Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia

Singh, H.; West, R.; Colavizza, G.

DOI
10.1162/qss_a_00105

Publication date
2021

Document Version
Final published version

Published in
Quantitative Science Studies

RESEARCH ARTICLE

Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia

Harshdeep Singh[1], Robert West[1], and Giovanni Colavizza[2]

[1]Data Science Laboratory, EPFL
[2]Institute for Logic, Language and Computation, University of Amsterdam

Keywords: citations, data, data set, Wikipedia

ABSTRACT

Wikipedia's content is based on reliable and published sources. To this date, relatively little is known about what sources Wikipedia relies on, in part because extracting citations and identifying cited sources is challenging. To close this gap, we release `Wikipedia Citations`, a comprehensive data set of citations extracted from Wikipedia. We extracted 29.3 million citations from 6.1 million English Wikipedia articles as of May 2020, and classified as being books, journal articles, or Web content. We were thus able to extract 4.0 million citations to scholarly publications with known identifiers—including DOI, PMC, PMID, and ISBN—and further equip an extra 261 thousand citations with DOIs from Crossref. As a result, we find that 6.7% of Wikipedia articles cite at least one journal article with an associated DOI, and that Wikipedia cites just 2% of all articles with a DOI currently indexed in the Web of Science. We release our code to allow the community to extend upon our work and update the data set in the future.

an open access journal

Check for updates

Citation: Singh, H., West, R., & Colavizza, G. (2020). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies*, *2*(1), 1–19. https://doi.org/10.1162/qss_a_00105

DOI: https://doi.org/10.1162/qss_a_00105

Received: 14 July 2020
Accepted: 23 November 2020

Corresponding Author: Giovanni Colavizza
g.colavizza@uva.nl

Handling Editor: Vincent Larivière

The MIT Press

1. INTRODUCTION

> *"Citations have several important purposes: to uphold intellectual honesty (or avoiding plagiarism), to attribute prior or unoriginal work and ideas to the correct sources, to allow the reader to determine independently whether the referenced material supports the author's argument in the claimed way, and to help the reader gauge the strength and validity of the material the author has used."*[1]

Wikipedia plays a fundamental role as a source of factual information on the Web: It is widely used by individual users as well as third-party services, such as search engines and knowledge bases (Lehmann, Isele et al., 2015; McMahon, Johnson, & Hecht, 2017).[2] Most importantly, Wikipedia is often perceived as a source of reliable information (Mesgari, Okoli et al., 2015). The reliability of Wikipedia's content has been debated, as anyone can edit it (Mesgari et al., 2015; Okoli, Mehdi et al., 2012). Nevertheless, the confidence that users and third-party services place on Wikipedia appears to be justified: Wikipedia's content is of general high quality and up to date, as shown by several studies over time (Colavizza, 2020; Geiger & Halfaker, 2013; Keegan,

[1] https://en.wikipedia.org/wiki/Citation (accessed January 3, 2020).
[2] https://en.wikipedia.org/wiki/Wikipedia:Statistics (accessed January 3, 2020).

Gergle, & Contractor, 2011; Kumar, West, & Leskovec, 2016; Okoli et al., 2012; Piscopo & Simperl, 2019; Priedhorsky, Chen et al., 2007).

To reach this goal, Wikipedia's verifiability policy mandates that "people using the encyclopedia can check that the information comes from a reliable source." A reliable source is defined, in turn, as a secondary and published, ideally scholarly, one.[3] Despite the community's best efforts to add all the needed citations, the majority of articles in Wikipedia might still contain unverified claims, in particular lower quality ones (Lewoniewski, Wecel et al., 2017). The citation practices of editors might also not be systematic at times (Chen & Roth, 2012; Forte, Andalibi et al., 2018). As a consequence, the efforts to expand and improve Wikipedia's verifiability through citations to reliable sources are increasing (Fetahu, Markert et al., 2016; Redi, Fetahu et al., 2019).

A crucial question to ask to improve Wikipedia's verifiability standards, as well as to better understand its dominant role as a source of information, is the following: *What sources are cited in Wikipedia?*

A high portion of citations to sources in Wikipedia refer to scientific or scholarly literature (Nielsen, Mietchen, & Willighagen, 2017), as Wikipedia is instrumental in providing access to scientific information and in fostering the public understanding of science (Heilman, Kemmann et al., 2011; Laurent & Vickers, 2009; Lewoniewski et al., 2017; Maggio, Steinberg et al., 2020; Maggio, Willinsky et al., 2017; Shafee, Masukume et al., 2017; Smith, 2020; Torres-Salinas, Romero-Frías, & Arroyo-Machado, 2019). Citations in Wikipedia are also useful for users browsing low-quality or underdeveloped articles, as they allow them to look for information outside of the platform (Piccardi, Redi et al., 2020). The literature cited in Wikipedia has been found to correlate positively with a journal's popularity, journal impact factor, and open access policy (Arroyo-Machado, Torres-Salinas et al., 2020; Nielsen, 2007; Teplitskiy, Lu, & Duede, 2017). Being cited in Wikipedia can also be considered as an "altmetric" indicator of impact in itself (Kousha & Thelwall, 2017; Sugimoto, Work et al., 2017). A clear influence of Wikipedia on scientific research has in turn been found (Thompson & Hanley, 2018), despite a general lack of reciprocity in acknowledging it as a source of information from the scientific literature (Jemielniak & Aibar, 2016; Tomaszewski & MacDonald, 2016). Nevertheless, the evidence on what scientific and scholarly literature is cited in Wikipedia is quite slim. Early studies point to a relative low coverage, indicating that between 1% and 5% of all published journal articles are cited in Wikipedia (Pooladian & Borrego, 2017; Priem, Piwowar, & Hemminger, 2012; Shuai, Jiang et al., 2013; Zahedi, Costas, & Wouters, 2014). These studies possess a number of limitations: They consider a by-now-dated version of Wikipedia, they use proprietary citation databases with limited coverage, or they only consider specific publishers (PLoS) and academic communities (computer science, library and information science). More recently, a novel data set has been released containing the edit history of all references in English Wikipedia, up till June 2019 (Zagovora, Ulloa et al., 2020). While the authors found a persistent increase of references equipped with some form of document identifier over time, they underline how relying on references with document identifiers is still not sufficient to capture all relevant publications cited from Wikipedia.

Answering the question of what exactly is cited in Wikipedia is challenging for a variety of reasons. First of all, editorial practices are not uniform, in particular across different language versions of Wikipedia: Citations are often given using citation templates somewhat liberally,[4] making it difficult to detect citations to the same source. Secondly, while some citations contain

[3] See respectively https://en.wikipedia.org/wiki/Wikipedia:Verifiability and https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources (accessed January 3, 2020).

[4] https://en.wikipedia.org/wiki/Wikipedia:Citation_templates (accessed January 3, 2020).

Figure 1. Example of citations in Wikipedia.

stable identifiers (e.g., DOIs), others do not. A recent study found that 4.42% of Wikipedia articles contain at least one citation with a DOI (Maggio et al., 2017): a low fraction that might indicate that we are missing a nonnegligible share of citations without identifiers. This is a significant limitation, as existing databases, such as Altmetrics, do provide Wikipedia citation metrics relying exclusively on citations with identifiers.[5] This in turn limits the scope of results relying on these data.

Our goal is to overcome these two challenges and expand upon previous work (Halfaker, Mansurov et al., 2018), by providing a data set of *all* citations from English Wikipedia, equipped with identifiers and including the code necessary to replicate and improve upon our work. The resulting data set is available on Zenodo (Singh, West, & Colavizza, 2020), while an accompanying repository contains code and further documentation.[6] By releasing a codebase that permits extraction of citations directly from Wikipedia data, we aim to address the following limitations found in previous work: the focus on specific publishers or scientific communities, the use of proprietary databases, and the lack of means to replicate and update results. Given how dynamic Wikipedia is, we deem it of importance to release a codebase to keep the `Wikipedia Citations` data set up to date for future reuse.

This article is organized as follows. We start by describing our pipeline focusing on its three main steps: citation template harmonization—to structure every citation in Wikipedia using the same schema; citation classification—to find citations to books and journal articles; and citation identifier look-up—to find identifiers such as DOIs. We subsequently evaluate our results, provide a description of the published data set, and conclude by highlighting some possible uses of the data set as well as ideas to improve it further.

2. METHODOLOGY

We start by briefly introducing Wikipedia-specific terminology:

- *Wikicode* : The markup language used to write Wikipedia pages; also known as *Wikitext* or *Wiki markup*.
- *Template*: A page that is embedded into other pages to allow for the repetition of information, following a certain Wikicode format.[7] Citation templates are specifically defined to embed citations.
- *Citation*: A citation is an abbreviated alphanumeric expression, embedded in Wikicode following a citation template, as shown in Figure 1; it usually denotes an entry in the

[5] https://help.altmetric.com/support/solutions/articles/6000060980-how-does-altmetric-track-mentions-on-wikipedia (accessed January 3, 2020). Identifiers considered by Altmetrics currently include DOI, URI from a domain white list, PMID, PMC ID, arXiv ID.
[6] https://github.com/Harshdeep1996/cite-classifications-wiki/releases/tag/0.2.
[7] https://en.wikipedia.org/wiki/Help:Template (accessed January 3, 2020).

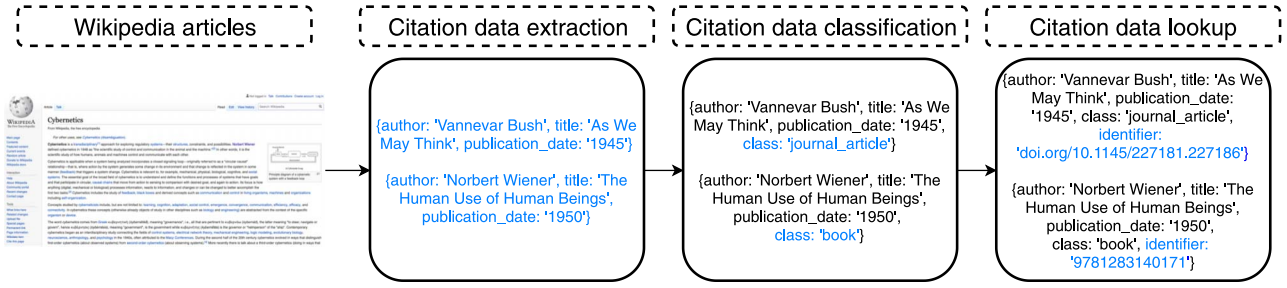


Figure 2. Overview of the citation data extraction pipeline. We highlight in blue/grey the outputs at every stage. These examples are illustrative simplifications from the actual data set.

References section of a Wikipedia page, but can be used anywhere on a page too (e.g., *Notes*, *Further work*).

2.1. Overview

Our process can be broken down into the following steps, as illustrated in Figure 2:

1. *Citation data extraction*: A Wikipedia dump is used to extract citations from all pages and considering various citation templates. The extracted citations are then mapped to a uniform set of key-value pairings.
2. *Citation data classification*: A classifier is trained to distinguish between citations to journal articles, books, or other Web content. The classifier is trained using a subset of citations already equipped with known identifiers or URLs, allowing us to label them beforehand. All the remaining citations are then classified.
3. *Citation data lookup*: All newly found citations to journal articles are labeled with identifiers (DOIs) using the Crossref API.

2.2. Citation Data Extraction

The citation data extraction pipeline is in turn divided into two steps, which are repeated for every Wikipedia article: *extraction* of all sentences that contain text in Wikicode format and *filtering* of sentences using the citation template Wikicode; and *mapping* of extracted citations to the uniform template and creation of a tabular data set. An example of Wikicode citations, extracted during step 1, is given in Table A1. The same citations after mapping to a uniform template are given in Table A2.

2.2.1. *Extraction and filtering*

We used the English Wikipedia XML dump from May 2020 and scraped it to get the content of each article/page. The number of unique pages is 6,069,685 after removing redirects, as they do not have any citations of their own.

Because we are restricting ourselves to citations that are given in Wikicode format, we used the `mwparserfromhell` parser,[8] which given as input a Wikipedia page, returns all text that is written in Wikicode format. Citations are generally present inside `<ref>` tags or between double curly brackets {{, as shown in Table A1. When multiple citations to the same source are given in a page, we consider only the first one. The number of extracted citations is 29,276,667.

[8] https://github.com/earwig/mwparserfromhell (version 0.6).

2.2.2. Mapping

Citation templates can vary, and different templates can be used to refer to the same source in different pages. Therefore, we mapped all citations to the same uniform template. For this step, we used the `wikiciteparser` parser.[9] This parser is written in Lua and it can be imported into Python using the lupa library.[10] The uniform template we use comprises 29 different keys. Initially, the `wikiciteparser` parser only supported 17 citation templates; thus we added support for an additional 18 of the most frequently used templates. More details on the uniform template keys and the extra templates we implemented can be found in the accompanying repository.

The resulting uniform key-value data set can easily be transformed in tabular form for further processing. In particular, this first step allowed us to construct a data set of citations with identifiers containing approximately 3.928 million citations. These identifiers—including DOI, PMC, PMID, and ISBN—allowed us to use such citations as training data for the classifier.

2.3. Citation Data Classification

After having extracted all citations and mapped them to a uniform template, we proceed to train a classifier to distinguish among three categories of cited sources: *journal articles*, *books*, and *Web content*. Our primary focus is journal articles, as those cover most citations to scientific sources. We describe here our approach to labeling a golden data set to use for training, the features we use for the classifier, and the classification model.

2.3.1. Labeling

We labeled available citations as follows:

- Every citation with a PMC or PMID was labeled as a *journal article*.
- Every citation with a PMC, PMID, or DOI and using the citation template for journals and conferences was labeled as a *journal article*.
- Every citation that had an ISBN was labelled as a *book*.
- All citations with their URL top-level domain belonging to the following: *nytimes*, *bbc*, *washingtonpost*, *cnn*, *theguardian*, *huffingtonpost*, *indiatimes* were labeled as *Web content*.
- All citations with their URL top-level domain belonging to the following: *youtube*, *rollingstone*, *billboard*, *mtv*, *metacritic*, *discogs*, *allmusic* were labeled as *Web content*.

After labeling, we removed all identifiers and the type of citation template as features, as they were used to label the data set. We also removed the fields *URL*, *work*, *newspaper*, *website*, and for the same reason. The final number of data points used for training and testing the classifier is given in Table 1, and was partially sampled to have a comparable number of journal articles, books and Web content.

2.3.2. Features

We next describe the features we used for the classification model:

- *Citation text*: The text of the citation, in Wikicode syntax.
- *Citation statement*: The text preceding a citation in a Wikipedia article, as it is known that certain statements are more likely to contain citations (Redi et al., 2019). We have used the 40 words preceding the first time a source is cited in an article.

[9] https://github.com/dissemin/wikiciteparser (version 0.1.1).
[10] https://pypi.org/project/lupa.

Table 1. Number of citations with a known class (* indicates a sampled subset)

Class label	Training data	Total known
Book	*951,991	2,103,492
Web content	*1,100,000	3,409,042
Journal article	*748,009	1,482,040
Total	2,800,000	6,994,574

- *Part of Speech (POS) tags*: POS tags in citation statements could also help qualify citations (Redi et al., 2019). These were generated using the NLTK library.[11]
- *Citation section*: The article section a citation occurs in.
- *Order of the citation* within the article, and *total number of words of the article*.

2.3.3. Classification model

The model that we constructed is a hybrid deep learning pipeline illustrated in Figure 3. The features were represented as follows:

- *Citation text*: The citation text in Wikicode syntax was fed to a character- level bidirectional LSTM (Schuster & Paliwal, 1997) on the dummy task of predicting whether the citation text is to a book/journal article or other Web content. The traintest split was done using a 90-10 ratio, yielding a 98.56% test accuracy. We used this dummy task to avoid the effects of vocabulary sparsity due to Wikicode syntax. The character-level embeddings are of dimension 300; we aggregated them for every citation text via summation and normalized the resulting vector to sum to one. We used character-level embeddings to deal with Wikicode syntax. The citation text embeddings were trained on the dummy task and frozen afterwards.
- *Citation statement*: The vocabulary for citation statements contains approximately 443,000 unique tokens, after the removal of tokens that appear strictly fewer than five times in the corpus. We used fastText to generate word-level embeddings for citation statements, using subword information (Bojanowski, Grave et al., 2017). FastText allowed us to deal with out of vocabulary words. We used the fastText model pretrained on English Wikipedia.[12]
- *POS tags*: The POS tags of citation statements were represented with a bag of words count vector. We were considering the top 35 tags by count frequency.
- *Citation section*: We used a one-hot encoding for the 150 most common sections within Wikipedia articles. The *order of the citation* within the article and *total number of words of the article* were represented as scalars.

Once the features had been generated, citation statements and their POS tags were further fed to an LSTM of 64 dimensions to create a single representation. All the resulting feature representations were concatenated and fed into a fully connected neural network with four hidden layers, as shown in Figure 3. A final Softmax activation function was applied on the output generated by the fully connected layers to map the output to one of the three categories of interest. We trained the model for five epochs using a train and test split of 90% and 10% respectively. For training, we

[11] https://www.nltk.org (version 3.4.1).
[12] https://fasttext.cc/docs/en/crawl-vectors.html.

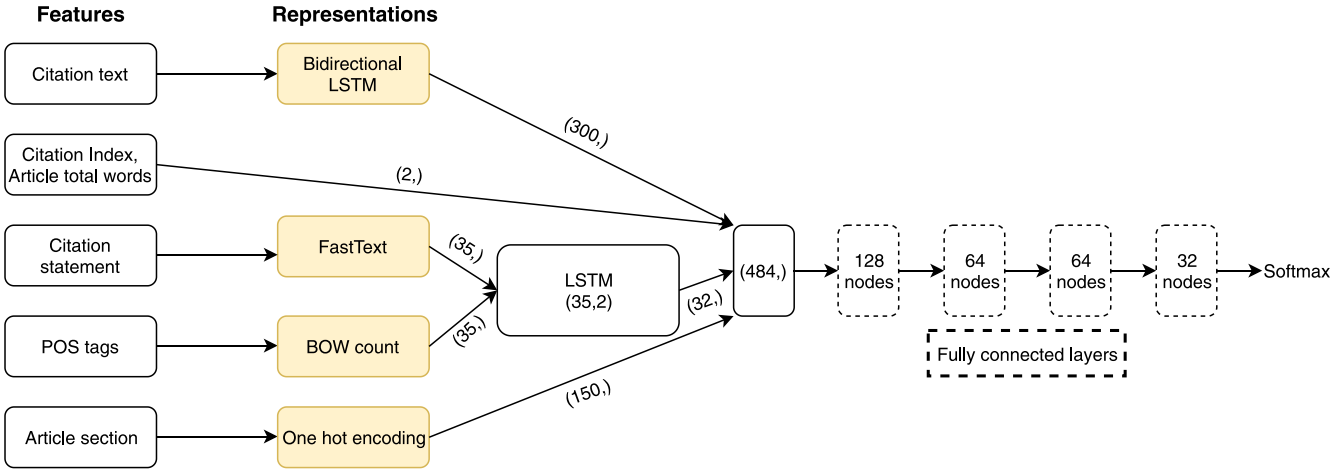


Figure 3. Citation classification model.

used the Adam optimizer (Kingma & Ba, 2014) and a binary crossentropy loss. The model's initial learning rate was set to 10^{-3}, and reduced minimally to 10^{-5} once the accuracy metric had stopped improving.

Previous work on citation classification has been based on textual and syntactic features (Dong & Schäfer, 2011) or link type analysis (Xu, Martin, & Mahidadia, 2013). Different stylistic or rhetorical cues have been also used as features (Di Marco, Kroon, & Mercer, 2006). We note that most of this previous work has focused on classifying the function or intent of a citation, rather than the typology of the cited object—that is, what a citation refers to, for example a book or journal.

2.4. Citation Data Lookup

The lookup task entails finding a permanent identifier for every citation missing one. We focused on journal articles for this final step, because they make up the bulk of citations to scientific literature found in Wikipedia for which a stable identifier can be retrieved. We used the Crossref API to get DOIs.[13] Crossref allows querying its API 50 times per second; we used the `aiohttp` and `asyncio` libraries to process requests asynchronously. For each citation query, we get a list of possible matches in descending ordered according to a Crossref confidence score. We kept the top three results from each query response.

3. EVALUATION

In this section we discuss the evaluation of the citation classification and lookup steps.

3.1. Classification Evaluation

After training the model for five epochs, we attained an accuracy of 98.32% on the test set. The confusion matrix for each of the labels is given in Table 2. The model is able to distinguish among the three classes very well.

The model was then used to classify all the remaining citations from the 29.276 million data set; that is to say approximately 22.282 million citations. Some examples of results from the

[13] https://www.crossref.org.

Table 2. Confusion matrix for citation classification. Results are based on a 10% held-out test set

Label	Book	Article	Web
Book	93,602 (98.32%)	1039	558
Article	961	73,682 (98.50%)	158
Web	1,136	180	108,684 (98.80%)

classification step are given in Table A3. The resulting total number of citations per class are given in Table 3.

3.2. Crossref Evaluation

For the lookup, we evaluated the response of the Crossref API to assess how to select results from it. We tested the API using 10,000 random citations with DOI identifiers and containing 9764 unique title-author pairs. We split this subset into a 80-20 split, tried out different heuristics on 80% of the data points and tested the best one on the remaining 20%. Table 4 shows the results for different heuristics, which confirms that the simple heuristic of picking the first result from Crossref works well.

This still leaves open the question of what Crossref confidence score to use. We picked the threshold for the confidence score to be 34.997 which gave us a precision of 70% and a recall of 67.55% to reach a balance between the two in the evaluation (Figure 4).

We finally tested the threshold using the 1,953 held-out examples, out of which 1,246 examples had the correct identifier with the first heuristic (out of 1,297) and the threshold, 646 examples gave a different result out of which 521 are over the threshold and only 10 requests were invalid for the API. Hence, the first metadata result is the best result from the Crossref API.

The lookup process was performed by extracting the title and the first author (if available) for all the potential journal articles and was queried against the CrossRef API to get the metadata. The top three results from the metadata were taken into account if they existed, and their DOIs and confidence scores were extracted. 260,752 citations were equipped with DOIs using the lookup step and 153,879 unique DOIs were found relating to each of these citations (selecting the DOI with highest Crossref score).

4. DATA SET

The resulting `Wikipedia Citations` data set is composed of three parts:

1. The main data set of 29.276 million citations from 35 different citation templates, out of which 3.928 million citations already contained identifiers (Table A4), and 260,752 out

Table 3. Number of newly classified citations per class

Label	New	Previously known	Total
Journal article	947,233	1,482,040	2,429,273
Book	3,243,364	2,103,492	5,346,856
Web content	18,091,496	3,409,042	21,500,538
Total	22,282,093	6,994,574	29,276,667

Table 4. Results for each heuristic tested on 80% of the subset

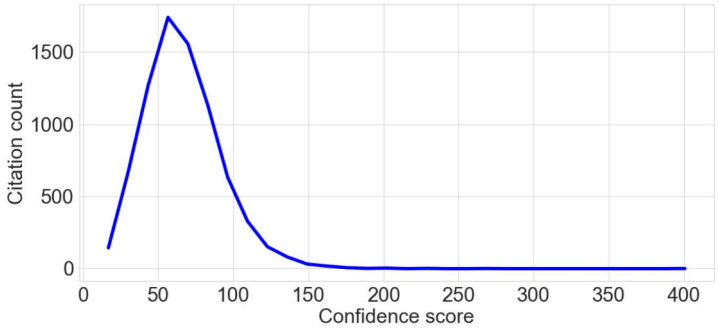
Heuristic	Matched	Not matched	Invalid request
1st result	5,258	2,510	43
2nd result	345	7,407	59
3rd result	96	7,647	67

of 947,233 newly-classified citations to journal articles were equipped with DOIs from Crossref.

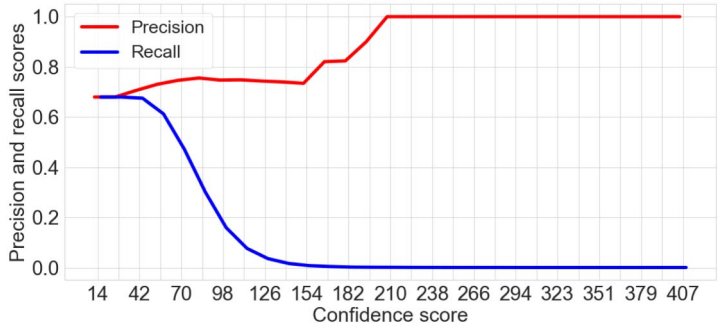
2. An example subset with the features for the classifier.
3. Citations classified as journal and their corresponding metadata/identifier extracted from Crossref to make the data set more complete.

4.1. Descriptive Analysis

We start by comparing our data set with previous work, which focused on citations with identifiers (Halfaker et al., 2018). The total number of citations per identifier type is found to be similar (Table 5). Minor discrepancies are likely due to the fact that we do not consider here all the edit history of every Wikipedia page, therefore missing changes between revisions, and that we consider a more recent dump. The total number of distinct identifiers across all Wikipedia, both previously known and newly found, are given in Table 6. Considering that in the Web of Science (WoS) (Birkle, Pendlebury et al., 2020) at the time there were 34,640,325 unique DOIs (version of June 2020; we only consider the typologies of "article," "review," "letter," and "proceedings



(a) Histogram of the Crossref API confidence scores over the validation set of the first result extracted from the lookup.



(b) Precision and recall for different Crossref API confidence score thresholds where the x-axis represents the scores returned by the Crossref API.

Figure 4. Evaluation of the Crossref API scores.

Table 5. Number of citations equipped with identifiers (excluding identifiers matched via lookup), per type and compared with (Halfaker et al., 2018). Note: A citation might be associated with two or more identifier types

Id.	Our data set	Previous work	Difference
DOI	1,442,177	1,211,807	230,370
ISBN	2,160,818	1,740,812	420,006
PMC	279,378	181,240	98,138
PMID	825,971	609,848	216,123
ArXiv	47,601	50,988	–3,387
Others	308,268	0	308,268
Total	4,755,945	3,794,695	961,250

paper"), Wikipedia is citing a volume of unique DOIs (1,157,571) corresponding to 3.3% of this total. Yet by doing an exact matching between Wikipedia DOIs and WoS DOIs, we can find 710,913 identifiers that are in common, or just 2% of the WoS total. This also entails that approximately 61% of unique DOIs in Wikipedia are indexed in the WoS. This result is in line with previous findings (Pooladian & Borrego, 2017; Priem et al., 2012; Shuai et al., 2013; Zahedi et al., 2014). The proportion of cited articles might seem low when compared to all of science, yet it is worth considering that an editorial selection takes place: Articles cited from Wikipedia are typically highly cited and published in visible journals (Arroyo-Machado et al., 2020; Colavizza, 2020; Nielsen, 2007; Teplitskiy et al., 2017). All in all, the relatively low fraction of scientific articles cited from Wikipedia over the total available does not *per se* entail a lack of coverage or quality in its contents: More work is needed to assess whether this might be the case.

We next consider the WoS subject categories for these 710,913 articles. We list the top 30 subject categories in Table A5, by number of distinct articles cited from Wikipedia. This ranking is dominated by Biochemistry & Molecular Biology (more than 11% of the articles) and Multidisciplinary Sciences (7%). The latter category accounts for megajournals such as *Nature*, *Science*, and *PNAS*. In general, the life sciences and biomedicine dominate. The top social science is Economics (1%) and the top humanities discipline is History (0.9%). To be sure, these results should be taken with caution, in particular when considering the arts, humanities, and social sciences. In this respect, the coverage of the WoS, and citation indexes more generally, is still wanting (Martín-Martín, Thelwall et al., 2020). Second, these proportions are not accounting for books, which are the primary means of publication in those fields of research.

We show in Figure 5 the number of citations to books and journal articles published over the period 2000 to 2020. This figure highlights how books appear to take longer to get cited in Wikipedia after publication. A similar plot, but considering a much wider publication time span (1500–2020) is given in Figure 6. Most published material in Wikipedia dates from the 1800s

Table 6. Number of distinct DOI and ISBN identifiers across Wikipedia

Category	Previously known	Newly found	Total
DOI	1,018,542	153,879	1,157,571
ISBN	901,639	–	901,639

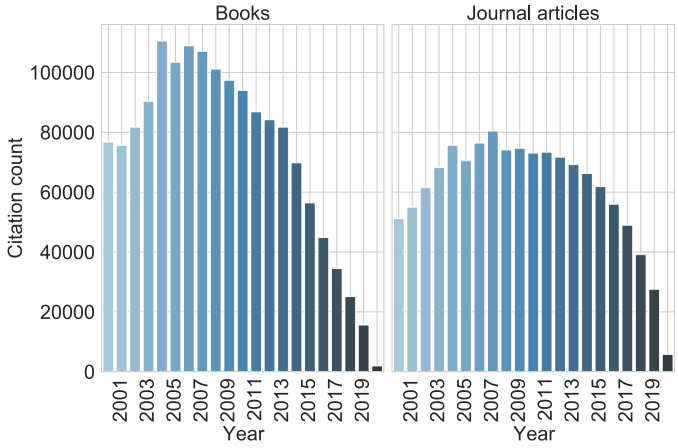


Figure 5. Publication years for *journal articles* and *books* for the period 2000–2020.

onward. We note that 89,098 journal article citations and 193,336 book citations do not contain a publication year.

Out of all the 28 template keys including the citation, most are not complete. For example, identifiers are present only in 13.42% of citations, whereas URLs are present in 85.25% of citations. This implies that many citations refer to Web content.

Out of 6,069,685 pages on Wikipedia, 405,359 have at least one or more citations with a DOI, that is about 6.7%; the proportion goes up to 12.84% for pages with at least one ISBN instead. This higher percentage of pages with DOIs, when compared to previously reported values (Maggio et al., 2017), is in large part due to our newly found identifiers from Crossref, which allowed us to equip with DOIs citations coming from Wikipedia pages with no previous presence of DOIs. We eventually considered the distribution of distinct DOIs per Wikipedia page and it was found that most of the pages have few citations with DOI identifiers, as shown in Figure 7. The top journals are listed in Table 7, and contain well-known megajournals (*Nature*, *Science*, *PNAS*) or other reputable venues (*Cell*, *JBC*).

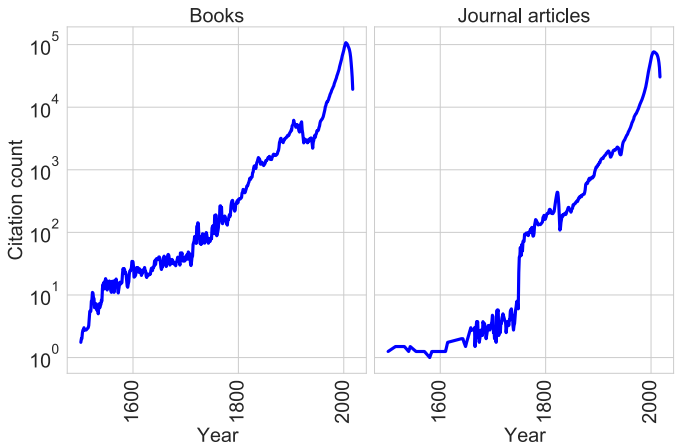


Figure 6. Number of citations per source publication year (1500–2020). A smoothing average using a window of 4 years is applied.

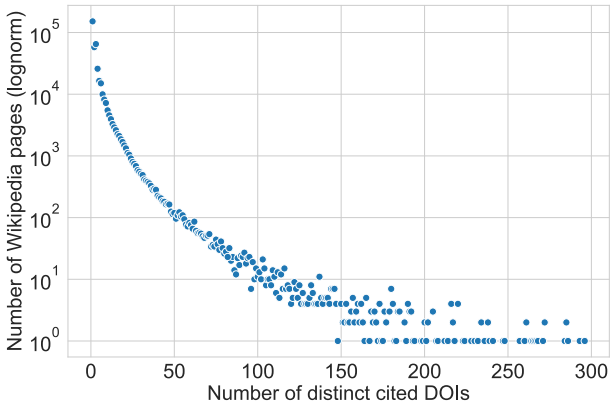


Figure 7. Number of distinct cited DOI per Wikipedia page.

4.2. Limitations and Research Opportunities

The `Wikipedia Citations` data set can be useful for research and applications in a variety of contexts. We suggest a few here, and also frame the limitations of our contribution as opportunities for future work.

4.2.1. Map of Wikipedia sources

What seems to us to be low-hanging fruit is creating a map of Wikipedia sources, following the science mapping and visualization methodologies (Börner, 2010; Chen, 2017; Shiffrin & Börner, 2004). Such work would allow us to comprehensively answer the question of what is cited from Wikipedia, from which Wikipedia articles, and how knowledge is reported and negotiated in Wikipedia. Importantly, such mapping should consider disciplinary differences in citations from Wikipedia, as well as books (5.3 million citations by our estimates) and nonscientific sources such as news outlets and other online media (21.5 million citations), which make up the largest share of Wikipedia citations. Answering these questions is critical to inform the community work on improving Wikipedia by finding and filling knowledge gaps and biases, at the same time guaranteeing the quality and diversity of the sources Wikipedia relies upon (Hube, 2017; Mesgari et al., 2015; Piscopo, Kaffee et al., 2017; Piscopo & Simperl, 2019; Wang & Li, 2020).

Table 7. Most cited journals

Journal name	Citations
Nature	36,136
Science	26,448
Journal of Biological Chemistry (JBC)	22,401
PNAS	21,347
The IUCN Red List of Threatened Species	10,082
Cell	9,329
Zootaxa	8,013
Genome Research	6,994

4.2.2. Citation reconciliation and recommendation

Link prediction in general, and citation recommendation in particular, have been explored for Wikipedia for some time (Fetahu et al., 2016; Paranjape, West et al., 2016; Wulczyn, West et al., 2016). Recent work has also focused on finding Wikipedia statements where a citation to a source might be needed (Redi et al., 2019). Our data set can further inform these efforts, in particular easing and fostering work on the recommendation of scientific literature to Wikipedia editors. The proposed citation classifier could also be reused for citation detection and reconciliation in a variety of contexts.

4.2.3. Citations as features

Citations from Wikipedia can be used as "features" in a variety of contexts. They have already been considered as altmetrics for research impact (Sugimoto et al., 2017), while they can also be used as features for machine learning applications, such as those focused on improving knowledge graphs, starting with Wikidata (Farda-Sarbas & Müller-Birn, 2019). It is our hope that more detail and novel use cases will also lead to a gradual improvement of the first version of the data set, which we release here.

4.2.4. Limitations

We highlight a set of limitations that constitute possible directions for future work. First of all, the focus on English Wikipedia can and should be rapidly overcome to include all languages in Wikipedia. Our approach can be adapted to other languages, provided that external resources (e.g., language models and lookup APIs) are available for them. Secondly, the data set currently does not account for the edit history of every citation from Wikipedia, which would allow us to study knowledge production and negotiation over time: Adding "citation versioning" would be important in this respect, as demonstrated by recent work (Zagovora et al., 2020). Thirdly, citations are used for a purpose, in a context; an extension of the `Wikipedia Citations` data set could include all the citation statements as well, to allow researchers to study the fine-grained purpose of citations. Furthermore, the classification of scientific publications that we use is limited. ISBNs, in particular, can refer to monographs, book series, book chapters, and edited books, which possess varying citation characteristics. Future work should extend the classification system to operate at such a finer-grained level. Lastly, the querying and accessibility of the data set is limited by its size; more work is needed to make Wikipedia's contents better structured and easier to query (Aspert, Miz et al., 2019).

5. CONCLUSION

We publish the `Wikipedia Citations` data set, consisting of 29.276 million citations extracted from 6.069 million articles from English Wikipedia. Citations are equipped with persistent identifiers such as DOIs and ISBNs whenever possible. Specifically, we extracted 3.928 million citations with identifiers—including DOI, PMC, PMID, and ISBN from Wikipedia itself, and further equipped an extra 260,752 citations with DOIs from Crossref. In so doing, we were able to raise the number of Wikipedia pages citing at least one scientific article equipped with a DOI from less than 5% to more than 6.7% (which corresponds to an additional 164,830 pages) and found that Wikipedia is citing just 2% of the scientific articles indexed in the WoS. We also release all our code to extend upon our work and update the data set in the future. Our work contributes to ongoing efforts (Halfaker et al., 2018; Zagorova et al., 2020) by expanding the coverage of Wikipedia citations equipped with identifiers, distinguishing between academic and nonacademic sources, and releasing a codebase to keep results up-to-date.

We highlighted a set of possible uses of our data set, from mapping the sources Wikipedia relies on, to recommending citations, and using citation data as features. The limitations of our contribution also constitute avenues for future work. We ultimately believe that `Wikipedia Citations` should be made available as data via open infrastructures (e.g., WikiCite[14] and OpenCitations[15]). We consider our work a step in this direction. It is therefore our hope that this contribution will start a collaborative effort by the community to study, use, maintain, and expand work on citations from Wikipedia.

AUTHOR CONTRIBUTIONS

All authors devised the study and wrote and reviewed the manuscript. HS: data collection and analysis. RW and GC: Study supervision.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding has been received for this research.

DATA AVAILABILITY

The data set is made available on Zenodo (Singh et al., 2020) and the accompanying repository contains all code and further documentation to replicate our results: https://github.com/Harshdeep1996/cite-classifications-wiki/releases/tag/0.2.

ACKNOWLEDGMENTS

The authors would like to thank Tiziano Piccardi (EPFL), Miriam Redi (Wikimedia Foundation), and Dario Taraborelli (Chan Zuckerberg Initiative) for their helpful advice. The authors also thank the Centre for Science and Technology Studies (CWTS), Leiden University, for granting access to the WoS.

REFERENCES

Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., & Romero-Frías, E. (2020). Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, *15*(2), e0228713. **DOI:** https://doi.org/10.1371/journal.pone.0228713, **PMID:** 32040488, **PMCID:** PMC7010282

Aspert, N., Miz, V., Ricaud, B., & Vandergheynst, P. (2019). A graph-structured dataset for Wikipedia Research. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 1188–1193), San Francisco. New York: ACM. **DOI:** https://doi.org/10.1145/3308560.3316757

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. **DOI:** https://doi.org/10.1162/qss_a_00018

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. **DOI:** https://doi.org/10.1162/tacl_a_00051

Börner, K. (2010). *Atlas of science: Visualizing what we know*. Cambridge, MA: MIT Press.

Chen, C. (2017). Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, *2*(2), 1–40. **DOI:** https://doi.org/10.1515/jdis-2017-0006

Chen, C.-C., & Roth, C. (2012). {{citation needed}}: The dynamics of referencing in Wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria. New York: ACM Press. **DOI:** https://doi.org/10.1145/2462932.2462943

[14] https://meta.wikimedia.org/wiki/WikiCite (accessed November 12, 2020).

[15] https://opencitations.net (accessed November 12, 2020).

Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies*, *1*(4), 1349–1380. **DOI:** https://doi.org/10.1162/qss_a_00080

Di Marco, C., Kroon, F. W., & Mercer, R. E. (2006). Using hedges to classify citations in scientific articles. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theory and applications* (pp. 247–263). Cham: Springer. **DOI:** https://doi.org/10.1007/1-4020-4102-0_19

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In H. Wang & D. Yarowsky (Eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 623–631).

Farda-Sarbas, M., & Müller-Birn, C. (2019). Wikidata from a research perspective—A systematic mapping study of Wikidata. *arXiv:1908.11153*. http://arxiv.org/abs/1908.11153

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

Forte, A., Andalibi, N., Gorichanaz, T., Kim, M. C., Park, T., & Halfaker, A. (2018). Information fortification: An on-line citation behavior. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork—GROUP '18* (pp. 83–92), Sanibel Island, FL. New York: ACM Press. **DOI:** https://doi.org/10.1145/3148330.3148347

Geiger, S. R., & Halfaker, A. (2013). When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, Hong Kong. ACM Press. **DOI:** https://doi.org/10.1145/2491055.2491061

Halfaker, A., Mansurov, B., Redi, M., & Taraborelli, D. (2018). Citations with identifiers in Wikipedia. *Figshare*. **DOI:** https://doi.org/10.6084/m9.figshare.1299540

Heilman, J. M., Kemmann, E., Bonert, M., Chatterjee, A., Ragar, B., … Laurent, M. R. (2011). Wikipedia: A key tool for global public health promotion. *Journal of Medical Internet Research*, *13*(1), e14. **DOI:** https://doi.org/10.2196/jmir.1589, **PMID:** 21282098, **PMCID:** PMC3221335

Hube, C. (2017). Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion—WWW '17 Companion* (pp. 717–721), Perth, Australia. New York: ACM Press. **DOI:** https://doi.org/10.1145/3041021.3053375

Jemielniak, D., & Aibar, E. (2016). Bridging the gap between Wikipedia and academia. *Journal of the Association for Information Science and Technology*, *67*(7), 1773–1776. **DOI:** https://doi.org/10.1002/asi.23691

Keegan, B., Gergle, D., & Contractor, N. (2011). Hot off the Wiki: Dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration—WikiSym '11*, Mountain View, CA. New York: ACM Press. **DOI:** https://doi.org/10.1145/2038558.2038577

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. https://arxiv.org/abs/1412.6980

Kousha, K., & Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, *68*(3), 762–779. **DOI:** https://doi.org/10.1002/asi.23694

Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 591–602), Montreal. New York: ACM Press. **DOI:** https://doi.org/10.1145/2872427.2883085

Laurent, M. R., & Vickers, T. J. (2009). Seeking health information online: Does Wikipedia matter? *Journal of the American Medical Informatics Association*, *16*(4), 471–479. **DOI:** https://doi.org/10.1197/jamia.M3059, **PMID:** 19390105, **PMCID:** PMC2705249

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., … Bizer, C. (2015). DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195. **DOI:** https://doi.org/10.3233/SW-140134

Lewoniewski, W., Wecel, K., & Abramowicz, W. (2017). Analysis of references across Wikipedia languages. In R. Damaševičius & V. Mikašyte (Eds.), *Information and software technologies* (Vol. 756, pp. 561–573). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-319-67642-5_47

Maggio, L. A., Steinberg, R. M., Piccardi, T., & Willinsky, J. M. (2020). Reader engagement with medical content on Wikipedia. *eLife*, *9*, e52426. **DOI:** https://doi.org/10.7554/eLife.52426, **PMID:** 32142406, **PMCID:** PMC7089765

Maggio, L. A., Willinsky, J. M., Steinberg, R. M., Mietchen, D., Wass, J. L., & Dong, T. (2017). Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia. *PLOS ONE*, *12*(12), e0190046. **DOI:** https://doi.org/10.1371/journal.pone.0190046, **PMID:** 29267345, **PMCID:** PMC5739466

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, September. **DOI:** https://doi.org/10.1007/s11192-020-03690-4, **PMID:** 32981987, **PMCID:** PMC7505221

McMahon, C., Johnson, I., & Hecht, B. (2017). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. Palo Alto, CA: AAAI Press.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, *66*(2), 219–245. **DOI:** https://doi.org/10.1002/asi.23172

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *First Monday*, *12*(8). https://firstmonday.org/article/view/1997/1872

Nielsen, F. Å., Mietchen, D., & Willighagen, E. (2017). Scholia, *Scientometrics* and Wikidata. In E. Blomqvist, K. Hose, H. Paulheim, A. Ławrynowicz, F. Ciravegna, & O. Hartig (Eds.), *The Semantic Web: ESWC 2017 Satellite Events* (Vol. 10577, pp. 237–259). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-319-70407-4_36

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2012). The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. *SSRN Electronic Journal*. **DOI:** https://doi.org/10.2139/ssrn.2021326

Paranjape, A., West, R., Zia, L., & Leskovec, J. (2016). Improving website hyperlink structure using server logs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 615–624), San Francisco, CA. New York: ACM Press. **DOI:** https://doi.org/10.1145/2835776.2835832, **PMID:** 28345077, **PMCID:** PMC5365094

Piccardi, T., Redi, M., Colavizza, G., & West, R. (2020). Quantifying engagement with citations on Wikipedia. In *Proceedings of The*

Web Conference 2020 (pp. 2365–2376), Taipei. New York: ACM. **DOI:** https://doi.org/10.1145/3366423.3380300

Piscopo, A., Kaffee, L. A., Phethean, C., & Simperl, E. (2017). Provenance information in a collaborative knowledge graph: An evaluation of Wikidata external references. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, & J. Heflin (Eds.), *The Semantic Web—ISWC 2017* (Vol. 10587, pp. 542–558). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-319-68288-4_32

Piscopo, A., & Simperl, E. (2019). What we talk about when we talk about Wikidata quality: A literature survey. In *Proceedings of the 15th International Symposium on Open Collaboration*, Skövde, Sweden. New York: ACM Press. **DOI:** https://doi.org/10.1145/3306446.3340822

Pooladian, A., & Borrego, Á. (2017). Methodological issues in measuring citations in Wikipedia: A case study in Library and Information Science. *Scientometrics*, *113*(1), 455–464. **DOI:** https://doi.org/10.1007/s11192-017-2474-z

Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Conference on Supporting Group Work*, Sanibel Island, FL. ACM Press. **DOI:** https://doi.org/10.1145/1316624.1316663

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. https://arxiv.org/html/1203.4745

Redi, M., Fetahu, B., Morgan, J., & Taraborelli, D. (2019). Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability. In *Proceedings of the World Wide Web Conference* (pp. 1567–1578), San Francisco, CA. New York: ACM Press. **DOI:** https://doi.org/10.1145/3308558.3313618

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. **DOI:** https://doi.org/10.1109/78.650093

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Häggström, A., & Heilman, J. (2017). Evolution of Wikipedia's medical content: Past, present and future. *Journal of Epidemiology and Community Health*, *71*, 1122–1129. **DOI:** https://doi.org/10.1136/jech-2016-208601, **PMID:** 28847845, **PMCID:** PMC5847101

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences*, *101* (Supplement 1), 5183–5185. **DOI:** https://doi.org/10.1073/pnas.0307852100, **PMID:** 14742869, **PMCID:** PMC387293

Shuai, X., Jiang, Z., Liu, X., & Bollen, J. (2013). A comparative study of academic and Wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries—JCDL '13*, Indianapolis. New York: ACM Press. **DOI:** https://doi.org/10.1145/2467696.2467746

Singh, H., West, R., & Colavizza, G. (2020). *Wikipedia citations: A comprehensive dataset of citations with identifiers extracted from English Wikipedia*. **DOI:** https://doi.org/10.5281/zenodo.3940692

Smith, D. A. (2020). Situating Wikipedia as a health information resource in various contexts: A scoping review. *PLOS ONE*, *15*(2), e0228786. **DOI:** https://doi.org/10.1371/journal.pone.0228786, **PMID:** 32069322, **PMCID:** PMC7028268

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. **DOI:** https://doi.org/10.1002/asi.23833

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, *68*(9), 2116–2127. **DOI:** https://doi.org/10.1002/asi.23687

Thompson, N., & Hanley, D. (2018). Science is shaped by Wikipedia: Evidence from a randomized control trial. *MIT Sloan Research Paper* 5238-17. **DOI:** https://doi.org/10.2139/ssrn.3039505

Tomaszewski, R., & MacDonald, K. I. (2016). A study of citations to Wikipedia in scholarly publications. *Science & Technology Libraries*, *35*(3), 246–261. **DOI:** https://doi.org/10.1080/0194262X.2016.1206052

Torres-Salinas, D., Romero-Frías, E., & Arroyo-Machado, W. (2019). Mapping the backbone of the humanities through the eyes of Wikipedia. *Journal of Informetrics*, *13*(3), 793–803. **DOI:** https://doi.org/10.1016/j.joi.2019.07.002

Wang, P., & Li, X. (2020). Assessing the quality of information on Wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology*, *71*(1), 16–28. **DOI:** https://doi.org/10.1002/asi.24210

Wulczyn, E., West, R., Zia, L., & Leskovec, J. (2016). Growing Wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 975–985). **DOI:** https://doi.org/10.1145/2872427.2883077, **PMID:** 27819073, **PMCID:** PMC5092237

Xu, H., Martin, E., & Mahidadia, A. (2013). Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*. New York: ACM.

Zagovora, O., Ulloa, R., Weller, K., & Flöck, F. (2020). 'I Updated the <ref>': The evolution of references in the English Wikipedia and the implications for altmetrics. *arXiv:2010.03083*. http://arxiv.org/abs/2010.03083

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, *101*(2), 1491–1513. **DOI:** https://doi.org/10.1007/s11192-014-1264-0

APPENDIX

Table A1. Different citation templates can be used to refer to the same source

Index	Extracted citation template
Citation 1	{{citation|**author**=John Smith| access-date=February 17, 2006}}
Citation 2	{{citation|**creator**=John Smith| access-date=September 15, 2006}}

Table A2. Citations are mapped to have the same keys

Index	Uniform citation template
Citation 1	{'author': 'John Smith', 'type': 'citation', 'access-date': 'February 17, 2006'}
Citation 2	{'author': 'John Smith', 'type': 'citation', 'access-date': 'September 15, 2006'}

Table A3. Example of newly classified citations

Label	Citation
Journal article	{'title': 'What is Asia?', 'author': 'Philip Bowring'}
Journal article	{'title': 'Right Ventricular Failure', 'journal': 'e-Journal of Cardiology Practice'}
Book	{'title': 'Histories of Anthropology Annual, Vol. I', 'author': 'HS Lewis'}
Book	{'title': 'The Art of the Sale', 'publisher': 'The Penguin Press'}
Web content	{'title': 'Barry White - Chart history (Hot R&B Hip-Hop Songs) Billboard', 'page_title': Let the Music Play (Barry White Album)'}
Web content	{'title': 'Sunday Final Ratings: Oscars Adjusted Up', 'work': 'TVbytheNumbers'}

Table A4. Presence of identifiers per citation for the 3.92 million citations with identifiers (with 0 = False and 1 = True). These counts sum up to 3,620,124, with an additional 308,268 citations associated with other identifiers such as OCLC, ISSN. The total adds up to 3,928,392 citations with identifiers

With DOI	With ISBN	With PMC	With PMID	With ARXIV	Total
0	0	0	0	1	4,447
0	0	0	1	0	41,417
0	0	0	1	1	7
0	0	1	0	0	829
0	0	1	1	0	11,261
0	0	1	1	1	5
0	1	0	0	0	2,119,545
0	1	0	0	1	192
0	1	0	1	0	223
0	1	1	0	0	13
0	1	1	1	0	8
1	0	0	0	0	592,557
1	0	0	0	1	35,824
1	0	0	1	0	501,176
1	0	0	1	1	5,101
1	0	1	0	0	4,241
1	0	1	0	1	3
1	0	1	1	0	261,173
1	0	1	1	1	1,265
1	1	0	0	0	35,706
1	1	0	0	1	756
1	1	0	1	0	3,794
1	1	0	1	1	1
1	1	1	0	0	40
1	1	1	1	0	540

Table A5. Web of Science 30 most-represented subject categories, by number of articles cited from Wikipedia. We consider the first subject category of each article, and discard the rest. The total number of articles which we could match in the Web of Science using DOIs is 710,913. The top 30 subject categories make up almost 60% of them

Web of Science subject category	Number of articles
Biochemistry & Molecular Biology	81,556
Multidisciplinary Sciences	51,368
Astronomy & Astrophysics	18,658
Medicine, General & Internal	17,134
Neurosciences	16,061
Cell Biology	14,411
Genetics & Heredity	13,922
Chemistry, Multidisciplinary	13,754
Microbiology	13,661
Oncology	13,606
Plant Sciences	13,050
Clinical Neurology	13,017
Immunology	11,231
Biotechnology & Applied Microbiology	10,980
Pharmacology & Pharmacy	10,351
Ecology	10,308
Zoology	10,081
Biology	8,979
Endocrinology & Metabolism	8,444
Geosciences, Multidisciplinary	7,653
Public, Environmental & Occupational Health	7,348
Biochemical Research Methods	7,337
Economics	7,211
Physics, Multidisciplinary	7,042
Paleontology	6,942
Behavioral Sciences	6,591
Environmental Sciences	6,316
Chemistry, Physical	6,292
History	6,174
Cardiac & Cardiovascular Systems	5,875
Total	425,353 (60%)