chem-bla-ics

The Chemical Object Identifier; or, the freedom to identify chemicals

Egon Willighagen

Published March 9, 2008

Citation

Willighagen, E. (2008). The Chemical Object Identifier; or, the freedom to identify chemicals. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/ed8nt-n6a25

Keywords

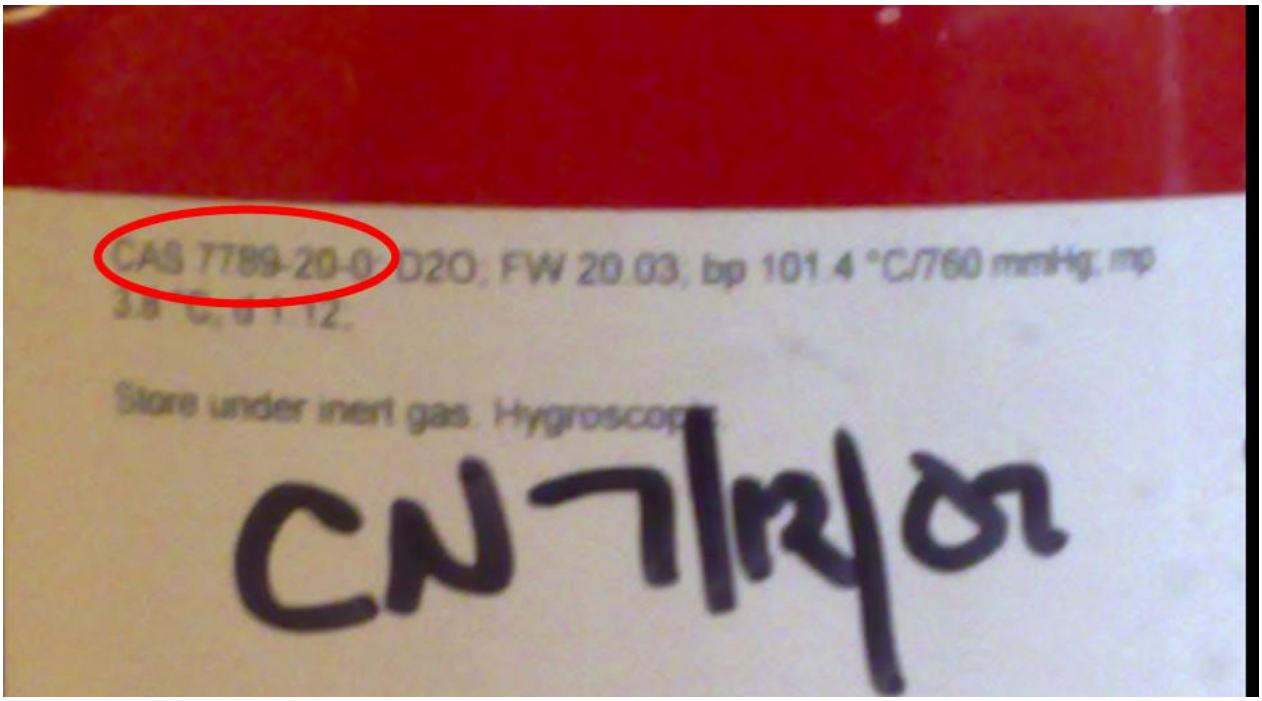
Cas, Cheminf

chem-bla-ics

IUPAC chemical names, SMILES and InChIs are too long. InChIKeys are not unique enough because of safety reasons (*you have a 1 in 10 billion chance of blowing up your building*; well, odds are actually much, much lower than getting hit by Osama or friends, let alone a car). Wikipedia URIs do not cover enough chemical space.

However, we need short identifier. Why, actually? Computers don't care about long identifiers. Systems can be integrated. A web link is easy to make. But we do. A bottle on the shelf does not have a HTML interface. And you do not have a scanner to read the chemical structure from a 2D barcode (see DOI:10.1021/ci049758i).

The CAS registry number has serviced this purpose for a long time. For example, as used on bottles visible in this picture (copyright: CC BY-SA, Science in the Open):



Now, when Anthony reported that CAS, the organization that builds the proprietary lookup service, which has done an amazing job in the past, that they do not wish to see CAS numbers in Wikipedia curated by means of the official database - it violates the *end user agreement* one has to sign before one can use the database - the blogging community reacted (here, here, here, here and here).

Personally, I agree with the CAS standpoint. It's been a proprietary database which people have been supporting financially for years, and thoughtfully signed the license agreement. So, don't complain afterwards. If you *really* want to, **end the agreement and object against the license**. I commented in the original blog:

> *In 1995 I started a Dutch website on organic chemistry [1] and the CAS number was as useful as it is now, and already then we knew we were not allowed to compose a database of CAS numbers. Not sure about the legal state of that, but our university had a license; not sure if students had access, but do not believe so. Anyway, building a substantial list of CAS number*

> *was not allowed. So, we looked for other means of identifying molecular structures, which led us to CML... this was around '96-'97 or so, at least before XML was released, and we started using CML actually when it was still in a more obscure SGML format :) Yeah, the XML recommendation was much appreciated!*
>
> *OK, so back to your blog item. You can imagine that the comment in WP by CAS does not surprise me at all; nothing really new. If they would allow this, it would set a precedence...*
>
> *The solution is, however, fairly easy. Use InChI(Key), PubChem CID, or ChemSpider CID; the latter two are on the same level as CAS numbers. CAS registry numbers are overrated. Not sure if they still hand out CAS numbers to mixture too... (I guess not).*
>
> *Oh, and I agree with Cpt. Renault... people should really abide to legal requirements. Period. If you don't like them, quit the legal agreement. As simple as that.*
>
> *1.http://www.woc.science.ru.nl/*

Here, I tend to disagree with Will who wrote that "*They are just numbers. i.e. descriptors*". The CAS number only makes sense with a (curated) look up table; making it tightly linked to the CAS database. While theoretically you may be allowed to copy numbers from that database, the license agreement strictly disagrees with that. Court would have to decide which right takes higher importance, but my vote is on the agreement, which you thoughtfully signed. So, I tend to agree with Joerg who wrote that CAS number are not public domain, are they?

An interesting bit in that blog item is the comment he left himself:

> *I just realized that Peter has also commented on it. And storing 10000 CAS numbers and structures is allowed? What happens, if a journal reaches this limit? Just imagine they publish 1000 papers with 100 CAS numbers for each article? I do not get this!*

Interesting indeed. This gets me back to a recent question I was confronted: *How would I use chemical literature in the current age?* Well, what about this hypothetical Taverna workflow:

- Node 1: get me a list of journals expected to contains CAS registry numbers (such as the JCIM)
- Node 2: for each, get me all publications of the last 25 years
- Node 3: process all articles and count cited CAS registry numbers per journal
- Node 4: complain if count_per_journal > 10000

Anyway. Common agreement seems to be that we can opt to do without the CAS registry number. The PubChem ID seems a reasonable candidate, and has been suggested here and here. The ChemSpider ID could be an option too, though ChemSpider content is periodically added to PubChem.

I'd also like to bring in the suggestion of having a *Chemical Object Identifier*: like the DOI, the COI is a simple alpha-numerical identifier, with a one-to-one connection to the InChI, and unlike the InChIKey unique as the InChI itself, but requiring a look up service. And the latter I

can offer: http://rdf.openmolecules.net/. It's a free (as in Open) resource, where we can provide this lookup service. It would be really easy to create a new COI when a InChI is passed it did not assign a COI yet. A PHP page to do the reverse lookup is easy too. Interested? I can have it going by the end of the month. It comes with full RDF support, so ready for the Web-NG.