Wikipedia Citations:
A comprehensive dataset 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

July 15, 2020

Abstract

Wikipedia’s contents are based on reliable and published sources. To this date, 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 dataset of citations extracted from Wikipedia. A total of 29.3M citations were extracted from 6.1M English Wikipedia articles as of May 2020, and classified as being to books, journal articles or Web contents. We were thus able to extract 4.0M citations to scholarly publications with known identifiers — including DOI, PMC, PMID, and ISBN — and further labeled an extra 261K 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. Scientific articles cited from Wikipedia correspond to 3.5% of all articles with a DOI currently indexed in the Web of Science. We release all our code to allow the community to extend upon our work and update the dataset in the future.

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.”\(^\dagger\)

\(^\dagger\)Corresponding author, g.colavizza@uva.nl

\[^\dagger^\]
https://en.wikipedia.org/wiki/Citation[accessed 2020-01-03].

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. Most importantly, Wikipedia is often perceived and relied upon as a source of “neutral” information.

The confidence that users and services place on Wikipedia has been found to be usually justified: Wikipedia’s content is of general high-quality and up-to-date.

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.

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. The citation practices of editors are also at times not systematic. As a consequence, the efforts to expand and improve Wikipedia’s verifiability through citations to sources are increasing.

A crucial question to ask in order 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, as Wikipedia is instrumental in providing access to scientific information and in fostering the public understanding of science. Citations in Wikipedia are also useful for users browsing low-quality or underdeveloped articles, as they allow them to look outside of the platform. The literature cited in Wikipedia has been found to positively correlate with the journal popularity, the journal impact factor and to its open access availability. Being cited in Wikipedia can also be considered as an ‘altmetric’ indicator of impact in itself. A clear influence of Wikipedia on scientific research has in turn been found, despite the general lack of acknowledgement of Wikipedia in the scientific literature.

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. Nevertheless, these studies either use of proprietary databases with limited coverage, or only consider specific publishers (PLoS) and academic communities (computer science).

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: citations are often given using citation templates somewhat liberally, making it difficult to detect citations to the same source. Secondly, while some citations contain stable identifiers (e.g., DOIs), others do not. A recent study found that 4.42% Wikipedia articles contain at least one citation with a DOI; a low number
which might indicate that we are missing a non-negligible fraction of citations without identifiers. This is a significant limitation since existing databases, such as Altmetrics, do provide Wikipedia citation metrics relying exclusively on citations with identifiers. This in turn limits the scope of results relying on these data.

Our goal here is to overcome these two challenges and expand upon previous work \cite{12}, by providing a dataset of all citations from the English Wikipedia, equipped with identifiers and including the code to replicate and improve upon our work. The dataset is available on Zenodo \cite{40} and the accompanying repository contains all code and further documentation to replicate our results.

This article is organized as follows. We start by describing our pipeline focusing on its three main steps: 1) citation template harmonization — to structure every citations in Wikipedia using the same schema; 2) citation classification — to find citations to books and journal articles; and 3) citation identifier look-up — to find identifiers such as DOIs. We subsequently evaluate our results, provide a description of the published dataset, and conclude by highlighting some possible uses of the dataset 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. 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 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:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Overview of the citation classification pipeline.}
\end{figure}

 Identifiers considered by Altmetrics currently include: DOIs, URIs from a domain white list, PMIDs, PMC IDs, arXiv IDs.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Overview of the citation classification pipeline.}
\end{figure}

Identifiers considered by Altmetrics currently include: DOIs, URIs from a domain white list, PMIDs, PMC IDs, arXiv IDs.
Figure 1: Example of citations in Wikipedia.

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 dataset.

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 known identifiers or URLs, allowing 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: 1) **extraction** of all sentences which contain text in Wikicode format, and **filtering** of sentences using the citation template Wikicode; 2) **mapping** of extracted citations to the uniform template and creation of a tabular dataset. An example of Wikicode citations, extracted during step 1, is given in Table 1. The same citations after mapping to a uniform template are given in Table 2.

#### 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 since they do not have any citations of their own.
Table 1: 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 2: 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’} |

Since we are restricting ourselves to citations which are given in Wikicode format, we used the mwparserfromhell parser[8] which given as input a Wikipedia page, it returns all text which is written in Wikicode format. Citations are generally present inside <ref> tags or between double curly brackets {{, as shown in Table 1. When multiple citations to the same source are given in a page, we only consider the first one. The number of extracted citations is 29,276,667.

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 dataset can easily be transformed in tabular form for further processing. In particular, this first step allowed us to construct a dataset 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.

[8] https://github.com/earwig/mwparserfromhell
[9] https://github.com/dissemin/wikiciteparser
[10] https://pypi.org/project/lupa
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 are journal articles, as those cover most citations to scientific sources. We describe here our approach to label a golden dataset to use for training, the features we use for the classifier, and the classification model.

2.3.1 Labeling

We labelled available citations as follows:

- Every citation with a non-null PMC or PMID was labeled as a journal article.
- Every citation with a non-null PMC, PMID or DOI and using the citation template for journals and conferences, was labeled as a journal article.
- Every citation which had a non-null 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 labelling, we removed all identifiers and the type of citation template as features, since they were used to label the dataset. We also removed the fields: URL, work, newspaper, website, for the same reason. The final number of data points used for training and testing the classifier is given in Table 3 and was partially sampled in order to have a comparable number of journal articles, books and Web content.

Table 3: Number of citations with a known class (* indicates a sampled subset).

| Class label     | Train 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   |

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 [35]. We have used the 40 words preceding the first time a source is cited in an article.

• **Part of Speech (POS) tags**: POS tags for citation statements could also be correlated to citations. [35]. 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 which 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 [36] on the dummy task of predicting whether the citation text is to a book/journal article or other Web content. The train-test split was done using a 90-10 ratio, yielding a 98.56% test accuracy. We used this dummy task in order 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 froze afterwards.

• **Citation statement**: The vocabulary for citation statements contains approximately 443,000 unique tokens, after the removal of tokens which appear strictly less than 5 times in the corpus. We used fastText to generate word-level embeddings for citation statements, using subword information [3]. FastText allowed us to deal with out of vocabulary words. We used the fastText model pre-trained 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.

---

[1] https://www.nltk.org [version 3.4.1].
[2] https://fasttext.cc/docs/en/crawl-vectors.html
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 used the Adam optimizer [17] and a binary cross-entropy loss. The model’s initial learning rate was set to 0.001, and reduced minimally to 0.00001 once the accuracy metric has stopped improving.

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, since 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.

Crossref allows to query 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.

[1] https://www.crossref.org/
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 4. The model is able to distinguish among the three classes very well.

Table 4: Confusion matrix for citation classification. Results are based on a 10% held-out test set.

| Label    | Book     | Article   | Web       |
|----------|----------|-----------|-----------|
| Book     | 93,602   | 1039      | 558       |
| Article  | 961      | 73,682    | 158       |
| Web      | 1,136    | 180       | 108,684   |

The model was then used to classify all the remaining citations from the 29.276 million dataset, that is to say approximately 22.282 million citations. Some examples of results from the classification step are given in Table 6. The resulting total number of citations per class are given in Table 5.

Table 5: 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 6: 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'} |
3.2 Crossref Evaluation

For the lookup, we evaluated the response of the Crossref API in order to assess how to pick 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 7 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 1953 held-out examples, out of which 1246 examples had the correct identifier with the first heuristic (out of 1297) 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
Table 7: Results for each heuristic tested on 80% of the subset.

| Heuristic | Matched | Not matched | Invalid request |
|-----------|---------|-------------|-----------------|
| 1st result | 5258    | 2510        | 43              |
| 2nd result | 345     | 7407        | 59              |
| 3rd result | 96      | 7647        | 67              |

CrossRef API to get the metadata. The top 3 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 320,887 unique DOIs were found relating to each of these citations.

4 Dataset

The resulting Wikipedia Citations dataset is composed of 3 parts:

1. The main dataset of 29.276 million citations from 35 different citation templates, out of which 3.928 million citations already contained identifiers (Table 8), and 260,752 out 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 dataset more complete.

4.1 Descriptive analysis

We start by comparing our dataset with previous work, which focused on citations with identifiers [12]. The total number of citations per identifier type is found to be similar (Table 9). 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 10. Considering that in the Web of Science there are 38,829,128 articles with a DOI (version of March 2020), Wikipedia is citing approximately 3.5% of them (1,347,893).

We show in Figure 5 the number of citations to books and journal articles published over the time 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 1800 onward. We note that a total of 89,098 journal article citations and 193,336 book citations do not contain a publication year.
Figure 5: Publication years for journal articles and books, for the period 2000-2020.

Figure 6: Number of citations per source publication year (1500-2020).
Table 8: 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.

| DOI Present | ISBN Present | PMC Present | PMID Present | ARXIV Present | Total |
|-------------|--------------|-------------|--------------|---------------|-------|
| 0           | 0            | 0           | 0            | 1             | 4,447 |
| 0           | 0            | 0           | 1            | 0             | 41,417|
| 0           | 0            | 1           | 0            | 1             | 7     |
| 0           | 1           | 0           | 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   |

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 contents.

Out of 6,069,685 pages on Wikipedia, 407,777 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 [23], 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
Table 9: Number of citations equipped with identifiers (not including the identifiers through lookup), per type and compared with [12]. Note: a citation might be associated with two or more identifier types.

| Id.  | Our dataset | Previous work [12] | 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    |

Table 10: Number of distinct DOI and ISBN identifiers across Wikipedia.

| Category | Previously known | Newly found | Total    |
|----------|------------------|-------------|----------|
| DOI      | 1,027,006        | 320,887     | 1,347,893|
| ISBN     | 901,639          | –           | 901,639  |

shown in Figure 7. The top journals are listed in Table 11, and contain well-known mega journals (Nature, Science, PNAS) or other reputed venues (Cell, JBC).

Table 11: Most cited journals.

| Journal Name                                           | Citations |
|--------------------------------------------------------|-----------|
| Nature                                                 | 36,136    |
| Science                                                | 26,448    |
| Journal of Biological Chemistry                        | 22,401    |
| PNAS                                                   | 21,347    |
| The IUCN Red List of Threatened Species                | 10,082    |
| Cell                                                   | 9329      |
| Zootaxa                                                | 8013      |
| Genome Research                                        | 6994      |

4.2 Research opportunities

The Wikipedia Citations dataset can be useful for research and applications in a variety of contexts. We suggest a few here.

4.2.1 Map of Wikipedia sources

What seems to us a low-hanging fruit is a map of Wikipedia sources, following the well-known science mapping and visualization methodologies [38, 4, 5]. Such
work would allow to comprehensively answer the question of what is cited from Wikipedia, from which Wikipedia articles, and how knowledge is reported and negotiated in Wikipedia. Answering these questions is critical to inform the community work on improving Wikipedia by finding and filling knowledge gaps and biases, all the same guaranteeing the quality and diversity of the sources Wikipedia relies upon [26, 31, 14, 32, 47].

4.2.2 Citation recommendation

Link prediction in general, and citation recommendation in particular, have been explored for Wikipedia since some time [9, 29, 48]. Recent work has also focused on finding Wikipedia statements where a citation to a source might be needed [35]. Our dataset can further inform these efforts, in particular easing and fostering work on the recommendation of scientific literature to Wikipedia editors.

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 [42], while they can also be used as features for machine learning applications such as those focused on improving knowledge graphs, starting with Wikidata [8]. It is our hope that more detail and novel use cases will also lead to a gradual improvement of the first version of the dataset which we release here.
5 Conclusion

We publish the Wikipedia Citations dataset, consisting of a total of 29.276M citations extracted from 6.069M articles from English Wikipedia. Citations are equipped with persistent identifiers such as DOIs and ISBNs whenever possible. Specifically, we extracted 3.928M 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 articles with at least one DOI from less than 5% to more than 6.7% (which is an additional 105,018 pages with a DOI) and found that Wikipedia is citing approximately 3.5% of the journal articles indexed in the Web of Science. We also release all our code to extend upon our work and update the dataset in the future.

A set of limitations are worth highlighting, which also 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 easily be adapted to other languages, provided that external resources (e.g., language models and lookup APIs) allow for them. Secondly, the dataset currently does not account for the edit history of every citation from Wikipedia: this would allow to study knowledge production and negotiation over time: adding ‘citation versioning’ would be important in this respect. Thirdly, citations are used for a purpose, in a context; our choice to focus on the citation network means that an extension of the dataset could include all the citation statements as well, in order to allow researchers to study the fine-grained purpose of citations. Lastly, the querying and accessibility of the dataset is limited by its size; more work is needed in order to make Wikipedia contents better structured and easier to query [2].

We highlighted a set of possible uses of our dataset, from mapping the sources Wikipedia relies on, to recommending citations and using citation data as features. It is our hope that this release will start a collaborative effort by the community to study, use, maintain and expand work on citations from Wikipedia.

Data availability

The dataset is made available on Zenodo [40] 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](https://github.com/Harshdeep1996/cite-classifications-wiki/releases/tag/0.2)

Acknowledgements

The authors would like to thank Tiziano Piccardi (EPFL), Miriam Redi (Wikimedia Foundation) and Dario Taraborelli (Chan Zuckerberg Initiative) for their helpful advice.
References

[1] Wenceslao Arroyo-Machado, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, 15(2):e0228713, February 2020. URL: [https://dx.plos.org/10.1371/journal.pone.0228713](https://dx.plos.org/10.1371/journal.pone.0228713), doi:10.1371/journal.pone.0228713.

[2] Nicolas Aspert, Volodymyr Miz, Benjamin Ricaud, and Pierre Vandergheynst. A Graph-structured Dataset for Wikipedia Research. *arXiv:1903.08597 [cs]*, 2019. arXiv: 1903.08597. URL: [http://arxiv.org/abs/1903.08597](http://arxiv.org/abs/1903.08597), doi:10.1145/3308560.3316757.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. URL: [http://arxiv.org/abs/1607.04606](http://arxiv.org/abs/1607.04606), arXiv:1607.04606.

[4] Katy Börner. *Atlas of science: visualizing what we know*. MIT Press, Cambridge, Mass., 2010.

[5] Chaomei Chen. Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, 2(2):1–40, 2017. URL: [http://content.sciendo.com/view/journals/jdis/2/2/article-p1.xml](http://content.sciendo.com/view/journals/jdis/2/2/article-p1.xml), doi:10.1515/jdis-2017-0006.

[6] Chih-Chun Chen and Camille Roth. {{citation needed}}: the dynamics of referencing in Wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria, 2012. ACM Press. URL: [http://dl.acm.org/citation.cfm?doid=2462932.2462943](http://dl.acm.org/citation.cfm?doid=2462932.2462943), doi:10.1145/2462932.2462943.

[7] Giovanni Colavizza. Covid-19 research in wikipedia. *bioRxiv*, 2020. URL: [https://www.biorxiv.org/content/early/2020/05/12/2020.05.10.087643](https://www.biorxiv.org/content/early/2020/05/12/2020.05.10.087643), doi:10.1101/2020.05.10.087643.

[8] Mariam Farda-Sarbas and Claudia Müller-Birn. Wikidata from a Research Perspective – A Systematic Mapping Study of Wikidata. *arXiv:1908.11153 [cs]*, 2019. arXiv: 1908.11153. URL: [http://arxiv.org/abs/1908.11153](http://arxiv.org/abs/1908.11153).

[9] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. Finding News Citations for Wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 337–346, Indianapolis, Indiana, USA, 2016. ACM Press. URL: [http://dl.acm.org/citation.cfm?doid=2983323.2983808](http://dl.acm.org/citation.cfm?doid=2983323.2983808), doi:10.1145/2983323.2983808.

[10] Andrea Forte, Nazanin Andalibi, Tim Gorichanaz, Meen Chul Kim, Thomas Park, and Aaron Halfaker. Information Fortification: An Online Citation Behavior. In *Proceedings of the 2018 ACM Conference on*
[11] R. Stuart Geiger and Aaron Halfaker. When the levee breaks: without bots, what happens to Wikipedia’s quality control processes? In Proceedings of the 9th International Symposium on Open Collaboration, pages 1–6, Hong Kong, China, 2013. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2491055.2491061 doi:10.1145/2491055.2491061

[12] Aaron Halfaker, Bahodir Mansurov, Miriam Redi, and Dario Taraborelli. Citations with identifiers in wikipedia, 2018. URL: https://figshare.com/articles/Citations_with_identifiers_in_Wikipedia/1299540/1 doi:10.6084/m9.figshare.1299540

[13] James M Heilman, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham M Beards, David J Iberri, Matthew Harvey, Brendan Thomas, Wouter Stomp, Michael F Martone, Daniel J Lodge, Andrea Vondracek, Jacob F de Wolff, Casimir Liber, Samir C Grover, Tim J Vickers, Bertalan Mesko, and Michael R Laurent. Wikipedia: A Key Tool for Global Public Health Promotion. Journal of Medical Internet Research, 13(1):e14, 2011. URL: http://www.jmir.org/2011/1/e14/ doi:10.2196/jmir.1589

[14] Christoph Hube. Bias in Wikipedia. In Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion, pages 717–721, Perth, Australia, 2017. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3041021.3053375 doi:10.1145/3041021.3053375

[15] Dariusz Jemielniak and Eduard Aibar. Bridging the gap between wikipedia and academia. Journal of the Association for Information Science and Technology, 67(7):1773–1776, July 2016. URL: http://doi.wiley.com/10.1002/asi.23691 doi:10.1002/asi.23691

[16] Brian Keegan, Darren Gergle, and Noshir Contractor. 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, California, 2011. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2038558.2038577 doi:10.1145/2038558.2038577

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Kayvan Kousha and Mike Thelwall. 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, 2017. URL: http://doi.wiley.com/10.1002/asi.23694 doi:10.1002/asi.23694
[19] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In Proceedings of the 25th International Conference on World Wide Web, pages 591–602, Montréal, Québec, Canada, 2016. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2872427.2883085 doi:10.1145/2872427.2883085

[20] M. R. Laurent and T. J. Vickers. Seeking Health Information Online: Does Wikipedia Matter? Journal of the American Medical Informatics Association, 16(4):471–479, 2009. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M3059 doi:10.1197/jamia.M3059

[21] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and others. DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2):167–195, 2015. URL: http://iospress.metapress.com/index/J0N4J44332PLU18.pdf doi:10.3233/SW-140134

[22] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. Analysis of References Across Wikipedia Languages. In Robertas Damaševičius and Vilma Mikšytė, editors, Information and Software Technologies, volume 756, pages 561–573. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-67642-5_47

[23] Lauren A Maggio, Ryan M Steinberg, Tiziano Piccardi, and John M Willinsky. Reader engagement with medical content on Wikipedia. eLife, 9:e52426, March 2020. URL: https://elifesciences.org/articles/52426 doi:10.7554/elife.52426

[24] Lauren A Maggio, John M Willinsky, Ryan M Steinberg, Daniel Mietrench, Joseph L Wass, and Ting Dong. Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia. PLOS ONE, 12(12):e0190046, 2019.

[25] Connor McMahon, Isaac Johnson, and Brent Hecht. 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. AAAI Press, 2017.

[26] Mostafa Mesgari, Chitu Okoli, Mohamad Mehd, Finn Nielsen, and Arto Lanamäki. “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, 2015. URL: http://doi.wiley.com/10.1002/asi.23172 doi:10.1002/asi.23172
[27] Finn Aarup Nielsen. Scientific Citations in Wikipedia. *First Monday*, 12, 2007.

[28] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. Scholia, Scientometrics and Wikidata. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Lawrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events*, volume 10577, pages 237–259. Springer International Publishing, Cham, 2017. URL: http://link.springer.com/10.1007/978-3-319-70407-4_36 doi:10.1007/978-3-319-70407-4_36

[29] Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. Improving Website Hyperlink Structure Using Server Logs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 615–624, San Francisco, California, USA, 2016. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2835776.2835832 doi:10.1145/2835776.2835832

[30] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. Quantifying Engagement with Citations on Wikipedia. In *Proceedings of The Web Conference 2020*, pages 2365–2376, Taipei Taiwan, April 2020. ACM. URL: https://dl.acm.org/doi/10.1145/3366423.3380300 doi:10.1145/3366423.3380300

[31] Alessandro Piscopo, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl. Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References. In Claudia d’Amato, Miriam Fernandez, Valentina Tamma, Freddy Lecue, Philippe Cudré-Mauroux, Juan Sequeda, Christoph Lange, and Jeff Heflin, editors, *The Semantic Web – ISWC 2017*, volume 10587, pages 542–558. Springer International Publishing, Cham, 2017. URL: http://link.springer.com/10.1007/978-3-319-68288-4_32 doi:10.1007/978-3-319-68288-4_32

[32] Alessandro Piscopo and Elena Simperl. 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, 2019. ACM Press. doi:10.1145/3306446.3340822

[33] Reid Priehorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Conference on supporting group work*, Sanibel Island, Florida, USA, 2007. ACM Press. URL: http://portal.acm.org/citation.cfm?doid=1316624.1316663 doi:10.1145/1316624.1316663

[34] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact, 2012. URL: https://arxiv.org/html/1203.4745
[35] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia’s Verifiability. In *Proceedings of the World Wide Web Conference*, pages 1567–1578, San Francisco, CA, USA, 2019. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3308558.3313618, doi:10.1145/3308558.3313618

[36] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[37] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, and James Heilman. Evolution of Wikipedia’s medical content: past, present and future. *Journal of Epidemiology and Community Health*, pages jech–2016–208601, August 2017. URL: http://jech.bmj.com/lookup/doi/10.1136/jech-2016-208601, doi:10.1136/jech-2016-208601

[38] Richard M. Shiffrin and Katy Börner. *Mapping knowledge domains*. National Acad Sciences, 2004. URL: http://www.pnas.org/content/101/suppl_1/5183.short

[39] Xin Shuai, Zhuoren Jiang, Xiaozhong Liu, and Johan Bollen. A comparative study of academic and Wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL ’13*, Indianapolis, Indiana, USA, 2013. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2467696.2467746, doi:10.1145/2467696.2467746

[40] Harshdeep Singh, Robert West, and Giovanni Colavizza. Wikipedia Citations: A comprehensive dataset of citations with identifiers extracted from English Wikipedia, January 2020. URL: https://doi.org/10.5281/zenodo.3940692, doi:10.5281/zenodo.3940692

[41] Denise A. Smith. Situating Wikipedia as a health information resource in various contexts: A scoping review. *PLOS ONE*, 15(2):e0228786, February 2020. URL: https://dx.plos.org/10.1371/journal.pone.0228786, doi:10.1371/journal.pone.0228786

[42] Cassidy R. Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. 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, 2017. URL: http://doi.wiley.com/10.1002/asi.23833, doi:10.1002/asi.23833

[43] Misha Teplitskiy, Grace Lu, and Eamon Duede. 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, 2017. URL: http://doi.wiley.com/10.1002/asi.23687, doi:10.1002/asi.23687.
[44] Neil Thompson and Douglas Hanley. Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial. *MIT Sloan Research Paper*, 5238(17), 2018. URL: [https://www.ssrn.com/abstract=3039505](https://www.ssrn.com/abstract=3039505), doi: 10.2139/ssrn.3039505.

[45] Robert Tomaszewski and Karen I. MacDonald. A Study of Citations to Wikipedia in Scholarly Publications. *Science & Technology Libraries*, 35(3):246–261, 2016. URL: [https://www.tandfonline.com/doi/full/10.1080/0194262X.2016.1206052](https://www.tandfonline.com/doi/full/10.1080/0194262X.2016.1206052), doi:10.1080/0194262X.2016.1206052.

[46] Daniel Torres-Salinas, Esteban Romero-Frías, and Wenceslao Arroyo-Machado. Mapping the backbone of the Humanities through the eyes of Wikipedia. *Journal of Informetrics*, 13(3):793–803, 2019. URL: [https://linkinghub.elsevier.com/retrieve/pii/S1751157718302955](https://linkinghub.elsevier.com/retrieve/pii/S1751157718302955), doi:10.1016/j.joi.2019.07.002.

[47] Ping Wang and Xiaodan Li. Assessing the quality of information on wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology*, 71(1):16–28, 2020. URL: [https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24210](https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24210), doi:10.1002/asi.24210.

[48] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985. International World Wide Web Conferences Steering Committee, 2016. URL: [http://dl.acm.org/citation.cfm?id=2883077](http://dl.acm.org/citation.cfm?id=2883077).

[49] Zohreh Zahedi, Rodrigo Costas, and Paul Wouters. How well developed are altmetrics? A cross-disciplinary analysis of the presence of ‘alternative metrics’ in scientific publications. *Scientometrics*, 101(2):1491–1513, 2014. URL: [http://link.springer.com/10.1007/s11192-014-1264-0](http://link.springer.com/10.1007/s11192-014-1264-0), doi:10.1007/s11192-014-1264-0.