chem-bla-ics

One Million IUPAC names #5: a new approach and 400k names

Egon Willighagen

Published May 2, 2026

Citation
Willighagen, E. (2026). One Million IUPAC names #5: a new approach and 400k names. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/gqtbx-jta57

Keywords
Iupac, Textmining, Xml, Europepmc

About fifteen months ago a new project started: One Million IUPAC names:

> *Thus, the idea came up, can we create a set of 1 million unique IUPAC names found in literature?*

We started out with using Europe PMC to get JATS XML files for the full texts of open access articles. Parsing the XML is easy and the text paragraphs are passed through OSCAR and OPSIN. That has not changed.

What did change last weekend is something I had long on my todo list (but life interfered). The first approach was to ask for named entities using the Europe PMC APIs. But I quickly realized that with OSCAR and OPSIN we could get more names out of the articles. The next step was to move from Google Colab to a command line script. That gave another boost, as explained in this second post in the series. We reached 200 thousand names in june 2025 but then things slowed down again in the growth. Two months later we only had 75 thousand more. However, plenty of discussion was happening and there turned out to be other, larger collections of IUPAC names under an open license. Millions of names, actually.

But another problem emerged. We were still using the Europe PMC API and were basically asking for open access articles between two dates. Practically, the API could answer requests between 1 and max 3 days. Beyond that, times outs and 404s became an issue. Moreover, because these dates are publications dates and not the dates on which the JATS were deposited, I had to got back to previous months and redo the queries. That gave another 5 thousand names since last August. Something had to change.

The new Approach

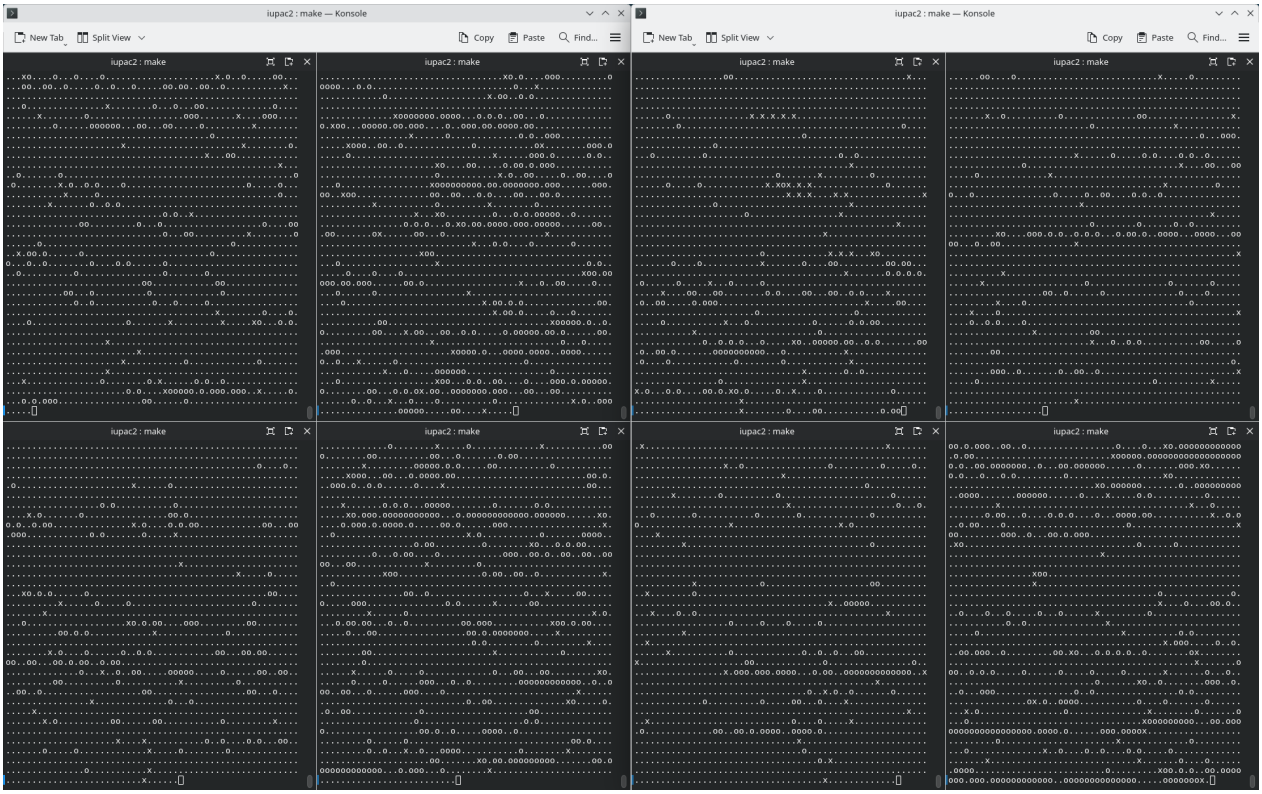
Europe PMC, however, also provides the JATS XML files as download on their FTP site. Already that august 2025 I had a prototype and knew it would change the game. These gzipped XML files are about 150 to 250 MB. Unzipped, about 1 GB each. Better, these files are based on Europe PMC identifiers, hopefully resolving the issue with using dates in the queries.

Now, parsing a 1 GB XML files is a total non-issue. I have done it plenty of times before. Just use a Simple API for XML (SAX) parser. This is a streaming parser giving you full control of how to parse things. It is ideal for this siutation: you just keep the current paragraph of text in memory and release that when done with that paragraph. That is, you do not have to read the full file in memory, just the bits you are interested in. I used this for my Chemical Markup Language patches for Jmol and JChemPaint back in the nineties.

Last weekend I finally made the jump. Use SAX to extract the **<p>** elements one by one, running OSCAR on them, filter with OPSIN, output that name, and clear the memory. Effectively, each gzipped file processes with a Groovy script in about 1 to 2 hours.

chem-bla-ics

The output is a mesmerizing stream of scientific literature (which I will use until someone points me to a Java CLI library that creates a Matrix-style falling letters equivalent), tho less so as a static image:



In this plot, an **x** means a new article to be processed. Each **.** and **o** that follows is a single <p> element and the difference is that an **o** means at least one IUPAC name was detected in the paragraph.

Each gzipped file gives 400 to 500 new IUPAC names. Indeed, going from 288 thousand to 300 thousand was a matter of a day and a half. And earlier this afternoon we passed the 400 thousand IUPAC names. With about 230 gzipped files. Now, I am going back in time, and the sizes of these files are shrinking: Another 500 files and the size has dropped to around 125 MB, so a rough estimate suggests that we will end up with 650 to 700 thousand names this way. This will be completed in a few weeks (and mostly because I need to focus first on other things again, because I can use our computing cluster do this).

Regarding the original goal, fortunately, we are still publishing at a higher rate every year, and more and more articles are available as open access. So, I still have good hopes we will reach the *1 million IUPAC names*. Also, keep in mind, we know how to boost this by simple name variations to several millions, even with the 400 thousand we have today.

Oh, and our next milestone will be in the pocket before I visit Christoph Steinbeck's cheminformatics team in Jena!