chem-bla-ics

New preprint: “Scholia Chemistry: access to chemistry in Wikidata”

Published May 25, 2025

Citation
Willighagen, E. (2025, May 25). New preprint: “Scholia Chemistry: access to chemistry in Wikidata”. *Chem-bla-ics*. https://doi.org/10.59350/zm558-pd424

Keywords
Wikidata, Scholia, Chemistry, Iccs

chem-bla-ics

Two week ago I uploaded a paper that has been in the works for some time. In fact, I first mention it as conference paper for the special issue of the 11th International Conference on Chemical Structures, you know, the meeting held in 2018, of which the 13th edition starts in 7 days. I had a poster at that conference which I described in this blog post.

In turn, that poster described work of at least three years, going back to adding identifiers in 2015 and chemical structures in early 2016. I started using scripts two months later. This helped a lot with migrating pKa data from a custom Semantic MediaWiki installation to Wikidata and with adding thousands of EPA CompTox identifiers in 2017.

But that 2018 conference paper never happened. Because Scholia did. And even on the ICCS poster, Scholia was used to visualize chemistry data in Wikidata. To be honest, not just that, of course. About a year ago I had a serious go at finishing the paper, and it was sent around to co-authors. But I realized at the time, that the paper was lacking some good suggestions how the peer review our actual contributions to Wikidata. I could hardly expect readers of the paper browse the individual histories of all, by then, 1.3 million chemical compounds. And during the holidays I collected a few tools, which I had lined up to add to the manuscript.

However, another thing happened, the COVID-19 pandemic. While all the experience helped a lot with getting knowledge together around SARS-CoV-2, it also made something else clear: the software behind Wikidata does not scale well (enough). This lead to plans to split the RDF graph representation into two separate SPARQL endpoints. And that breaks many, if not most, of Scholia's SPARQL queries, including those for the chemistry aspects. The situation in Summer 2024 was that there was a significant chance Scholia would not survive the split. And the *Scholia Chemistry* paper had to wait. You cannot publish an article of which the website is gone before it is formally accepted.

Let me make clear, this graph split is not solved and the risk is not gone. But a serious of unfunded, weekend hackathons allows us to refactor Scholia to give us a chance. It started with making Scholia more configurable. We had the first hackathons in October and November, and I had four more hackathon weekends this April.

The graph split into a main graph and a scholarly graph happened on May 9. Currently, we have been granted extra time and can use a legacy server with the full graph, but a lot less hardware, so slower. A final patch, merged in last week, allows us to define which SPARQL endpoint a query should run. So, each time we port a SPARQL query, we can directly update Scholia, making the migration somewhat more manageable.

But, with those uncertainties out of the way, it was time to finish the Scholia Chemistry paper!

The preprint (doi:10.26434/chemrxiv-2025-53n0w) brings 10 years of research together, and describes details of the used methods not formally peer-reviewed before. We describe in detail how chemical structures are added, the choices of Wikidata on how to represent chemical structures, how we curate the quality, and how we visualize chemical structures and data with

Scholia. As you can expect, the Chemistry Development Kit has an important role, along with the InChI.

The paper introduces three new Scholia *aspects* for chemicals, chemical classes, and elements. Each aspect is a template for a page with information about molecular entities and chemical substances, compound classes (like *fatty acids*), and elements (like carbon). Each template provides relevant information. Of course, any compound, class, or element can also still be opened in the Scholia "topic" aspect, listing relevant literature.

With this paper we aim to show that Wikidata is a innovative platform that meets the needs for a chemical structure database, with detailed data provenance, and scalable community curation.

I welcome your strongest peer review on the preprint. I don't liking settling for anything less. Here's the abstract:

> *Sharing knowledge on chemicals in the digital age has been the playground of databases such as the Chemical Abstract Services and PubChem. Wikipedia complements this field by providing context to chemicals aimed at a broad audience, but is not easily read by machines. Wikidata was started as a database service to improve the machine readability of the knowledge captured in Wikipedia. Wikidata has an open license, application programming interfaces, and a strong provenance model. Scholia uses the features to provide access to chemical knowledge. This study reviews the chemistry in Wikidata, shows how thousands of new chemicals were added, extends Wikidata with new properties for chemical representation and external links to additional databases, and shows how we extended Scholia to represent the chemistry in Wikidata.*

Thanks to Finn, Denise, Daniel, and Adriano!