New paper: “Wikidata subsetting: approaches, tools, and evaluation”

Egon Willighagen

Published February 13, 2024

Citation

Willighagen, E. (2024). New paper: “Wikidata subsetting: approaches, tools, and evaluation”. *Chem-bla-ics*. https://doi.org/10.59350/57rv7-5m756

Keywords

Wikidata, Scholia

Requires Full Dump	Live Subsetting	Supports Massive Data	Supports Qualifiers	Supports References	Graph Traversal	Further Output Transforms	Analytics
-	+	-	+	+	+	-	-
-	-	+	+	+	-	-	-
+	-	+	+	+	-	-	+
+	-	+	WIP	WIP	+	-	-
+	-	+	+	+	-	-	-
+	-	+	+	WIP	+	+	+
-	+	-	+	+	-	+	-

Copyright

Just before the end of the year, the *Wikidata subsetting: approaches, tools, and evaluation* paper by Seyed Amir Hosseini Beghaeiraveri *et al.* got published (doi:10.3233/SW-233491). I am really excited our group (i.e. Ammar and Denise) has been able to contribute to this. I think it also is a great example of the power of hackathons to bring together people.

To me, subsetting of Wikidata (or any large knowledge graph) is important for a couple of reasons. First, there can be practical reasons. Scholia, for example, is computationally expensive, and the idea we explore in the Alfred P. Sloan Foundation grant for Scholia (doi: 10.3897/rio.5.e35820) was that a subset of Wikidata would make it more performant and potentially more environmental-friendly.

A second reason is more about the scientific process. When doing an analysis and when you want to make the reasoning transparent, you want to share the analyzed data as part of the research output (basically, the "data"). For example, the data may have undergone some curation, or you combined data from two or more different sources. And you will want to share this as part of the scientific process. Resharing a full dump of the larger knowledge base would not be practical for at least two reasons: duplication of huge data, and a lot of unrelated content makes it hard for peers to find the bits of interest to the study.

Subsetting may be useful here. This paper evaluates a number of different subsetting approaches. Myself, I am particularly excited about the idea that we can take a shape expression (e.g. ShEx) as input. I still love the idea that I take the SPARQL queries in my analyses, convert that into shapes automatically, and then get a subet that returns the exact same results as the query would on the full dataset.