chem-bla-ics

Adding disclosures to Wikidata with Bioclipse

Egon Willighagen

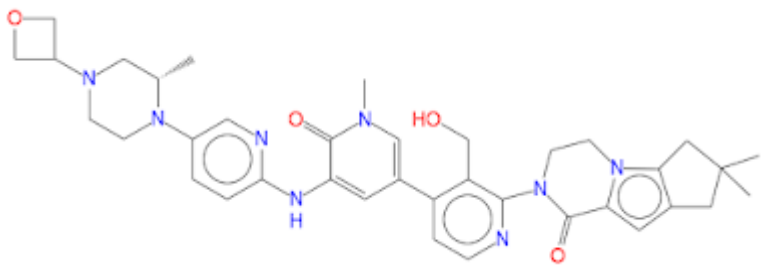
Published March 20, 2016

Citation

Willighagen, E. (2016, March 20). Adding disclosures to Wikidata with Bioclipse. *Chem-bla-ics*. https://doi.org/10.59350/k8jnz-7fb76

Keywords

Acs, Bioclipse, Chembl, Inchi, Pubchem



chem-bla-ics

Last week the huge, bi-annual ACS meeting took place (#ACSSanDiego), during which commonly new drug (leads) are disclosed. This time too, like this one tweeted by Bethany Halford:

Because getting this information out in the open is important, I think it's a good idea to add them to Wikidata (see doi:10.3897/rio.1.e7573). So, with Bioclipse (doi:10.1186/1471-2105-8-59) I redrew the structure:

I previously blogged about how to add chemicals to Wikidata , but I realized that I wanted to also use Bioclipse to automate this process a bit. So, I wrote this script to generated the SMILES, InChI, InChIKey, double check the compound is not already in Wikidata (using the Wikidata SPARQL endpoint), an look up the PubChem compound identifier (example SMILES).

```
smiles = "CCCC"

mol = cdk.fromSMILES(smiles)
ui.open(mol)

inchiObj = inchi.generate(mol)
inchiShort = inchiObj.value.substring(6)
key = inchiObj.key // key = "GDGXJFJBRMKYDL-FYWRMAATSA-N"

sparql = """
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT ?compound WHERE {
  ?compound wdt:P235 "$key" .
}
"""

if (bioclipse.isOnline()) {
  results = rdf.sparqlRemote(
    "https://query.wikidata.org/sparql", sparql
  )
  missing = results.rowCount == 0
} else {
  missing = true
}

formula = cdk.molecularFormula(mol)

// Create the Wikidata QuickStatement,
// see https://tools.wmflabs.org/wikidata-todo/quick_statements.php

item = "LAST" // set to Qxxxx if you need to append info,
```

```
              // e.g. item = "Q22579236"

pubchemLine = ""
if (bioclipse.isOnline()) {
  pcResults = pubchem.search(key)
  if (pcResults.size == 1) {
    cid = pcResults[0]
    pubchemLine = "$item\tP662\t\"$cid\""
  }
}

if (!missing) {
  println "===================="
  println "Already in Wikidata as " + results.get(1,"compound")
  println "===================="
} else {
  statement = """
    CREATE

    $item\tDen\t\"chemical compound\"
    $item\tP233\t\"$smiles\"
    $item\tP274\t\"$formula\"
    $item\tP234\t\"$inchiShort\"
    $item\tP235\t\"$key\"
    $pubchemLine
  """

  println "===================="
  println statement
  println "===================="
}
```

The output of this script is a QuickStatement for Magnus Manske's tool (IMPORTANT: it's not meant to automate editing Wikidata! I only automate creating the input, which I carefully check (e.g. checking all stereochemistry is defined)! Note, how Bioclipse opens up the structure in a viewer with ui.open()), which is a list of commands to create and edit entries in Wikidata. You need to enable it first, but if you have an account, this is not too hard. Of course, the advantage is that it is a lot quicker. I have similar script to create QuickStatements starting with only a ChEMBL identifier.

The QuickStatement for GDC-0853 looks like:

```
    CREATE

    LAST Den "chemical compound"
```

chem-bla-ics

```
    LAST P233 "O=C1C(=CC(=CN1C)c2ccnc(c2CO)N4C(=O)c3cc5c(n3CC4)CC(C)
(C)C5)Nc6ncc(cc6)N7CCN(C[C@@H]7C)C8COC8"
    LAST P274 "C37H44N8O4"
    LAST P234 "1S/C37H44N8O4/
c1-23-18-42(27-21-49-22-27)9-10-43(23)26-5-6-33(39-17-26)40-30-13-25(19-41(4)35(30)47)28-
h5-8,13-14,17,19,23,27,46H,9-12,15-16,18,20-22H2,1-4H3,(H,39,40)/t23-/m0/s1"
    LAST P235 "WNEODWDFDXWOLU-QHCPKHFHSA-N"
    LAST P662 "86567195"
```

The first line creates a new Wikidata item, while the next ones add information about this compound. GDC-0853 is now also Q23304817. The label I added manually afterwards. Note how the Bioclipse script found the PubChem identifier, using the InChIKey. I also use this approach to add compounds to Wikidata that we have in WikiPathways.