**WikiFactMine for Phytochemistry**

**Mining the scientific literature for facts**

**Tom Arrow**, 1, **Charles Matthews**1, **Jenny Molloy**1,2, **Ross Mounce**1,2, **Peter Murray-Rust**1,3, **Richard Smith-Unna**1,2, **Lars Willighagen**1

1The ContentMine, Cambridge, CB4 2HY, 2Dept of Plant Sciences, University of Cambridge, 3Dept of Chemistry, University of Cambridge

**Introduction**

Understanding phytochemical diversity and metabolism can answer many important scientific questions and provide economically important information; forming the foundation for metabolic engineering of plant compounds. Phytochemical database resources exist but much information on their association with species, enzymes and places without the standardised format and metadata required to enable machine analysis. In some cases it is painstakingly extracted manually, but this approach is not scalable. Semi-automated extraction of phytochemical data across the full-text open access literature is anticipated to significantly extend previous abstract-only coverage. Here we present an open source pipeline and preliminary results for terpene data mining.

**ContentMine and Wikidata**

Wikidata is “Wikipedia for machines” and supports ContentMine’s FullContent search of the bioscience literature. We go beyond keywords to automatically generated structured dictionaries with thousands of terms and aliases. FullContent means not just words, but structured documents, tables and diagrams. We (and you) can search the whole literature (via EuropePMC or Crossref) every day automatically or retrospectively for your sub-areas of interest.

Example:

Find facts about terpenes emitted by conifers in Indonesia. We autogenerate 3 large dictionaries for all terpenes, conifers and Indonesian place/ island names in Wikidata.

**INTELLIGENT QUERIES**

Reusable WikitFactMine Dictionaries. We expand the Wikidata term terpene automatically to ~450 items (such as carvone) giving >1000 precise search terms and data. Similarly in a few seconds we can generate dictionaries of {conifers (1899); and Indonesian islands (6344) making broad queries precise.

**Search Strategies.**

(A) Daily search. All new open publications (300-1000) on EuropePMC are downloaded to WikimediaLabs, indexed by dictionaries, and the extracted facts (dictionary hits) stored in Zenodo (CERN’s Open repository). Each paper may have hundreds or thousands of facts.

(B) On-Demand. A researcher, especially those doing systematic reviews, creates a fairly general query in her field with a range of dates, journals, etc. and downloads papers (getpapers and quickscrape). The papers are filtered locally with a much more precise query (norma/ami).

**INTELLIGENT CONTENT**

(Bio)chemical transformations

Phylogenetics

- Pyromabacter piscolens: 195, *Jupatelea anthrisci* E3; *Synergistes jonesii* 201. (OCA: Tisserat) - And Imageanalysis (Norma)

- Diagrams of Chemical and biochemical reactions can be automatically extracted from PDFs into the Researcher’s filestore.

- Phylogenetic trees can be automatically extracted from bitmap diagrams or PDFs, and species names verified. Mounce, Murray-Rust, Wills: http://doi.org/10.3897/r.o.3.e13589

- Tables and graphs can be automatically extracted into researcher’s filestore and turned into CSV tables or spectra. Designed for re-use with your favourite tools (R, Python, etc.)

All software (Apache2) and Data (CC0) are Open. http://github.com/ContentMine. ContentMine.org is a not-for-profit UK company.

We thank The Shuttleworth Foundation for a Fellowship to PMR and The Wikimedia Foundation for funding for TA and CM. Contact peter@contentmine.org.