the WINNOWER

CHEMISTRY

New Edition! Getting CAS registry numbers out of WikiData

EGON WILLIGHAGEN [1]
1. *Maastricht University*

CORRESPONDENCE:
egon.willighagen@gmail.com

DATE RECEIVED:
January 08, 2016

DOI:
10.15200/winn.145228.82018

ARCHIVED:
January 08, 2016

KEYWORDS:
CAS registry number, wikidata

CITATION:
Egon Willighagen, New Edition! Getting CAS registry numbers out of WikiData, *The Winnower* 3:e145228.82018 , 2016 , DOI: 10.15200/winn.145228.82018

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.

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 loose 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):

```
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:

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

— John Irwin Chemistry (@chem4biology) December 22, 2015

PubChem compound 100 million comes from ZINC! https://t.co/lCoYZ7P34e -> https://t.co/vJ1qX5cUNB #zinc15 #win #watchoutcas #woot

— John Irwin Chemistry (@chem4biology) December 17, 2015

Willighagen, E., 2015. Getting CAS registry numbers out of WikiData. The Winnower.
URL http://dx.doi.org/10.15200/winn.142867.72538