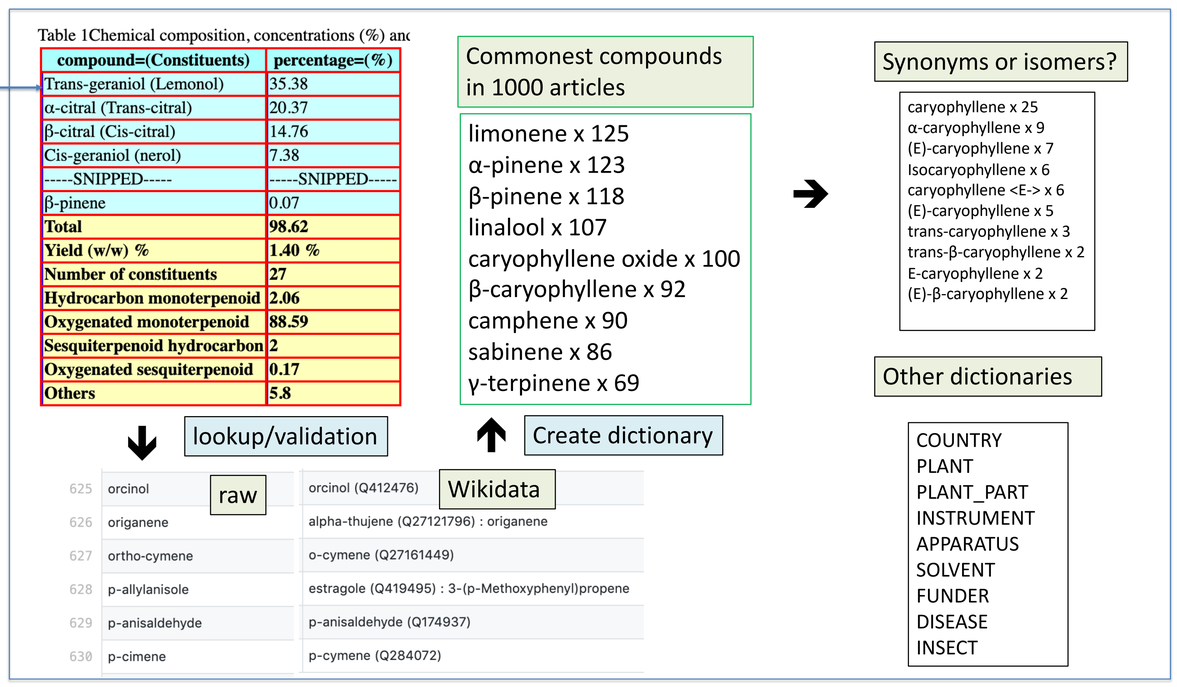
`Q408883 P1582 Q148668` (thymol **isNaturalProductOf** *T. vulgaris*)

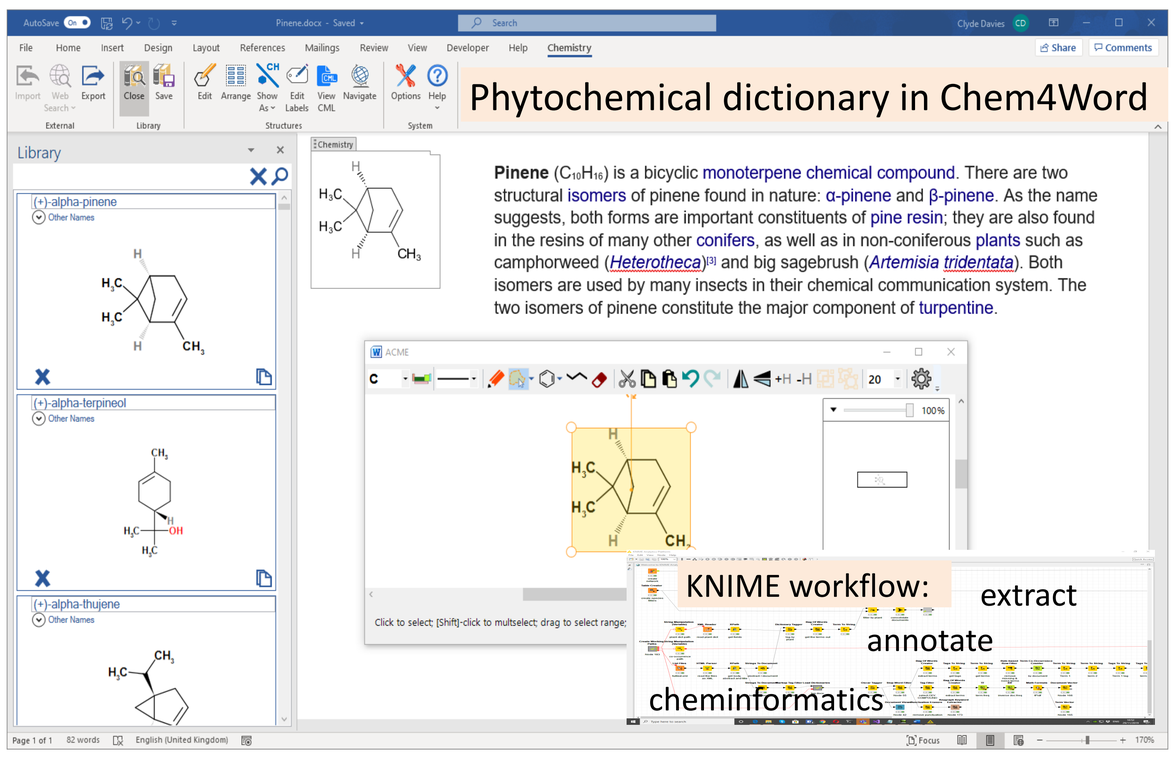
`SEARCH "((essential oil) AND ((antimicrobial) OR (antibacterial) OR (antioxidant) OR (anthelmintic)))"`
`using ContentMine's "getpapers" we got 16006 open access results in 2019-11(EuPMCVersion: 6.2)`

We use ContentMine dictionaries to support text-mining …



… and text-mining to populate Dictionaries!





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?

a: (PMR's IDs) [Wikidata] : [Q908710] (ORCID) (0000-0003-3386-3972)