Uniwersytet im. Adama Mickiewicza
w Poznaniu

BIBLIOTEKA 2023, NR 27 (36)
DOI: 10.14746/b.2023.27.12

The Most Cited Scientific Information Sources in Wikipedia Articles Across Various Languages[1]

Najczęściej cytowane źródła informacji naukowej w artykułach Wikipedii w różnych językach

Włodzimierz Lewoniewski
Uniwersytet Ekonomiczny w Poznaniu
https://orcid.org/0000-0002-0163-5492
e-mail: wlodzimierz.lewoniewski@ue.poznan.pl

Abstract. Ensuring the accuracy of information in Wikipedia articles requires the use of reliable sources that can be assessed by encyclopedia readers. However, the determination of source reliability can be subjective, with variations based on the language and subject matter of the article. Consequently, each language version of Wikipedia may have its own guidelines for judging the trustworthiness of sources. Some Wikipedia citations lead to scientific resources, which are usually deemed more reliable than websites due to their rigorous peer-review procedures and publication by esteemed academic publishers. This implies that the data presented in these scientific sources has been meticulously examined by specialists in the relevant area, providing a higher level of precision and trustworthiness. In this study, 332 million references from 61 million Wikipedia articles across 309 language versions were analyzed to identify citations to scientific publications. Additionally, OpenAlex was used to find unified metadata of important sources of information of multilingual Wikipedia.

Keywords: Wikipedia, references, sources of information, wiki, scientific sources, data science, data exploration, OpenAlex.

[1] This paper was originally presented at the First Poznań Conference on Bibliometric Analyses in the Evaluation of Open Science organized by the Scientometrics Research Unit at the University Library in Poznań, 18–19 April, 2023.

1. Introduction

Verifiability is one of Wikipedia's main quality criteria. This principle means that every piece of information in Wikipedia articles should originate from reliable sources. Reliable references for Wikipedia include scholarly articles that have undergone peer-review, books from credible publishing houses, mainstream media, and governmental reports. Furthermore, sources cited within other trustworthy resources can also be used in Wikipedia's articles.

It is crucial to understand that Wikipedia is not meant for original research or personal viewpoints. Every detail shared on Wikipedia needs to be grounded in reliable references and depicted from an unbiased perspective.

The trustworthiness of a source hinges on several factors, such as the publisher's or author's reputation, the rigor of the peer-review process, and the precision of the shared information. When selecting sources to be used in Wikipedia, the editors are urged to opt for those that are recognized as reliable and esteemed within their corresponding domains.

Scientific references are typically viewed as more dependable than other types of information since they undergo a rigorous peer review process. Here, other experts evaluate and critique research before its publication. Conversely, other information sources, like news stories, blogs, or social media posts, may be rooted in personal beliefs or unverified assertions, without the extensive scrutiny that scientific research faces.

This implies that information from scientific sources has been meticulously assessed by experts in the field, thus ensuring greater accuracy and trustworthiness. This study analyses the most commonly cited scientific sources of information in Wikipedia in different language versions.

The main research questions of this article are the following:

1. How often are scientific sources of information used within selected language versions of Wikipedia?
2. Which scientific information sources are most commonly cited in Wikipedia articles in selected languages?

2. Related work

Reference analysis of Wikipedia articles is a research area addressed by various works. One of the first studies found that the number of references correlates with the quality of Wikipedia articles.[2] Features related to reference count can be

[2] J.E. Blumenstock, *Size matters: word count as a measure of quality on Wikipedia*, "Proceedings of the 17th international conference on World Wide Web", 2008, pp. 1095–1096.

used to assess the millions of Wikipedia articles in different languages.[3] Such methods were implemented in various approaches, such as WikiRank[4] and the ORES score.[5] Relevant research[6] has proposed 10 models for sources assessment in the multilingual Wikipedia. The approach is also implemented in BestRef.[7]

Some research studies have focused on analyzing the metadata of sources cited in Wikipedia references. One such study attempted to standardize references using ISBN and DOI identifiers to determine how similar sources were between different language editions of Wikipedia.[8]

Teplitskiy et al.[9] examined the impact of open access on Wikipedia's scientific references, discovering that open access journals have a 47% higher chance of being cited on the English Wikipedia. The platform thus amplifies the diffusion of open access scientific literature to a wider audience.

Research conducted by Kousha and Thelwall[10] has assessed the value of Wikipedia citations for scholarly articles and books, discovering that 5% of articles and 33% of monographs have at least one citation from Wikipedia. The results suggest that Wikipedia citations can provide extra impact evidence for academic monographs. By analyzing 24,331 Public Library of Science articles, other study writted by Priem et al.[11] explored altmetrics as complementary measures to traditional citation metrics. This study found that altmetrics indicators track related but distinct impacts, capturing different types of influence on different audiences.

---

[3] W. Lewoniewski, K. Węcel & W. Abramowicz, *Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics*, "Computers", 8, 2019.

[4] WikiRank. *Quality and Popularity Assessment of Wikipedia Articles*, https://wikirank.net [Accessed 19 June 2023].

[5] A. Halfaker & R.S. Geiger, *ORES: Lowering barriers with participatory machine learning in Wikipedia*, "Proceedings of the ACM on Human-Computer Interaction", 4, 2020, pp. 1–37.

[6] W. Lewoniewski, K. Węcel & W. Abramowicz, *Modeling Popularity and Reliability of Sources in Multilingual Wikipedia*, "Information", 11, 2020, p. 263.

[7] BestRef. *Popularity and Reliability Assessment of Wikipedia Sources*, https://bestref.net [Accessed 19 June 2023].

[8] W. Lewoniewski, K. Węcel & W. Abramowicz, *Analysis of references across Wikipedia languages*, in: *International Conference on Information and Software Technologies*, 2017, pp. 561–573.

[9] M. Teplitskiy, G. Lu & E. Duede, *Amplifying the impact of open access: Wikipedia and the diffusion of science*, "Journal of the Association for Information Science and Technology", 68, 2017, pp. 2116–2127.

[10] K. Kousha & M. Thelwall, *Are Wikipedia citations important evidence of the impact of scholarly articles and books?*, "Journal of the Association for Information Science and Technology", 68, 2017, pp. 762–779.

[11] J. Priem, H.A. Piwowar & B.M. Hemminger, *Altmetrics in the wild: Using social media to explore scholarly impact*, 2012. DOI: https://doi.org/10.48550/arXiv.1203.4745

The study conducted by Arroyo-Machado et al.[12] presented an overview of science from Wikipedia's perspective by analyzing co-citation networks among scientific papers. It found a significant presence of medicine and biochemistry papers, with those with a high-impact factor and multidisciplinary journals being more likely to be cited. Another work written by Tomaszewski and MacDonald[13] examined the increasing trend in Wikipedia citations in scholarly publications from 2002 to 2015 using the Web of Science (WoS) database. The authors assessed the assumption that Wikipedia was being cited mostly in lower-quality journals or institutions and provided insights into the disciplines and types of journals that accept Wikipedia as an authoritative reference source.

The work written by Piccardi et al.[14] investigated user interactions with citations on Wikipedia, finding that overall engagement was low. Factors such as shorter pages and lower quality pages increased the likelihood of reference clicks, revealing how users consulted references when Wikipedia lacks the desired information. Other research by Chen and Roth[15] investigated the referencing process in Wikipedia articles, showing that referencing is associated with substantial editing when articles reach a certain level of maturity. References are often contributed by more frequent and substantial editors, suggesting a subset of qualified or committed editors. Comparing WikiProject Medicine (WPM) articles to other Wikipedia articles, another related study by Maggio et al.[16] found that WPM articles are longer, possess more external links, and have higher visitation rates. WPM readers are more likely to view footnotes but less likely to visit hyperlinked sources.

The study research conducted by Jemielniak et al.[17] ranked the most influential medical journals based on their representation in English-language Wikipedia medical pages and evaluated the time taken for articles to be cited

---

[12] W. Arroyo-Machado, D. Torres-Salinas, E. Herrera-Viedma & E. Romero-Frías, *Science through Wikipedia: A novel representation of open knowledge through co-citation networks*, "PloS one", 15, 2020, e0228713.

[13] R. Tomaszewski & K.I. MacDonald, *A study of citations to Wikipedia in scholarly publications*, "Science & technology libraries", 35, 2016, pp. 246–261.

[14] T. Piccardi, M. Redi, G. Colavizza & R. West, *Quantifying engagement with citations on Wikipedia*, in: *Proceedings of The Web Conference 2020*, New York 2020, pp. 2365–2376.

[15] C.-C. Chen & C. Roth, *{{Citation needed}}: The dynamics of referencing in Wikipedia in Proceedings of the eighth annual international symposium on wikis and open collaboration*, New York 2012, pp. 1–4.

[16] L.A. Maggio, R.M. Steinberg, T. Piccardi & J.M. Willinsky, *Reader engagement with medical content on Wikipedia*, "Elife", 9, 2020, e52426.

[17] D. Jemielniak, G. Masukume & M. Wilamowski, *The most influential medical journals according to Wikipedia: quantitative analysis*, "Journal of medical Internet research", 21, 2019, e11429.

in Wikipedia. The findings suggest that traditional high-impact medical and multidisciplinary journals are extensively cited, and it takes about 90 days for a citation to be used in Wikipedia. Other research by Yang and Colavizza[18] explored the relationship between Wikipedia articles and scientific journal articles, discovering that most journal articles cited belong to STEM fields. Wikipedia's biographies are noted to play an important role in connecting STEM fields with the humanities.

The paper authored by Nicholson et al.[19] analyzed more than 100 million references to over 800,000 scientific articles. Authors of the research measured the quality of scientific citations in Wikipedia using machine learning algorithms, and found wide variability in contradicting or supporting evidence found in Wikipedia articles.

Focusing on medical citations, other research by Maggio et al.[20] explored Wikipedia's role as a gateway to biomedical research, providing benchmarks for the relative distribution and referral rate of citations, which could be used to assess strategies for changing referral patterns and enhancing learning. Another study conducted by Koppen et al.[21] compared references from drug-related Wikipedia articles and a drug information database. It concluded that Wikipedia may not be a reliable, up-to-date resource for drug safety information.

In the work written by Singh et al.[22] authors presented a comprehensive dataset of 29.3 million citations extracted from 6.1 million English Wikipedia articles, classifying them into books, journal articles, or web content. With four million citations having known identifiers, they found that 6.7% of Wikipedia articles cited at least one journal article with a DOI and that only 2% of articles with a DOI were cited by Wikipedia. The authors released their code and data for further community analysis and usage.

---

[18] P. Yang & G. Colavizza, *A Map of Science in Wikipedia*, in: *Companion Proceedings of the Web Conference 2022*, New York 2022, pp. 1289–1300.

[19] J.M. Nicholson et al., *Measuring the quality of scientific references in Wikipedia: an analysis of more than 115M citations to over 800 000 scientific articles*, "The FEBS journal", 288, 2021, pp. 4242–4248.

[20] L.A. Maggio et al., *Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia*, "PloS one", 12, 2017, e0190046.

[21] L. Koppen, J. Phillips & R. Papageorgiou, *Analysis of reference sources used in drug-related Wikipedia articles*, "Journal of the Medical Library Association: JMLA", 103, 2015, p. 140.

[22] 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, pp. 1–19.

Those and other works showed that including scientific sources in references within Wikipedia articles is becoming a more prevalent practice.[23] References often lead to open-access publications[24] and newly published articles in scientific journals.[25] Some publications described the results of research on information sources in Wikipedia articles on a specific topic. For example, one of the studies focused on scientific research related to COVID19 discovered that only approximately 2% of the scientific works published during that time were used as a source of information.[26] Another work provided an analysis of the changes in the importance of sources during the COVID-19 pandemic.[27]

In the case of news websites, such sources of information are frequently used in Wikipedia, and there is a method which suggests new references automatically for a selected piece of information.[28] Particularly popular are references to web sources about recent content or life events.[29]

In a number of publications,[30] the authors addressed the issue of finding and updating news citations for Wikipedia statements using a two-stage supervised approach. They propose a classifier to determine the type of citation needed and develop a news citation algorithm. Their evaluation on 20 million real-world news articles shows promising results for precision and scalability.

Another study conducted by Redi et al.[31] provided an empirical characterization of how Wikipedia cites external sources to comply with its own verifiability guidelines. The authors construct a taxonomy of citation reasons, crowdsource a dataset of annotated Wikipedia sentences, and design algorithmic models to determine if a statement requires a citation and predict

---

[23] W. Lewoniewski, K. Węcel & W. Abramowicz, *Modeling Popularity...*

[24] M. Teplitskiy, G. Lu & E. Duede, op. cit.

[25] D. Jemielniak, G. Masukume & M. Wilamowski, op., cit.

[26] G. Colavizza, *COVID-19 research in Wikipedia*, "Quantitative Science Studies", 1, 2020, pp. 1349–1380.

[27] W. Lewoniewski, K. Węcel & W. Abramowicz, *Reliability in Time: Evaluating the Web Sources of Information on COVID-19 in Wikipedia across Various Language Editions from the Beginning of the Pandemic* presented at Wiki WorkShop 2022 (held virtually at The Web Conference 2022) on 25 April 2022.

[28] B. Fetahu, K. Markert, W. Nejdl & A. Anand, *Finding news citations for Wikipedia*, in: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, New York 2016, pp. 337– 346.

[29] T. Piccardi, M. Redi, G. Colavizza & R. West, op. cit.

[30] B. Fetahu, K. Markert, W. Nejdl & A. Anand, op. cit.

[31] M. Redi, B. Fetahu, J. Morgan & D. Taraborelli, *Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability*, in: *The World Wide Web Conference*, 2019, pp. 1567–1578.

reason for the citation. Their models are evaluated across different article classes and external datasets.

Other research[32] evaluated the adherence to Wikipedia's citation policy in a subset of history articles, revealing a lack of verifiability and limited reference diversity. The authors suggest raising awareness of the social context of text production and use among a new generation of Wikipedia editors and users to improve citation quality.

In the study conducted by Piscopo et al.[33] the relationship between Wikidata and Wikipedia was explored by analyzing the external references used. The results showed that while only a small number of sources are directly reused, references often point to the same domain, and Wikidata uses less Anglo-American-centered sources. The authors suggest diversifying sources and greater involvement from various language communities.

Thompson and Hanley[34] demonstrated that incorporating ideas into Wikipedia leads to those ideas being used more in the scientific literature. They provide correlational and causal evidence through a randomized control trial, showing that new Wikipedia articles result in changes in related scientific journal articles and that adding scientific articles as references to Wikipedia increases their academic citations.

Another related research by Pooladian and Borrego[35] examined the methodological issues of measuring Wikipedia citations in Library and Information Science literature. The authors of the study found a low citation rate (less than 3%) and highlighted the limitations of using Wikipedia citations for research evaluation due to the lack of standardization and incompleteness of references. They also observed that less than one-third of Wikipedia citations link to open access sources, potentially underestimating their availability. Another work by Kaffee and Elsahar[36] focused on the process of selecting references for new Wikipedia articles from an editor's perspective, mapping out their workflow.

---

[32] B. Luyt & D. Tan, *Improving Wikipedia's credibility: References and citations in a sample of history articles*, "Journal of the American Society for Information Science and Technology", 61, 2010, pp. 715–722.

[33] A. Piscopo et al., *What do wikidata and wikipedia have in common? An analysis of their use of external references*, in: *Proceedings of the 13th International Symposium on Open Collaboration*, New York 2017, pp. 1–10.

[34] N. Thompson & D. Hanley, *Science is shaped by Wikipedia: evidence from a randomized control trial*, 2018.

[35] A. Pooladian & Á. Borrego, *Methodological issues in measuring citations in Wikipedia: a case study in Library and Information Science*, "Scientometrics", 113, 2017, pp. 455–464.

[36] L.-A. Kaffee & H. Elsahar, *References in Wikipedia: The editors' perspective*, in: *Companion Proceedings of the Web Conference 2021*, New York 2021, pp. 535–538.

Another relevant study conducted by Hernandez et al.[37] investigated the references used in Wikipedia's featured articles across English, French, Portuguese, and Spanish languages. Its authors discovered cultural differences in editorial practices and citation patterns, identifying a coverage gap in the Citoid tool for automatically generating citations from URLs. The authors attempted to estimate the width and nature of this gap before and after the implementation of the Web2cit tool.

Baigutanova et al.[38] focused on analyzing the reliability of Wikipedia through its references. It introduced a "Citation Detective" tool which can calculate the reference need score and has observed a decrease in the score over the past decade. It proposes a strategy to pair novice and experienced editors to enhance reference quality.

Another study by Alnajrani et al.[39] introduces a rule-based approach to classify Wikipedia references, aiding researchers in searching for reliable references. The proposed approach achieved an average accuracy of 97.5%.

Another recent publication written by Zagorova et al.[40] presents a dataset of the history of all references in the English Wikipedia until June 2019. The study observed a persistent effort to improve the quality of references and suggests its importance in the design of relevance indexes for altmetrics.

Liu et al. in the recent study[41] proposed a method for generating Wikipedia articles as a multi-document summarization of source documents. The paper introduces a decoder-only architecture that can handle very long sequences and generate fluent, coherent multi-sentence paragraphs and entire Wikipedia articles.

Work by Petroni et al.[42] introduced an AI-based system to identify Wikipedia citations that are unlikely to support their claims and recommend better ones from the web. The system, trained on existing Wikipedia references,

---

[37] N.A. Hernández, G. del Rio & D. de la Hera, *Insights on the references of Wikipedia's featured articles in English, French, Portuguese and Spanish*, https://wikiworkshop.org/2022/papers/WikiWorkshop2022_paper_18.pdf [Accessed 19 June 2023].

[38] A. Baigutanova et al., *Longitudinal Assessment of Reference Quality on Wikipedia*, in: *Proceedings of the ACM Web Conference 2023*, New York 2023, pp. 2831–2839.

[39] B. Alnajrani, A. Alghamdi, M. Alotaibi, S.A. Atta-Ur-rahman & M. Nabil, *A Novel Approach to Wikipedia References Classification*, "ICIC Express Letters, Part B: Applications", 13, 2022, pp. 1321–1330.

[40] O. Zagorova, R. Ulloa, K. Weller & F. Flöck, *"I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics*, "Quantitative Science Studies", 3, 2022, pp. 147–173.

[41] P.J. Liu et al., *Generating Wikipedia by summarizing long sequences*, 2018.

[42] F. Petroni et al., *Improving Wikipedia Verifiability with AI*, 2022.

showed promising results, indicating that AI could be used to improve the verifiability of Wikipedia.

A 2023 study[43] discusses a method for identifying reliable internet sources about companies based on Wikipedia articles and their references. It highlights the existence of universal sources shared by many languages, and each language having its own specific sources of information.

This study focused on Wikipedia references that are related to scientific publications. Such analysis were conducted for 310 language versions of Wikipedia.

3. References extraction

A reference within an article typically consists of two components. Initially, segments of text that are derived from or directly quoted from an external source are indicated with an in-text citation, often manifested as a superscript number. Subsequently, a comprehensive reference list is needed, supplying full, formatted information about the source.

Some individual item of this list of references may be used more than once in the content of a Wikipedia article. This means that one Wikipedia article can have several references to the same source. Therefore, each reference (even to the same source within a single source) will count in the statistics shown in this section.

The list of references usually has a separate section within the Wikipedia article (in the English Wikipedia it is called "References", in Polish – "Przypisy"), which is usually placed on the bottom of the article.

However, it should be added that within some Wikipedia articles there may be other sections that have a list of sources of information related to the described entity. However, such sections do not point to specific places within the Wikipedia article where these sources appear to support specific information. Therefore, the presence of some sources on a given list may not be directly related to the content of the Wikipedia article. Therefore, such sources of information do not have textual references, and therefore they were not taken into account in the analysis.

To extract references from 310 Wikipedia languages, this work used Wikimedia dumps of Wikipedia in HTML as of June 2023.[44] During the

[43] W. Lewoniewski, K. Węcel & W. Abramowicz, *Companies in Multilingual Wikipedia: Articles Quality and Important Sources of Information*, in: *Information Technology for Management: Approaches to Improving Business and Society*, eds. E. Ziemba, W. Chmielarz & J. Wątróbski, Cham 2023, pp. 48–67.

[44] *Wikimedia Enterprise HTML Dumps*, https://dumps.wikimedia.org/other/enterprise_html/ [Accessed 19 June 2023].

extraction process some data quality issues were detected. For example, information about some articles were missing or appeared several times (was repeated). Therefore, an additional analysis of these data in terms of consistency was additionally performed. Information about missing Wikipedia articles was taken from other sources (such as the Wikipedia API[45]).

309 language versions of Wikipedia contained 61,218,277 Wikipedia articles in June 2023. However, not all Wikipedia articles contained references: 69.23% of them had at least one reference. The total number of references that were found in those articles is 332,424,439.

In order to find scientific references, the presence of digital object identifiers (DOI) in the metadata of those references was analyzed. As a result, 10,191,992 scientific references were found. The total number of publications with unique DOI identifiers in these references is 1,974,052.

4. References analysis

Analysis for some of the most developed language versions of Wikipedia is presented in the subsections below. Results for other languages are available in the supplementary materials.[46]

Each of the following subsections contained also list of the 10 most important scientific sources based on the number of their appearances in the Wikipedia references in the considered language version. To obtain metadata for such sources (scientific publications), OpenAlex was used. It is a free and open catalog of the world's scholarly papers, researchers, journals, and institutions [42]. However, metadata about some important scientific publications were unavailable in this open catalog and some of the missing data were extracted from Wikipedia articles or directly from publisher websites. More extensive lists of important sources are also available in the supplementary materials.

4.1. Arabic Wikipedia

There were 1,207,115 articles in the Arabic Wikipedia, and 91.3% of them had at least one reference. These articles had in total 7,469,900 references and 3.82% of them had DOI identifiers. Overall, there were 160,694 unique scientific references with DOIs in the Arabic Wikipedia. Table 1 presents the top 10 most important scientific sources in the Arabic Wikipedia.

[45] English Wikipedia. *MediaWiki API help*, https://en.wikipedia.org/w/api.php [Accessed 19 June 2023].

[46] data.lewoniewski.info. *Supplementary materials for this research*, https://data.lewoniewski.info/importantsources [Accessed 19 June 2023].

Table 1. The most important scientific sources in Arabic Wikipedia

Publications	Refs
J. R. Masiero et al. (2011). **MAIN BELT ASTEROIDS WITH WISE /NEOWISE. I. PRELIMINARY ALBEDOS AND DIAMETERS**. *The Astrophysical Journal*. DOI: 10.1088/0004-637x/741/2/68	10,646
(2016). **IOC World Bird List 6.3**. DOI: 10.14344/ioc.ml.6.3	6,722
(2010). **Allgemeines Künstlerlexikon Online**. DOI: 10.1515/akl	1,494
Christopher S. Kochanek et al. (2001). **The K-Band Galaxy Luminosity Function**. *The Astrophysical Journal*. DOI: 10.1086/322488	1,319
Carl von Linné (1754). **Genera plantarum :eorumque characteres naturales secundum numerum, figuram, situm, et proportionem omnium fructificationis partium /by Caroli Linnæi**. DOI: 10.5962/bhl.title.746	1,276
Christopher P. Ahn et al. (2012). **THE NINTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY: FIRST SPECTROSCOPIC DATA FROM THE SDSSIII BARYON OSCILLATION SPECTROSCOPIC SURVEY**. *Astrophysical Journal Supplement Series*. DOI: 10.1088/0067-0049/203/2/21	1,125
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	837
(2017). **IOC World Bird List 7.2**. DOI: 10.14344/ioc.ml.7.2	750
(2016). **IOC World Bird List 6.4**. DOI: 10.14344/ioc.ml.6.4	667
(2009). **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III**. *Botanical Journal of the Linnean Society*. DOI: 10.1111/j.1095-8339.2009.00996.x	540

4.2. German Wikipedia

There were 2,805,540 articles in the German Wikipedia, and 61.19% of them had at least one reference. These articles had in total 14,128,057 references and 1.9% of them had DOI identifiers. Overall, there were 140,792 unique scientific references with DOIs in the German Wikipedia. Table 2 presents the top 10 most important scientific sources in the German Wikipedia.

Table 2. The most important scientific sources in German Wikipedia

Publications	Refs
P. P. Rohrlach (2018). **Historisches Ortslexikon für die Altmark**. *Veröffentlichungen des Brandenburgischen Landeshauptarchivs*. DOI: 10.35998/9783830522355	3,207
Laurence N. Warr (2021). **IMA–CNMNC approved mineral symbols**. *Mineralogical Magazine*. DOI: 10.1180/mgm.2021.43	1,812

Publications	Refs
Lotte Burkhardt (2018). **Verzeichnis eponymischer Pflanzennamen – Erweiterte Edition. Index of Eponymic Plant Names – Extended Edition. Index de Noms éponymiques des Plantes – Édition augmentée**. DOI: 10.3372/epolist2018	1,185
Kramida, A., Ralchenko, Yu., Reader, J. (2022). **NIST Atomic Spectra Database**. *National Institute of Standards and Technology, Gaithersburg, MD*. DOI: 10.18434/t4w30f	545
GBIF Secretariat (2022). **GBIF Backbone Taxonomy**. *Checklist dataset*. DOI: 10.15468/39omei	416
Eli Eisenberg, Erez Y. Levanon (2003). **Human housekeeping genes are compact**.*Trends in Genetics*. DOI: 10.1016/s0168-9525(03)00140-9	330
P. Ellerholz, H. Lodemann, H. von Wedell (1879). **1. Band: Das Königreich Preussen, Lfg.1: Die Provinz Brandenburg**. *Humboldt-Universität zu Berlin eBooks*. DOI: 10.18452/377	281
Nadja Korotkova et al. (2021). **Cactaceae at Caryophyllales.org – a dynamic online species-level taxonomic backbone for the family**. *Willdenowia*. DOI: 10.3372/wi.51.51208	258
(2006). Ärzte **Lexikon**. *Springer eBooks*. DOI: 10.1007/978-3-540-29585-3	251
Mikio Ishiwatari, Federica Ranghieri (2014). **Learning from Megadisasters: Lessons from the Great East Japan Earthquake**. DOI: 10.1596/978-1-4648-0153-2	243

4.3. English Wikipedia

There were 6,662,867 articles in the English Wikipedia, and 84.19% of them had at least one reference. These articles had in total 77,836,866 references and 4.6% of them had DOI identifiers. Overall, there were 1,473,305 unique scientific references with DOIs in the English Wikipedia. Table 3 presents the top 10 most important scientific sources in the English Wikipedia.

Table 3. The most important scientific sources in English Wikipedia

Publications	Refs
Gaia Collaboration et al. (2018). **Gaia Data Release 2**. DOI: 10.1051/0004-6361/201833051	11,961
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	10,218
Ethel M. Doidge (1950). **Titles of Works, not included in the General bibliography**. *Bothalia*. DOI: 10.4102/abc.v5i1.1880	7,705

Publications	Refs
Erik H. Anderson, Charles Francis (2011). **XHIP: An extended hipparcos compilation**. *Astronomy Letters*. DOI: 10.1134/s1063773712050015	6,707
World Spider Catalog Version 20.0. Natural History Museum Bern. DOI: 10.24436/2	6,097
A. Vallenari, A.G.A. Brown, T. Prusti (2023). **Gaia Data Release 3. Summary of the content and survey properties**. *Astronomy and Astrophysics*. DOI: 10.1051/00046361/202243940	5,240
A. K. Mainzer et al. (2011). **NEOWISE STUDIES OF SPECTROPHOTOMETRICALLY CLASSIFIED ASTEROIDS: PRELIMINARY RESULTS**. *The Astrophysical Journal*. DOI: 10.1088/0004-637x/741/2/90	3,952
Fumihiko Usui et al. (2011). **Asteroid Catalog Using AKARI: AKARI/IRC Mid-Infrared Asteroid Survey**. *Publications of the Astronomical Society of Japan*. DOI: 10.1093/pasj/63.5.1117	3,677
A. G. A. Brown et al. (2021). **Gaia Early Data Release 3**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361/202039657	3,272
J. R. Masiero et al. (2012). **PRELIMINARY ANALYSIS OF WISE /NEOWISE 3BAND CRYOGENIC AND POST-CRYOGENIC OBSERVATIONS OF MAIN BELT ASTEROIDS**. *The astrophysical journal*. DOI: 10.1088/2041-8205/759/1/l8	2,382

4.4. Spanish Wikipedia

There were 1,805,077 articles in the Spanish Wikipedia and 79.31% of them had at least one reference. These articles had in total 12,164,596 references and 2.18% of them had DOI identifiers. Overall, there were 150,256 unique scientific references with DOIs in the Spanish Wikipedia. Table 4 presents the top 10 most important scientific sources in the Spanish Wikipedia.

Table 4. The most important scientific sources in Spanish Wikipedia

Publications	Refs
Daniel A. Potter et al. (2007). **Phylogeny and classification of Rosaceae**. *Plant Systematics and Evolution*. DOI: 10.1007/s00606-007-0539-9	1,679
Kevin J. Burns et al. (2014). **Phylogenetics and diversification of tanagers (Passeriformes: Thraupidae), the largest radiation of Neotropical songbirds**. *Molecular Phylogenetics and Evolution*. DOI: 10.1016/j.ympev.2014.02.006	478
Jan I. Ohlson et al. (2013). **Phylogeny and classification of the New World suboscines (Aves, Passeriformes)**. *Zootaxa*. DOI: 10.11646/zootaxa.3613.1.1	300

Publications	Refs
Elizabeth P. Derryberry et al. (2011). **LINEAGE DIVERSIFICATION AND MORPHOLOGICAL EVOLUTION IN A LARGE-SCALE CONTINENTAL RADIATION: THE NEOTROPICAL OVENBIRDS AND WOODCREEPERS (AVES: FURNARIIDAE)**. *Evolution*. DOI: 10.1111/j.1558--5646.2011.01374.x	270
Kevin J. Burns, Philip Unitt, Nicholas A. Mason (2016). **A genus-level classification of the family Thraupidae (Class Aves: Order Passeriformes)**. *Zootaxa*. DOI: 10.11646/zootaxa.4088.3.2	241
Jose G. Tello et al. (2009). **Phylogeny and phylogenetic classification of the tyrant flycatchers, cotingas, manakins, and their allies (Aves: Tyrannides)**. *Cladistics*. DOI: 10.1111/j.1096-0031.2009.00254.x	189
F. Keith Barker et al. (2015). **New insights into New World biogeography: An integrated view from the phylogeny of blackbirds, cardinals, sparrows, tanagers, warblers, and allies**. *The Auk*. DOI: 10.1642/auk-14-110.1	151
Morton L. Isler, Gustavo A. Bravo, Robb T. Brumfield (2013). **Taxonomic revision of Myrmeciza (Aves: Passeriformes: Thamnophilidae) into 12 genera based on phylogenetic, morphological, behavioral, and ecological data**. *Zootaxa*. DOI:10.11646/zootaxa.3717.4.3	141
(2012). **Compendium of botanicals reported to contain naturally occuring substances of possible concern for human health when used in food and food supplements**. *EFSA Journal*. DOI: 10.2903/j.efsa.2012.2663	137
Rosie Woodroffe, Joshua R. Ginsberg (1999). **Conserving the African wild dog Lycaon pictus. I. Diagnosing and treating causes of decline**. *Oryx*. DOI: 10.1046/j.13653008.1999.00052.x	136

4.5. French Wikipedia

There were 2,526,237 articles in the French Wikipedia and 67.05% of them had at least one reference. These articles had in total 19,026,898 references and 1.88% of them had DOI identifiers. Overall, there were 162,247 unique scientific references with DOIs in the French Wikipedia. Table 5 presents the top 10 most important scientific sources in the French Wikipedia.

Table 5. The most important scientific sources in French Wikipedia

Publications	Refs
Daniel Joly et al. (2010). **Les types de climats en France, une construction spatiale**. *Cybergeo*. DOI: 10.4000/cybergeo.23155	28,990
J. R. Masiero et al. (2011). **MAIN BELT ASTEROIDS WITH WISE /NEOWISE. I. PRELIMINARY ALBEDOS AND DIAMETERS**. *The Astrophysical Journal*. DOI: 10.1088/0004-637x/741/2/68	13,181

Publications	Refs
Gaia Collaboration et al. (2018). **Gaia Data Release 2**. DOI: 10.1051/0004-6361/201833051	4,919
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	2,790
A. P. Mahtessian (1998). **Groups of galaxies. III. Some empirical characteristics**. *Astrophysics*. DOI: 10.1007/bf03036100	1,411
J. R. Masiero et al. (2012). **PRELIMINARY ANALYSIS OF WISE /NEOWISE 3BAND CRYOGENIC AND POST-CRYOGENIC OBSERVATIONS OF MAIN BELT ASTEROIDS**. *The astrophysical journal*. DOI: 10.1088/2041-8205/759/1/l8	960
Erik H. Anderson, Charles Francis (2011). **XHIP: An extended hipparcos compilation**. *Astronomy Letters*. DOI: 10.1134/s1063773712050015	752
A. G. A. Brown et al. (2021). **Gaia Early Data Release 3**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361/202039657	628
N. N. Samus et al. (2017). **General catalogue of variable stars: Version GCVS 5.1**. *Astronomy Reports*. DOI: 10.1134/s1063772917010085	616
(2009). **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III**. *Botanical Journal of the Linnean Society*. DOI: 10.1111/j.1095-8339.2009.00996.x	568

4.6. Italian Wikipedia

There were 1,812,783 articles in the Italian Wikipedia and 53.19% of them had at least one reference. These articles had in total 8,637,915 references and 2.54% of them had DOI identifiers. Overall, there were 85,613 unique scientific references with DOIs in the Italian Wikipedia. Table 6 presents the top 10 most important scientific sources in the Italian Wikipedia.

Table 6. The most important scientific sources in Italian Wikipedia

Publications	Refs
(2016). **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV**. *Botanical Journal of the Linnean Society*. DOI: 10.1111/boj.12385	3,093
N. Tetzlaff, Ralph Neuhäuser, Markus M. Hohle (2010). **A catalogue of young runaway Hipparcos stars within 3 kpc from the Sun**. *Monthly Notices of the Royal Astronomical Society*. DOI: 10.1111/j.1365-2966.2010.17434.x	435
S. van den Bergh (1966). **A study of reflection nebulae**. *The Astronomical Journal*. DOI: 10.1086/109995	336
Gila Fatran, Naftali Greenwood (1994). **The "Working Group"**. *Holocaust and Genocide Studies*. DOI: 10.1093/hgs/8.2.164	308

Publications	Refs
Frank C. Hawthorne et al. (2012). **Nomenclature of the amphibole supergroup**. *American Mineralogist*. DOI: 10.2138/am.2012.4276	302
Stewart Sharpless (1959). **A Catalogue of H II Regions**. *Astrophysical Journal Supplement Series*. DOI: 10.1086/190049	245
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	224
Donald R. Davis, Patricia Gentili (2003). **Andesianidae, a new family of monotrysian moths (Lepidoptera: Andesianoidea) from austral South America**. *Invertebrate Systematics*. DOI: 10.1071/is02006	206
Leo Blitz, M. Fich, Antony A. Stark (1982). **Catalog of CO radial velocities toward galactic H II regions**. *Astrophysical Journal Supplement Series*. DOI: 10.1086/190795	203
Mark W. Chase et al. (2015). **An updated classification of Orchidaceae**. *Botanical Journal of the Linnean Society*. DOI: 10.1111/boj.12234	198

4.7. Japanese Wikipedia

There were 1,375,674 articles in the Japanese Wikipedia and 67.05% of them had at least one reference. These articles had in total 14,188,621 references and 1.93% of them had DOI identifiers. Overall, there were 118,507 unique scientific references with DOIs in the Japanese Wikipedia. Table 7 presents the top 10 most important scientific sources in the Japanese Wikipedia.

Table 7. The most important scientific sources in Japanese Wikipedia

Publications	Refs
Japanese National Railways (1966). *List of Stations as of March 1966*. DOI: 10.11501/1873236	1,545
Shin-ichiro Kawada et al. (2018). **Sekai Honyurui Hyojun Wamei Mokuroku**. *Honyurui Kagaku (Mammalian Science)*. DOI: 10.11238/mammalianscience.58.s1	414
Javier Ortega-Hernández, Ralf Janssen, Graham E. Budd (2017). **Origin and evolution of the panarthropod head – A palaeobiological and developmental perspective**. DOI: 10.1016/j.asd.2016.10.011	293
Gaia Collaboration et al. (2018). **Gaia Data Release 2**. DOI: 10.1051/0004-6361/201833051	290
Peter Van Roy, Allison C. Daley, Derek E. G. Briggs (2015). **Anomalocaridid trunk limb homology revealed by a giant filter-feeder with paired flaps**. *Nature*. DOI: 10.1038/nature14256	284

Publications	Refs
J. Moysiuk, Jean-Bernard Caron (2019). **A new hurdiid radiodont from the Burgess Shale evinces the exploitation of Cambrian infaunal food sources**. *Proceedings of The Royal Society B: Biological Sciences*. DOI: 10.1098/rspb.2019.1079	271
J. Moysiuk, Jean-Bernard Caron (2022). **A three-eyed radiodont with fossilized neuroanatomy informs the origin of the arthropod head and segmentation**. *Current Biology*. DOI: 10.1016/j.cub.2022.06.027	263
GBIF Secretariat (2022). **GBIF Backbone Taxonomy**. *Checklist dataset*. DOI: 10.15468/39omei	238
Cédric Aria et al. (2020). **Fossils from South China redefine the ancestral euarthropod body plan**. *BMC Evolutionary Biology*. DOI: 10.1186/s12862-019-1560-7	219
J. Moysiuk, Jean-Bernard Caron (2021). **Exceptional multifunctionality in the feeding apparatus of a mid-Cambrian radiodont**. *Paleobiology*. DOI: 10.1017/pab.2021.19	218

4.8. Polish Wikipedia

There were 1,569,798 articles in the Polish Wikipedia and 65.84% of them had at least one reference. These articles had in total 8,619,932 references and 1.67% of them had DOI identifiers. Overall, there were 45,739 unique scientific references with DOIs in the Polish Wikipedia. Table 8 presents the top 10 most important scientific sources in the Polish Wikipedia.

Table 8. The most important scientific sources in Polish Wikipedia

Publications	Refs
R. Barbon et al. (1999). **The Asiago Supernova Catalogue - 10 years after**. *Astronomy & astrophysics. Supplement series*. DOI: 10.1051/aas:1999404	23,385
Michael T. Ruggiero et al. (2015). **A Higher Level Classification of All Living Organisms**. *PLOS ONE*. DOI: 10.1371/journal.pone.0119248	13,285
David J. Mabberley (2017). **Mabberley's Plant-book**. DOI: 10.1017/9781316335581	1,410
(2016). **A community-derived classification for extant lycophytes and ferns**. *Journal of Systematics and Evolution*. DOI: 10.1111/jse.12229	616
(1998). **Flowering Plants · Monocotyledons**. *Springer eBooks*. DOI: 10.1007/978-3662-03533-7	526
Maarten J. M. Christenhusz et al. (2011). **A new classification and linear sequence of extant gymnosperms**. *Phytotaxa*. DOI: 10.11646/phytotaxa.19.1.3	398

Publications	Refs
Sônia A. Casari (2008). **A phylogenetic study of the subtribe Dicrepidiina (Elateridae, Elaterinae, Ampedini)**. *Revista Brasileira De Entomologia*. DOI: 10.1590/s00855262008000200002	389
Klaas Dijkstra et al. (2014). **Redefining the damselfly families: a comprehensive molecular phylogeny of Zygoptera (Odonata)**. *Systematic Entomology*. DOI: 10.1111/syen.12035	248
Frank M Hull (1973). **Bee flies of the world: the genera of the family Bombyliidae**. DOI: 10.5962/bhl.title.48406	217
Patrice Bouchard et al. (2011). **Family-Group Names In Coleoptera (Insecta)**. *ZooKeys*. DOI: 10.3897/zookeys.88.807	194

4.9. Portuguese Wikipedia

There were 1,103,469 articles in the Portuguese Wikipedia and 81.62% of them had at least one reference. These articles had in total 7,458,232 references and 2.67% of them had DOI identifiers. Overall, there were 105,864 unique scientific references with DOIs in the Portuguese Wikipedia. Table 9 presents the top 10 most important scientific sources in the Portuguese Wikipedia.

Table 9. The most important scientific sources in Portuguese Wikipedia

Publications	Refs
Timothy D. Morton et al. (2016). **FALSE POSITIVE PROBABILITIES FOR ALL KEPLER OBJECTS OF INTEREST: 1284 NEWLY VALIDATED PLAN-ETS AND 428 LIKELY FALSE POSITIVES**. *The Astrophysical Journal*. DOI: 10.3847/0004-637x/822/2/86	1,187
Gaia Collaboration et al. (2018). **Gaia Data Release 2**. DOI: 10.1051/0004-6361/201833051.	925
Jason F. Rowe et al. (2014). **VALIDATION OF KEPLER 'S MULTIPLE PLANET CANDIDATES. III. LIGHT CURVE ANALYSIS AND ANNOUNCEMENT OF HUNDREDS OF NEW MULTI-PLANET SYSTEMS**. *The Astrophysical Journal*. DOI: 10.1088/0004-637x/784/1/45	740
A. G. A. Brown et al. (2021). **Gaia Early Data Release 3**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361/202039657	450
Andrea L. Lacaita, Sergio Ortolani, Valerio Nascimbeni (2015). **Age consistency between exoplanet hosts and field stars**. *Astronomy and Astrophysics*. DOI: 10.1051/00046361/201527297	348
Peter P. Eggleton, Andrei Tokovinin (2008). **A catalogue of multiplicity among bright stellar systems**. *Monthly Notices of the Royal Astronomical Society*. DOI: 10.1111/j.13652966.2008.13596.x	250

Publications	Refs
(2009). **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III**. *Botanical Journal of the Linnean Society*. DOI: 10.1111/j.1095-8339.2009.00996.x	190
N. Tetzlaff, Ralph Neuhäuser, Markus M. Hohle (2010). **A catalogue of young runaway Hipparcos stars within 3 kpc from the Sun**. *Monthly Notices of the Royal Astronomical Society*. DOI: 10.1111/j.1365-2966.2010.17434.x	189
Vítor de Q. Piacentini et al. (2015). **Annotated checklist of the birds of Brazil by the Brazilian Ornithological Records Committee / Lista comentada das aves do Brasil pelo Comitê Brasileiro de Registros Ornitológicos**. DOI: 10.1007/bf03544294	168
Raquel Varela (2012). **"Um, dois, três MFA...": o Movimento das Forças Armadas na Revolução dos Cravos - do prestígio à crise**. DOI: 10.1590/s0102-01882012000100019	150

4.10. Russian Wikipedia

There were 1,919,023 articles in the Russian Wikipedia and 71.97% of them had at least one reference. These articles had in total 15,531,208 references and 1.86% of them had DOI identifiers. Overall, there were 117,081 unique scientific references with DOIs in the Russian Wikipedia. Table 10 presents the top 10 most important scientific sources in the Russian Wikipedia.

Table 10. The most important scientific sources in Russian Wikipedia

Publications	Refs
R. Brent Tully, Hélène M. Courtois, Jenny G. Sorce (2016). **COSMICFLOWS-3**. *The Astronomical Journal*. DOI: 10.3847/0004-6256/152/2/50	4,151
R. Brent Tully (2015). **GALAXY GROUPS: A 2MASS CATALOG**. *The Astronomical Journal*. DOI: 10.1088/0004-6256/149/5/171	3,089
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	2,535
Pierre Kervella et al. (2019). **Stellar and substellar companions of nearby stars from Gaia DR2**. DOI: 10.1051/0004-6361/201834371	1,441
Aidan C. Crook et al. (2007). **Groups of Galaxies in the Two Micron All Sky Redshift Survey**. *The Astrophysical Journal*. DOI: 10.1086/510201	1,358
N. N. Samus et al. (2017). **General catalogue of variable stars: Version GCVS 5.1**. *Astronomy Reports*. DOI: 10.1134/s1063772917010085	1,348
(2021). **IOC World Bird List 11.2**. DOI: 10.14344/ioc.ml.11.2	940
(2010). **Allgemeines Künstlerlexikon Online**. DOI: 10.1515/akl	814

Publications	Refs
A. K. Mainzer et al. (2011). **NEOWISE STUDIES OF SPECTROPHOTOMETRICALLY CLASSIFIED ASTEROIDS: PRELIMINARY RESULTS**. *The Astrophysical Journal*. DOI: 10.1088/0004-637x/741/2/90	769
Gaia Collaboration et al. (2018). **Gaia Data Release 2**. DOI: 10.1051/0004-6361/201833051	705

4.11. Chinese Wikipedia

There were 1,357,954 articles in the Chinese Wikipedia and 65.39% of them had at least one reference. These articles had in total 7,853,665 references and 1.21% of them had DOI identifiers. Overall, there were 48,471 unique scientific references with DOIs in the Chinese Wikipedia. Table 11 presents the top 10 most important scientific sources in the Chinese Wikipedia.

Table 11. The most important scientific sources in Chinese Wikipedia

Publications	Refs
F.E. van Leeuwen (2007). **Validation of the new Hipparcos reduction**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361:20078357	1,398
(2022). **IOC World Bird List 12.1**. DOI: 10.14344/ioc.ml.12.1	902
Jason F. Rowe et al. (2014). **VALIDATION OF KEPLER 'S MULTIPLE PLANET CANDIDATES. III. LIGHT CURVE ANALYSIS AND ANNOUNCEMENT OF HUNDREDS OF NEW MULTI-PLANET SYSTEMS**. *The Astrophysical Journal*.DOI: 10.1088/0004-637x/784/1/45	737
Gaia Collaboration et al.(2018). **Gaia Data Release 2.** DOI: 10.1051/0004-6361/201833051	386
N. A. M. Rodger (2004). **Nelson, Horatio, Viscount Nelson (1758–1805), naval officer**. *Oxford Dictionary of National Biography*. DOI: 10.1093/ref:odnb/19877	332
A. G. A. Brown et al. (2021). **Gaia Early Data Release 3**. *Astronomy and Astrophysics*. DOI: 10.1051/0004-6361/202039657	278
Kakuzo Morimoto (1933). Nihon Nengo Taikan. Meguro Bookstore. DOI: 10.11501/1688	202
Zhu Xinxin et al. (2019). **Synopsis of Aristolochia L. and Isotrema Raf. (Aristolochiaceae) in China**. *Shengwu duoyangxing*. DOI: 10.17520/biods.2019183	198
Ricardo Betancur-R. et al. (2017). **Phylogenetic classification of bony fishes**. *BMC Evolutionary Biology*. DOI: 10.1186/s12862-017-0958-3	173
(2015). **Complex crises call for adaptable and durable capabilities**. *The military balance*. DOI: 10.1080/04597222.2015.996334	171

5. Limitations and Discussion

This research was able to analyze tens of millions of articles using Wikipedia backups for different language versions in HTML format. This makes it possible to analyze the references that the reader sees in the final version of each Wikipedia article. This means that you can include all the references in an article regardless of how they are placed. However, this approach is not without its drawbacks. Sometimes, analyzing the source code of an article in wiki notation can show which references were placed directly into the Wikipedia article in question, and which were placed automatically due to the inclusion of a template (which may appear in a large number of articles on the same topic). That is, in this case, when placing or replacing a reference (information source) in one template automatically changes the references in all Wikipedia articles that have such a template. Nevertheless, the most important thing is that users who read Wikipedia articles see such references and, in case of doubt, can correct such sources or submit a proposal to replace them.

Some references may be included automatically in Wikipedia articles. This is possible through specialized bots that any Wikipedia user can create, but their launch requires some form of consent from the community of a specific language version of this encyclopedia. This may mean that a relatively unpopular or new source of information may quickly gain a large number of citations. In other words, the situation is similar to that with templates, which was described in the previous paragraph. However, it is worth paying attention to the fact that even if a reference was posted automatically, Wikipedia readers may notice and correct such a reference. Moreover, to check which references were posted by bots, it is necessary to analyze the history of edits, which requires a much more complex analysis of a much larger data set, which is not the focus of this study.

As part of the analysis, it turned out that the most cited scientific sources of information in various language versions are publications focused on a similar area (primarily biology). Therefore, it would be worthwhile to carry out a deeper analysis using the division of articles into topics and the use of various models to evaluate sources (e.g. taking into account the popularity of articles containing references). These and other elements are planned for future research.

6. Conclusions and Future Work

This study analyzed 332 million references from 61 million Wikipedia articles across 309 language versions. The first goal of this research was to analyze how scientific sources of information are used within selected language versions of Wikipedia. Based on this, DOI identifiers were analyzed to identify citations of scientific publications. As a result, 10 million scientific references were found and the total number of unique scientific publications was almost 2 million. Depending on the language version, the share of scientific information sources varied. For example, in the English version of Wikipedia the share of such sources was 4.6%; in the Polish version – 1.67%.

A second goal of this study was to identify which sources of scientific information are most often cited in Wikipedia articles in selected languages. The research showed that scientific sources are a common thread across different language versions of Wikipedia. However, there is considerable variability in their use. This suggests that each language community may have unique norms and standards for sourcing information, reflecting broader sociolinguistic and cultural patterns.

The study also showed the most cited scientific publications in each of the considered language versions of Wikipedia. In order to collect more complete metadata about scientific publications, OpenAlex was used and proved its efficiency in analyzing large-scale data sets, such as Wikipedia. However, information about some important publications was unavailable in this open catalog and some of the missing data were extracted from Wikipedia articles or directly from publishers' websites. Additionally, the most cited scientific sources of information in various language versions were publications focused on a similar area (primarily biology).

The study provides a foundation for further research into the specifics of source usage across Wikipedia. These might include investigations into which scientific disciplines are most frequently cited, or how source usage varies across different types of articles. It is planned to used different models[47] to assess the importance of scientific sources of information on Wikipedia across topics and languages.

In future research, a more complex analysis of the use of scientific information sources on Wikipedia is planned. For this purpose, more complex source popularity assessment models can be used, which take into account not only the number of references in articles, but also the various quality measures

[47] W. Lewoniewski, *Identification of Important Web Sources of Information on Wikipedia across various Topics and Languages*, "Procedia Computer Science", 207, 2022, pp. 3290–3299.

utilized in Wikipedia articles. Additionally, it is planned to use the division of articles into topics in order to find important sources in various areas. In addition, the use of open knowledge bases such as OpenAlex may allow for a deeper analysis of the characteristics of the sources that are used in Wikipedia articles.

References

Alnajrani B., A. Alghamdi, M. Alotaibi, S.A. Atta-Ur-rahman & M. Nabil, *A Novel Approach to Wikipedia References Classification*, "ICIC Express Letters, Part B: Applications", 13, 2022, pp. 1321–1330.

Arroyo-Machado W., D. Torres-Salinas, E. Herrera-Viedma & E. Romero-Frías, *Science through Wikipedia: A novel representation of open knowledge through co-citation networks*, "PloS one", 15, 2020, e0228713. DOI: https://doi.org/10.1371/journal.pone.0228713

Baigutanova A. et al., *Longitudinal Assessment of Reference Quality on Wikipedia*, in: *Proceedings of the ACM Web Conference 2023*, New York: Association for Computing Machinery, 2023, pp. 2831–2839. DOI: https://doi.org/10.1145/3543507.3583218

BestRef. *Popularity and Reliability Assessment of Wikipedia Sources*, https://bestref.net [Accessed 19 June 2023].

Blumenstock J.E., *Size matters: word count as a measure of quality on Wikipedia*, "Proceedings of the 17th international conference on World Wide Web", 2008, pp. 1095–1096. DOI: https://doi.org/10.1145/1367497.1367673

Chen C.-C. & C. Roth, *{{Citation needed}}: The dynamics of referencing in Wikipedia*, in: *Proceedings of the eighth annual international symposium on wikis and open collaboration*, New York: Association for Computing Machinery, 2012, pp. 1–4. DOI: https://doi.org/10.1145/2462932.2462943

Colavizza G., *COVID-19 research in Wikipedia*, "Quantitative Science Studies", 1, 2020, pp. 1349–1380. DOI: https://doi.org/10.1162/qss_a_00080

data.lewoniewski.info. *Supplementary materials for this research*, https://data.lewoniewski.info/importantsources [Accessed 19 June 2023].

English Wikipedia. *MediaWiki API help*, https://en.wikipedia.org/w/api.php [Accessed 19 June 2023].

Fetahu B., K. Markert, W. Nejdl & A. Anand, *Finding news citations for Wikipedia*, in: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, New York: Association for Computing Machinery, 2016, pp. 337–346. DOI: https://doi.org/10.1145/2983323.2983808

Halfaker A. & R.S. Geiger, *ORES: Lowering barriers with participatory machine learning in Wikipedia*, "Proceedings of the ACM on Human-Computer Interaction", 4, 2020, pp. 1–37. DOI: https://doi.org/10.1145/3415219

Hernández N.A., G. del Rio & D. de la Hera, *Insights on the references of Wikipedia's featured articles in English, French, Portuguese and Spanish*, https://wikiworkshop.org/2022/papers/WikiWorkshop2022_paper_18.pdf [Accessed 19 June 2023].

Jemielniak D., G. Masukume & M. Wilamowski, *The most influential medical journals according to Wikipedia: quantitative analysis*, "Journal of medical Internet research", 21, 2019, e11429. DOI: https://doi.org/10.2196/11429

Kaffee L.-A. & H. Elsahar, *References in Wikipedia: The editors' perspective*, in: *Companion Proceedings of the Web Conference 2021*, New York: Association for Computing Machinery, 2021, pp. 535–538. DOI: https://doi.org/10.1145/3442442.3452337

Koppen L., J. Phillips & R. Papageorgiou, *Analysis of reference sources used in drug-related Wikipedia articles*, "Journal of the Medical Library Association: JMLA", 103, 2015, p. 140. DOI: https://doi.org/10.3163/1536-5050.103.3.007

Kousha K. & M. Thelwall, *Are Wikipedia citations important evidence of the impact of scholarly articles and books?*, "Journal of the Association for Information Science and Technology", 68, 2017, pp. 762–779. DOI: https://doi.org/10.1002/asi.23694

Lewoniewski W., *Identification of Important Web Sources of Information on Wikipedia across various Topics and Languages*, "Procedia Computer Science", 207, 2022, pp. 3290–3299. DOI: https://doi.org/10.1016/j.procs.2022.09.387

Lewoniewski W., Węcel K. & W. Abramowicz, *Analysis of references across Wikipedia languages*, in: *International Conference on Information and Software Technologies*, 2017, pp. 561–573. DOI: https://doi.org/10.1007/978-3-319-67642-5_47

Lewoniewski W., K. Węcel & W. Abramowicz, *Companies in Multilingual Wikipedia: Articles Quality and Important Sources of Information*, in: *Information Technology for Management: Approaches to Improving Business and Society*, eds. E. Ziemba, W. Chmielarz & J. Wątróbski, Cham: Springer Nature Switzerland, 2023, pp. 48–67. DOI: https://doi.org/10.1007/978-3-031-29570-6_3

Lewoniewski W., K. Węcel & W. Abramowicz, *Modeling Popularity and Reliability of Sources in Multilingual Wikipedia*, "Information", 11, 2020, p. 263. DOI: https://doi.org/10.3390/info11050263

Lewoniewski W., K. Węcel & W. Abramowicz, *Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics*, "Computers", 8, 2019. DOI: https://doi.org/10.20944/preprints201905.0144.v2

Lewoniewski W., K. Węcel & W. Abramowicz, *Reliability in Time: Evaluating the Web Sources of Information on COVID-19 in Wikipedia across Various Language Editions from the Beginning of the Pandemic* presented at Wiki WorkShop 2022 (held virtually at The Web Conference 2022) on 25 April 2022.

Liu P.J. et al., *Generating Wikipedia by summarizing long sequences*, 2018. DOI: https://doi.org/10.48550/arXiv.1801.10198

Luyt B. & D. Tan, *Improving Wikipedia's credibility: References and citations in a sample of history articles*, "Journal of the American Society for Information Science and Technology", 61, 2010, pp. 715–722. DOI: https://doi.org/10.1002/asi.21304

Maggio L.A. et al., *Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia*, "PloS one", 12, 2017, e0190046. DOI: https://doi.org/10.1371/journal.pone.0190046

Maggio L.A., R.M. Steinberg, T. Piccardi & J.M. Willinsky, *Reader engagement with medical content on Wikipedia*, "Elife", 9, 2020, e52426. DOI: https://doi.org/10.7554/eLife.52426

Nicholson J.M. et al., *Measuring the quality of scientific references in Wikipedia: an analysis of more than 115M citations to over 800 000 scientific articles*, "The FEBS journal", 288, 2021, pp. 4242–4248. DOI: https://doi.org/10.1111/febs.15608

OpenAlex. *OpenAlex: The open catalog to the global research system*, https://openalex.org [Accessed 19 June 2023].

Petroni F. et al., *Improving Wikipedia Verifiability with AI*, 2022. DOI: https://doi.org/10.21203/rs.3.rs-2116541/v1

Piccardi T., M. Redi, G. Colavizza & R. West, *Quantifying engagement with citations on Wikipedia*, in: *Proceedings of The Web Conference 2020*, New York: Association for Computing Machinery, 2020, pp. 2365–2376. DOI: https://doi.org/10.1145/3366423.3380300

Piscopo A. et al., *What do wikidata and wikipedia have in common? An analysis of their use of external references*, in: *Proceedings of the 13th International Symposium on Open Collaboration*, New York: Association for Computing Machinery, 2017, pp. 1–10. DOI: https://doi.org/10.1145/3125433.3125445

Pooladian A. & Á. Borrego, *Methodological issues in measuring citations in Wikipedia: a case study in Library and Information Science*, "Scientometrics", 113, 2017, pp. 455–464. DOI: https://doi.org/10.1007/s11192-017-2474-z

Priem J., H.A. Piwowar & B.M. Hemminger, *Altmetrics in the wild: Using social media to explore scholarly impact*, 2012. DOI: https://doi.org/10.48550/arXiv.1203.4745

Redi M., B. Fetahu, J. Morgan & D. Taraborelli, *Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability*, in: *The World Wide Web Conference*, 2019, pp. 1567–1578. DOI: https://doi.org/10.1145/3308558.3313618

Singh H., R. West & G. Colavizza, *Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia*, "Quantitative Science Studies", 2, 2021, pp. 1–19. DOI: https://doi.org/10.1162/qss_a_00105

Teplitskiy M., G. Lu, & E. Duede, *Amplifying the impact of open access: Wikipedia and the diffusion of science*, "Journal of the Association for Information Science and Technology", 68, 2017, pp. 2116–2127. DOI: https://doi.org/10.1002/asi.23687

Thompson N. & D. Hanley, *Science is shaped by Wikipedia: evidence from a randomized control trial*, 2018. DOI: https://doi.org/10.2139/ssrn.3039505

Tomaszewski R. & K.I. MacDonald, *A study of citations to Wikipedia in scholarly publications*, "Science & technology libraries", 35, 2016, pp. 246–261. DOI: https://doi.org/10.1080/0194262X.2016.1206052

Wikimedia Dumps. *Wikimedia Enterprise HTML Dumps*, https://dumps.wikimedia.org/other/enterprise_html/ [Accessed 19 June 2023].

WikiRank. *Quality and Popularity Assessment of Wikipedia Articles*, https://wikirank.net [Accessed 19 June 2023].

Yang P. & G. Colavizza, *A Map of Science in Wikipedia*, in: *Companion Proceedings of the Web Conference 2022*, New York: Association for Computing Machinery, 2022, pp. 1289–1300. DOI: https://doi.org/10.1145/3487553.3524925

Zagorova O., R. Ulloa, K. Weller & F. Flöck, *"I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics*, "Quantitative Science Studies", 3, 2022, pp. 147–173. DOI: https://doi.org/10.1162/qss_a_00171

Włodzimierz Lewoniewski

Najczęściej cytowane źródła informacji naukowej w artykułach Wikipedii w różnych językach

Streszczenie. W celu zapewnienia wiarygodności informacji zawartych w Wikipedii niezbędne jest wykorzystanie rzetelnych i godnych zaufania źródeł, które mogą być poddawane ocenie przez czytelników tej encyklopedii. Określenie rzetelności źródła może być jednak subiektywne, a dodatkowo zależne od przyjętej wersji językowej czy obranego tematu artykułu. W konsekwencji każda wersja językowa może być oparta na specyficznych dla autora wytycznych czy wskazówkach oceny wiarygodności źródeł. Część cytowań i odnośników umieszczanych w Wikipedii odwołuje się do zasobów naukowych, które zwykle uznawane są za bardziej wiarygodne niż strony internetowe, a to za sprawą ich rygorystycznych procedur recenzji naukowej czy publikacji przez uznanego wydawcę akademickiego. Tym samym można założyć, że przedstawione w tych źródłach dane zostały skrupulatnie sprawdzone przez specjalistów, co zapewnia wyższy poziom precyzyjności i wiarygodności. Celem badania przedstawionego w niniejszym artykule jest identyfikacja cytowań i odwołań do publikacji naukowych na podstawie analizy przypisów w artykułach Wikipedii. Analizie poddano 332 mln odwołań referencyjnych zamieszczonych w 61 mln artykułach Wikipedii w 309 wersjach językowych. Ponadto w badaniu wykorzystano otwarty katalog autorów i publikacji naukowych OpenAlex w celu identyfikacji ujednoliconych metadanych istotnych źródeł informacji wielojęzycznej Wikipedii.

Słowa kluczowe: Wikipedia, materiał źródłowy, wiki, naukowe źródła danych, data science, eksploracja (analiza) danych, OpenAlex.

Tekst wpłynął do Redakcji 20 czerwca 2023 roku.