Assessing the Reliability and Scientific Rigor of References in Wikidata

Hannah Schuster[1,2,*,†], Amin Anjomshoaa[1,†] and Axel Polleres[1,2,†]

[1]Vienna University of Economics and Business, 1020 Vienna, Austria
[2]Complexity Science Hub Vienna, 1080 Vienna, Austria

Abstract

Wikidata is a rapidly growing user-edited open knowledge graph that provides easy access to structured data. Since Wikidata allows contradictory information, references are crucial for supporting statements and tracking the source of information. Consequently, investigating the use, types, and scientific value of references within Wikidata is essential. In this paper, we will first conduct a heuristic evaluation of Wikidata references using a sampling method. Subsequently, we will focus on a specific category of references, Digital Object Identifiers (DOIs), known for citing scientific publications. Our sampled Wikidata statements analysis indicates widespread adoption of the DOI system within Wikidata. To assess the quality of scholarly resources referenced in Wikidata, we used percentile metrics derived from the OpenAlex platform. Additionally, h-index indicators from OpenAlex were employed to evaluate the credibility of these sources and determine whether the Wikidata citations originated from reputable sources or publishers. Our findings show that papers in the social and physical sciences tend to perform better in Wikidata compared to OpenAlex. Moreover, while top-tier journals dominate citations in OpenAlex—particularly in the health and life sciences— Wikidata shows a higher citation rate for mid-tier and emerging journals. This indicates a broader representation of scholarly contributions within Wikidata.

Keywords

Wikidata, Data Quality, Scholarly Citations.

1. Introduction

Wikidata is an entity-oriented database designed to represent items related to various topics, concepts, and objects, along with detailed claims and qualifiers describing these entities. Similar to Wikipedia pages, the Wikidata Knowledge Graph relies on references to support the claims and statements it entails. These references should point to the source of the provided statement, with statements supported by and linked to at least one source according to internal Wikidata guidelines.

Since its rapid expansion from 42.3 million items in 2017 to over 113 million items in 2024 [1], Wikidata has placed greater emphasis on ensuring the quality of its data. The platform aims to cover a wide range of topics through user collaborations. As the content is primarily created and edited by users, references on Wikidata should be relevant, authoritative, and accessible according to their policies. Additionally, referenced sources should provide context and supportive arguments for statements. The evaluation of references is the responsibility of the Wikidata user community.

Several researchers and practitioners have investigated different features and characteristics of Wikidata, with much of the work focusing on the quality of the sources. Adequate, relevant, and trustworthy references are increasingly important for improving the reputation of Wikimedia projects. In contrast, missing or inappropriate references can affect reliability and hinder the reuse of data.

In this paper, we first examine the quality and structure of external references within Wikidata, with a particular emphasis on their scientific character and background. Next, we focus on a specific category of references: scientific publications identified by Digital Object Identifiers (DOIs) and examine how

4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, 12 November 2024 - Baltimore, MD, USA

*Corresponding author.
†These authors contributed equally.
schuster@csh.ac.at (H. Schuster); Amin.Anjomshoaa@wu.ac.at (A. Anjomshoaa); Axel.Polleres@wu.ac.at (A. Polleres)
0000-0003-3032-1959 (H. Schuster); 0000-0001-6277-742X (A. Anjomshoaa); 0000-0001-5670-1146 (A. Polleres)

these DOIs are utilized within Wikidata. We compare the performance and impact of these referenced papers in Wikidata with their performance in the broader scientific community, using data from the OpenAlex dataset.

2. Background and Related Works

In Wikidata's data structure, a claim combines a property with at least one value and optional qualifiers to provide information about items. When this property-value pair is enriched with additional information (such as references or ranks), it becomes a statement. A claim without a qualifier is called a snak, representing a basic triple consisting of an item and a property-value pair. References or other qualifiers, which can be appended to statements, are connected to the value, allowing for multiple references for a single statement. Wikidata accommodates contradicting statements to reflect controversial, uncertain, or debatable information, requiring references to support the entire statement.

Albeit, there have been numerous studies on Wikidata and its references, no study has solely examined the scientific background of sources in more detail with a special focus on the used properties and identifiers. The research conducted by [2] compares the use of external references between Wikipedia and Wikidata. The paper does not tackle the topic regarding the scientific character of external references within Wikidata. The authors also analyzed the relevance and authoritativeness of Wikidata references which are the only requirements for sources. However, this work does not directly tackle the topic of the scientific character of the exported references.

The authors in [3] developed a tool named 'Scholia'. The purpose of the tool is to create on-the-fly scholarly profiles for researchers, organizations, journals, publishers, individual scholarly works, and for research topics using bibliographic and other information in Wikidata. Besides the functionality, the basic structure of Wikidata and the contained references and author information is described.

Another study [4] presents a comprehensive dataset of citations extracted from English Wikipedia. It highlights that some references to scientific articles and publications lacked corresponding DOIs. Therefore, identifiers are a reliable indication for scientific publications. However, this principle does not work vice versa meaning that if a reference does lack an identifier it is not a scientific citation. Consequently, the authors recommend an approach that goes beyond the used identifiers of scientific databases like Crossref [5] and most famously Altmetric [6].

A recent Wikidata analysis [7], explores the linked Wikidata resources, including external datasets and ontologies. However, the paper does not include external Wikidata references or identifiers regarding their scientific character. Another work in this context [8] focuses on analyzing Wikipedia references across different languages and includes some identifiers for scientific sources.

Another paper [9] proposes a reference quality assessment framework to enhance the quality of Wikidata references. Quality, in this context, means that the reference is accessible and verifies the statement to which it is connected. Additionally, a reference suggestion framework is introduced to propose references for Wikidata claims. However, the scientific rigor of the references was not considered among the metrics in this research.

3. Quality of Reference Data

Over the past few years, the number of Wikidata items classified as scholarly articles has significantly increased. As of September 2024, Wikidata contains over 44 million scholarly articles, reflecting an increase of more than 6 million in the past two years. All of these articles are instances of a single entity, Q13442814, in Wikidata.

We analyzed the properties and their frequencies within reference nodes. As of September 2022, we identified more than 5,267 distinct reference properties, encompassing a total of 335,960,448 records. The top 10 reference properties represent 89.2% of the total population and are illustrated in Figure 1. The two top most used reference properties (P248 and P813) are used more than 80M times. Those

properties can be universally used and reference to either external or internal reference values. There are 65,151,203 reference URLs in Wikidata which point towards external references.

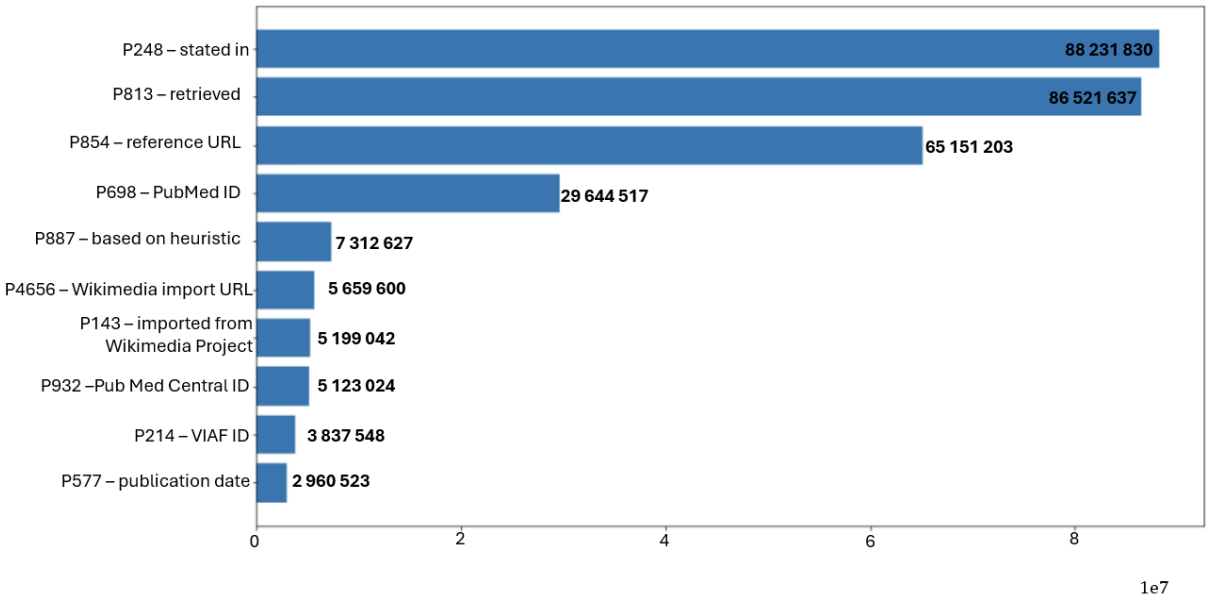

Figure 1: Top 10 Wikidata reference properties.

PubMed ID with 29,644,517 and PubMed Central ID with 5,123,024 reference counts show that there are many references linked with those identifiers. One possible reason these IDs are more prevalent in Wikidata is the presence of a bot that links the IDs to Wikidata items. This may also explains the large number of data properties in three categories: 'stated in', 'retrieved', and 'reference URL', as shown in 1.

The most popular identifiers for scientific references are PubMedID and PMCID, which indicates that in the fields of biomedical and life sciences, Wikidata entries link to more scientific sources than in other areas. The distribution of the top 10 Wikidata reference properties for scientific identifiers excluding PubMed ID and PMC ID is depicted in Figure 2.

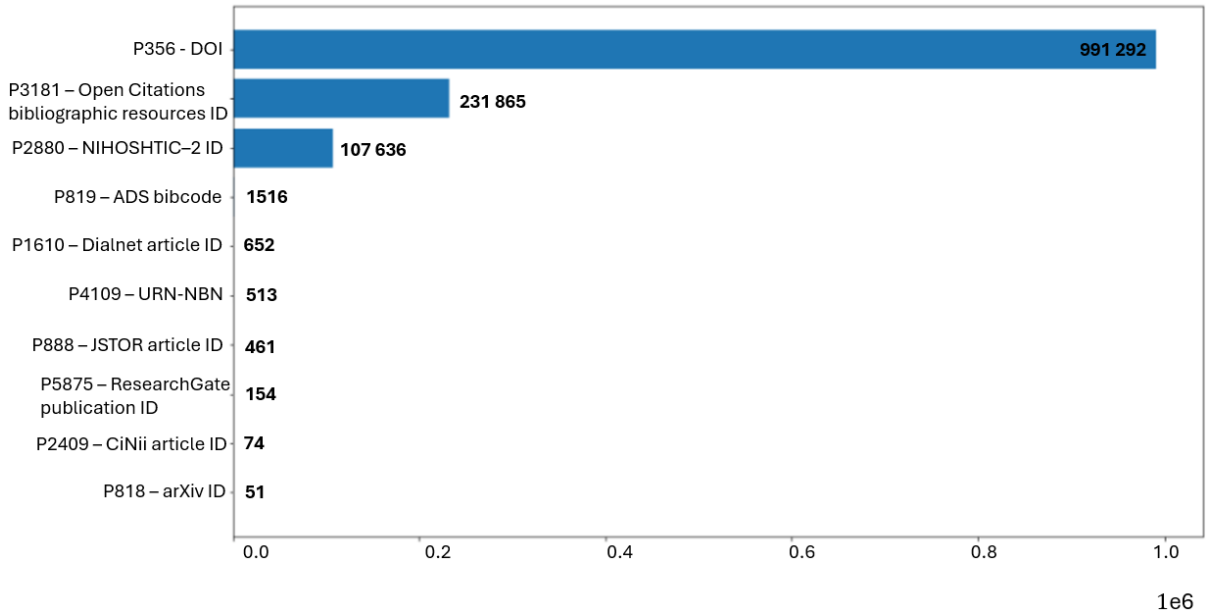

Figure 2: Top 10 Wikidata reference properties for scientific identifiers excluding PubMed ID and PMC ID

4. Implementing Performance Measures for Wikidata References

Building on our examination of reference quality, we now focus on systematically measuring the scientific impact and credibility of references within Wikidata. Given Wikidata's diverse user contributions, it is crucial to assess whether the cited papers are scientifically robust and well-regarded. In this section, we implement performance measures using citation counts and journal H-index metrics from OpenAlex to evaluate the scholarly merit of these references.

We integrated two data sources for this analysis. The first dataset from Wikidata provides DOI-based citations, reflecting citation activity within the Wikidata community but not directly measuring scholarly impact. To complement this, we use a second dataset from OpenAlex, which provides detailed information on scientific papers, including DOI, publication year, citations, domain categorization, and journal metrics like the H-index. While OpenAlex also provides percentile metrics, their calculation methodology remains unspecified, leading us to exclude these metrics from our analysis. The complete dataset is publicly available and can be accessed on Figshare [10].

Our analytical approach begins with the H-index of journals as provided by OpenAlex, a widely accepted bibliometric metric that quantifies the impact of a journal based on the citation performance of its most frequently cited articles. Specifically, a journal's H-index is defined as the number h, where the journal has h articles that have each received at least h citations over time. To facilitate comparison between Wikidata and OpenAlex sources, we categorized journals into four tiers based on their H-index: Top-tier (H-index $>$ 100), Well-regarded (50 $\leq$ H-index $<$ 100), Mid-tier (20 $\leq$ H-index $<$ 50), and Emerging (H-index $<$ 20).

Despite the advantage older journals might have in citation accumulation, the H-index remains a valuable measure of a journal's acceptance and impact within the scientific community. This tiered categorization allows us to better assess citation patterns across different journal quality levels and identify trends in scholarly referencing within Wikidata and OpenAlex.

In Figure 3, we observe that in OpenAlex the percentage of citations from top-tier journals is significantly higher than that of the other three categories, especially in the domains of health sciences and life sciences. Notably, the physical sciences exhibit an increase in citations from well-regarded journals, while life sciences have the lowest proportion of top-tier journals, though it remains above 50%. Conversely, the percentage of citations from mid-tier and emerging journals is very low across all four domains, with the highest percentage found in the social sciences.

When examining the distribution of citations in Wikidata within the same figure, we see a different trend. In all domains, the percentage of top-tier journal citations is lower. This is particularly evident in the social sciences, where only approximately one-third of citations are from top-tier journals. Additionally, emerging journals are cited more frequently in Wikidata than in OpenAlex. Overall, while top-tier journals still dominate in some categories, Wikidata shows a higher citation rate for mid-tier and emerging journals, indicating broader usage.

In the next step, we implemented a percentile measure in OpenAlex to gain deeper insights. We calculated the percentile rank of citations as follows:

$$Percentile\ Rank = \frac{Rank - 1}{Total\ number\ of\ citations - 1} \times 100, \tag{1}$$

where the *Rank* is the position of a paper's citation count within its field, publication year, and open access status, and the *Total number of citations* is the number of papers within the same grouping. Next, we translated the percentile ranking from OpenAlex to Wikidata by applying the percentile rank calculated for each paper in OpenAlex to the number of citations recorded in Wikidata. This allowed us to place the Wikidata citation counts within the same percentile framework used in OpenAlex. Using a common percentile ranking system enables a more accurate comparison of paper performance, as it avoids the creation of two independent ranking systems, which could lead to inconsistencies in interpretation.

In Figure 4, we can see that especially in the social sciences, the box for Wikidata is much higher, and the whisker ends before the box of the OpenAlex part ends, indicating that the interquartile range of

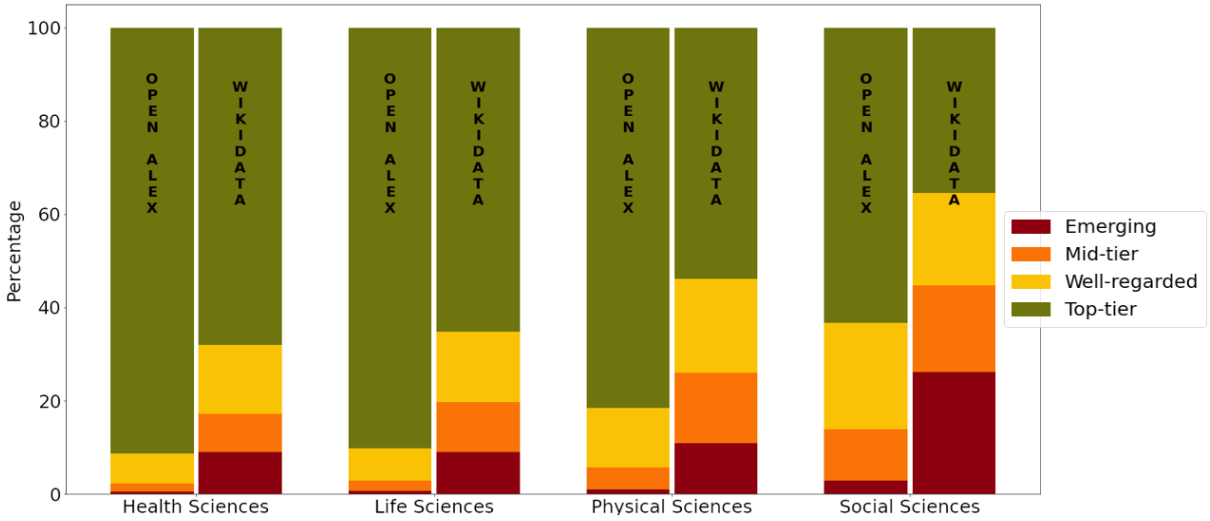

Figure 3: Distribution of citations from different journal tiers across domains.

the percentiles for Wikidata is higher. We can also see that the box for Wikidata in physical sciences is higher than the box for OpenAlex, but the whiskers are of the same length, suggesting a higher median but similar variability. The differences in the other two categories are not as pronounced; the boxes in life sciences and health sciences are slightly bigger for the Wikidata one.

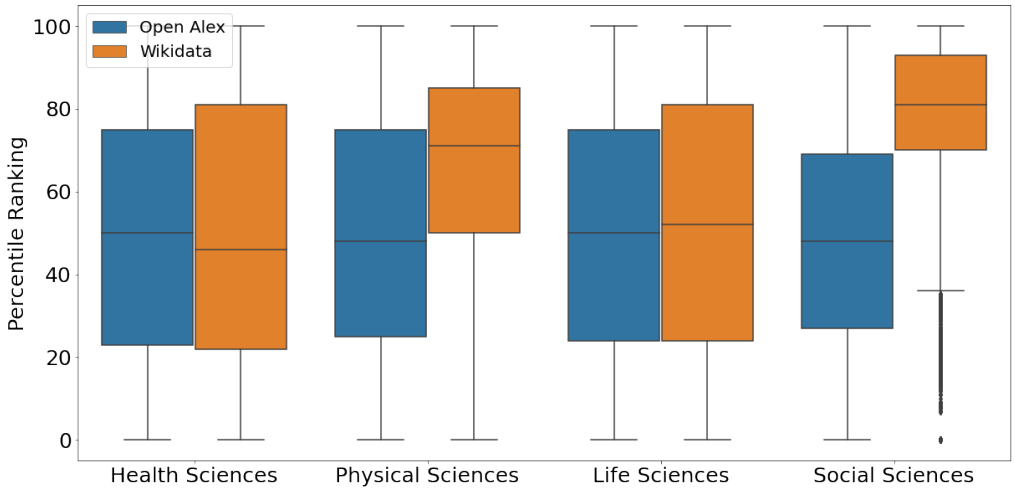

Figure 4: Comparing the percentiles grouped into the 4 domains based on OpenAlex.

In Figure 5, we wanted to see how the performance according to the OpenAlex percentile ranking changes when we separate them into the different journal categories we defined. The most apparent difference is that the percentile ranking of Wikidata references from emerging journals is much higher than that of the OpenAlex references, as we can see that the boxes are nearly at opposite ends of the scale. Also, in the mid-tier category, the percentile rankings in Wikidata are much higher than in OpenAlex. We can also see that the percentile rankings in the well-regarded category are higher for Wikidata than in OpenAlex. In the top-tier category, they are nearly the same.

We observed that, especially in the social sciences, papers cited in Wikidata generally exhibit higher citation counts compared to those in OpenAlex. This trend is somewhat visible across other domains as well, suggesting that Wikidata may include a wider range of journals in its citations. Additionally,

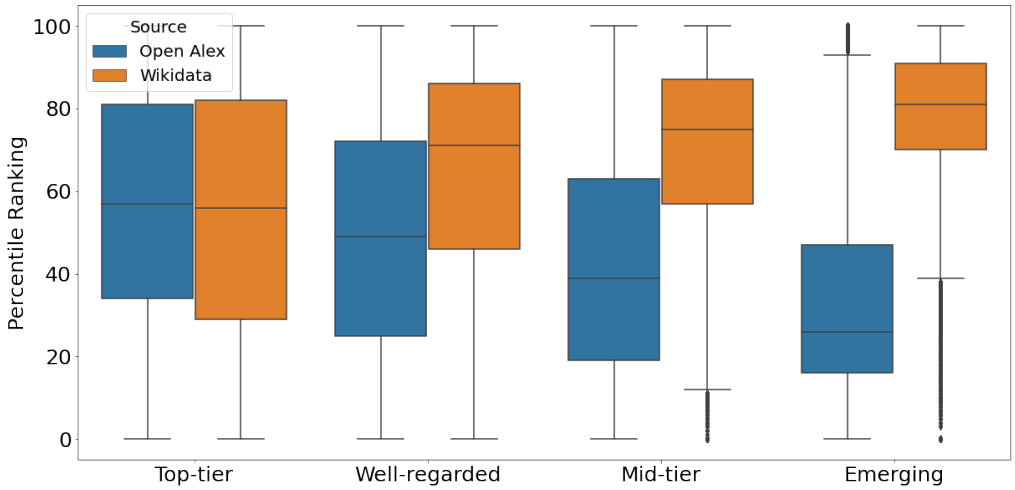

Figure 5: Comparing the percentiles grouped into the 4 journal categories based on the h-index.

while top-tier journals continue to receive a substantial number of citations on both platforms, Wikidata shows a more varied distribution of citations across different journal tiers. This could indicate a more inclusive referencing approach within Wikidata, potentially encompassing a broader spectrum of scientific contributions. However, it's also possible that this broader distribution reflects efforts by authors to increase the visibility of their work by citing it in Wikidata.

5. Results and Discussion

Our study examines the quality of scientific papers cited within Wikidata, a rapidly expanding open knowledge graph that supports diverse contributions. By September 2022, we identified over 5,000 distinct reference properties across 335,960,448 records. While Wikidata has seen significant growth in scholarly references, most references rely on a few dominant properties, particularly PubMed ID and PMC ID in the biomedical sciences. This highlights a strong reliance on external, credible sources but suggests potential gaps in citation diversity across other fields.

Even though the reference system in Wikidata has improved over the past few years, this alone does not provide insight into the quality of the cited sources. To evaluate the scholarly merit of these references, we used OpenAlex metrics, such as the H-index, to assess a journal's reputation. Our analysis reveals that papers cited in Wikidata, particularly in the social sciences, often show stronger performance metrics compared to those in OpenAlex. This trend suggests a greater diversity in the journal sources referenced within Wikidata.

However, it is crucial to acknowledge several factors that may influence these findings. One limitation is the use of the h-index as a ranking metric, which heavily depends on citation counts and may not fully capture journal quality or impact. Alternative metrics, such as the SCImago Journal Rank (SJR), could potentially offer a more nuanced evaluation. Additionally, our analysis is constrained by the absence of certain sources in OpenAlex that are present in Wikidata, impacting the comparability of the datasets.

The differences in citation distributions between Wikidata and OpenAlex may also arise from the nature of contributions to Wikidata, including potential self-citations by authors seeking to increase their work's visibility.

Despite these challenges, our findings suggest that Wikidata has a broader citation distribution. However, further research is needed to understand the underlying reasons for these differences and

their implications for scholarly communication. Limitations in our methodology, such as the reliance on the H-index and the potential omission of influential sources, should be considered when interpreting our results and generalizing them to broader contexts.

Acknowledgments

We would like to express our gratitude to Marco Marsoner for his valuable contributions to this work. We acknowledge support from the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility and Technology (BMK) via the ICT of the Future Program - FFG No 887554.

References

[1] Wikimedia Foundation, Wikidata:statistics, https://www.wikidata.org/wiki/Wikidata:Statistics, Last Updated on July 14, 2024. Accessed: September 2024.

[2] A. Piscopo, L.-A. Kaffee, C. Phethean, E. Simperl, Provenance information in a collaborative knowledge graph: an evaluation of wikidata external references, in: International semantic web conference, Springer, 2017, pp. 542–558.

[3] F. Å. Nielsen, D. Mietchen, E. Willighagen, Scholia, scientometrics and wikidata, in: The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28–June 1, 2017, Revised Selected Papers 14, Springer, 2017, pp. 237–259.

[4] H. Singh, R. West, G. Colavizza, Wikipedia citations: A comprehensive data set of citations with identifiers extracted from english wikipedia, Quantitative Science Studies 2 (2021) 1–19.

[5] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, Quantitative Science Studies 1 (2020) 414–427.

[6] J. Priem, D. Taraborelli, P. Groth, C. Neylon, Altmetrics: A manifesto (2011).

[7] A. Haller, A. Polleres, D. Dobriy, N. Ferranti, S. J. Rodríguez Méndez, An analysis of links in wikidata, in: European Semantic Web Conference, Springer, 2022, pp. 21–38.

[8] W. Lewoniewski, K. Węcel, W. Abramowicz, Analysis of references across wikipedia languages, in: Information and Software Technologies: 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017, Proceedings, Springer, 2017, pp. 561–573.

[9] S. A. Hosseini Beghaeiraveri, Towards automated technologies in the referencing quality of wikidata, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 324–328.

[10] A. Anjomshoaa, Wikidata 2023 References to Scientific Publications (2024). URL: https://figshare.com/articles/dataset/Wikidata_2023_References_to_Scientific_Publications/27028582. doi:10.6084/m9.figshare.27028582.v1.