One Million IUPAC names #4: a lot is happening

Published August 9, 2025

Citation

V2lsbGlnaGFnZW4sIEUuICgyMDI1LCBBdWd1c3QgOSkuIE9uZSBNaWxsaW9uIElVUEFDIG5hbWVz
ICM0OiBhIGxvdCBpcyBoYXBwZW5pbmcuIDxpPkNoZW0tYmxhLWljczwvaT4uIGh0dHBzOi8vZG9p
Lm9yZy8xMC41OTM1MC9rcnc5bi1kdjQxNw==

Keywords
Iupac, Beilstein, Chembl

Abstract
A lot is happening. If you have been following this project more closesly, you may have already seen some interesting updates, but I will post it here too. First, a quick recap. In March I started a new Blue Obelisk project to collect CCZero IUPAC names from primary literature (paper still pending). It turned out we can automate that, while legally not violating any laws or licenses.

Copyright
Copyright © None 2025. 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.

A lot is happening. If you have been following this project more closesly, you may have already seen some interesting updates, but I will post it here too. First, a quick recap. In March I started a new Blue Obelisk project to collect CCZero IUPAC names from primary literature (paper still pending). It turned out we can automate that, while legally not violating any laws or licenses. In April I reported on some tweaks boosting the efficiency of the use of the API. I also reported on some possible further steps, including how to use the extracted names to create a larger set. Indeed, in June I could report to have passed the 200k IUPAC names, which with the idea from April gave us more than 1M IUPAC names.

In this post I want to give an update.

275k IUPAC names

I have continued running the scripts to detect new IUPAC names in full text, open access papers in Europe PMC, but something more awesome actually did much more since the June post: in July I received a pull request from mnietfeld with more than 40 thousand unique and new IUPAC names from the Beilstein Journal of Organic Chemistry (see also their LinkedIn post or this archived version that doesn't require an account). While Europe PMC provides these articles too (and actually one of the first I analyzed), a lot of these names come from supplementary information, not provided by Europe PMC. Thanks!

This is focusing on names from primary literature, but there is more happening. Because I want to restrict the above project to names from primary literature (and supplementary information is still that), I have not been sure what to do with other collections yet, and they have been coming in. I have been taking notes in the project issue tracker, for future reference (like now, here). I have not forgotten about these!

Other large collections of IUPAC names

4M, CCZero

Let's start with the news yesterday. The Chemical Biology Services team released 4 million IUPAC names from patent literature as CCZero! The CCZero license/waiver makes it compatible with our list. Their Zenodo release:

> *... contains IUPAC names text-mined from patents (US, WIPO, EPO, Chinese, Japanese).*

The post also includes a nice example of the complexity of IUPAC names which makes the counting of unique names tricky: `O-methylphenol` and `o-methylphenol`. Thanks, Noel and the rest of the EMBL-EBI team!

2.3 million, CC-BY

And then Haydn Jones was one of the earliest to coin in, and released 2.3 million IUPAC names under the CC-BY license.

chem-bla-ics

850k, CCZero

Wikidata also turnes out to have many IUPAC names. Adriano found more than 850 thousand IUPAC names, see this project.

Next week I will do some comparisons of the datasets with a clear Creative Commons license.

Even more

Beyond these five data releases, there is more. PubChem and other databses have millions of names, but often these are generated by proprietary software. These IUPAC name collections may be under some license agreement, and thus not compatible with Open Science. This is why it is so important that we very clearly know where these names are coming from.

5-6 million, license unclear

I also learned about ChemPile about which Adrian Mirza explained me it has about 5-6 million IUPAC names. But the source of this list of names is not yet clear to me.

Names from PhD theses and preprints

I also want to give a shout out to Peter Murray-Rusts proposal to start extracting IUPAC names from PhD theses. There have been projects to extract chemistry from PhD thesis in the past, and this will yield a lot of unique names. Please ping Peter, if you want to get involved in his idea!

What's next

I am so excited with all these efforts and very grateful with the contribution by Beilstein. I really hope more Open Science publishers will follow, like perhaps the Royal Society of Chemistry for which it should be easy, with their Project Prospect background!

I am also excited by the release by ChEMBL under CCZero. That will allow the WikiProject Chemistry use this for Wikidata!

So, I have one week left to write the article about the work we started in March. The outlook is bright. I played last week with the Europe PMC full text downloads and can confirm that should yield thousands of additional names from the full texts. A single download file gave me more than two thousand new unique names. I think the 500k IUPAC names is absolutely in reach with purely the full texts from Europe PMC.

This brings us to the end of 2025. By then, we should have a many millions of openly-licensed IUPAC names. And by March 2026, I hope we reached the 1M IUPAC names extracted from primary literature. That will require some creativity and enthusiasm, but sounds feasible!