The Wikidata Query Service Split and its Impact on the Scholarly Graph

Tiago Lubiana[1,*,†], Lane Rasberry[2,†] and Daniel Mietchen[3,†]

[1] *Wikimedia Brasil, São Paulo, SP, Brazil*

[2] *School of Data Science, University of Virginia, Charlottesville, VA, United States of America*

[3] *FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Berlin, Germany*

Abstract

Wikidata, the open knowledge graph sister to Wikipedia, is undergoing major changes in 2025 as the Wikimedia Foundation splits its data into two graphs. One of the split pieces is essentially the data of the WikiCite project, which is an initiative expanding the use of Wikidata as a platform for scholarly information. Due to WikiCite's success, it accounts for over 50% of the triples on the Wikidata graph. This split was motivated by challenges in scaling up the Wikidata SPARQL Query Service, which relies on Blazegraph, a technology unsuited for Wikidata's current scale and growth rate. While the split has been prepared since 2021, this large infrastructure change has wide implications both for community tools which use WikiCite data, such as the Scholia platform, as well as for core functionalities of Wikidata, such as systems that detect duplicate items and constraint violations. In this paper, we present an overview of the infrastructure available on Wikidata for querying scholarly information and how it serves the community endeavours related to the WikiCite initiative. In particular, we focus on the Wikidata Query Service split, its motivations, and its impacts for those intending to use Wikidata as a source of semantic scholarly data. We present the alternatives for rewriting or redirecting broken queries, making explicit the rules of the graph split, and when federated queries are now required, guiding stakeholders on how to adapt to the new infrastructure.

Keywords

Wikidata, Scholia, WikiCite, scholarly knowledge graph, SPARQL endpoint, Blazegraph, scalability

1. Introduction

Wikidata has matured as a core resource in the Linked Open Data ecosystem for academic information. As of 2025, the platform has become part of workflows in biocuration and scholarmetrics, collecting identifiers, disambiguating names, and providing a Query Service to gather information from a single access point.

The WikiCite project has scaled the capacity of Wikidata to host metadata about publications, bringing semantics to the handling of citations in the Wikimedia ecosystem [1, 2]. The CiteQ template on Wikipedia, for example, streamlined structured generation of references and is now evaluated as part of the Altmetric credit pipeline [3]. Even inside Wikidata, the scholarly information curated through WikiCite supports provenance for statements [4].

The statements on Wikidata are retrievable in different ways. Data is openly available with CC0 dedication, and access options range from complete dumps [5] to on-demand requests via the MediaWiki Action API and the recently developed Wikibase REST API. One of the arguably most powerful ways is via SPARQL queries, which enable precise and complex requests for information [4].

Wikidata hosts an official endpoint with a graphical interface, known as the "Wikidata Query Service" (WDQS). The web interface of the query service, available at https://query.wikidata.org, facilitates the writing of queries by providing autocompletion, syntax highlighting, and showcasing examples. The

SEMANTiCS Conference 2025, September 03–05, 2025, Vienna, Austria

[*] Corresponding author.

[†] These authors contributed equally.

✉ tiagolubiana@gmail.com (T. Lubiana); lr2ua@virginia.edu (L. Rasberry); daniel.mietchen@fiz-karlsruhe.de (D. Mietchen)
🌐 https://tiago.bio.br (T. Lubiana); https://bluerasberry.com/ (L. Rasberry); https://scholia.toolforge.org/author/Q20895785 (D. Mietchen)
iD 0000-0003-2473-2313 (T. Lubiana); 0000-0002-9485-6146 (L. Rasberry); 0000-0001-9488-1870 (D. Mietchen)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

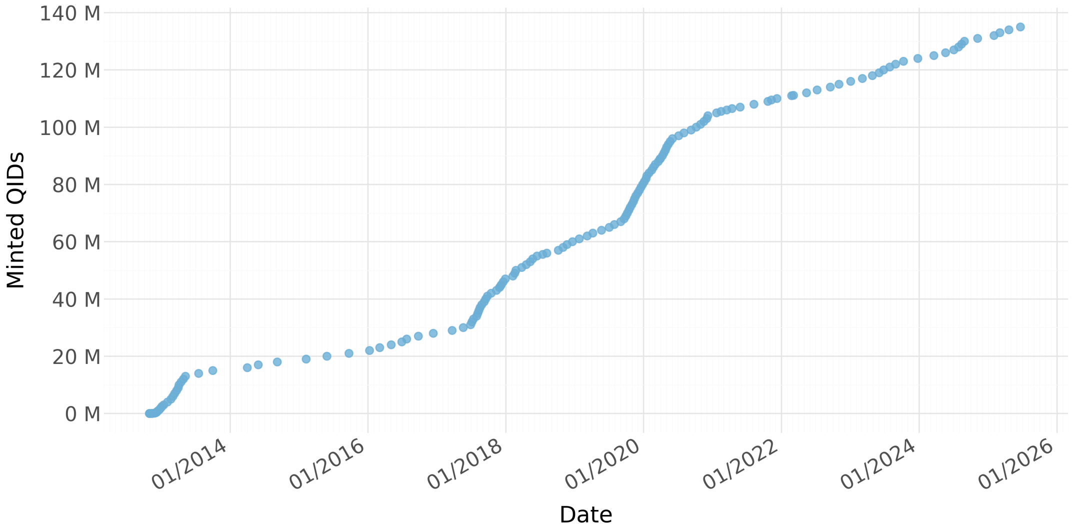


Figure 1: The growth of Wikidata, as seen by the number of minted QIDs. An acceleration is seen from 2017-2021 largely due to the creation of items for scholarly publications (https://w.wiki/EdfY).

WDQS also provides a rich set of default visualizations, including image grids, interactive maps, bar charts, and timelines, which can be embedded directly into web pages using <iframe> tags. The WDQS is updated nearly in real time, with query results usually reflecting changes within minutes or even seconds of the information coming to the platform. [6]

Powerful as it is, the community widely uses the Wikidata Query Service, and embedding its results is common practice. Of note, the Scholia platform for presenting and curating scholarly information relies heavily on WDQS and its standardized, plug-and-play visualizations [7]. The Scholia-style display of rich bibliometric information via the Query Service was also adopted by other projects, such as Vitrine NeuroMat [8] and the SARS-CoV-2 Query Book [9].

In 2015, the query service was launched by the Discovery Department at the Wikimedia Foundation [10] using Blazegraph as a database, out of many possible competing technologies [11]. The Blazegraph software, however, was soon left unmaintained and increasingly unable to keep up with the demands of a growing Wikidata, including the tens of millions of articles added in the context of WikiCite (Figure 1). By 2022, the Wikimedia Foundation had already undergone several rounds of research towards a replacement, but no clear winner emerged [12]. They also analyzed the nature of the scholarly content of Wikidata, at the time around 6.4 billion triples, half the size of the database [13]. As an emergency measure to avoid a catastrophic failure of the service, the Search Platform Team (which replaced the Discovery Department at the Foundation in 2017) decided to divide the SPARQL Endpoint into two: one for just the scholarly articles and another one for the rest of Wikidata.

Even if carefully planned, such a change brings a number of challenges to Wikidata. Here, we describe the graph split from a WikiCite perspective, documenting it with a focus on the academic community relying on the Wikidata infrastructure. A more complete version of the discussion is documented on a living Meta-Wiki page [14]. There are also other perspectives published by the Wikidata community, including a 2024 op-ed on Wikipedia's newspaper, The Signpost [15], and a 2025 benchmark report of alternative backends [16].

2. Handling The Split

2.1. Which items are included in each part of the split?

As of August 2025, the technical rules for the split [17] determine which items are considered scholarly articles or not. Then, the processing pipeline separates the triples where scholarly articles are subjects into a separate graph, which is loaded into the scholarly SPARQL endpoint. Every time a change happens in Wikidata, the pipeline updates the triples served by each endpoint.

Included in the scholarly graph are all triples for items that either (1) contain a (non-deprecated) statement with the property "publication type of scholarly work" (P13046) or (2) have a (non-deprecated) value for "instance of" (P31) that matches any one out of 49 scholarly QIDs. This list includes "scholarly article" (Q13442814) and 24 subclasses of it, "thesis" (Q1266946) and 15 subclasses of it, plus "erratum" (Q1348305), "dissertation" (Q1385450), "comment" (Q58897583), "research report" (Q59387148), "field study report" (Q1402850), "conference poster" (Q54670950), "scientific note" (Q114613919) and, perhaps surprisingly, "Bachelor of Literature" (Q112585758). These QIDs were selected by the Search Platform team from a list containing over 4000 values [18]. The triples that do not match the rules are directed to the main graph, and there is no duplication of content.

There are around 77 million items on the main graph and 45 million items on the scholarly graph. The information on these scholarly works is particularly rich, corresponding to 8.7 billion triples, while the main endpoint hosts slightly less, with around 8.4 billion.

2.2. How to query for scholarly information

It is often not trivial to know in which endpoint a particular query should be run, as query results may be split across different graphs. To aid users in rewriting queries, we developed a simple online checker for verifying queries across the Wikimedia-hosted endpoints [19]. We also designed a simple decision graph, towards helping SPARQL query writers in adapting to the change (Figure 2). This workflow is being used to support the rewriting of queries in the Scholia platform, many of which depend on information from both endpoints [7].

For SPARQL queries needing both graphs, internal federation is necessary. Its usage is being documented collaboratively in the Internal Federation Guide [20]. An example of basic syntax for retrieving items from one graph and labels from another is shown in Query 1. Note that while the article type for Q41799194 is retrieved from the scholarly graph, the labels for the types are only retrieved if calling the wikibase:label service from the main endpoint.

Query 1: Wikidata Query Service internal federation example

```
SELECT ?articleType ?articleTypeLabel WHERE {
  SERVICE wdsubgraph:scholarly_articles {
    wd:Q41799194 wdt:P31 ?articleType   }
  SERVICE wikibase:label {bd:serviceParam wikibase:language "en".}}
```

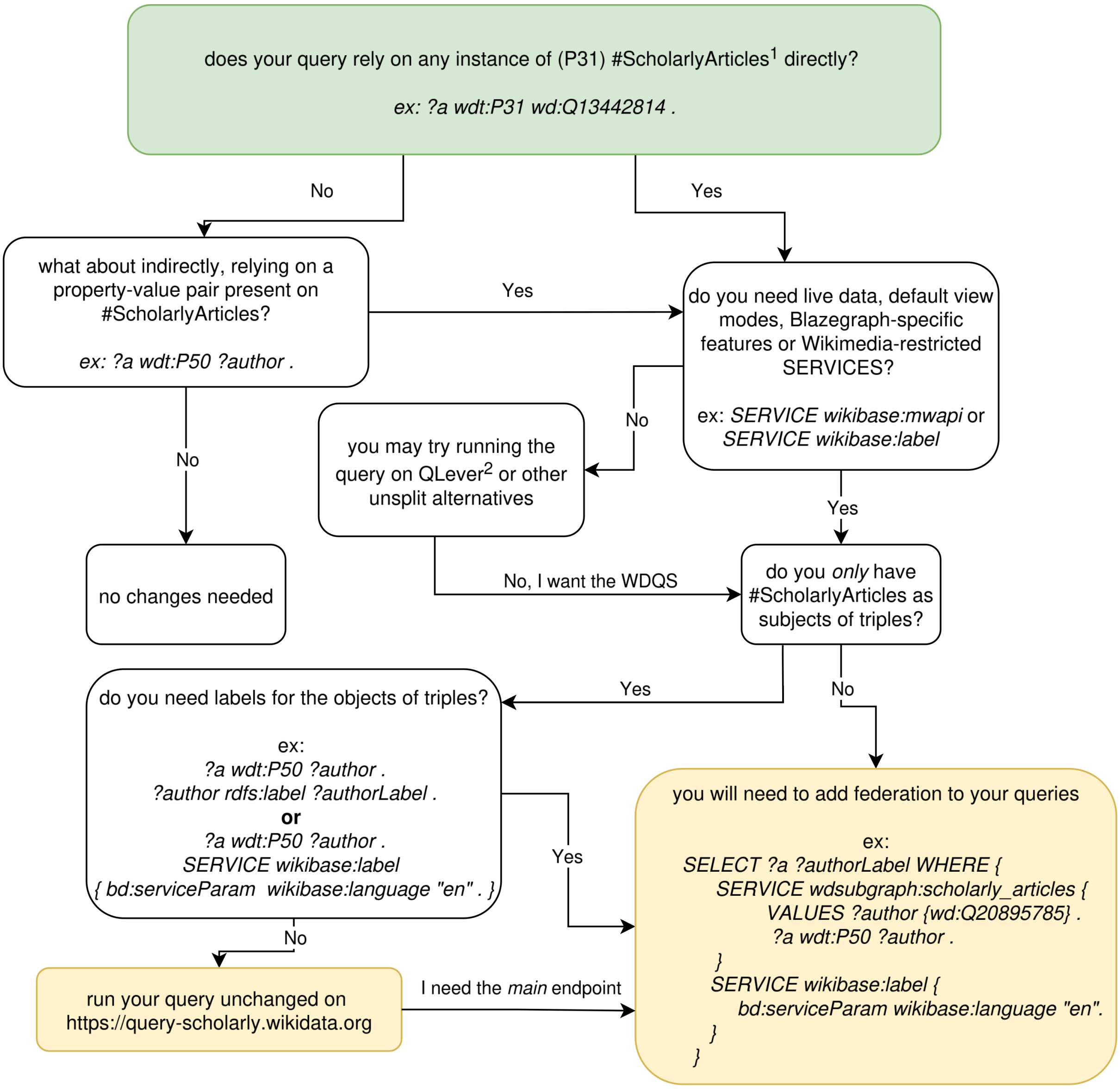
2.3. Does the split affect anything other than the Wikidata Query Service? For example, the Action API?

The split does not affect functionalities of Wikidata that are independent of the Wikidata Query Service. For example, requests via the REST API or the Action API should not be changed.

There are assessments of growth on multiple parts of Wikidata, including discussions on the limits of the platform and on extra regulation of mass edits [21]. While allocating scholarly information in a separate Wikibase instance may happen in the future, to the best of our knowledge, as of 2025, there are no plans to split any other infrastructures beyond the SPARQL query service.

3. Conclusion

This paper discusses the Wikidata Query Service graph split from a WikiCite perspective, providing the community with resources to navigate the changes. A longer, multilingual version of the presented FAQ is available on the Meta page for WikiCite [14]. With that, we hope to help semantic web researchers use the scholarly graph on Wikidata and join the conversation towards a sustainable future for this major Linked Open Data infrastructure.


does your query rely on any instance of (P31) #ScholarlyArticles[1] directly?
ex: ?a wdt:P31 wd:Q13442814 .
No
Yes
what about indirectly, relying on a property-value pair present on #ScholarlyArticles?
ex: ?a wdt:P50 ?author .
Yes
do you need live data, default view modes, Blazegraph-specific features or Wikimedia-restricted SERVICES?
ex: SERVICE wikibase:mwapi or SERVICE wikibase:label
No
you may try running the query on QLever[2] or other unsplit alternatives
No
no changes needed
Yes
No, I want the WDQS
do you only have #ScholarlyArticles as subjects of triples?
Yes
No
do you need labels for the objects of triples?
ex:
?a wdt:P50 ?author .
?author rdfs:label ?authorLabel .
or
?a wdt:P50 ?author .
SERVICE wikibase:label
{ bd:serviceParam wikibase:language "en" . }
Yes
you will need to add federation to your queries
ex:
SELECT ?a ?authorLabel WHERE {
SERVICE wdsubgraph:scholarly_articles {
VALUES ?author {wd:Q20895785} .
?a wdt:P50 ?author .
}
SERVICE wikibase:label {
bd:serviceParam wikibase:language "en".
}
}
No
run your query unchanged on https://query-scholarly.wikidata.org
I need the main endpoint

[1] See https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_graph_split/Rules

[2] https://qlever.cs.uni-freiburg.de/wikidata

Figure 2: A decision graph for adapting WDQS queries to the graph split.

4. Acknowledgments

This work was supported in part by Wikimedia CH and Wikimedia Deutschland.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly, ChatGPT in order to: Grammar and spelling check, Formatting assistance. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Taraborelli, J. Dugan, L. Pintscher, D. Mietchen, C. Neylon, WikiCite 2016 Report, 2016. doi:10.6084/m9.figshare.4042530.v2.

[2] M. Lemus-Rojas, J. D. Odell, Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia 6 (2018). URL: https://jlsc-pub.org/articles/10.7710/2162-3309.2272/galley/201/download/. doi:10.7710/2162-3309.2272.

[3] Altmetric, Wikipedia tracking enhancement: supporting the Cite Q template, 2025. URL: https://updates.altmetric.com/announcements/wikipedia-tracking-enhancement-supporting-the-cite-q-template.

[4] T. S. Shafee, D. Mietchen, T. Lubiana, D. Jemielniak, A. Waagmeester, Ten quick tips for editing Wikidata 19 (2023) e1011235. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011235. doi:10.1371/journal.pcbi.1011235.

[5] Wikidata: Database download, 2025. URL: https://www.wikidata.org/wiki/Wikidata:Database_download.

[6] A. Shorland, Wikidata query service updater evolution, 2022. URL: https://addshore.com/2022/04/wikidata-query-service-updater-evolution/.

[7] F. Nielsen, D. Mietchen, E. Willighagen, Scholia, Scientometrics and Wikidata, in: The Semantic Web: ESWC 2017 Satellite Events, Lecture Notes in Computer Science, Springer, 2017, pp. 237–259. doi:10.1007/978-3-319-70407-4_36.

[8] NeuroMat Research Group, Vitrine NeuroMat, 2025. URL: https://vitrine.numec.prp.usp.br/.

[9] Addshore, D. Mietchen, E. L. Willighagen, Wikidata Queries around the SARS-CoV-2 virus and pandemic, Zenodo, 2020. URL: https://egonw.github.io/SARS-CoV-2-Queries/. doi:10.5281/zenodo.3977414.

[10] D. Garry, Announcing the release of the Wikidata Query Service, 2015. URL: https://lists.wikimedia.org/hyperkitty/list/wikidata@lists.wikimedia.org/thread/N2HPRCYIWGLM2IDTNCHQLNY574H5ZEQR/, mailing list message.

[11] Wikimedia Foundation, Wikidata query pick backend (comparison spreadsheet), 2016. URL: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Benchmark_for_the_backend.

[12] A. Westerinen, WDQS Search Team, WDQS Backend Alternatives: The Process, Details and Results - Version 1.1, https://commons.wikimedia.org/wiki/File:WDQS_Backend_Alternatives_working_paper.pdf, 2022. Version 1.1, March 29, 2022.

[13] A. Khatun, Wikidata Scholarly Articles Subgraph Analysis, https://wikitech.wikimedia.org/wiki/User:Aisha_Khatun/Wikidata_Scholarly_Articles_Subgraph_Analysis, 2021.

[14] WikiCite/WDQS graph split — FAQ, 2025. URL: https://meta.wikimedia.org/wiki/WikiCite/WDQS_graph_split.

[15] L. Rasberry, Wikidata to split as sheer volume of information overloads infrastructure, https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2024-05-16/Op-Ed, 2024.

[16] P. F. P. Schneider, Scaling Wikidata/Benchmarking/Final Report, https://www.wikidata.org/wiki/Wikidata:Scaling_Wikidata/Benchmarking/Final_Report, 2025.

[17] Wikidata, Wikidata:SPARQL query service/WDQS graph split/Rules, 2025. URL: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_graph_split/Rules.

[18] T. Lubiana, Snapshot — WikiCite subclass hierarchy for the Wikidata scholarly graph split , 2025. doi:10.5281/zenodo.16884396.

[19] T. Lubiana, Wikidata Graph Split Simple SPARQL Benchmark, 2025. URL: http://tiago.bio.br/query-split-tester/.

[20] Wikidata SPARQL query service – internal federation guide, 2025. URL: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_graph_split/Internal_Federation_Guide.

[21] L. Pintscher, Wikidata:Requests for comment/Mass-editing policy, https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Mass-editing_policy, 2025.