



# Open access WebAPI registration and retrieval of ontology concepts

InChI Trust 2019, August 23



# ... a data scientists dream

Typical question: *Which tetrazoles have been published and what is their use?*



**uspto**

**WIPO**



**PubChem**



ZINC

**PHAROS**



GINAS  
Global  
Ingredient  
Archival  
System



Drug Central



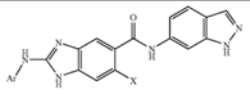


## Aggregating our knowledge on molecules

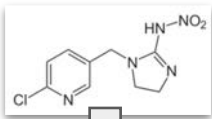
- OntoChem is extracting compounds and their properties & data from patents, scientific publications, books, web pages, databases - **text, images, tables**
- Semiotic data normalization - considering syntactic, semantic and pragmatic dimensions:
  - finding the **same**
  - registering the **new**
  - normalizing - adding a unique ontology concept id ( **OCID** )
- Make data available:
  - with FAIR principles - findable, accessible, interoperable and reusable

... helping to build an internet of molecules

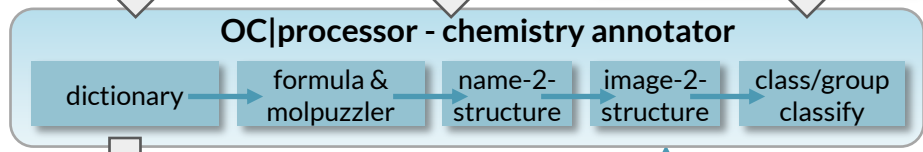
TABLE 4



Ex.	Ar	X	MS (m/z)
49	2-Isopropylphenyl	Morpholio-4-yl	496
50	2-Trifluoromethylphenyl	4-Methylpiperazin-1-yl	535
51	3,5-Dichlorophenyl	4-Methylpiperazin-1-yl	503
52	2,4-Dichlorophenyl	4-Methylpiperazin-1-yl	536
53	Thiazol-2-yl	4-Methylpiperazin-1-yl	474
54	2-Trifluoromethylphenyl	Morpholio-4-yl	522
55	3,5-Dichlorophenyl	Morpholio-4-yl	490
56	2,4-Dichlorophenyl	Morpholio-4-yl	522
57	2-Trifluoromethylphenyl	Piperidin-1-yl	520
58	3-Methylpiperidin-2-yl	4-Methylpiperazin-1-yl	482
59	Thiazol-2-yl	Morpholio-4-yl	461
60	3-Methylpiperidin-2-yl	Morpholio-4-yl	469
61	1-Isopropyl-1H-imidazol-2-yl	4-Methylpiperazin-1-yl	499
62	1-Cyclopropyl-1H-imidazol-2-yl	4-Methylpiperazin-1-yl	525
63	1-Isopropyl-1H-imidazol-2-yl	Morpholio-4-yl	486



# Compound Registration



... composition comprising a neonicotinoid such as **imidacloprid** <190000000809> and a ...

**190000000809**

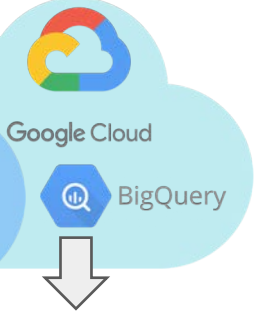
deliver OCID and data

**WebAPI**

is it known?

if no register

compound store

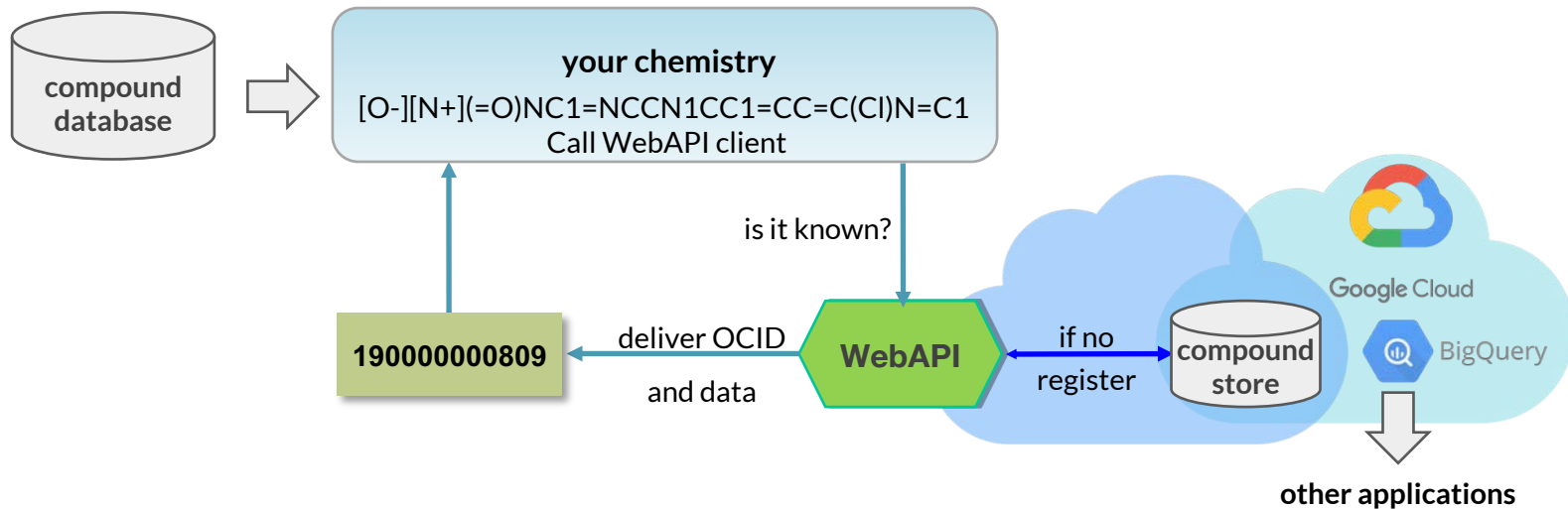


other applications



- provides chemistry registration system based on InChI
- unique, stable OCID
- substructure searchable (JChem SQL Database)
- synonyms & classification connected to OCID
- can be used for local compounds in DB's or documents as well

# Compound Registration API



Get the client & description at our public FTP [transfer.ontochem.com](https://transfer.ontochem.com)  
User: Mievoogh pwd: phae9Goo

# Classification

- Find known tetrazoles using **compound\_classes**, **compound\_parent**, **ontochem** tables:

921.428

The screenshot shows the Google Cloud Platform BigQuery interface. The query editor contains the following SQL code:

```

1 Select * from `sciwalker-open-data.chemistry_compounds.ontochem` where ocid in (
2 select ocid FROM `sciwalker-open-data.chemistry_compounds.compound_parents` where parentid in (
3     SELECT ocid FROM `sciwalker-open-data.ontologies.compound_name` where name='tetrazoles'
4 )
5 )
6
7

```

The query results table is displayed below, showing columns for `update_time` and `part_date`. The first few rows are:

update_time	part_date
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-14T16:03:32	2019-07-16 09:12:09.291 UTC
2019-05-31T08:05:32	2019-07-16 09:12:09.291 UTC
2019-05-17T08:05:48	2019-07-16 09:12:09.291 UTC

The interface also shows a sidebar with a resource tree for `sciwalker-open-data`, including tables like `ChEBI`, `ChEMBL`, `chemistry_compounds`, `chembl`, `chembl_25`, `compound_classes`, `compound_classes_pare...`, `compound_parents`, `drugcentral`, `dsstox`, `ontochem`, and `pubchem`. The bottom of the screen indicates the query processed 278.2 GB and returned 1 - 100 of 921428 rows.

# Sequence Registration



OC|processor - sequence extraction

DRVYIHPFGC

is it known?

17000002341

deliver OCID  
and data

WebAPI

if no  
register

sequence  
store

Google Cloud



BigQuery

other applications

- provides sequence registration system based on string
- unique, stable OCID
- BLAST searchable
- synonyms & classification connected to OCID



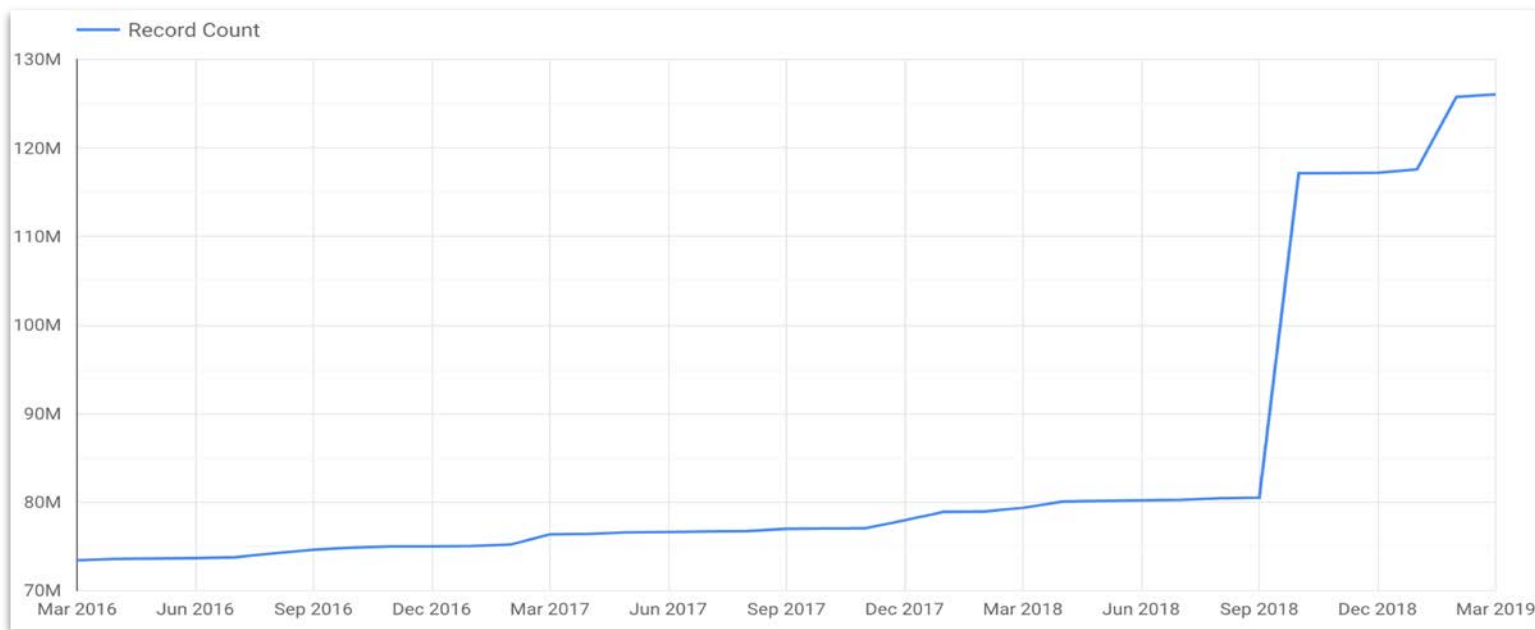
# “SciWalker Open Data” project in Google BigQuery

- **Ontologies - public OCID and hierarchy**
  - anatomy, biomarker, chemistry, clinicalTrials, compound\_classes, cosmetology, drugs, effects, herbal\_drugs, human\_genes, inorganic\_materials, institutions, magnitudes, methods, natural\_products, nutrition, proteins, regions, species, substances, toxicity
- **Molecules: about 130 million unique InChI compounds from OntoChem registration server, 74 million nucleotide and 11 million peptide sequences = 215 million unique molecules**
  - OntoChem, PubChem, ChEMBL, Zinc, DssTox
  - Nucleotide sequences, protein and peptide sequences
- **Genes & Proteins**
  - GWAS - Genome wide association studies, ClinVar
- **Clinical & Drug Data**
  - ClinicalTrials.gov, Drug Central, Drug labels



# Growth of registered unique compounds

About 50.000 new compounds per month from new patents and articles



- Query history
- Saved queries
- Job history
- Transfers
- Resources + ADD DATA
- Search for your tables and datasets
- sciwalker-open-data
  - ChEMBL
  - chemistry
  - clinical\_trials\_aact
  - clinical\_trials\_eudract
  - clinical\_trials\_who
  - cloud\_annotation
  - DrugCentral
  - exports
  - genome\_association\_studies
  - ontologies
  - pubchem
  - sequences
    - nucleotides
    - proteins
- bigquery-public-data
- biomarker
- patents-public-data

Unsaved query Edited HIDE EDITOR

```
1 SELECT * FROM `sciwalker-open-data.sequences.nucleotides` WHERE insert_time < TIMESTAMP("2019-03-31") LIMIT 1000
```

Run
Save query
Save view
More
This query will process 31.9 GB when run.

nucleotides 
[QUERY TABLE](#)
[COPY TABLE](#)
[DELETE TABLE](#)
[EXPORT](#)

This is a partitioned table. [Learn more](#)

Schema [Details](#) Preview

**Description** **Labels**   
 None None

**Table info**

Table ID	sciwalker-open-data.sequences.nucleotides
Table size	31.95 GB
Number of rows	74,664,941
Created	Mar 28, 2019, 2:52:53 PM
Table expiration	Never
Last modified	Mar 29, 2019, 3:00:06 PM
Data location	US
Table type	Partitioned
Partitioned by	Day
Partitioned on field	insert_time
Partition filter	Not required
Clustered by	ocid, seq_hash

# Finding the same

We need agreed criteria for what is the same, e.g.

Compounds

InChI

Reactions

RInChI

RNA, DNA, Protein Sequences

Sequence String

*(+defined hash?)*

Compositions

*MInChI*

Diseases...

*Normalized name + hierarchical criteria*

# Data on aspirin ?

SciWalker + BigQuery API

compound link-outs

SciWalker open data

Search Chemistry Knowledge Help

Hello Guest Enterprise Login

Match: Full Learn more  
 \* Similarity threshold: 0.9  
 Maximum number of hits: 10  
 Maximum time: No limit

Search in structure DB

Help for structure editor

(De)select: Select all Search in documents SMI file Export

9 structure(s) found.  
 (1 of 1)

#	Select	Structure	Names	OCID	More	Ext. links
1	<input checked="" type="checkbox"/>		11126-35-5 11126-37-7 156865-15-5 Ioxr 2-(Acetyl-Oxy)Benzoic Acid 2-(acetyloxy)-Benzoic acid 2-(acetyloxy)-Benzoicacid 2-(acetyloxy)benzoic acid 2-Acetoxybenzenecarboxylic acid 2-Acetoxybenzoate 2-Acetylsalicylic acid 2-Carboxyphenyl acetate	190000021540	Compound details Find this compound in documents	PubChem DrugCentral ChEMBL EPA DSSTox DrugBank Human Metabolome Database

# OCID

identifler.org EMBL-EBI

links to find  
information on a OCID

e.g. sitagliptin  
ocid:190000005275

<https://www.sciwalker.com/sciwalker/faces/ociddata.xhtml?ocid=190000005275>

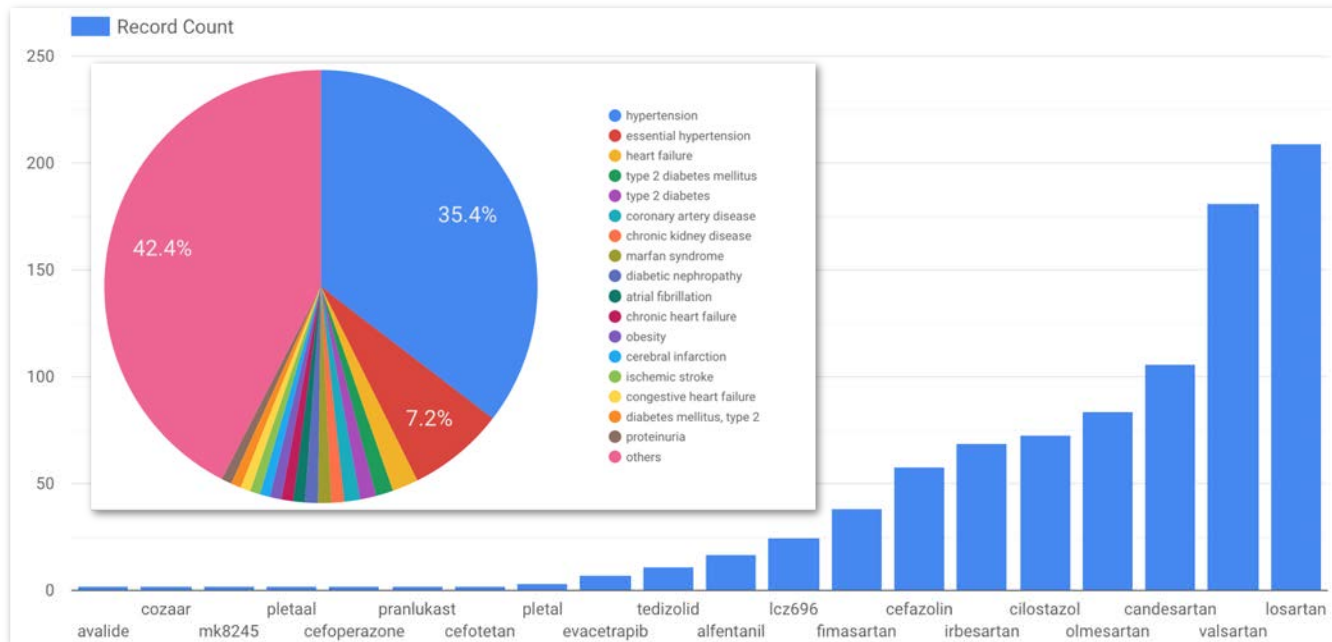
The screenshot shows a web browser window with the URL `sciwalker.com/sciwalker/faces/ociddata.xhtml?ocid=190000005275`. The page features a green navigation bar with the SciWalker logo and menu items like Search, Chemistry, Knowledge, and Help. Below the navigation bar, there are two search input fields: one for 'Annotated concept term' containing 'Sitagliptin' and another for 'Concept OCID'. Both fields have orange 'Search' buttons. The main content area is titled 'Concept info:' and contains a table with the following data:

OCID:	190000005275
Domain:	Chemistry
Main term:	<b>Sitagliptin</b>
Synonyms:	(3R)-3-Amino-1-[3-(trifluoromethyl)-5,6,7,8-tetrahydro-1,2,4-triazolo[4,3-a]pyrazin-7-yl]-4-(2,4,5-trifluorophenyl)butan-1-one (3R)-3-amino-1-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl]-4-(2,4,5-trifluorophenyl)butan-1-one (3R)-3-amino-1-[3-(trifluoromethyl)-5H,6H,7H,8H-[1,2,4]triazolo[4,3-a]pyrazin-7-yl]-4-(2,4,5-trifluorophenyl)butan-1-one (3R)-3-amino-1-[3-(trifluoromethyl)-6,8-dihydro-5H-[1,2,4]triazolo[3,4-c]pyrazin-7-yl]-4-(2,4,5-trifluorophenyl)butan-1-one (3R)-3-amino-1-[3-(trifluoromethyl)-6,8-dihydro-5H-[1,2,4]triazolo[4,3-a]pyrazin-7-yl]-4-(2,4,5-trifluorophenyl)butan-1-one (R)-4-oxo-4-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl]-1-(2,4,5-trifluorophenyl)butan-2-amine 1x70 486460-32-6 7-((3R)-3-amino-1-oxo-4-(2,4,5-trifluorophenyl)butyl)-5,6,7,8-tetrahydro-3-(trifluoromethyl)-1,2,4-Triazolo(4,3-a)pyrazine 790712-60-6 A25516 AB0006583 AB1004599 ABP000233 CID190000005275 ACT02665 AK172047 AMX10120 CHEMBL1422 CID4369359

# Answering pharma related questions

**Q: What tetrazole containing drug candidates are in how many clinical trials for which diseases ?**

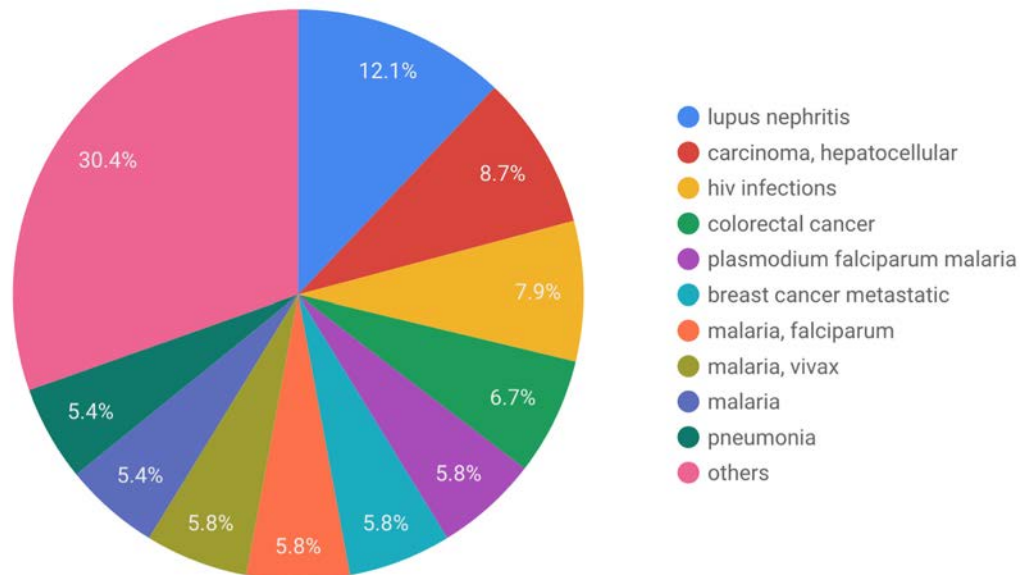
**A: from 0.92 million published tetrazoles, 35 compounds were in 926 trials:**



# Answering pharma related questions

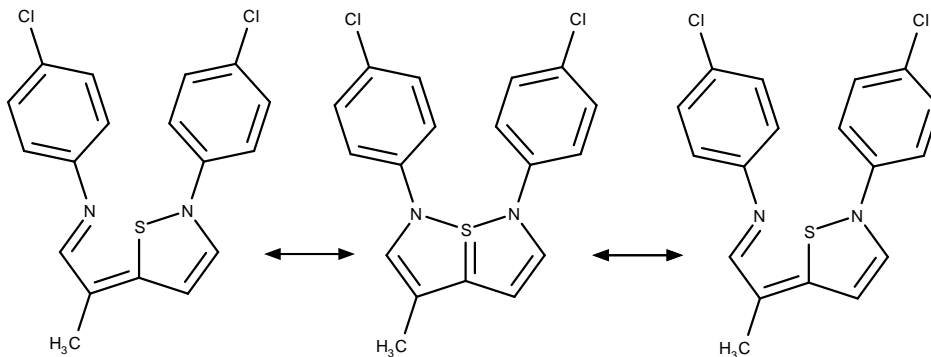
## Q: Which diseases are in trials with sesquiterpenes ?

inputs are chemistry & disease ontology + normalized clinical trials from NIH:



## Next Steps ?

- RInChI: open access publication of published chemical reactions from text documents
- MInChI: published compositions, e.g. tablets or alloys
- Polymers, Tautomers and valence tautomers:  
Weber, L.; Schulze, B.; Szargan, R.; Mühlstädt, M. "Nitrogen-15 NMR, 2D NMR and ESCA Characterization of a New Stable 6a-Thia(SIV)-1,6-diazapentalene" Magn. Reson. Chem. 1990, 28, 419-22.







# Thanks to

Stephen Boyer, [Collabra](#)  
Ian Wetherbee, [Google](#)  
Team, [OntoChem](#)

