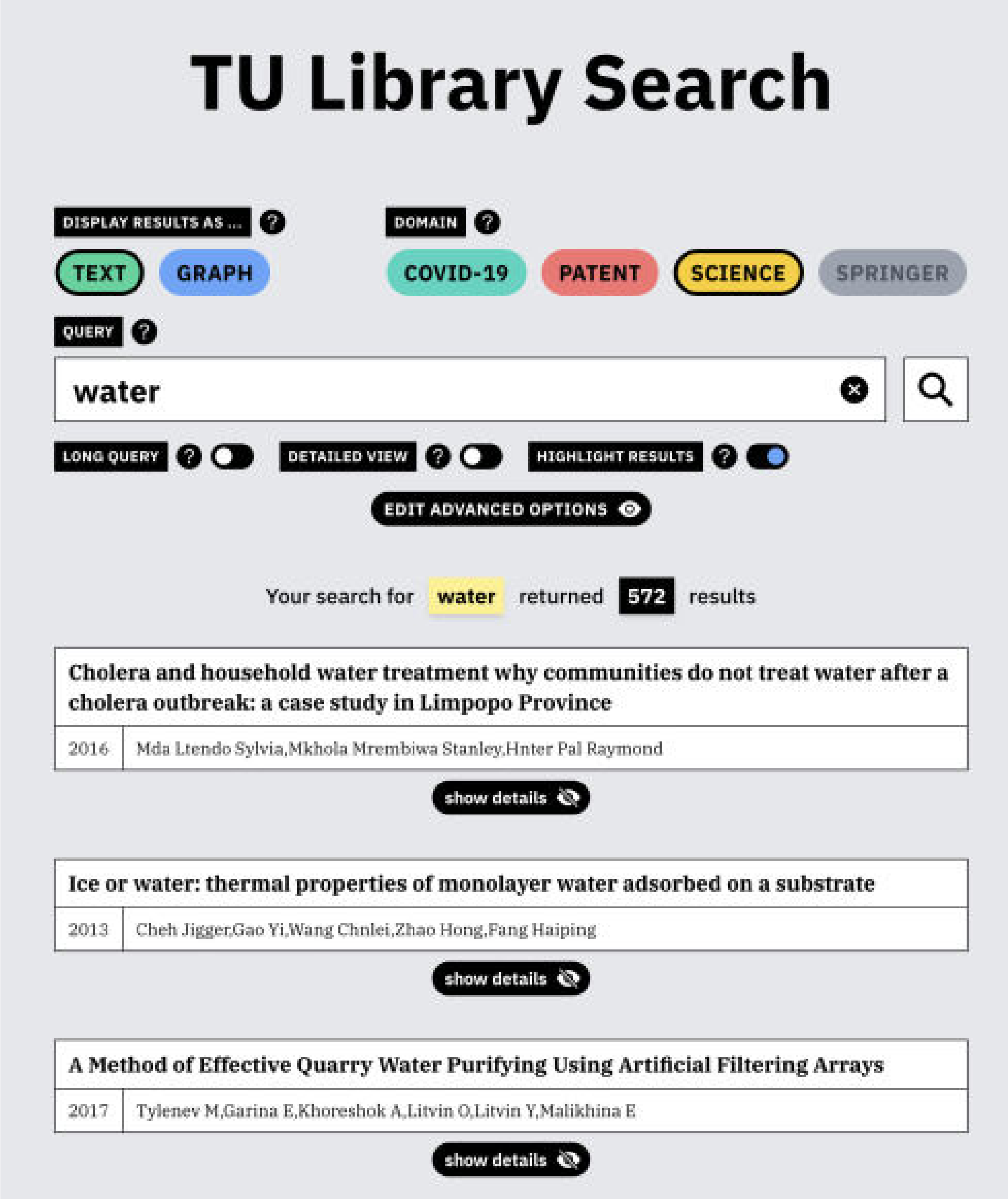
TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Artificial Researcher

# AR-Science: From Raw Text Data to Structured Semantic Representation

# Text Search Interface



## TU Library Search

water

Your search for **water** returned **572** results

**Cholera and household water treatment why communities do not treat water after a cholera outbreak: a case study in Limpopo Province**

2016 | Mda Ltendo Sylvia,Mkhola Mrembiwa Stanley,Hnter Pal Raymond

**Ice or water: thermal properties of monolayer water adsorbed on a substrate**

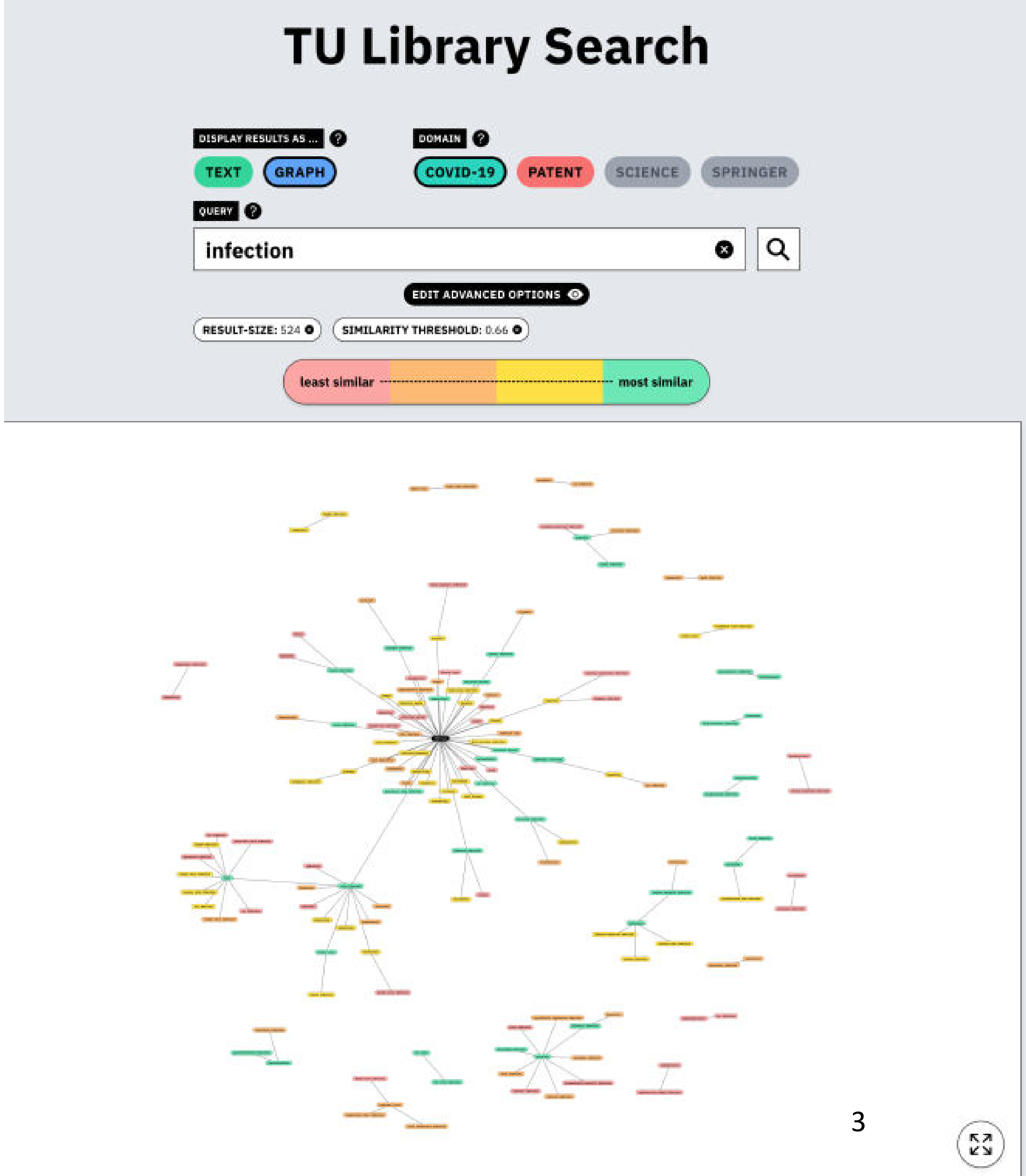
2013 | Cheh Jigger,Gao Yi,Wang Chnlei,Zhao Hong,Fang Haiping

**A Method of Effective Quarry Water Purifying Using Artificial Filtering Arrays**

2017 | Tylenev M,Garina E,Khoreshok A,Litvin O,Litvin Y,Malikhina E

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

# Graph Search Interface



TU Library Search

DISPLAY RESULTS AS ... ?

TEXT [✓ GRAPH]

DOMAIN ?

[✓ COVID-19] PATENT SCIENCE SPRINGER

QUERY ?

infection

EDIT ADVANCED OPTIONS

RESULT-SIZE: 524

SIMILARITY THRESHOLD: 0.66

least similar ---------------- most similar

# Requirements for Scientific Search

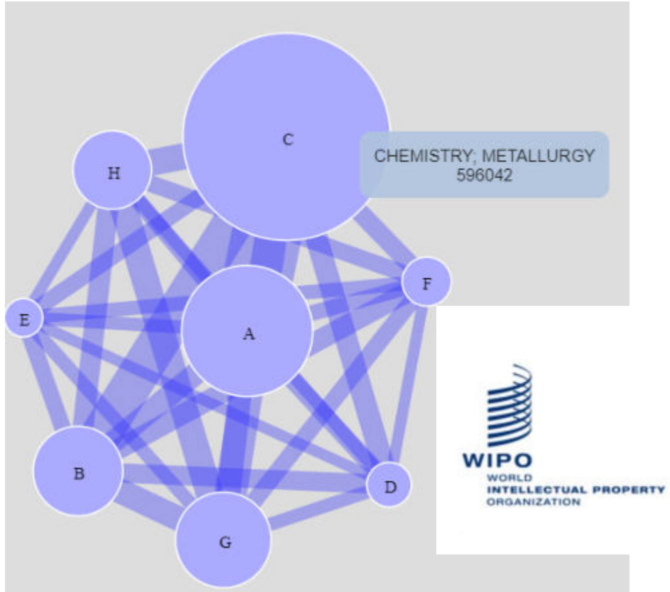
- Meta-data
    - Bibliographic data
    - References and citations
    - Resource sharing (code/data)
- Full-text
    - Identification of domain-specific terminology and concepts
    - Identification category scientific discipline
    - Identification of related terms in discipline context
    - Identification  text paragraph
    - The essence of the article in terms of claim, method, result, conclusion

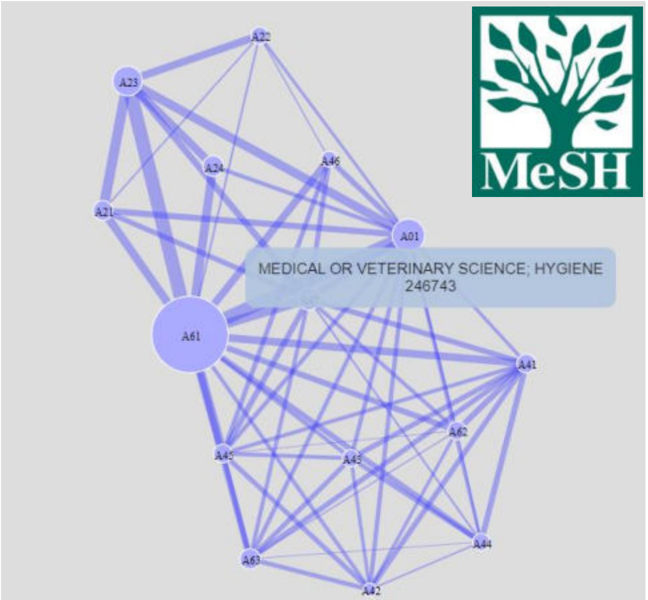
# From Text to Structured Semantic Representation of Scientific Publications

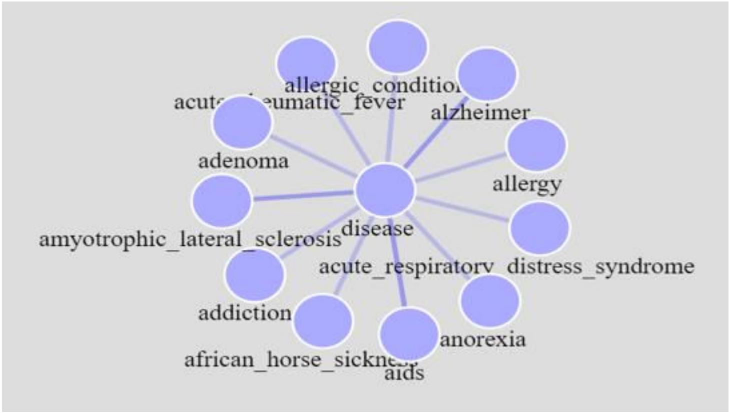
```
<!DOCTYPE article
PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Archiving and Interchange
DTD with MathML3 v1.2 20190208//EN" "JATS-archivearticle1-
mathml3.dtd</p></sec><sec><title>Objective</title><p> For
se                       t system segmentation from CTX data, we
pr                       e</italic> approach, which is an extension
of                       cale vesselness&#x0201d; method and
in                       ce. It all                         of
ro                       several <                          m in
po                       Additiona                          italic>
is                       tation of                          , but
is                       ge roots w                         an 40
voxels. </p></sec><sec><title>Results<                      me data
of full-grown bean plants as well as t                      growth
studies of cassava plants, <italic>RootForce</italic> produces similar
(and much faster) results compared to manual segmentation of the
regarded root architectures. Furthermore, <italic>RootForce</italic>
enables                  raits not possible to
before,                  olume (<italic>V</ital
total r                  </italic><sub>root</su
over de                  es
(<itali                  sub>min</sub>,
<italic                  ub>mean</sub>, and
```

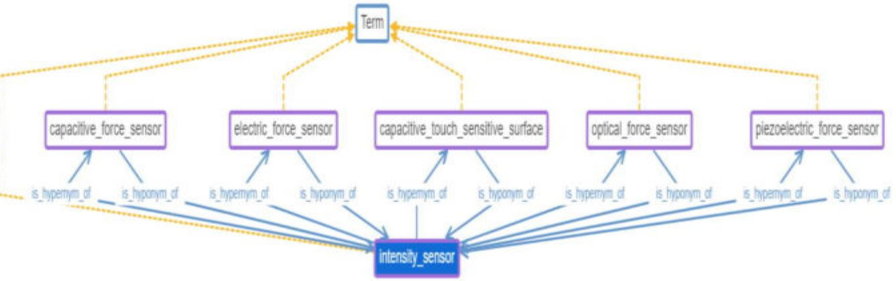
United States Design Patent
Context and Structure in Automated Full-Text Information Access
United States Patent
Exploiting Hierarchy in Text Categorization

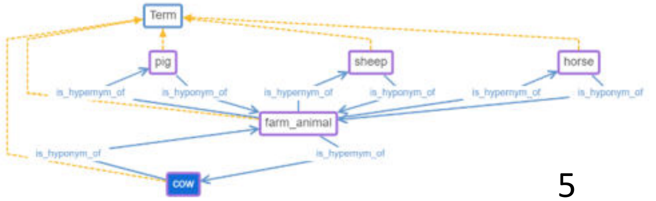
International Patent Classification (IPC)







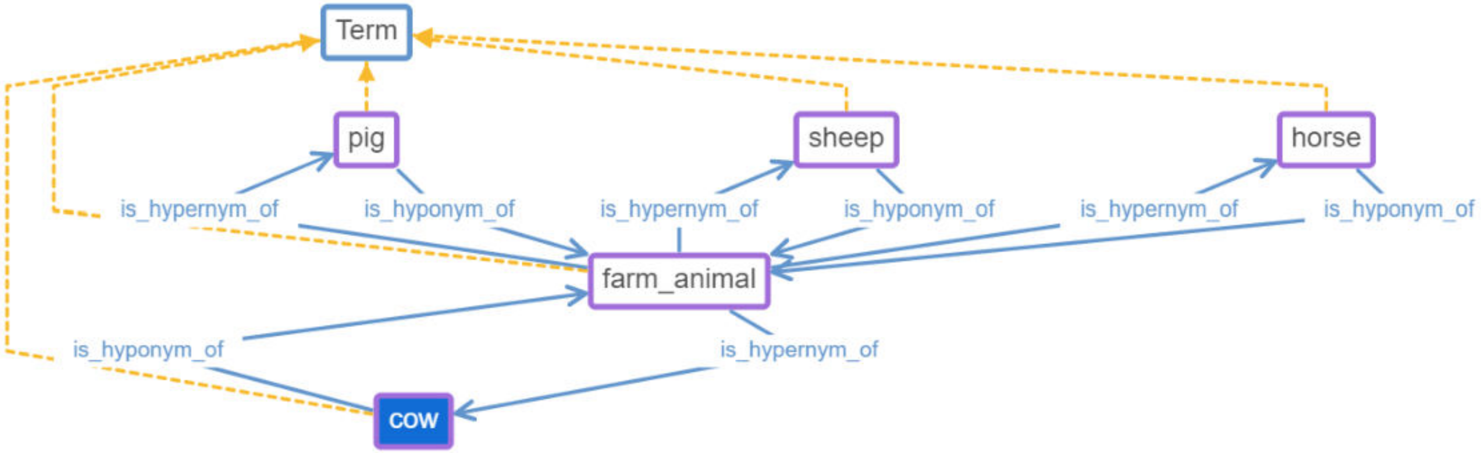




# Lexical Relations

Hyponym relations:

- part of
- kind of
- member of
- type of
- domain specific relation

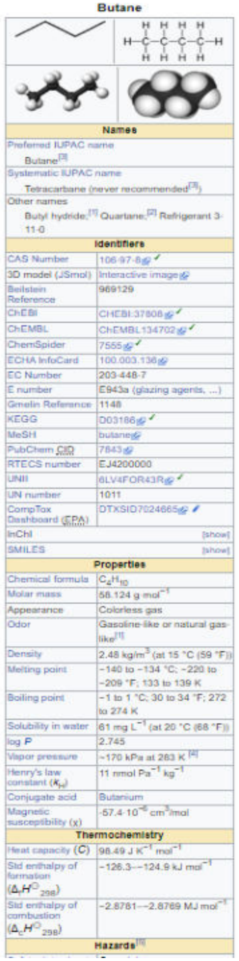


Hypernym, **a broad term**,
e.g. farm animals a hypernym of pig, cow, sheep, horse

Hyponym, **a narrow term**
e.g. horse is a hyponym of farm animals

# Integrating Different Resources

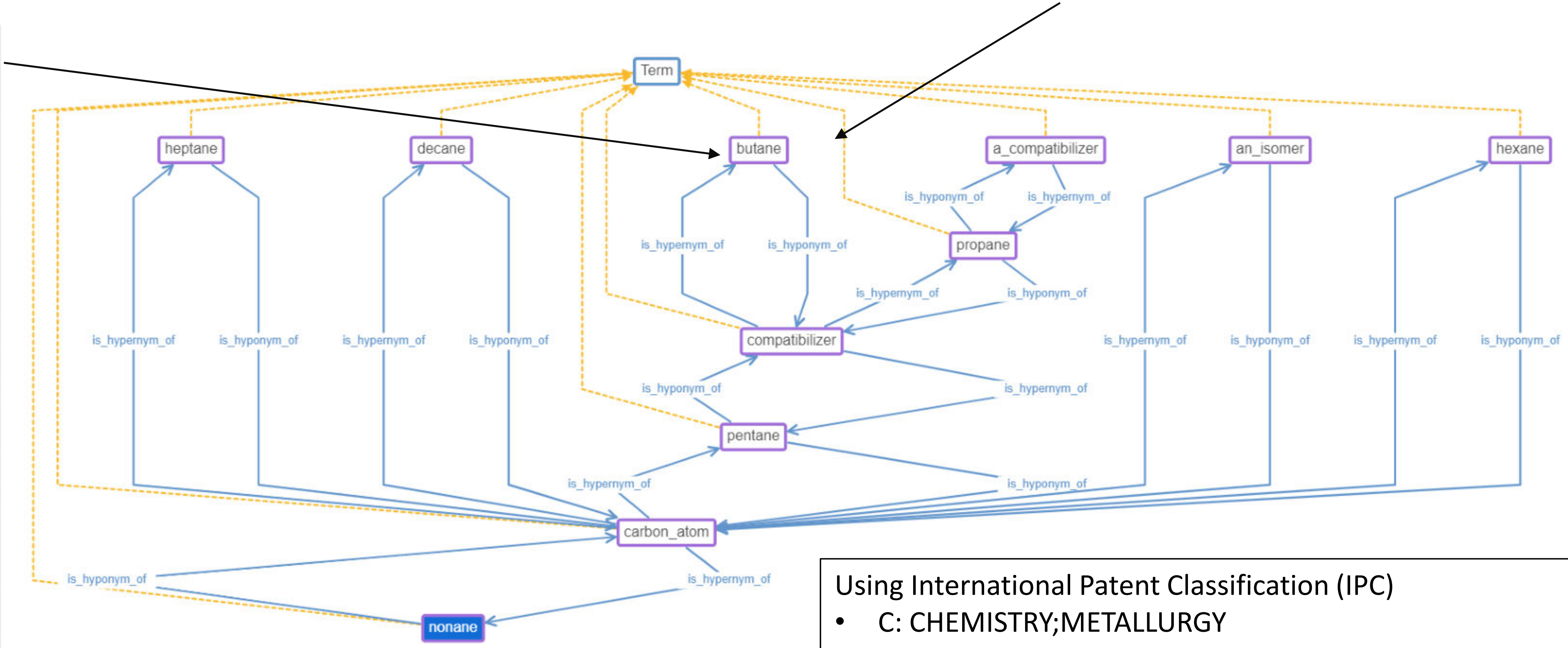
Wikipedia



Butane

| | |
|---|---|
| **Names** | |
| Preferred IUPAC name<br>Butane[3] | |
| Systematic IUPAC name<br>Tetracarbane (never recommended[3]) | |
| Other names<br>Butyl hydride;[1] Quartane;[2] Refrigerant 3-11-0 | |
| **Identifiers** | |
| CAS Number | 106-97-8 ✓ |
| 3D model (JSmol) | Interactive image |
| Beilstein Reference | 969129 |
| ChEBI | CHEBI:37808 ✓ |
| ChEMBL | ChEMBL134702 ✓ |
| ChemSpider | 7555 ✓ |
| ECHA InfoCard | 100.003.136 |
| EC Number | 203-448-7 |
| E number | E943a (glazing agents, ...) |
| Gmelin Reference | 1148 |
| KEGG | D03186 ✓ |
| MeSH | butane |
| PubChem CID | 7843 |
| RTECS number | EJ4200000 |
| UNII | 6LV4FOR43R ✓ |
| UN number | 1011 |
| CompTox Dashboard (EPA) | DTXSID7024665 |
| InChI | [show] |
| SMILES | [show] |
| **Properties** | |
| Chemical formula | $C_4H_{10}$ |
| Molar mass | 58.124 g mol$^{-1}$ |
| Appearance | Colorless gas |
| Odor | Gasoline-like or natural gas-like[1] |
| Density | 2.48 kg/m$^3$ (at 15 °C (59 °F)) |
| Melting point | −140 to −134 °C; −220 to −209 °F; 133 to 139 K |
| Boiling point | −1 to 1 °C; 30 to 34 °F; 272 to 274 K |
| Solubility in water | 61 mg L$^{-1}$ (at 20 °C (68 °F)) |
| log *P* | 2.745 |
| Vapor pressure | ~170 kPa at 283 K [4] |
| Henry's law constant ($k_H$) | 11 nmol Pa$^{-1}$ kg$^{-1}$ |
| Conjugate acid | Butanium |
| Magnetic susceptibility (χ) | -57.4·10$^{-6}$ cm$^3$/mol |
| **Thermochemistry** | |
| Heat capacity (*C*) | 98.49 J K$^{-1}$ mol$^{-1}$ |
| Std enthalpy of formation ($\Delta_f H^\circ_{298}$) | −126.3–−124.9 kJ mol$^{-1}$ |
| Std enthalpy of combustion ($\Delta_c H^\circ_{298}$) | −2.8781–−2.8769 MJ mol$^{-1}$ |
| **Hazards**[5] | |

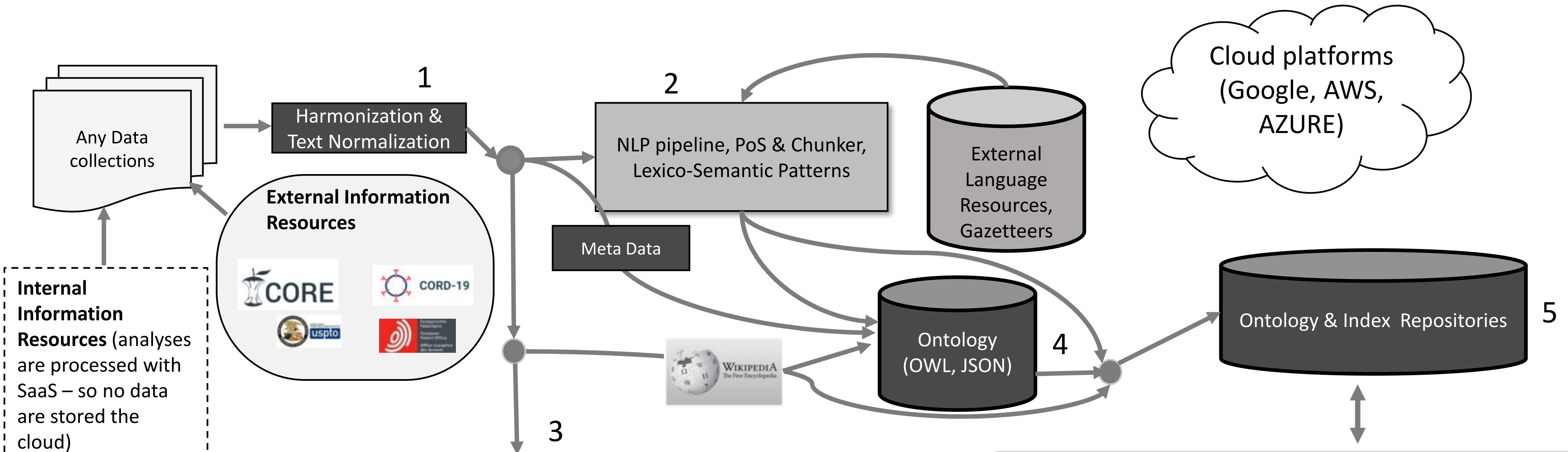
Label: Liquefied petroleum gas



Using International Patent Classification (IPC)

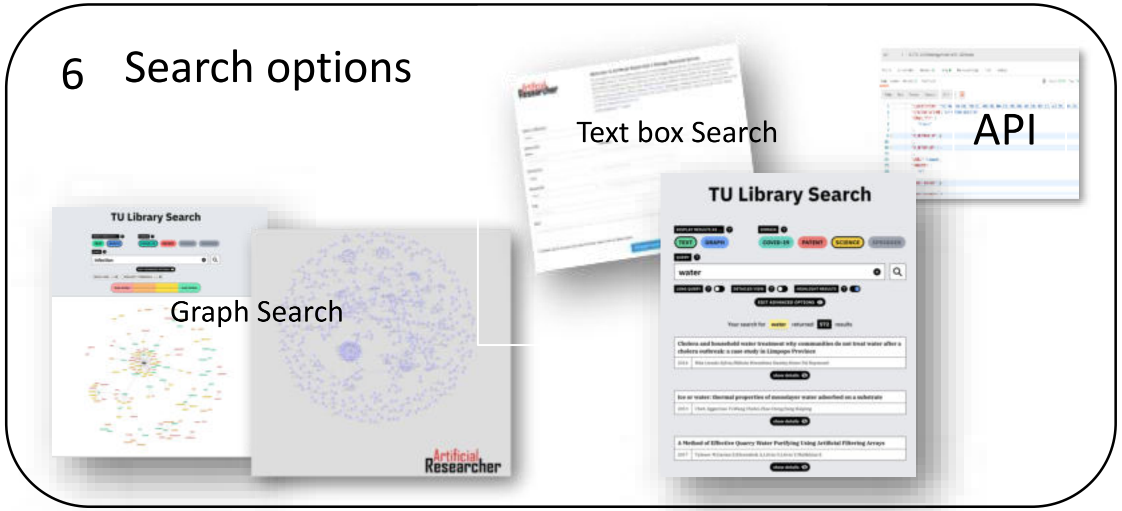
- C: CHEMISTRY;METALLURGY
  - C01: INORGANIC CHEMISTRY
    - C01B; NON-METALLIC ELEMENTS; COMPOUNDS

# Assembly Line for Creating Indices and Ontologies


Any Data collections
1
Harmonization & Text Normalization
2
NLP pipeline, PoS & Chunker, Lexico-Semantic Patterns
External Language Resources, Gazetteers
Cloud platforms (Google, AWS, AZURE)
External Information Resources
CORE
CORD-19
uspto
Meta Data
Internal Information Resources (analyses are processed with SaaS – so no data are stored the cloud)
Wikipedia The Free Encyclopedia
Ontology (OWL, JSON)
4
Ontology & Index Repositories
5
3

**Artificial Researcher NLP-Toolkit**

| Word Similarity Model | Context Similarity Model | Task Model |
|---|---|---|

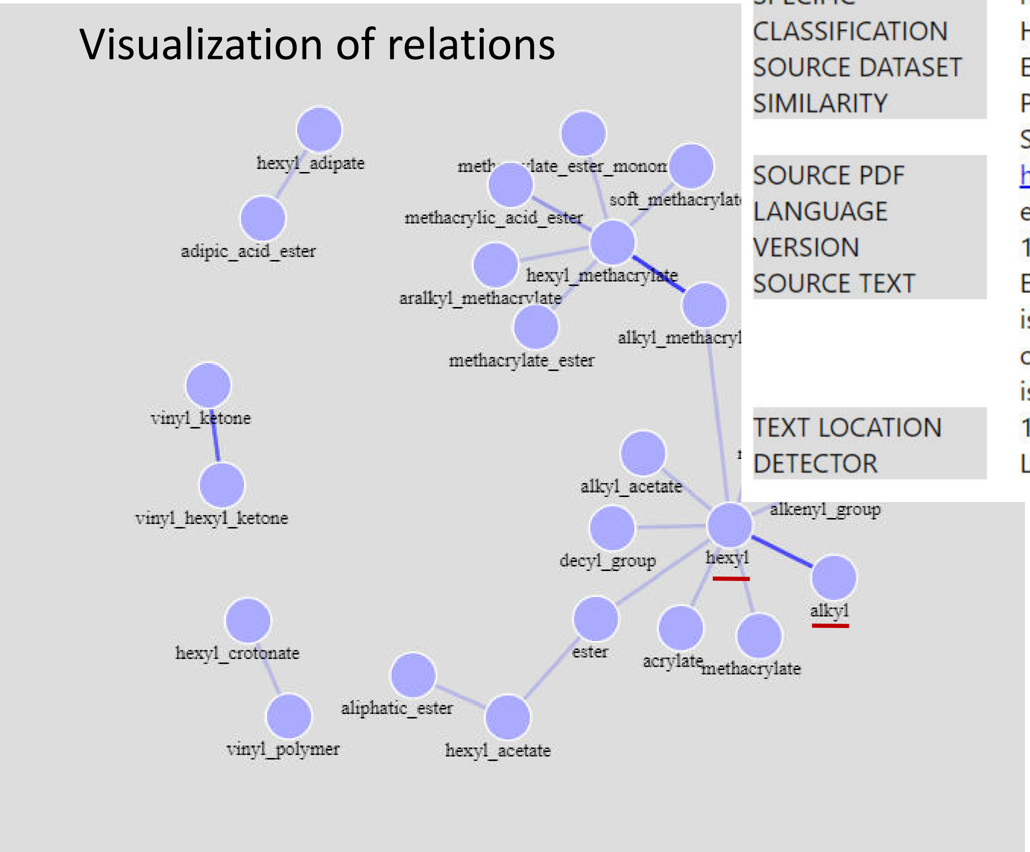


Documents are processed through several text analysis modules:

1. Text Normalization and harmonization (including PDF converter)
2. NLP and relation identification
3. Artificial Researcher NLP-Toolkit - Technical Terms
4. Packaged as enhanced XML/JSON
5. Software AR Ontology Services and AR Passage Retrieval
6. End user applications rest APIs, desktop script, GUI

# Visual display and direct access to original resource

Visualization of relations



| | |
|---|---|
| GENERAL | alkyl |
| SPECIFIC | hexyl |
| CLASSIFICATION | H04R, |
| SOURCE DATASET | EPO |
| SIMILARITY | PatBERT: 0.1342575,<br>SciBERT: 0.83727044, |
| SOURCE PDF | https://data.epo.org/publication-server/pdf-document?cc=EP&pn=1701583&ki=A1&pd=2006-09-13 |
| LANGUAGE | en |
| VERSION | 1 |
| SOURCE TEXT | Examples of alkyl include hexyl , heptyl , octyl , isooctyl , nonyl , isononyl , decyl , isodecyl , undecyl , dodecyl , tridecyl , isotridecyl , tetradecyl , isotetradecyl , pentadecyl , isopentadecyl , hexadecyl , isohexadecyl , heptadecyl , isoheptadecyl , ocadecyl , nonadecyl , isononadecyl , eicocyl , isoeicocyl , henicosyl , isohenicosyl , docosyl , isodocosyl , tricosyl , isotricosyl , tetracosyl and isotetracosyl . |
| TEXT LOCATION | 1701583_DESCR_0 |
| DETECTOR | LSP_PATTERN_DETECTOR |

Access to the Meta data such as taxonomy (IPC), and context semantic similarities

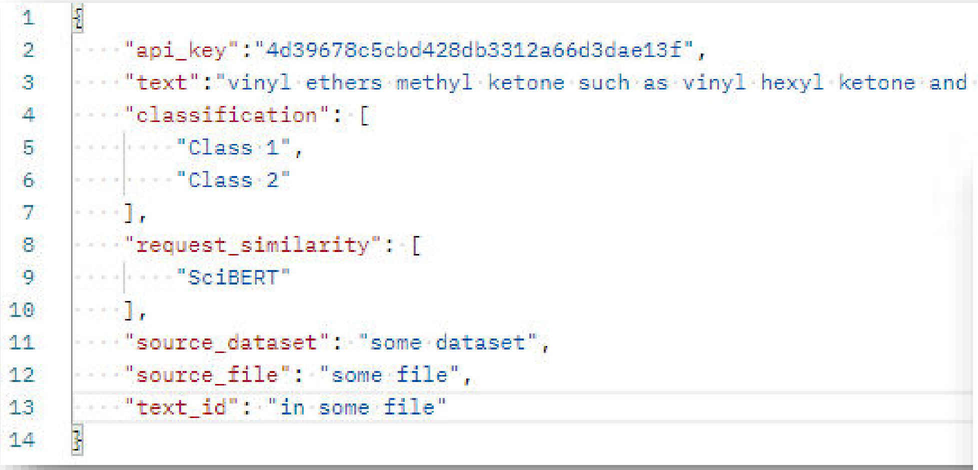
Direct access to the source and original sentence

## Query formulation Boolean

(ester OR hexyl OR acrylate OR methyl OR alkyl OR methacrylate OR alkyls OR diol OR radical) OR ("decyl group"~4 OR "alkyl methacrylate"~4 OR "alkyl acetate"~4 OR "alkenyl group"~4 OR "substituted alkyl"~4 OR "aliphatic alcohol"~4 OR "ethylene glycol"~4 OR "methacrylic acid"~4 OR "crotonic acid"~4 OR "alkyl radical"~4 OR "suitable alkyl acrylate"~5 OR "normal alkyl radical"~5 OR "lower alkyls"~4 OR "c alkyl"~4 OR "selenite ester"~4 OR "branched chain"~4 OR "alkoxyalkyl group"~4 OR "ester linkage"~4)

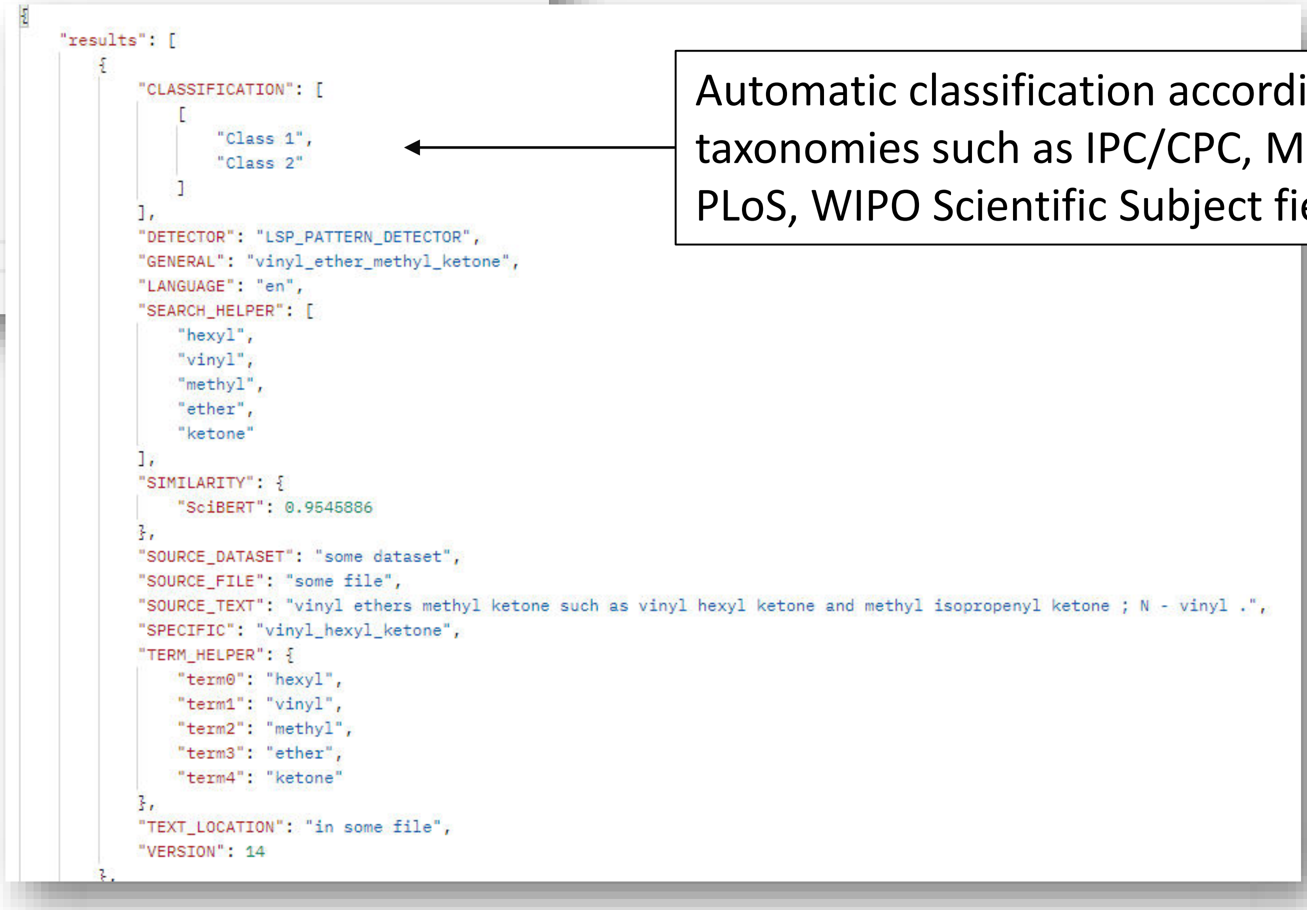
# Extracting semantic relation from natural language text

Input



```
{
    "api_key":"4d39678c5cbd428db3312a66d3dae13f",
    "text":"vinyl ethers methyl ketone such as vinyl hexyl ketone and methyl isopropenyl ketone ; N-vinyl.",
    "classification": [
        "Class 1",
        "Class 2"
    ],
    "request_similarity": [
        "SciBERT"
    ],
    "source_dataset": "some dataset",
    "source_file": "some file",
    "text_id": "in some file"
}
```

Output



```
{
    "results": [
        {
            "CLASSIFICATION": [
                [
                    "Class 1",
                    "Class 2"
                ]
            ],
            "DETECTOR": "LSP_PATTERN_DETECTOR",
            "GENERAL": "vinyl_ether_methyl_ketone",
            "LANGUAGE": "en",
            "SEARCH_HELPER": [
                "hexyl",
                "vinyl",
                "methyl",
                "ether",
                "ketone"
            ],
            "SIMILARITY": {
                "SciBERT": 0.9545886
            },
            "SOURCE_DATASET": "some dataset",
            "SOURCE_FILE": "some file",
            "SOURCE_TEXT": "vinyl ethers methyl ketone such as vinyl hexyl ketone and methyl isopropenyl ketone ; N - vinyl .",
            "SPECIFIC": "vinyl_hexyl_ketone",
            "TERM_HELPER": {
                "term0": "hexyl",
                "term1": "vinyl",
                "term2": "methyl",
                "term3": "ether",
                "term4": "ketone"
            },
            "TEXT_LOCATION": "in some file",
            "VERSION": 14
        },
```

Automatic classification according to taxonomies such as IPC/CPC, MeSH, PLoS, WIPO Scientific Subject fields

# Better than Distributional Semantics

Using semantically enriched data can retrieve up to 60% more related concepts than traditional pre-trained contextual models (a.k.a. Neural Network-based methods, BERT-family).

## Artificial Researcher's text mining technology

**Technical terms**

Related concept

**brake pedal:**

- vehicle operating pedal,
- conventional hydraulic brake system
- pedal devices
- position brake actuating member
- brake actuating member
- hydraulically-assisted rack pinion steering gear
- brake operating member
- conventional braking system
- pair pedals

**accelerator pedal**

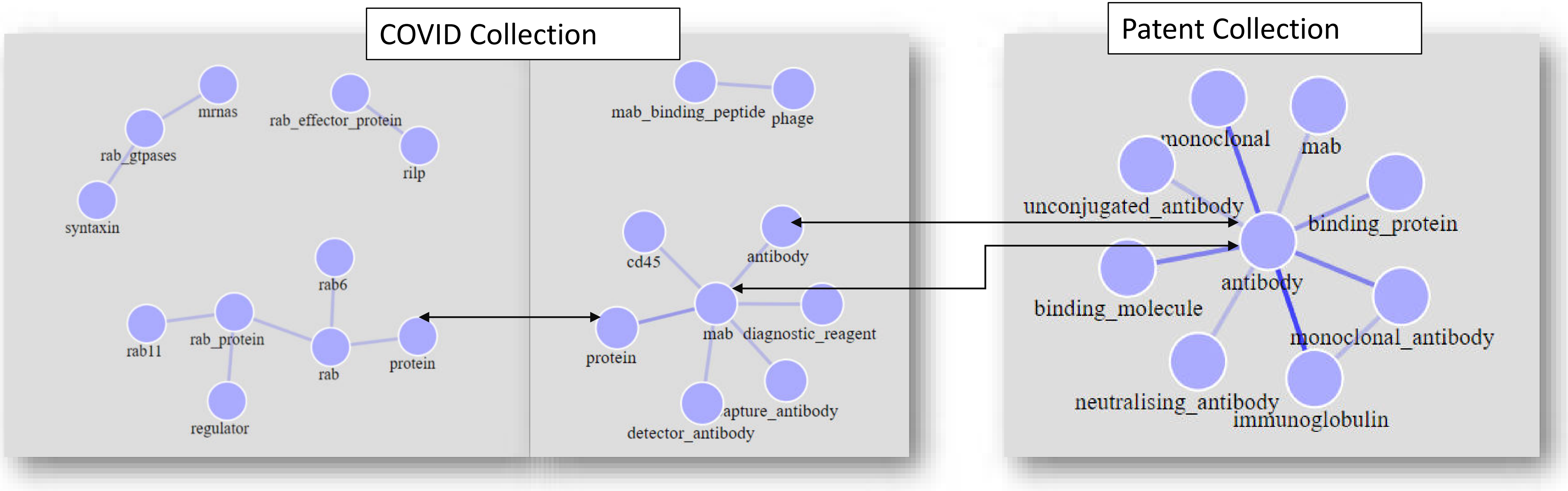
- case pedal device
- pedal device

## Extraction using pre-trained contextual models only

| **brake pedal** | PatBERT | SciBERT |
|---|---|---|
| conventional hydraulic brake system | 0.69 | 0.89 |
| hydraulically-assisted rack pinion steering gear | 0.49 | 0.80 |
| conventional braking system | 0.66 | 0.84 |

Artificial Researcher

# Connecting Different Collections and Domains



# Verified Use Cases

Artificial Researcher in Open Access

**Intellectual property Industry**

Cloud-based architecture

- Scientific publication
- Patent data

***Pre-Seed 2019-2020***

Artificial Researcher in Science

**Next generation scientific search tool**

Onsite architecture

- Open access
- Internal resources

***FemPower IKT 2019-2022***

EUROPEAN OPEN SCIENCE CLOUD

**iFAIR**
Identifying datasets in scientific publications. (2021)

**AR-ONTO-COVID**
A Knowledge-Based Resource for Covid-19 (2020)

# Pain points

- Data quality
- Data quality
- Data quality

– descriptions missing
– no statistics about the data
– wrong content (OCRs, classification symbols)
– …

Artificial Researcher

# Thank you for your attention

artificialresearcher.com

florina.piroi@artificialresearcher.com

linda.andersson@artificialresearcher.com

austria wirtschafts service aws

Bundesministerium Verkehr, Innovation und Technologie

vienna business agency A service offered by the City of Vienna

EUROPEAN OPEN SCIENCE CLOUD