WikiFactMine for Phytochemistry

Mining the scientific literature for facts

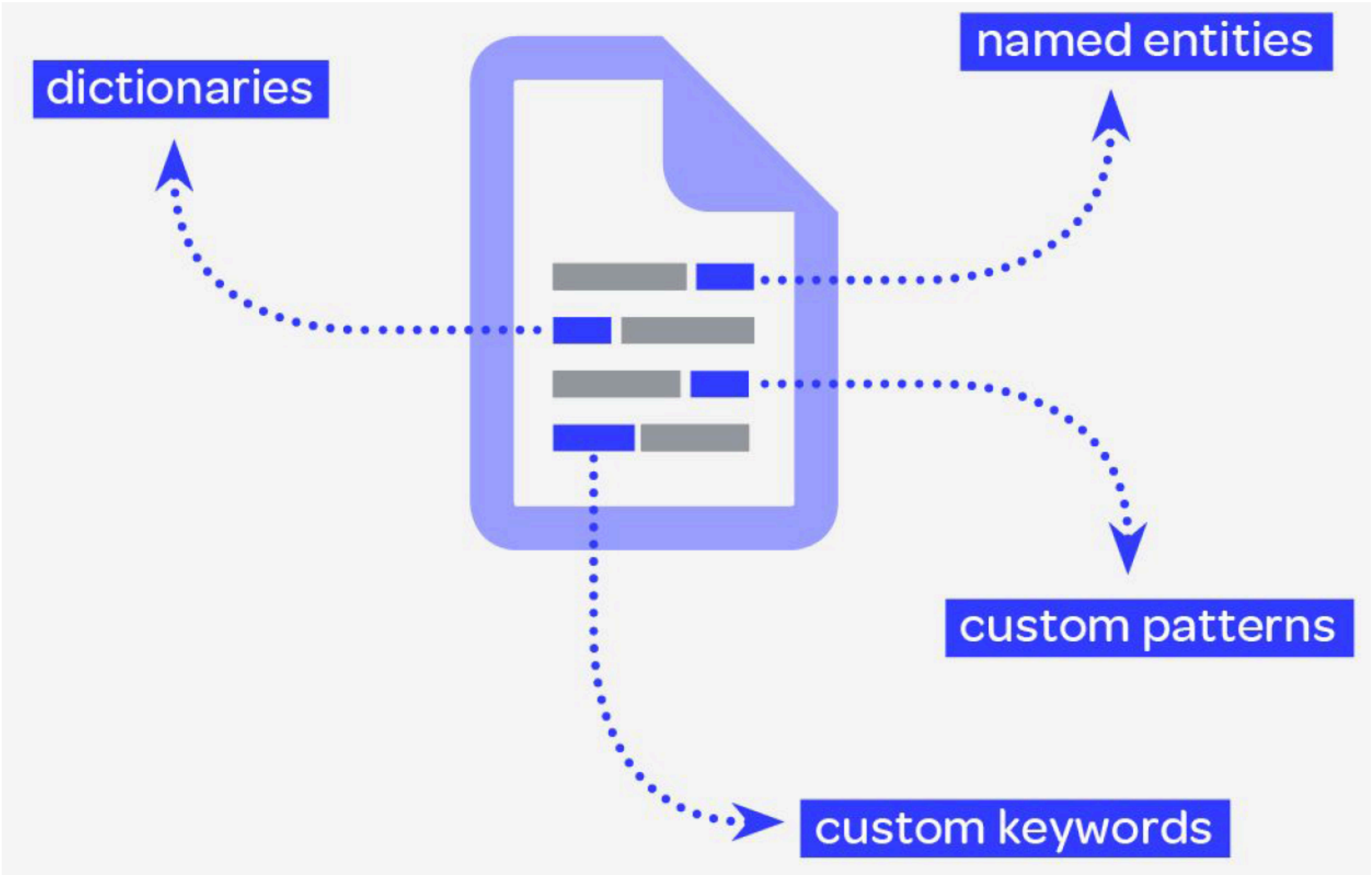
Tom Arrow[1], Charles Matthews[1], Jenny Molloy[1,2], Ross Mounce[1,2] Peter Murray-Rust[1,3], Richard Smith-Unna[1,2], Lars Willighagen[1]

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

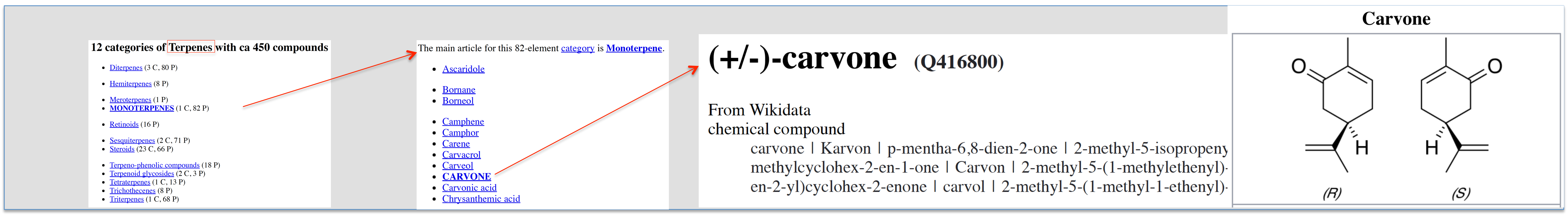
Wikidata is "Wikipedia for machines" and supports ContentMine's **FullContent search** of the Bioscience literature. We go beyond keywords to **automatically generated structured dictionaries** with thousands of terms and aliases. **FullContent** means not just words, but structured documents, tables 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



12 categories of Terpenes with ca 450 compounds

- Diterpenes (3 C, 80 P)
- Hemiterpenes (8 P)
- Meroterpenes (1 P)
- **MONOTERPENES** (1 C, 82 P)
- Retinoids (16 P)
- Sesquiterpenes (2 C, 71 P)
- Steroids (23 C, 66 P)
- Terpeno-phenolic compounds (18 P)
- Terpenoid glycosides (2 C, 3 P)
- Tetraterpenes (1 C, 13 P)
- Trichothecenes (8 P)
- Triterpenes (1 C, 68 P)

The main article for this 82-element category is **Monoterpene**.

- Ascaridole
- Bornane
- Borneol
- Camphene
- Camphor
- Carene
- Carvacrol
- Carveol
- **CARVONE**
- Carvonic acid
- Chrysanthemic acid

(+/-)-carvone (Q416800)

From Wikidata

chemical compound

carvone | Karvon | p-mentha-6,8-dien-2-one | 2-methyl-5-isopropeny methylcyclohex-2-en-1-one | Carvon | 2-methyl-5-(1-methylethenyl)-en-2-yl)cyclohex-2-enone | carvol | 2-methyl-5-(1-methyl-1-ethenyl)-

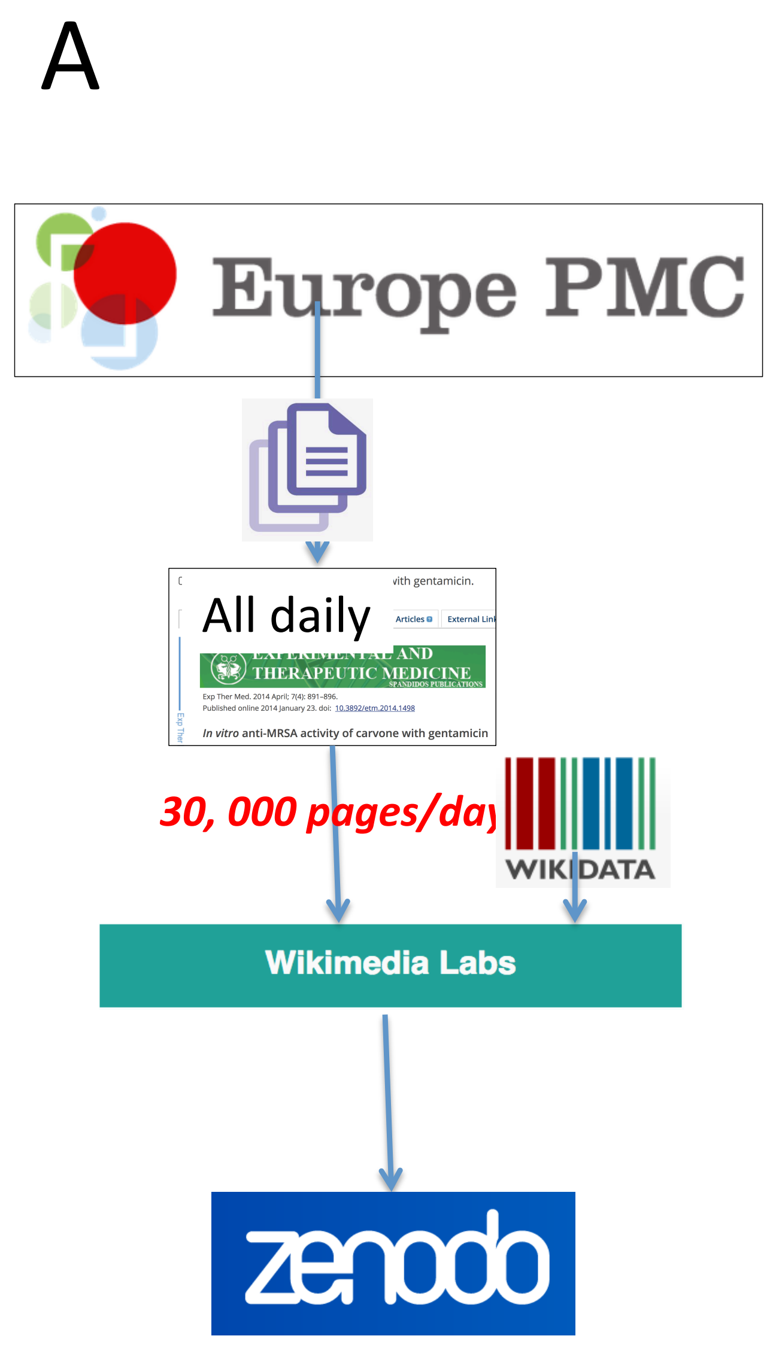
Carvone

Reusable WikiFactMine 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.

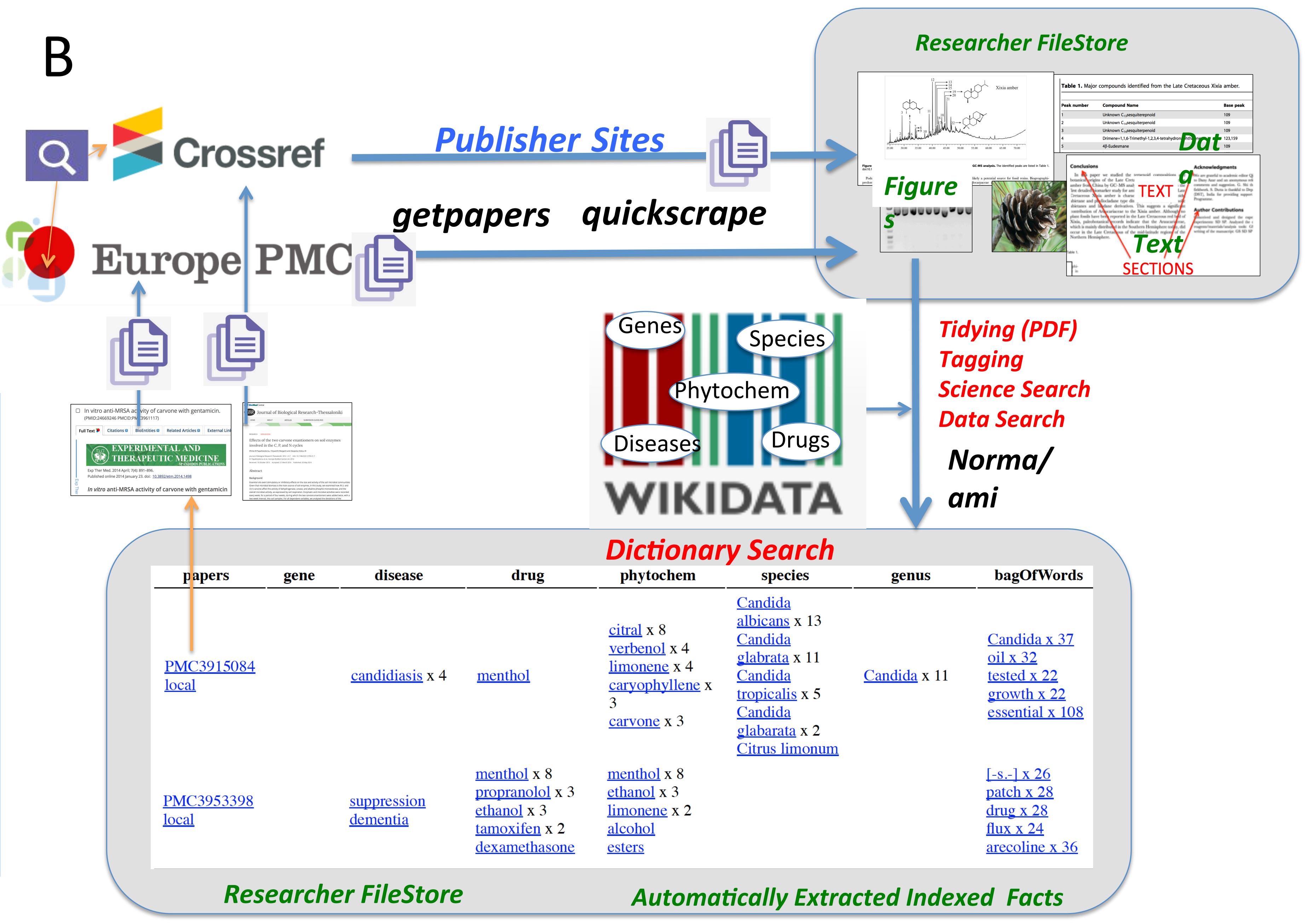
(A) Daily search. All new Open publications (300-1000) on EuropePMC are downloaded to WikimediaLabs, indexed by dictionaries, and the extracted facts (dictionary hits) stored in Zenodo (CERN's Open repository) . Each paper may have hundreds or thousands of facts.

(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**).



A. All EPMC papers are downloaded every day and the facts are extracted into Zenodo and made publicly available.

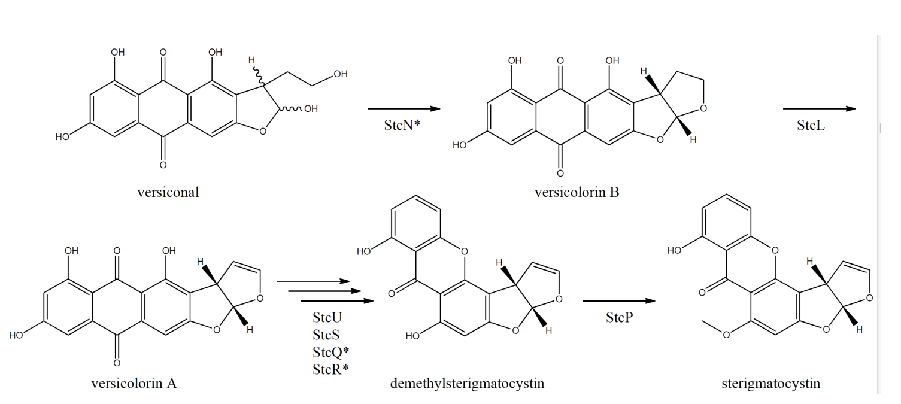
B. Researcher searches repositories and also scrapes publisher sites for whatever chunk of the literature she wants. She runs local dictionaries and saves the results to disk where they can be further analyzed. She can add any papers she has legal access to and re-run whenever required. E.g. Bag Of Words is a powerful tool for classifying papers



papers	gene	disease	drug	phytochem	species	genus	bagOfWords
PMC3915084 local		candidiasis x 4	menthol	citral x 8 verbenol x 4 limonene x 4 caryophyllene x 3 carvone x 3	Candida albicans x 13 Candida glabrata x 11 Candida tropicalis x 5 Candida glabarata x 2 Citrus limonum	Candida x 11	Candida x 37 oil x 32 tested x 22 growth x 22 essential x 108
PMC3953398 local		suppression dementia	menthol x 8 propranolol x 3 ethanol x 3 tamoxifen x 2 dexamethasone	menthol x 8 ethanol x 3 limonene x 2 alcohol esters			[-s.-] x 26 patch x 28 drug x 28 flux x 24 arecoline x 36

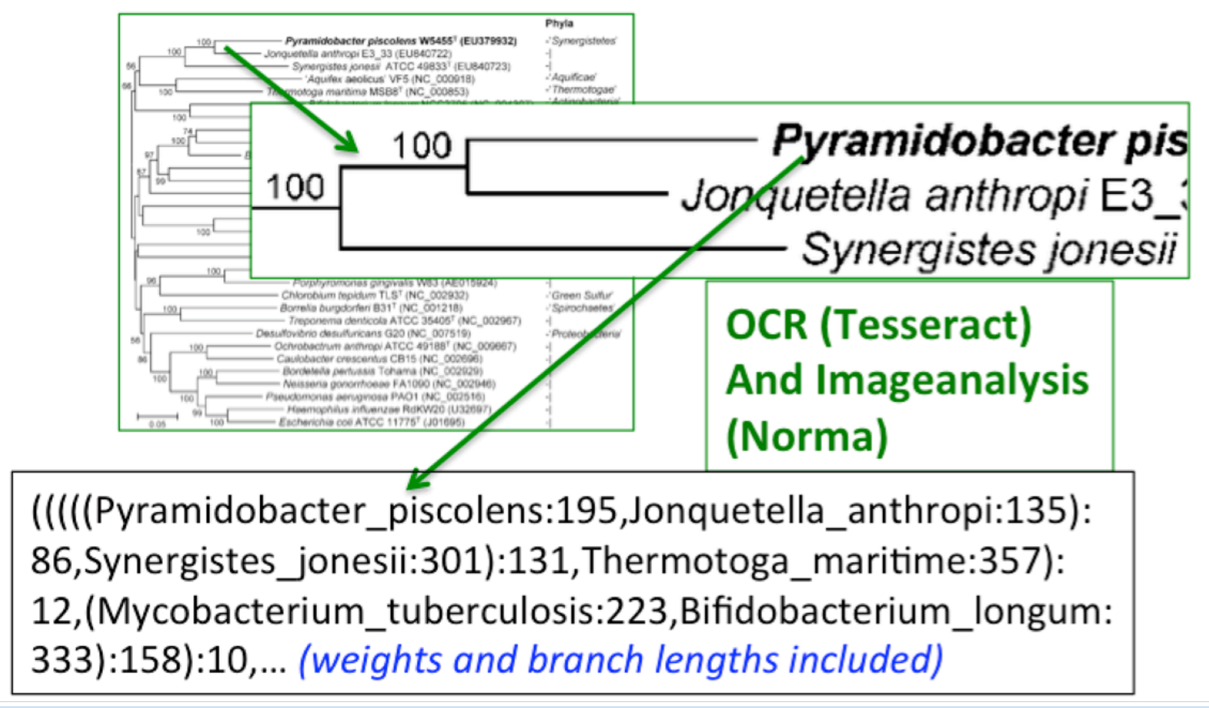
INTELLIGENT CONTENT

(Bio)chemical transformations



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

Phylogenetics



(((((Pyramidobacter_piscolens:195,Jonquetella_anthropi:135):86,Synergistes_jonesii:301):131,Thermotoga_maritime:357):12,(Mycobacterium_tuberculosis:223,Bifidobacterium_longum:333):158):10,... *(weights and branch lengths included)*

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

Tables and graphs

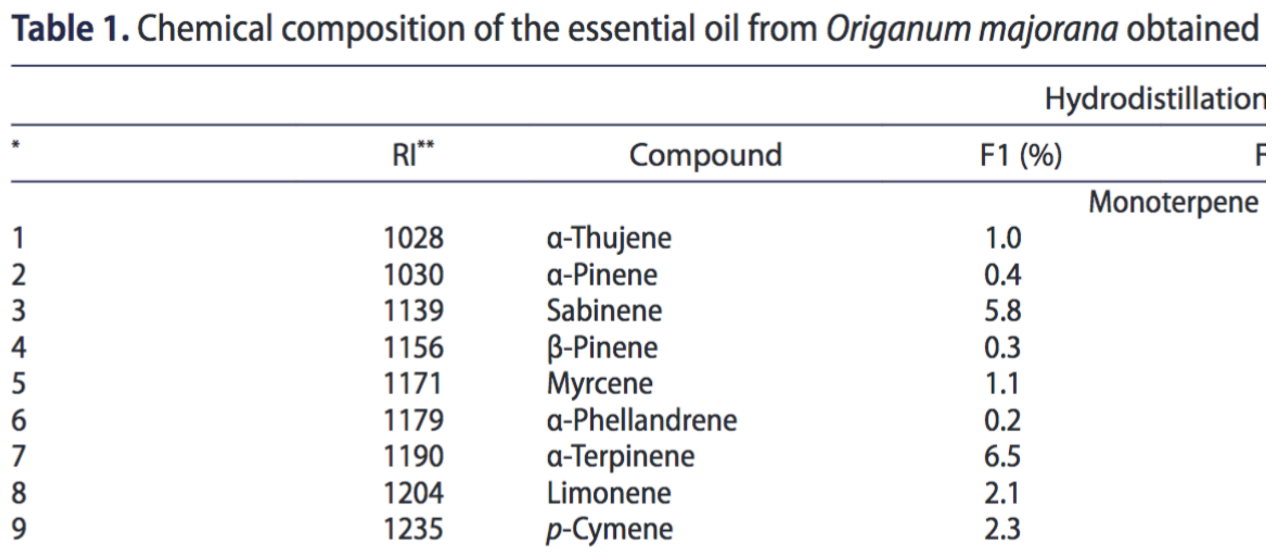


Table 1. Chemical composition of the essential oil from *Origanum majorana* obtained

			Hydrodistillation	
*	RI**	Compound	F1 (%)	F
				Monoterpene
1	1028	α-Thujene	1.0	
2	1030	α-Pinene	0.4	
3	1139	Sabinene	5.8	
4	1156	β-Pinene	0.3	
5	1171	Myrcene	1.1	
6	1179	α-Phellandrene	0.2	
7	1190	α-Terpinene	6.5	
8	1204	Limonene	2.1	
9	1235	*p*-Cymene	2.3	

C. 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 tp PMR and The Wikimedia Foundation for funding for TA and CM. *Contact* ***peter@contentmine.org***