



Scientific articles and beyond in Wikidata

The big picture

Lydia Pintscher
@nightrose
[Q18016466](#)

With data analysis by Aisha Khatun and Manuel Merz



{ } wikicite

Where are we?

The good, the bad and the ugly



The Good

The vision side

- Knowledge integrity inside and outside the Wikimedia projects is crucial and only becoming more important
- A very dedicated community has come together
- Great tooling has been built (e.g. Scholia and Author Disambiguator) and innovation for Wikidata as a whole pushed forward



The Bad

The technical side

- Wikidata Query Service is reaching its limits, limiting maintenance and reuse
- Dumps are unusable for the majority of reusers we care about



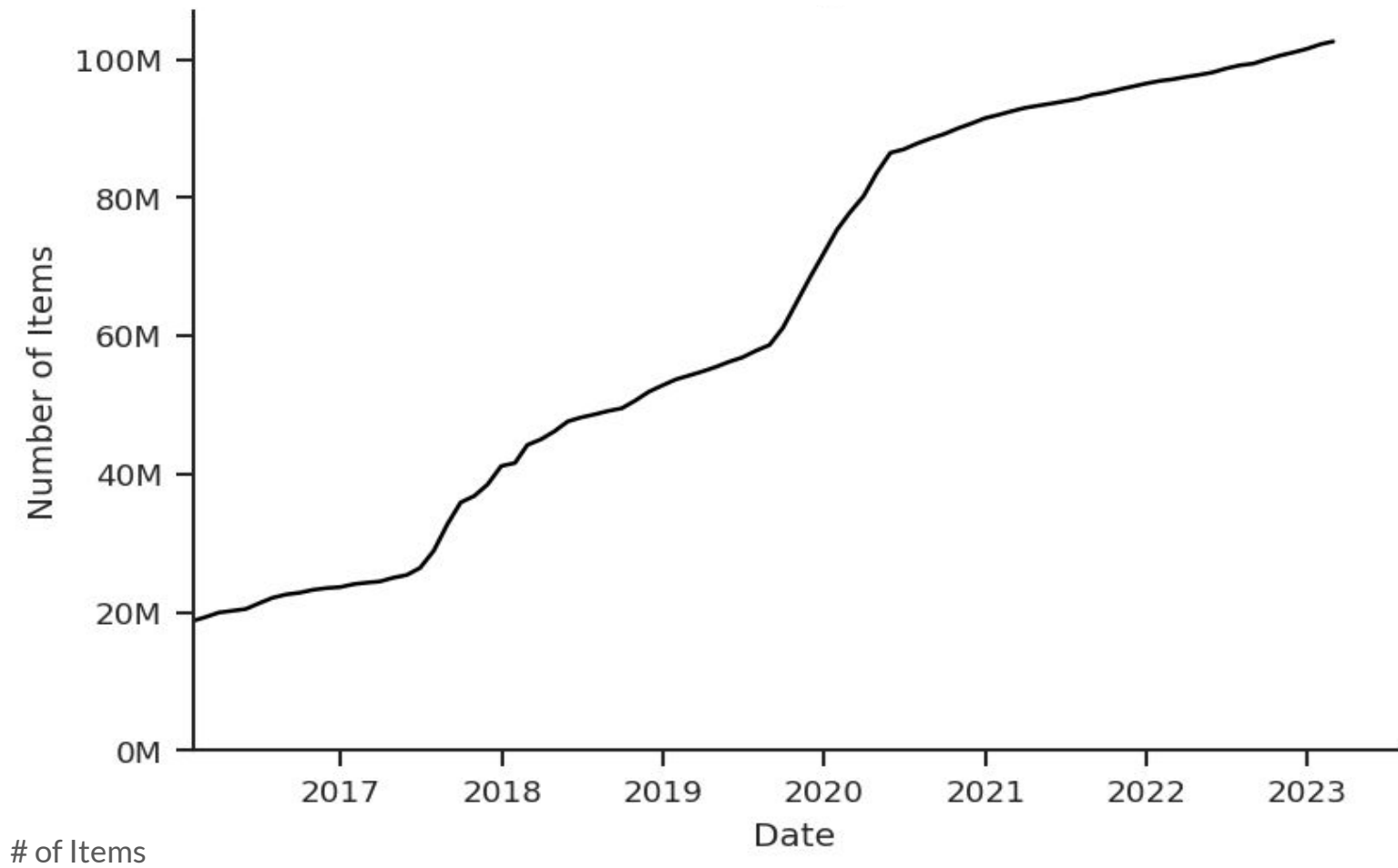
The Ugly

The social side

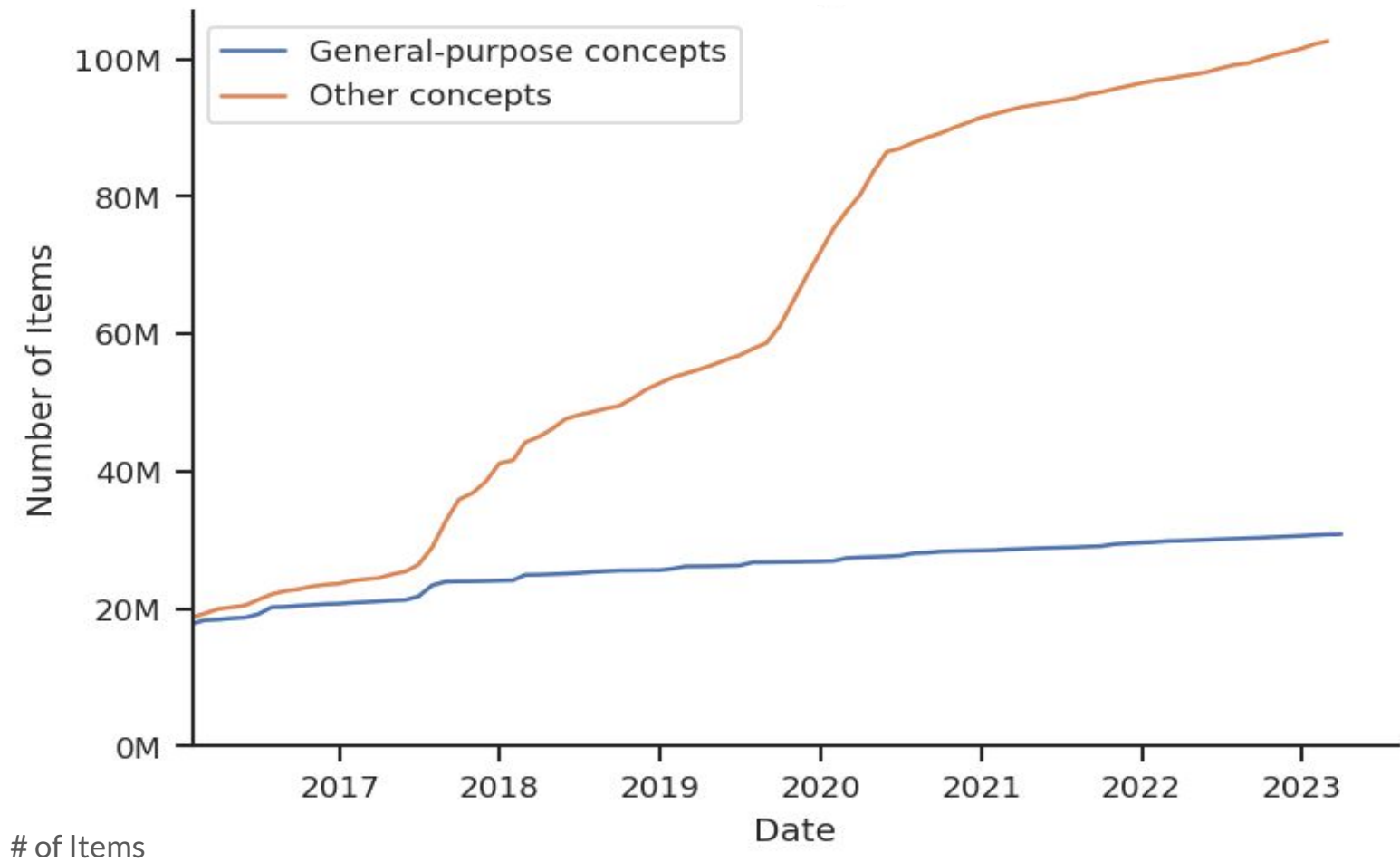
Scientific paper data

- makes up an outsized part of Wikidata, shifting it from a general-purpose knowledge graph to one about scientific papers
- is highly underused (e.g. ~2% of queries, > 60% of which is from Wikidata Integrator)
- is heavily biased towards content that is available in machine-readable form already
- is different from the general purpose Items in significant ways (how it's added, how it's structured, how it's maintained, ...)

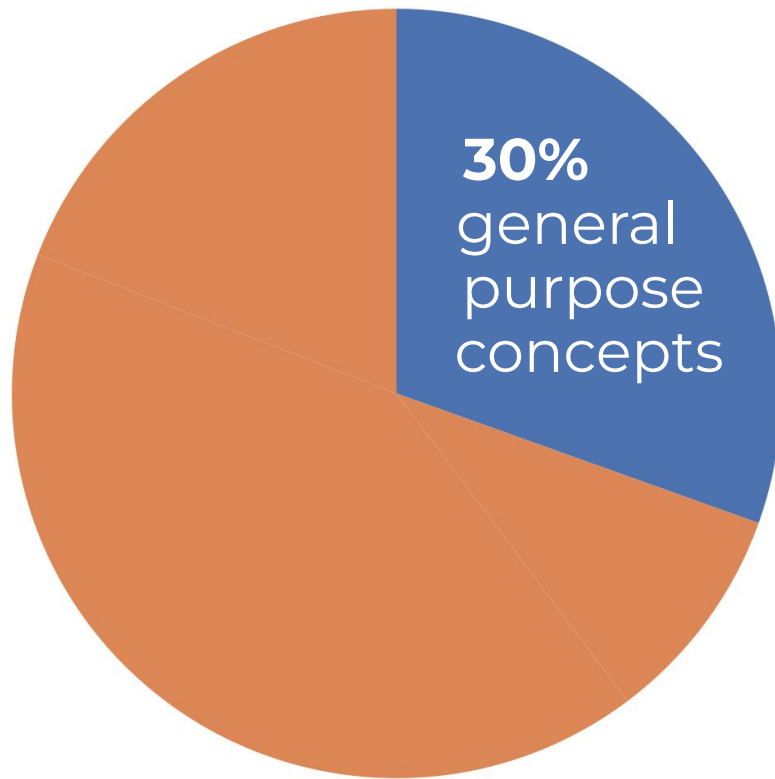
What does this look like in detail?



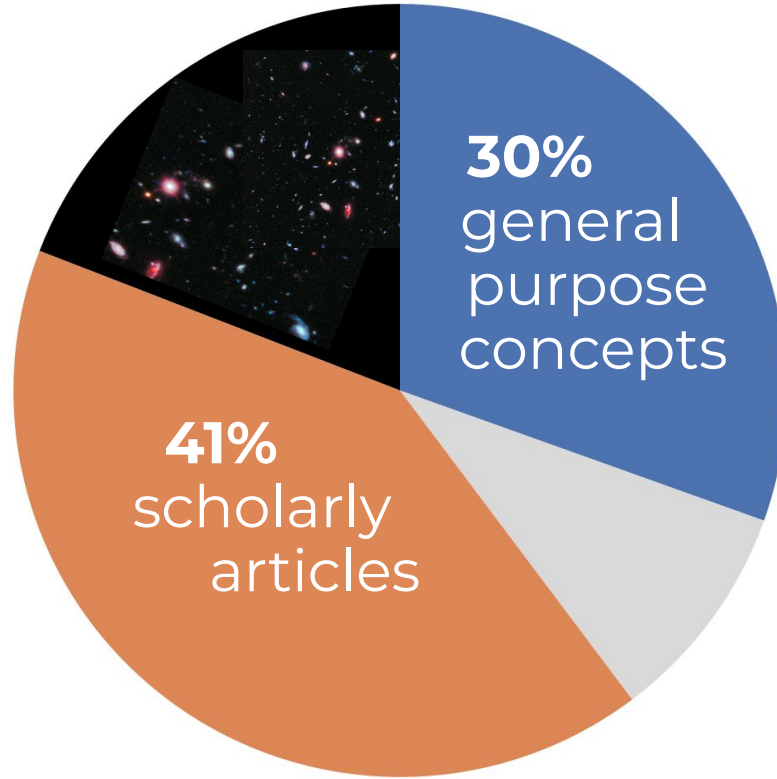
of Items



of Items



% of Items as of today

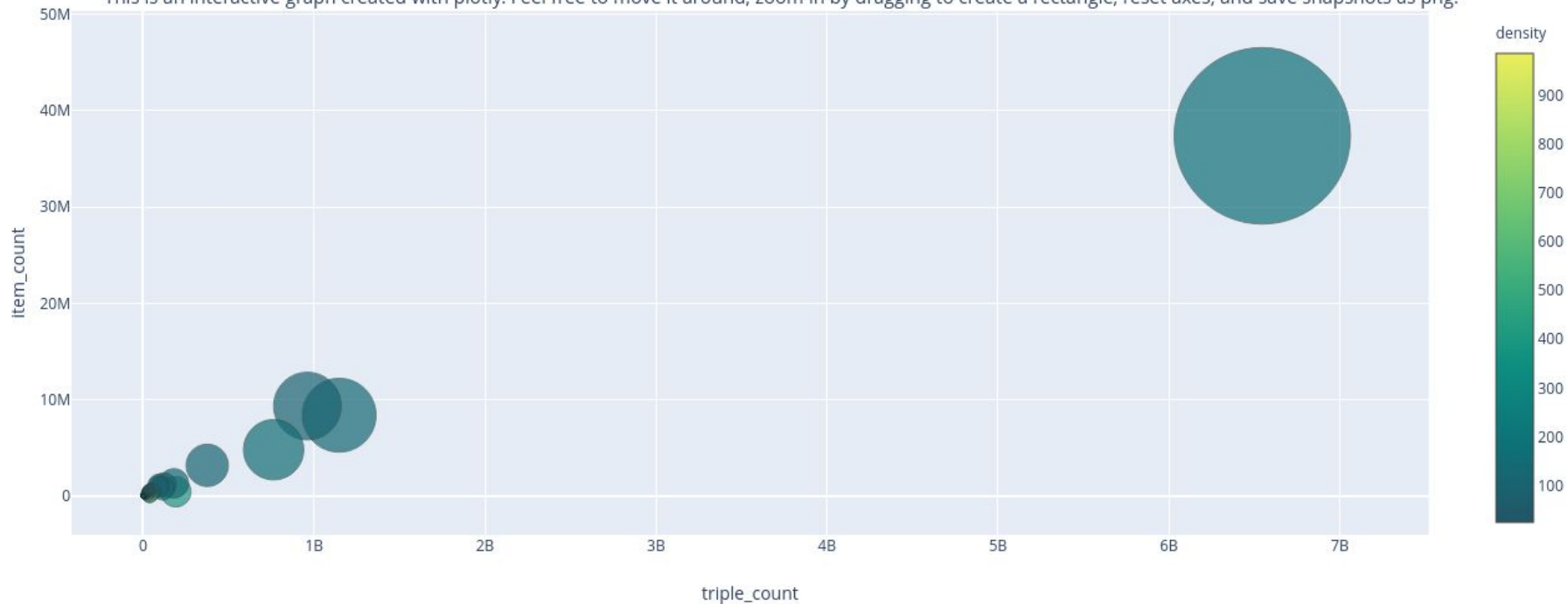


% of Items as of today

Top 340 Subgraphs in Wikidata

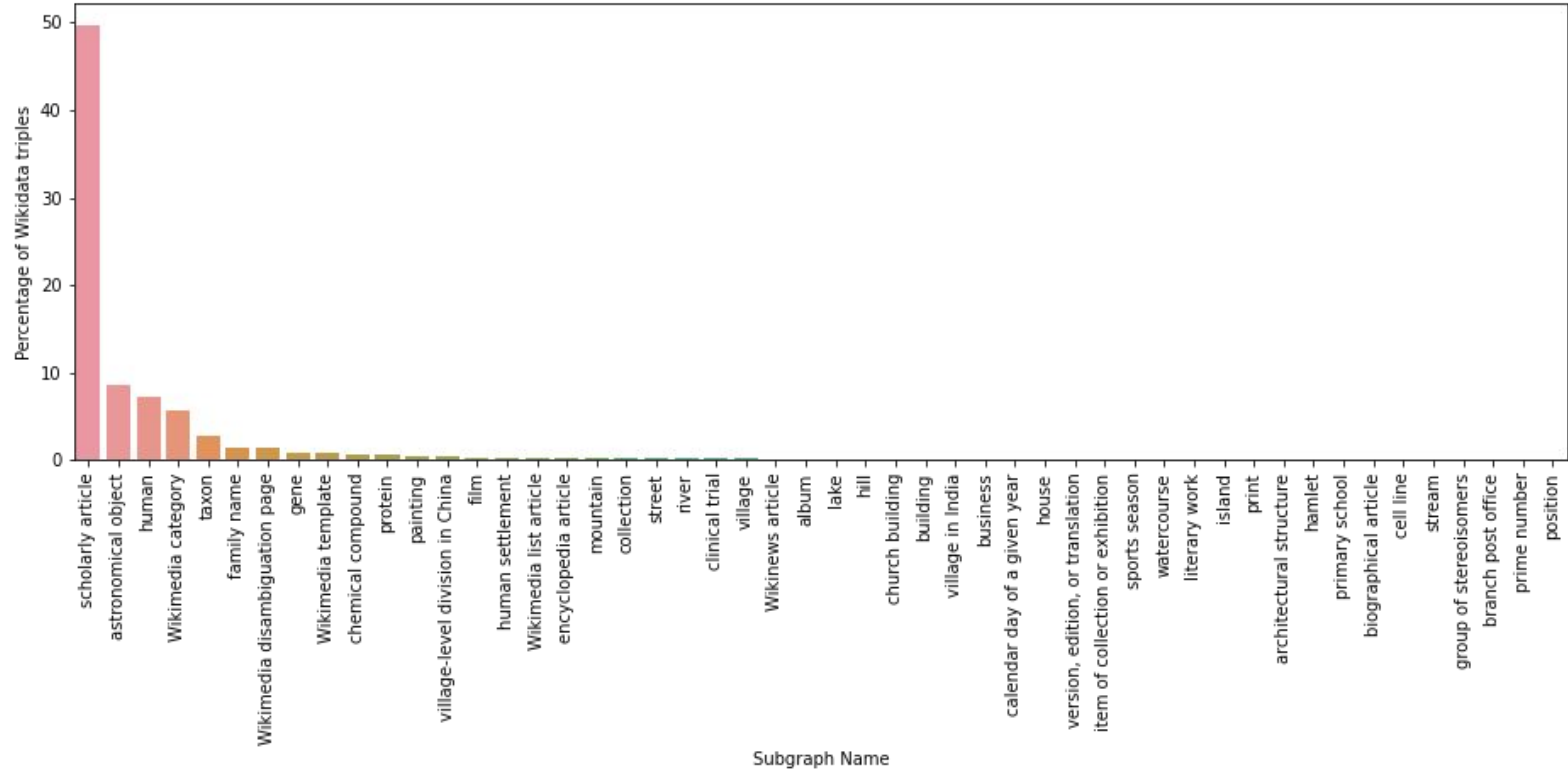
These form 90% of all Wikidata triples. Density = triple_count/item_count

This is an interactive graph created with plotly. Feel free to move it around, zoom in by dragging to create a rectangle, reset axes, and save snapshots as png.



[Top subgraphs in Wikidata](#)

Size (in percents) of the top 50 subgraphs in Wikidata
in order of subgraph size



[Percentage of triples in the top 50 subgraphs](#)



[# of active editors](#)

> 8.400

Number of Items one active editor (in theory!) is maintaining

> 750

Number of active Items one active editor (in theory!) is maintaining

What needs to happen?

**We need a decision about
what our vision and goal is and
then commit to it.**






We have options.

They determine how to move forward!

Just some of them:

- A database of all sources cited in Wikimedia projects
- A database for curated bibliographic corpora
- The “bibliographic commons”
- ...



Depending on our vision and goal, we can commit to a way forward.

- Keep data in Wikidata but reduce scope significantly
- Move data (mostly) to a dedicated Wikibase instance and go wild
- ...

The [risks and benefits from 2018](#) largely still hold true.

How do we make this decision?
