

New Edition! Getting CAS registry numbers out of WikiData

Egon Willighagen 

Published December 22, 2015

Citation

Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. <https://doi.org/10.59350/vt6nr-28g72>

Keywords

Cas, Wikidata, Chemistry

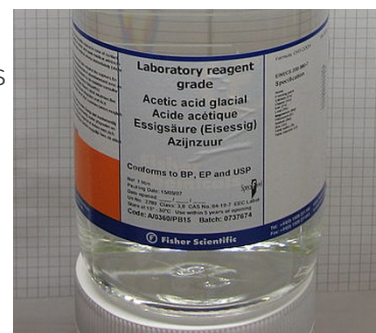
Copyright

Copyright © Egon Willighagen 2015. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

April this year I blogged about an important SPARQL query for many chemists: getting CAS registry numbers from Wikidata. This is relevant for two reasons:

1. CAS works together with Wikimedia on a large, free CAS-to-structure database
2. Wikidata is CCZero

The original effort validated about eight thousand registry numbers, made available via Wikipedia and the Common Chemistry website. However, the effort did not stop there, and Wikipedia now contains many more CAS registry numbers. In fact, Wikidata picked up many of these and now lists almost twenty thousand CAS numbers. That well exceeds what databases are allowed to aggregate and make available.



Names
IUPAC name
Acetic acid ^{[3][4]}
Systematic IUPAC name
Ethanoic acid ^[5]
Other names
Vinegar (when dilute); Hydrogen acetate; Methanecarboxylic acid ^{[1][2]}
Identifiers
CAS Number
64-19-7 ✓

Source: Wikipedia. CC-BY-SA

Since the post in April, Wikidata put online a new SPARQL end point and created "direct" property links. This way, you lose the provenance information, but the query becomes simpler:

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT ?compound ?id WHERE {
  ?compound wdt:P231 ?id .
}
```

The other thing that changed since April is that others and I requested the creation of more compound identifiers, and here's an overview along with the current number of such identifiers in Wikidata:

- CAS registry number (P231): 19420
- PubChem ID (CID) (P662): 16616
- InChI (P234): 14312
- ChemSpider ID (P661): 11566
- ChEBI ID (P683): 4313
- KEGG ID (P665): 3983
- Drugbank ID (P715): 2518
- KnapSack ID (P2064): 9
- HMDB ID (P2057): 6
- ZINC ID (P2084): 4
- LIPID MAPS ID (P2063): 3
- Leadscope ID (P2083): 3

Clearly, some identifiers are not well populated yet. This is what bots are for, like those used by the Andrew Su team.

Because there is also a predicate for SMILES, we can also create a query that puts the CAS registry number alongside to the SMILES (or any other identifier):

chem-bla-ics

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT ?compound ?id ?smiles WHERE {
  ?compound wdt:P231 ?id ;
             wdt:P233 ?smiles .
}
```

Of course, then the question is, [are these SMILES string valid...](#) And, importantly, this is nothing compared to the number of chemical compounds we know about, which currently is in the order of 100 million, of which a quarter can be readily purchased:

Chemistry

4

Biology

John Irwin Chemistry
@chem4biology · [Follow](#)

X

refreshed! [@pubchem](#) now contains 25,758,525 purchasable molecules from ZINC15 [#docking](#) [#chemoinformatics](#)

3:54 PM · Dec 22, 2015



 [Reply](#)

 [Copy link](#)

[Read 1 reply](#)

Chemistry 4 Biology **John Irwin Chemistry**
@chem4biology · [Follow](#)



PubChem compound 100 million comes from ZINC!
pubchem.ncbi.nlm.nih.gov/compound/10000... ->
zinc15.docking.org/substances/ZIN... #zinc15 #win
#watchoutcas #woot



pubchem.ncbi.nlm.nih.gov
N-[(2S)-2-(1-benzofuran-2-yl)-2-hydroxypropyl]-3,5-
dimethyl-1,2-oxazole-4-sulfonamide

9:57 PM · Dec 17, 2015



1



Reply



Copy link

[Read more on X](#)