"I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics

Zagorova, Olga; Ulloa, Roberto; Weller, Katrin; Flöck, Fabian

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

RESEARCH ARTICLE

"I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics

Olga Zagorova, Roberto Ulloa, Katrin Weller, and Fabian Flöck

GESIS-Leibniz Institute for the Social Sciences

Keywords: altmetrics, data quality, data set, edit histories, Wikipedia editors, Wikipedia references

ABSTRACT

With this work, we present a publicly available data set of the history of all the references (more than 55 million) ever used in the English Wikipedia until June 2019. We have applied a new method for identifying and monitoring references in Wikipedia, so that for each reference we can provide data about associated actions: creation, modifications, deletions, and reinsertions. The high accuracy of this method and the resulting data set was confirmed via a comprehensive crowdworker labeling campaign. We use the data set to study the temporal evolution of Wikipedia references as well as users' editing behavior. We find evidence of a mostly productive and continuous effort to improve the quality of references: There is a persistent increase of reference and document identifiers (DOI, PubMedID, PMC, ISBN, ISSN, ArXiv ID) and most of the reference curation work is done by registered humans (not bots or anonymous editors). We conclude that the evolution of Wikipedia references, including the dynamics of the community processes that tend to them, should be leveraged in the design of relevance indexes for altmetrics, and our data set can be pivotal for such an effort.

1. INTRODUCTION

The collaborative online encyclopedia Wikipedia incorporates one of the largest reference repositories in existence. This is primarily due to the guidelines that Wikipedia has put in place to strongly encourage its users to make all article content verifiable. Enabling verifiability is achieved by providing a pointer to a reliable source that supports the statements or facts presented in the article text[1]. These pointers are added in the form of in-text citations that lead to reference lists. Thus, many Wikipedia articles include reference lists created and maintained by the community of users who are also collaboratively writing the Wikipedia articles. Every Wikipedia article text, its cited references, and reference lists are dynamic and can be modified or removed by users, with all changes being tracked in the article's revision history. Over the course of time, the revision history of the entire English Wikipedia has documented more than 55 million different sources[2]. Cited sources can be different types of publications, including, for example, formally published scientific papers, books, and news media articles, but also links to websites or any other type of web documents (Lewoniewski, Węcel, & Abramowicz, 2017).

[1] https://en.wikipedia.org/wiki/Wikipedia:Verifiability

[2] This comprises all references ever generated, but not necessarily still present, as of June 2019; see details about the data set in Section 5 and Zagovora, Ulloa et al. (2020).

an open access journal

Check for updates

Citation: Zagorova, O., Ulloa, R., Weller, K., & Flöck, F. (2022). "I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics. *Quantitative Science Studies*, *3*(1), 147–173. https://doi.org/10.1162/qss_a_00171

DOI: https://doi.org/10.1162/qss_a_00171

Peer Review: https://publons.com/publon/10.1162/qss_a_00171

Supporting Information: https://doi.org/10.1162/qss_a_00171

Received: 13 October 2020
Accepted: 26 October 2021

Corresponding Author:
Olga Zagorova
olga.zagovora@gesis.org

Handling Editor:
Ludo Waltman

The MIT Press

These references are exposed to an enormous readership, as Wikipedia is accessed by a wide audience around the world. With more than 250 million page views per day for the English Wikipedia alone [3], it is one of the top 15 most visited websites in the world[4]. While recent studies seem to indicate that a large number of users do not fully engage with references by visiting links or retrieving the referenced document otherwise (Piccardi, Redi et al., 2020), references still make statements more credible simply by appearing alongside them; and they are actively being interacted with more than 32 million times a month (measured by mouse-hovering over the reference footnote [Piccardi et al., 2020]). In addition, Wikipedia content, including its references, is incorporated into other data sources and projects, and thus reaches even wider audiences. For instance, Wikipedia content is used as a source for the collaborative knowledge base WikiData[5], which is again also used by other platforms. Scholia[6], for instance, creates scholarly profile pages based on WikiData.

Given its appeal to the general public, Wikipedia has also attracted a lot of attention in the scientific community, where it has become a subject of research itself. The research about Wikipedia includes, among others, the examination of recommendations and pitfalls when it comes to the analyses of its content (Bayliss, 2013; Denning, Horning et al., 2005; Eijkman, 2010; Luyt & Tan, 2010), studies that evaluate the accuracy of articles (Holman Rector, 2008) and of references (Bould, Hladkowicz et al., 2014), as well as efforts to attribute ownership of content to editors, such as WikiWho[7] (Flöck & Acosta, 2014).

Wikipedia has also become an object of interest in the field of *altmetrics*, an area of research dedicated to studying ways of measuring the impact of scientific work outside of traditional scholarly citation schemes, and often based on social media interactions (Kousha & Thelwall, 2017; Priem, Taraborelli et al., 2010). Altmetrics research is looking into different ways in which users of online platforms may interact with scientific publications (e.g., including the link to a publication in a tweet or saving a reference on a bookmarking platform), as these kinds of actions might indicate which publications have some sort of impact in a specific user community. The term *altmetrics* may also refer to a line of practical applications and tools that assign new types of indicators to rate publications' performance or impact by the interactions they receive through social media or other online platforms, typically based on the quantity of mentions of a publication.

Wikipedia data is considered in altmetrics data implementations (and sold) by aggregators in the field. Currently the most prominent are Altmetric.com[8], PlumX[9], CrossRef[10], and Lagotto[11]. Their indicators are applied in different settings, such as publishers' sites or repositories (e.g., institutional or discipline-specific publication databases), and they are used to advertise "impactful" publications (based on quantitative measures from user interactions). The metrics behind these indicators vary substantially between different aggregators. There is, for example, no standard for detecting or aggregating Wikipedia references, although it can be assumed that the use of document identifiers (DIDs), such as PubMed Identifiers

[3] https://tools.wmflabs.org/siteviews/?sites=en.wikipedia.org, as of March 10, 2021.
[4] https://www.alexa.com/topsites, as of February, 15 2020.
[5] https://www.wikidata.org/
[6] https://tools.wmflabs.org/scholia/
[7] https://www.wikiwho.net/
[8] https://www.altmetric.com/explorer/
[9] https://plumanalytics.com
[10] https://www.crossref.org/
[11] https://www.lagotto.io/docs/api/

(PMIDs) or Document Object Identifiers (DOIs), is a common practice among aggregators[12] (Haustein, 2016). The specific procedures are not transparent and altmetrics aggregators must be viewed as black boxes that could be subject to manipulations (Kousha & Thelwall, 2017), such as researchers adding references to their own publications into Wikipedia articles[13], or even strategic campaigns to insert publications from a specific publisher into Wikipedia articles[14].

In the broad context of altmetrics research and applications, the assumed unique value of Wikipedia as a data source is that it provides an immense repository of literature curated by a large editor community and likely legitimated as important sources by these "Wikipedians." With the self-control mechanisms and guidelines applied within this community, Wikipedia references are expected to meet basic quality standards (Lewoniewski, Węcel, & Abramowicz, 2020). At the very least, they are presumed to be topically relevant and ideally, they represent a comprehensive, up-to-date, and balanced collection of the most relevant sources. Given the dynamic nature of Wikipedia, it might also be possible to opportunely detect novel and trending publications through the additions and changes to the community-created repository of references. Overall, a (scientific) publication being cited in a Wikipedia article is considered an indicator of some form of impact for this publication (Kousha & Thelwall, 2017).

However, despite the academic interest in Wikipedia references and their practical implementation in some altmetrics indicators, relatively little is known about the origins of Wikipedia references and about their creators. With this paper, we want to illustrate that a better understanding about the nature of Wikipedia references can help to clarify their role as potential indicators for the general public's view of important sources. For this, it needs to be acknowledged that the dynamic nature of Wikipedia and the ability of users to perform and undo changes highly shapes Wikipedia's content and references, leading to various practical challenges in working with Wikipedia data and technical challenges in identifying and tracking references.

To illustrate some of the challenges in incorporating Wikipedia references into reliable altmetrics indicators, we will take a closer look at a particular example publication and how it is cited across articles in the English Wikipedia (as identified by our extraction method and data set that we will introduce below). Our example is based on several "Wikipedia references"[15] across different Wikipedia article pages pointing to (and thus citing) the publication "*Roy et al.*

[12] For example, Altmetric.com is collecting data using the following identifiers https://help.altmetric.com/support/solutions/articles/6000234171-how-outputs-are-tracked-and-measured, and CrossRef's collection uses DOI and landing page URLs https://www.crossref.org/services/event-data/.

[13] Wikipedia's guidelines about Conflict of Interest include a section on "Citing yourself," which allows self-citations within certain boundaries: see https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest. To the best of our knowledge, there are no studies that investigate in detail how common self-citations are in Wikipedia or that aim to identify misconduct in the area of self-promoting scientific articles through Wikipedia.

[14] One example can be found at: https://web.archive.org/web/20200323131800/https://annualreviewsnews.org/2020/02/25/seeking-a-wikipedian-in-residence/.

[15] Wikipedia's terminology related to references is not always consistent with the distinct definitions of citations and references existing in the field of information science. In the context of this paper, a "reference" is technically defined as the content included inside a Wikipedia <ref> tag, which is content pointing to some external sources (and thus conceptually citing them). This means that in the example, the Roy et al. (2001) publication is receiving citations from different Wikipedia article pages, as these pages have incorporated the respective pointers to the paper in <ref> tags (i.e., as Wikipedia *references*). We will use the term *references* thus in the remainder of this paper. See Section 3 for technical details on capturing Wikipedia references via <ref> tags.

(2001) Structure and function of south-east Australian estuaries. Estuarine, Coastal and Shelf Science 53(3): 351–384." The first reference citing this publication was added to a Wikipedia article in August 2012 (Figure 1, blue line), more than 10 years after the paper's release. Nine months later (June 1, 2013), there were already 53 articles that cited this publication. All of these articles received the reference to this publication from the same editor (Editor A). However, none of the references included the publication's existing DOI. The corresponding DOI to this publication was added to the different existing Wikipedia references during the first quarter of 2014 (Figure 1, orange line), and this was mostly done by one single editor in March 2014 (Editor C). In November 2018, another editor (Editor D) removed 27 instances (50%) of the references, although some of them were quickly reinstated (Figure 1, blue line).

This basic example illustrates several issues that motivated our work and that are largely overlooked, despite the widespread popularity and importance of Wikipedia in general and the use of Wikipedia data in the altmetrics field (mainly via altmetrics aggregators) as outlined above. First, the example highlights a weakness of mining Wikipedia references based only on document identifiers (orange line), which potentially misses numerous references, that led us to create an alternative method that uses the entire text of the reference (blue line); DOI-based approaches would miss the reference for the first 2 years of its existence. Second, it shows the impact that a single editor can have on the visibility of a reference by systematically adding or removing it from different articles—which at least challenges the concept of viewing publications that receive high numbers of citations from Wikipedia as being recommended by a community of users. Third, it exposes the general lack of understanding about Wikipedia editors as the creators and curators of Wikipedia and their impact of references being implemented. Fourth, it illustrates different editing activities (creation, modification, deletion, reinsertion) that affect the countable numbers of references, making Wikipedia a somewhat dynamic data source for altmetrics.

At the same time, the example captures the value of our investigation as an important step to close the gap in understanding the nature and quality of Wikipedia references in altmetrics.

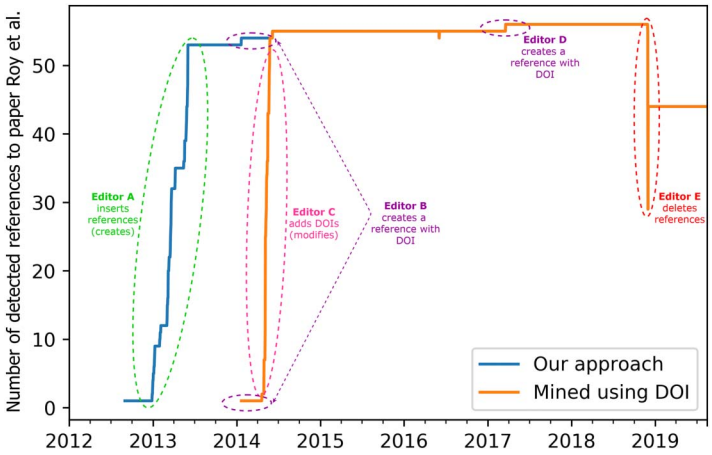


Figure 1. Wikipedia references in the English Wikipedia pointing to one example paper, as identified by our approach (blue line) and by approaches based only on document identifiers (orange line). Areas highlighted by circles correspond to edits made by one specific Wikipedia user: The green circle indicates an editor adding instances of the reference without any document identifier, the pink circle indicates an editor who modified existing references (e.g., by adding a DOI), violet indicates editors who create new references with DOI identifiers and red indicates editors who deleted references from articles.

It suggests that anomalies in the activity around Wikipedia references can be disclosed by tracking their *origin and evolution* within the articles, and that many of the collaborative negotiation processes that govern the inclusion, modification, and deletion of references can reveal information about the editor community responsible for the maintenance of this asset.

With this in mind, we present a novel data set (Zagovora et al., 2020) that contains individual revision histories of all Wikipedia references ever created in the English Wikipedia until June 2019. The data set is created by leveraging WikiWho (Flöck & Acosta, 2014), a service that tracks the additions, changes, and reinsertions of words (tokens) written in Wikipedia. Our evaluation (with crowdworkers) demonstrates its high accuracy at tracking references. To show the value of the data set, we investigate research questions in the following specific areas:

1. ***Insights into reference evolution over time***. The ongoing transformation and expansion of Wikipedia content affects the potential (measured) impact of cited sources by dynamically increasing or decreasing the number of reference instances that point to them. Therefore, Wikipedia presents a scenario that is different from other settings in citation analysis in altmetrics. Although the altmetrics field often deals with fluid types of data sources, as they include dynamic material[16] such as tweets or Facebook posts that might be deleted or modified, Wikipedia is unique as it relies on consensuses between members that can take time to reach an equilibrium, and which might be perturbed again as new information becomes available. References may be added by one person, removed by another, and then reinserted or edited again. These processes can repeat indefinitely, and little is known about how this has affected Wikipedia's references in the past and how many editing activities are performed on references overall. This leads to our first two research questions:

    - (RQ1) *How do Wikipedia references evolve over time?* We examine the fluctuation of all references of Wikipedia by analyzing the number of actions performed on them, providing the first longitudinal study of the evolution of references across all revisions in the English Wikipedia.
    - (RQ2) *What is the current and past coverage of references that include document identifiers (DIDs)?* For practical reasons, altmetrics indicators typically use DIDs for the detection of publications and references that lack DIDs are simply missed by methods that rely solely on them. We will tackle this question by estimating, at different points in time, the proportion of references that include DIDs, and by using current knowledge from our 2019 data set to calculate which references lacked DIDs in the past.

2. ***Insights about the editors of Wikipedia references***. We are interested in getting a better understanding of *who* adds, modifies, or deletes Wikipedia references. Learning more about the people who produce social media contents is just in its beginnings (Holmberg, 2015; Imran, Akhtar et al., 2018). We, therefore set out to answer the following:

    - (RQ3) *Who creates and maintains Wikipedia references, and in which way?* This question pertains to the characterization of the Wikipedia editor base engaging in different reference-related activities (e.g., automated bots or occasional users), and

[16] Although social media content containing altmetrics indicators (e.g., Facebook posts) is deleted to some extent after the initial altmetrics detection, we are not aware of aggregators' metrics that take these deletions into account. To the best of our knowledge, most aggregators are removing only deleted Tweets as per terms of Twitter data usage.

to the discovery of patterns of interaction with references exhibited by editors (e.g., focusing on reference maintenance). This more fine-grained picture of possible roles of editors in the reference ecosystem can help to understand the editor community that is responsible for the activity around the Wikipedia references.

The rest of this paper is organized as follows: Section 2 will offer an overview of the related work relevant for Wikipedia references and altmetrics, Section 3 is dedicated to the description of methods to build the data set, Section 4 presents an evaluation of our methods and the quality of the data set we provide, Section 5 presents general statistics of the Wikipedia references and main findings regarding our research questions, and Sections 6 and 7 conclude and summarize our findings.

2. RELATED WORK

The most comparable data set to the one we provide is presented by Halfaker, Mansurov et al. (2019) and Redi and Taraborelli (2018), which also includes a form of *historical* data about references in Wikipedia. However, the work differs from our approach because it relies on the presence of standardized DIDs as part of the reference—whereas our method does not—and thus is not capturing all references and is assigning editors and timestamps of origin to references according to the Wikipedia revision in which the identifier was included, even if in fact the reference as such was created earlier (cf. Figure 1). Lastly, modifications and deletions done to the references after the inclusion of the identifiers were not tracked. While the data set has been publicly shared with the community and was used (e.g., to study topics of citations), to the best of our knowledge it was not used to study the evolution of references or editing behavior related to references.

Other works only provide *static* (nonhistorical) snapshots of references in Wikipedia language editions, such as Nielsen (2008)[17] or Singh, West, and Colavizza (2021), that were created for specific tasks. Nielsen (2008) used the "cite journal" template from references to create a data set of journal papers that were cited in Wikipedia pages. This data set was then used to cluster Wikipedia pages and corresponding scientific journals into distinct research topics. Singh et al. (2021) created a data set of references and classified them into three groups: journal articles, books, and other Web content.

Recently, research has started to look more closely at how Wikipedia *readers* interact with references. With Wikipedia references being actionable items that users can click on, they have been described as a "bridge to the next layer of academic resources" (Grathwohl, 2011). However, recent studies (Piccardi et al., 2020; Redi, 2018) show that not all references are being equally visited by Wikipedia readers. Piccardi et al. (2020) conclude that, regarding references, "readers are more likely to use Wikipedia as a gateway on topics where Wikipedia is still wanting and where articles are of low quality and not sufficiently informative." They found that in most cases where Wikipedia articles are of high quality, readers do not follow the references but stay at the Wikipedia article as the "final destination" of their information journey (Piccardi et al., 2020). This kind of work gives us more insights into the *consumer perspective* of Wikipedia references, which adds to the general perspective of how Wikipedia is used (e.g., how Wikipedia articles are read or how people are citing from Wikipedia articles: Bould et al., 2014; Okoli, Mehdi et al., 2014).

[17] The data set is available via https://hendrix.imm.dtu.dk/services/wikipedia/citejournalminer.html.

To the best of our knowledge, there are only a few studies focusing on editors as the creators of references in Wikipedia and thus contributing to the *producer perspective*. With a comparatively small data sample (~5,000 articles), Chen and Roth (2012) showed that "a reference occurs when a set of committed and qualified editors are attracted to the article." Huvila (2010) conducted a survey of Wikipedia editors, also including questions broadly related to reference editing. Specifically, the survey enabled them to differentiate editors based on their information behavior and the sources the editors were using for editing articles. The results indicate a preference for sources that are available online. Kaffee and Elsahar (2021) extended the previous study by surveying editors about tools they use to create articles and to add corresponding references. There is also some specific, ongoing research on other and more general perspectives on the producer side of Wikipedia (e.g., on who edits Wikipedia), general editing patterns (Flöck, Erdogan, & Acosta, 2017), who becomes a power editor (Panciera, Halfaker, & Terveen, 2009), or how editors collaborate (Kittur, Suh et al., 2007; Murić, Abeliuk et al., 2019).

Furthermore, in the field of *altmetrics* research, a certain focus has been placed on untangling the relations between references in Wikipedia and the scientific publications they are citing. For example, altmetrics researchers have scrutinized the relevance of scientific publications mentioned on Wikipedia (Kousha & Thelwall, 2017; Sugimoto, Work et al., 2017). Shuai, Jiang et al. (2013) found that papers, authors, and topics that were covered by Wikipedia references have higher citation counts than those that were not mentioned. At the same time, only a narrow set of influential scientific works is cited on Wikipedia (Kousha & Thelwall, 2017). Nielsen (2007) showed that citations from Wikipedia are correlated with the total number of journal citations, whereas the correlation was weak with the journal impact factor. Yet, according to Nielsen (2007), Wikipedia editors tend to cite articles from high-impact journals such as *Nature*, *Science*, or *New England Journal of Medicine*. Teplitskiy, Lu, and Duede (2017) conducted a similar experiment with a newer data set and found that impact factor increases not only the probability of a paper being mentioned on Wikipedia but also open access principles. According to Mesgari, Okoli et al. (2015), the *quality of content* and of referenced sources was one of the major study objects on Wikipedia. For example, Lewoniewski et al. (2017) studied the similarity of sources from different Wikipedia language editions. They found that URLs in references shared many domain names between language versions, but there were not many cases of exact matches of URLs in references across languages. Lin and Fenner (2014) showed that ecology and evolution are better covered with references from *PLOS* than other subjects. Nevertheless, these results might not show the full picture when references are reported as incomplete and accompanied by the lack of standardization (Pooladian & Borrego, 2017).

The altmetrics community has investigated whether a citation in Wikipedia articles indicates that a scientific publication has an impact on the nonscientific audience (Lin & Fenner, 2013; Thelwall, 2016). Lin and Fenner (2013) argue that Wikipedia references might capture a "discussion" group, one of the engagement types with research publications. Our data set can enable a finer analysis of the revisions of references that can help to detect potential disruptions (e.g., sudden appearance of the same reference across various articles, or highly active individual editors who are responsible for large numbers of new references).

Zahedi and Costas (2018) and Ortega (2018) have started to compare different altmetrics aggregators to illustrate potential challenges for data quality. Differences start with coverage by aggregators. In the context of Wikipedia, this means that references appearing on Wikipedia make up from 2% of publications tracked by Altmetric.com up to 5.1% of those tracked by Lagotto. Those differences are due to the aggregator's methodology and the data sets of

publications they are tracking (Zahedi & Costas, 2018). These studies also observe different mean values for how often publications are mentioned on Wikipedia: Publications in the Altmetric.com collection are on average cited by 1.7 Wikipedia pages, publications in the Lagotto collection are on average cited by 2.9 Wikipedia pages, and publications in CrossRef Event Data are on average cited by 15.7 Wikipedia pages (Zahedi & Costas, 2018). We assume that these wide differences are not only due to the diverse sets of publications covered by the aggregators but also due to their distinct methods of tracing Wikipedia references, which are prone to various errors considering the challenges inherent to Wikipedia data. Besides the difficulties of keeping track of continuous changes in Wikipedia where references may be modified or removed, one important source of coverage errors (Sen, Flöck et al., 2021) is the reliance on standard document identifiers to trace publications (Ortega, 2018). Similarly, other approaches that rely on explicit bibliographic information, such as title and first author name (Kousha & Thelwall, 2017) fail to identify references that do not specify this information in the provided fields (Pooladian & Borrego, 2017). Given the quality of our data set, it has the potential to serve as an external base for comparing different data collection approaches used by altmetrics aggregators, giving them the opportunity to increase their coverage and impact indexes by looking at different points in time of the revision history.

3. CREATING THE REFERENCE HISTORIES DATA SET

In this section, we describe the central concepts and methodological details of the text mining process, extended by further information in Appendix A in the Supplementary material.

The resulting data set[18] is based on all revisions of all articles in the English Wikipedia edition since their origin until June 2019. It contains the change history of all 55,503,998 individual references ever created until this point in time, no matter if they contain a document identifier such as a DOI, ISBN, etc. or not. References are pointers to external sources (which may be any type of document) and are inserted into Wikipedia in a standardized way. They appear as "inline citations"[19] in the main body of the article, immediately after the statements they support, and are formatted by `<ref>` … `</ref>` tags in Wiki markup language. For our work we consider all such inline citations marked by ref tags as Wikipedia references[20].

In the following subsections, we explain our reference tracing and matching approach and how we extract document identifiers (DIDs) for those references that are assigned one at any point in time.

3.1. Extracting the Revision History of Individual References

The main content corpus of the Wikipedia encyclopedia is organized in articles. Each article A consists of an ordered list of revisions R (i.e., $A = [R_0, \ldots, R_n]$), where each revision is a new version of the text that was contributed by editor e at timestamp z. For the front-end HTML

[18] We also provide a Python notebook with examples on how to process the data, and the code can be directly executed on the GESIS Notebooks server. More details on the data format are in Zagovora et al. (2020).

[19] The Wikipedia community utilizes the term *inline citation,* which broadly speaking corresponds to the "in-text citation" as known from bibliometrics. See more details here https://en.wikipedia.org/wiki/Wikipedia:Inline_citation.

[20] Additionally, some references can be added automatically by dedicated templates. We are not considering materials that are not referenced as inline citations (e.g., publications from the "Additional reading" section), as the guidelines recommend to include references via <ref> tags (inline citations) as the standard (https://en.wikipedia.org/wiki/Wikipedia:Citing_sources).

Table 1. Type of actions that can be applied to a reference. The first column indicates the name of the actions that can be applied to references, and the second column the description of such an action

Action	Description
Creation (*)	First time the reference appears in an article
Modification	Changes to tokens of the reference (e.g., by correcting the name of an author)
Deletion	Complete removal of the reference
Reinsertion	Complete addition of a reference that was previously removed

* Note that only one creation per reference is possible.

representation, text inside the `<ref>` ... `</ref>` tags is converted by a Wikitext parser into a readable reference, placed at the bottom of the Wikipedia article in a dedicated reference section.

The revision history of a reference is given by the article revisions in which it was added and changed, either in its entirety or partially. As each revision within an article is associated with exactly one editor e (see Section 5.2 for a typology of editors), so is each action (see Table 1) performed on a specific reference through that revision.

Identifying the specific revisions in which the changes of Table 1 are applied to a given uniquely identified reference in Wikipedia presents two major challenges:

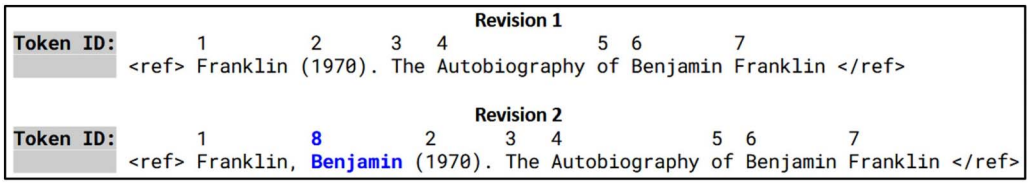
1. Tracking changes of any target text sequence is often error prone in Wikipedia (Flöck & Acosta, 2014). In these instances, standard text difference algorithms lose track of sequences and erroneously assign them as new content or as deleted[21].
2. Even if all changes to a reference are correctly tracked, deciding if a reference corresponds to another reference in two consecutive article revisions is nontrivial. For example, a large part of the reference might have been replaced or key tokens such as the title might have been modified.

To address these issues, we take advantage of WikiWho, an algorithmic approach that solves the change attribution problem at a token level with over 95% accuracy (Flöck & Acosta, 2014). Each token ever inserted in an article has been assigned a token ID that uniquely identifies it through all revisions. Figure 2 illustrates the allocation of token IDs for the two first revisions of a reference.

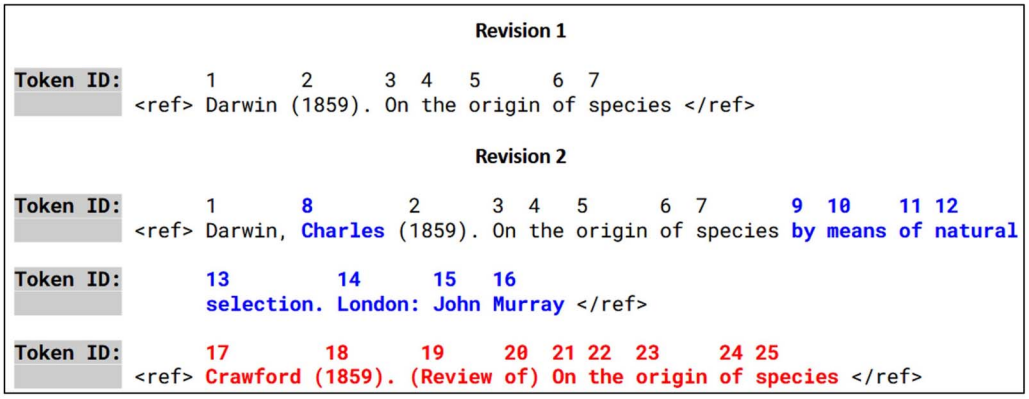
Our data set of references is organized per Wikipedia article, and we do not—for this work—match references across articles. Formally, for each article A, the data set contains a list of tuples $H_f = [<a_{f_0}, t_{f_0}, r_{f_0}, h_{f_0}, e_{f_0}, z_{f_0}>, ..., <a_{f_n}, t_{f_n}, r_{f_n}, h_{f_n}, e_{f_n}, z_{f_n}>]$ that represents the history of actions a_{f_i} ("creation," "insertion," "deletion," or "reinsertion") performed over reference f, where:

- t_{f_i} is the list of WikiWho token IDs that were part of the reference in revision r_{f_i},
- h_{f_i} is a hash value calculated of t_{f_i},
- e_{f_i} is an editor that performed an action a_{f_i} at time z_{f_i}, and
- H_f is sorted according to time z_f, where z_{f_0} is the oldest reference.

[21] Cf. Wikipedia diffs (https://en.wikipedia.org/wiki/Help:Diff) and Flöck and Acosta (2014) for a more general discussion.



Example **a)**



Example **b)**

Figure 2. Examples of token ID assignments before and after an edit. Example (a): For Revision 1, we assume the reference to be already existing and having been assigned token IDs 1–7. In Revision 2, "*Benjamin*" (blue) is inserted and WikiWho assigns token ID 8. Note how the older instance of "*Benjamin*" (ID 6) is tracked as a distinct token**.** Example (b): In Revision 2, "*Charles*" and "*by means of*…" (blue) are inserted and new token IDs (8–24) are assigned. Another reference "*Crawford (1859)* …" is added in Revision 2 with the tokens identified as new.[22]

To build this data set, we mine all inline citations of all Wikipedia revisions using the WikiWho token IDs that correspond to the string tags `<ref>` … `</ref>` and `<ref name=…>` … `</ref>`; the void tags (i.e., the one-sided tags `<ref name=… />`) are excluded because they correspond to duplications of existent references. For each revision R_i in each article A, we then have a list of references that belong to that revision $G_i = [f_0, ..., f_m]$, where each reference f_j is a tuple $<t_j, h_j, e_j, z_j>$.

The next step is to associate the references in G_i to those in G_{i+k}, so that two references are added to H_f if they are *equivalent* (i.e., they refer to the same publication).

In trivial cases, a reference f does not change between article revisions so we use the hash values to match all identical references across all G, and we store the matched references of f in H_f. For now, each H_f is incomplete, as there could be two reference histories H_f and H_g that belong together, because with this procedure, even a small modification is enough to change the hash value. Therefore, all actions a_f are tagged as "unknown."

In the nontrivial cases, the references have been modified between two consecutive revisions. We then rely on the Jaccard similarity between the lists of WikiWho token IDs of the references. The core idea is that a reference f' is considered the successor of the reference f if the Jaccard similarity between f' and f is higher than 0.2 (see the evaluation in Section 4), or if the token IDS of f are all contained in f' (i.e., $t_f \subset t_{f'}$); f' is not already the successor of another reference; and the revision $r_{f'}$ happened after the revision r_f (i.e., $z_{f'} > z_f$). Also, if f' is a successor of f, then the action is considered a modification if the revision $r_{f'}$ happened

[22] These toy examples do not track punctuation for simplicity, while WikiWho does so in practice.

immediately after the revision r_f (i.e., there is no revision between r_f and $r_{f'}$). Otherwise, a **deletion** occurred in revision r_f and a **reinsertion** in $r_{f'}$. The exact details of the procedure applied to each reference *f* is presented in Figure A1 of Appendix A in the Supplementary material.

3.2. Tracking of DID References

The content of a reference may include different types of document identifiers (DID) that have been assigned to the referenced source during its publication process (e.g., a Digital Object Identifier: DOI). DIDs can easily be used to trace individual references unambiguously, both within Wikipedia and outside of it. While with our approach and data set we extract and monitor all references in a Wikipedia article, we take a closer look at the subset of references containing DIDs for two reasons: First, this enables comparisons with previous works, which have relied exclusively on document identifiers to extract references for Wikipedia articles. Second, Wikipedia includes references to publications that range from strictly refereed and well-reputed scientific outlets to everyday blogs, Twitter profiles, and Reddit posts, and we aim to utilize DIDs to put one focus of our investigation on such publications relevant to altmetrics and the academic community and compare them to the complete set of references. Although DIDs can be an indicator that a reference is academic[23], we are mindful that references with DIDs are not necessarily academic works. Yet, they provide a viable filter to concentrate on references relevant in the context of this work.

Therefore, an important aspect of the evolution of Wikipedia references is the *point in time* at which DIDs are added to references in the version history. A reference that currently has a DID could have been missing it in the past. By using the present information and by looking back into the past, we can estimate how many references were lacking DIDs, and thus would have been omitted by approaches that rely solely on the presence of DIDs for identifying and counting Wikipedia references.

We distinguish between several types of references based on DID information (Table 2). The term *DID-Reference* (DID-R) corresponds to references that by the time of our data collection (June 2019) had a DID. If the DID was immediately included when the reference was created, we refer to it as *DID-Born Reference* (DBorn). Otherwise, if the DID was added after the reference was created—usually because the referenced work had been assigned a DID at a later point in time or it was erroneously omitted upon reference creation—we call it *DID-Lagged Reference* (DLag). Their counterparts (i.e., references that by the data collection date did not have a DID) are called *No-DID References*. Note that this classification depends on the time of data collection, as some of the DID-Lagged References would have been classified as No-DID References in previous years and current No-DID References may still receive a DID at a future point in time.

After we trace the history of all references for each Wikipedia article as explained in the previous subsection, we proceed to extract the DIDs for all the versions of each reference. We used modified versions of regular expressions based on Halfaker et al. (2019) to extract the following DIDs: Digital Object Identifier (DOI), International Standard Book Number (ISBN), PubMed Identifier (PMID), PubMed Central identifier (PMCID), International Standard Serial Number (ISSN) and arxiv.org Identifiers (ArXiv ID). Once we extract the DIDs (see Figure D2 in Appendix D of the Supplementary material for distributions), we can retroactively recognize the DLag references and their content (t_f), as our data set already contains historical information

[23] We use the term *academic* instead of *scientific* to indicate the inclusion of all works not only from "harder" sciences but also from social sciences and humanities. This is in line with Halfaker et al. (2019).

Table 2. Types of references according to if and when a DID was added. The first and second columns indicate the names that we use to identify the type and subtype of reference respectively. The third column describes the subtype of references based on when the DID was added

Type	Subtype	Description
DID Reference (DID-R)	**DID-Born Reference (DBorn)**	References that already included a DID when they were created.
	DID-Lagged Reference (DLag)	References that did not include a DID when they were created, but were assigned a DID at a later point, before the time of our data collection.
No-DID Reference (No-DID)		References that did not include a DID by the time of data collection. These *might* receive a DID later (after our data collection) if a DID in fact exists for the referenced publication.

for each reference (H_f). Our method properly handles cases in which a reference has two identifiers (e.g., correction of a DID, or one DOI and one ISBN). We keep the timestamp (z_{f_i}) and editor (e_{f_i}) that introduced or modified the DID, so that we can further analyze the dynamics of creation and addition of the DIDs.

4. EVALUATION OF THE REFERENCE CHANGE TRACKING METHOD

In this section, we evaluate the performance of our method for tracking version histories for references. We describe a gold standard data set that we created for evaluation purposes using crowdworkers, present the overall performance, and compare our method to a baseline relying on cosine similarity.

4.1. Gold Standard Data Set

To make sure that our method correctly identifies references in different forms across histories, we created a gold standard data set of 952 pairs of references, in which each pair looks similar to the example in Figure 2(a). The pairs are labeled as *Equivalent* or *Distinct,* depending on whether each pair corresponds to the same bibliographical resource or not. Each pair of references was judged by at least three FigureEight[24] crowdworkers. Each worker indicated if the pair corresponds to the same resource or different resources, or if it was not clear. See Appendix B in the Supplementary material for the instructions we provided for FigureEight crowdworkers, an example question, and a note on fair payment (Zaldivar, Tomlinson et al., 2018).

If the agreement[25] between the workers fell below the limit of 0.7, additional crowdworkers were assigned to the task until the agreement reached the required limit (0.7), or until at least five individuals had made judgments. Prior to the task, each worker was trained with a selection taken from 115 examples that illustrated different cases, and they had to correctly label at least five out of six test pairs of references. All the answers from a given worker were discarded (and a new worker assigned) if their accuracy fell below 0.8. Training and test items have been prelabeled by the authors of this paper.

[24] https://www.figure-eight.com (formerly known as Dolores Labs, CrowdFlower) was acquired and renamed by Appen as of April 8, 2020.

[25] We adopted the "confidence score of the row" of the FigureEight platform. This value describes the level of agreement between multiple contributors, where the sum of the contributors' trust scores of the most common answer is divided by the sum of the trust scores of respondents to that question. See details here: https://success.appen.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score.

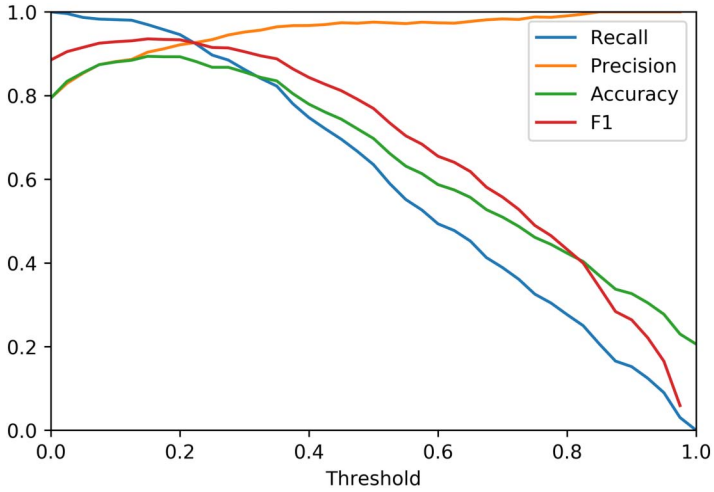


Figure 3. Performance metrics for identifying equivalent references. The *x*-axis shows the threshold of Jaccard similarity between pairs and the *y*-axis shows the Precision (Blue), Recall (Orange), Accuracy (Green), and F1 (Red) scores.

One thousand items were presented to the workers, out of which 952 were labeled as either Equivalent or Distinct. No final annotation was reached for 48 pairs of references (i.e., five assigned workers did not agree above the 0.7 limit[26]).

The set of 1,000 items was taken using a stratified random sample from all the references in Wikipedia revisions (Appendix C in the Supplementary material). The set consists of eight strata with similarities from 0 to 1 with 0.125 steps, and 125 pairs of references per stratum. Therefore, we make sure that our sample covers the full range of the Jaccard similarity scores used in our method, as with a pure random sampling most pairs of references would have fallen into the extreme values of similarity (i.e., 0 or 1) and would have constituted mostly trivial examples.

4.2. Performance

We compared the 952 pairs of references labeled by the crowdworkers against the labels assigned using our method. Figure 3 illustrates the performance metrics for different Jaccard similarity thresholds in our method. Based on this data, we selected a threshold of 0.2 as a trade-off between precision and recall.

To find the overall performance metrics for our method we resampled our stratified sample so that it is representative of the original distribution of Jaccard similarities (calculated with a 100,000 sample) of pairs of references extracted in the same fashion as described in Section 3.1. Table 3 presents the micro-average performance metrics for the identification of the same and different references between revisions.

Upon labeling the 48 cases—in which the crowdworkers could not agree—ourselves, we found that in 30 cases our method was able to decide appropriately based on the contextual information that is encoded in the WikiWho data model. Overall, our method maps identical references between revisions with very high confidence.

[26] One of the researchers closely inspected these cases and confirmed that the low agreement score stemmed from the ambiguity of the items. The inspection was done using contextual information from the text surrounding the references in previous revisions, testing URLs, and external resources (e.g., search engines, archive.org).

Table 3. Micro-average performance metrics for the labeling of pairs of references. The three metrics are calculated so that they represent the original distribution of Jaccard similarities in the method by resampling from the stratified sample. Each evaluated pair of references contributes equally to the score (regardless of the strata they belong to)

Precision	Recall	F1 score
0.96	0.96	0.96

4.3. Baseline Comparison

To our knowledge, there is currently no other approach that maps references over Wikipedia revisions, and thus no direct comparison for our approach. Therefore, we implemented a straightforward baseline that maps references using cosine similarity between Bag of Words representations of the strings of the Gold Standard reference pairs. We then resampled using the distribution of cosine similarities calculated in the original data. To estimate the distribution we used the same procedure of random sampling (Section 4.1) but we assume that the buckets have an infinite size (Appendix C in the Supplementary material, Step 1), and stop after 100,000 pairs of references have been sampled. Figure 4 shows how our method, leveraging WikiWho and Jaccard similarities, outperforms the alternative based on cosine similarity between reference strings through all possible thresholds.

5. DATA SET COMPOSITION AND ANALYSIS

Our data set contains the references of 6,073,708 nonredirect[27] articles in the English Wikipedia. It comprises 55,503,998 references with 164,530,374 actions. The actions consist of 33.73% creations, 31.3% modifications, 23.15% deletions, and 11.81% reinsertions. We find that 77.21% of the articles (4,690,046) have at least one reference (median = 4, μ = 11.83, max = 12,797). But out of those articles, 78.42% do not yet have any DID-Rs (3.68 million; i.e., 60.54% of total articles, Figure D1 in Appendix D in the Supplementary material). The rest of the articles (1,012,289) have at least one DID-R, and 50,615 (5%) articles contain more than 50% DID-Rs. More than 88% of the DIDs currently used to track the references correspond to ISBNs and DOIs (Figure D2 in Appendix D in the Supplementary material).

As of June 2019, only 7.11% (3,943,984) of all references include one of the identifiers we were tracking. The distribution of articles according to the number of references when either all of them are included or when only DID-Rs are included, suggests a power law distribution; however, the distribution is smoother for all references (α = 1.66) compared to only DID-Rs (α = 2.38); see Figure D3 in Appendix D.

About 10% of all DID References are *DID-Lagged References* (i.e., they did not have DIDs in their early Wikipedia article revisions; Table 2). By now—and in the future—this number will likely be higher, as DIDs can still be added to the references that were classified as *No-DID References* in our 2019 data set. We also observe that 12.1% of actions on the DID References occurred during the initial revisions in which the references did not yet have a DID; hence, this information would not be considered in any approach that relies only on DIDs for identifying and monitoring references.

[27] We excluded Wikipedia pages that are redirects. Redirects are Wikipedia pages that automatically send visitors to another page and do not have their own content. Example: https://en.wikipedia.org/w/index.php?title=Symbiont&redirect=no.

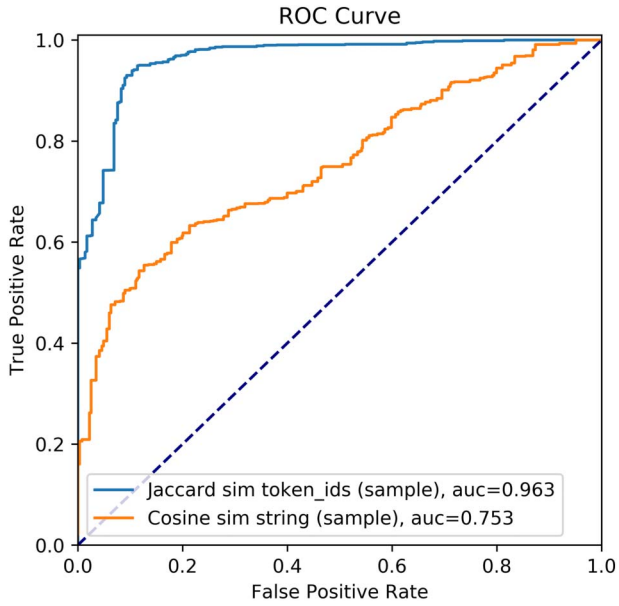


Figure 4. ROC curves to compare our method and a simple method based on cosine similarity. The light blue line shows the ROC curve for our method based on Jaccard similarity over WikiWho token IDs, and the orange line the ROC curve for a method based on cosine similarity of strings. Each data point is calculated for each possible threshold in the sample data.

In the following section, we will take a closer look at the data to find answers to our research questions. We will first look at the temporal evolution of different types of references (based on the presence of DIDs), and second at the editors who are creating and editing the references.

5.1. Wikipedia References Over Time

The first reference in an article of the English Wikipedia edition was introduced in December 2005. Since then, more and more references have been added yearly (Figure 5). There was an initial steep increment of new references per year until 2010, in which more than 4 million references (which corresponds to 7.4% of all references) were created. After that, the increment of yearly created references continued more moderately, and it seems to have settled in 2017 and 2018: about 5.58 million (10.05%) and 5.64 million (10.15%) of all references were added in the respective years.

After references have been created, some of them have never changed in any way, while others have been either deleted or modified at least once. According to our data, modifications are the most common action (~51.5 million) that happens to references after their creation. The number of modifications per year has not grown monotonically as we have seen for creations; for example, there is a peak of modifications between 2016 and 2018: 6.41 million in 2016, 8.10 million in 2017, and back to 6.71 million in 2018. We suspect that the increase in modifications 2016–2018 is due to the WikiCite[28] project and a sequence of editing events that started in 2016. The ratio of modifications to creations has been increasing, but during that period (2016–2018) the ratio went above 1 (i.e., there were more modifications than creations), reaching 1.4 in 2017 (Appendix N in the Supplementary material).

[28] https://meta.wikimedia.org/wiki/WikiCite_2016

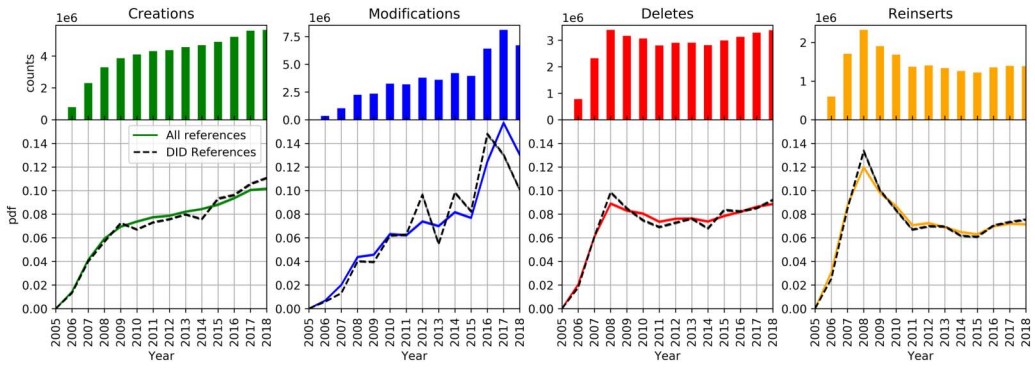


Figure 5. Distribution of actions over time. Each of the four plots depicts the dynamics of one of the actions: creations, modifications, deletions, and reinsertions. On the top subplot of each action, bars represent the number of actions (*y*-axis) performed over all references per year (*x*-axis). For example, around 2.3 million references were created in 2007. On the bottom subplot of each action, the solid lines represent the proportion of actions (*y*-axis) that occurred yearly (*x*-axis) for all references. The dashed lines represent the proportion of actions that occurred yearly (*x*-axis) for only the DID References (DID-Rs). For example, around 8.9% of all deletions were done in 2008, whereas for DID-Rs around 9.9% of deletions were done in 2008.

Apart from 2005 and 2006 (years with small reference counts), the proportion of deletions has shown a decreasing trend until 2014. This was most likely due to cleanup efforts of initial reference additions, plus high volatility (e.g., because of disagreements), also shown in the high reinsertion counts until 2010, which are, by definition, a reaction to previous deletions.[29] Starting at its high count in 2008, the number of reinsertions dropped unevenly from 2.33 million (11.98%) actions in 2008 to 1.39 million (7.14%) actions in 2018.

One might expect the same distribution of actions across years for DID References (i.e., that they would be treated by editors in the same way as general references). Yet, there are some differences between general references and DID-Rs. The most distinct patterns are noticeable in the creations and modifications of references (the dashed and solid lines in Figure 5):

- Until 2009 the number of creations of DID-Rs was aligned with creations of all references (overlap of the dashed and continued line). However, between 2010 and 2014 fewer (than expected) DID-Rs were created, and after 2015 the trend was reversed. For instance, in 2018, around 11.06% of new DID-Rs (versus 10.15% of general references) have been added to Wikipedia articles.
- There is no clear trend in the modifications of DID-Rs (the second plot from the left in Figure 5), as the plot shows multiple peaks and troughs across the years. We observe fewer modifications of DID-Rs in 2007–2009, 2013, 2017, and 2018; and more modifications in 2012 and 2014–2016. The highest number of modifications was reached in 2016 (1.02 million actions or 14.79%) and 2017 (0.9 million actions or 13.03%).
- The relatively small differences in deletions of some years (2008, 2010–2012, 2014, and 2015 in Figure 5) do not necessarily mean that their presence ended in those years (because they can be reinserted). However, we found that DID-Rs have a higher survival rate: They are deleted (without further reinsertions) at a lower rate than the rest of the references at any point in time (see Figure E1 of Appendix E in the Supplementary

[29] The years 2006–2010 in the English Wikipedia have been pointed out as a highly volatile period before (Flöck et al., 2017).

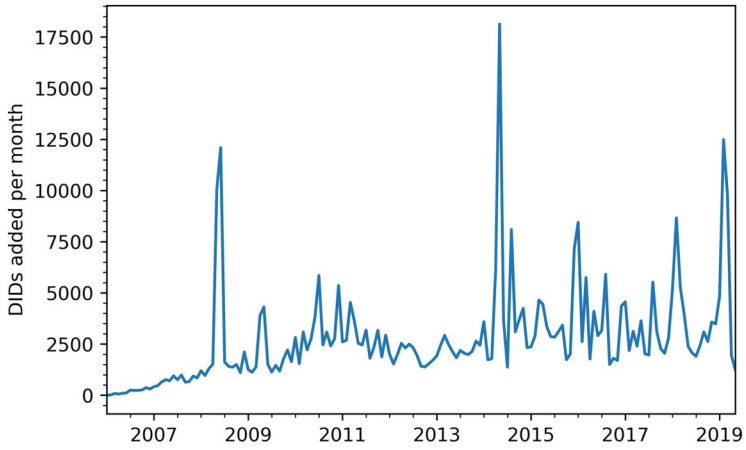


Figure 6. Monthly total of modifications that added a DID to existing references. The *x*-axis displays the year and *y*-axis the number of modifications—no matter which year the reference was originally added (e.g., in 2019, references from several years earlier were changed along with references created the same year). Only modifications in which a DID was added to a reference are considered.

material). As of June 2019, around 31.8% (17.02 million) references had been deleted (without further reinsertions) between 2005 and 2019; 0.97 million of them are DID-Rs, representing only 25.7% of all DID-Rs. This speaks to a higher value of these references to the editor community, possibly because of their perceived trustworthiness.

We observed in Figure 5 (second subplot from the left) that there are differences in the overall number of modifications, and the number of modifications of DID references. Some of these modifications are of particular interest because they are the ones in which DIDs are added to already existing references (DLag, Table 2). Therefore, we have closely investigated these modifications (Figure 6). The highest peaks of newly added DIDs occurred during (a) May and June 2008 with 22,126 DIDs added during two months, (b) May 2014 with 18,131 DIDs added, and (c) February 2019 with 12,486 DIDs added. This indicates the presence of campaigns (or individual editors' efforts, with the help of scripts or bots) that targeted missing DIDs. Based on information until June 2019, these three peaks correspond, respectively, to (a) 19–26%, (b) 17%, and (c) 56% of references that at the time should have had a DID (see Appendix F in the Supplementary material for statistics of other peaks). Putting it the other way around, 44–83% of the references remained without a DID even after pronounced waves of DID additions.

The reported percentages of missing DIDs will be even higher in the future (after June 2019), as more DIDs will be added to references that existed at those peaks. Hence, we also analyze how long it takes for the reference to be attributed with DIDs. Figure 7 presents the distribution of time spans between reference creation and DID introduction for references created in three different years (see Appendix G in the Supplementary material for all the years). In 2006, it took between 500 and 1,000 days for most of the references to gain their DID. In contrast in 2018, it took less than 10 days for most of the references to get a DID. There are clear peaks in the plots corresponding to 2006 and 2012 (Figure 7), around 500 and 750 days after the reference was created. These peaks can be associated with the spikes of DID additions in May and June 2008 and May 2014 (in Figure 6). However, the spike of February 2019 (Figure 6) can barely be observed in the 2018 plot (~300 days, Figure 7), indicating that most of the modifications of 2019's spike corresponded to references created before 2018. This

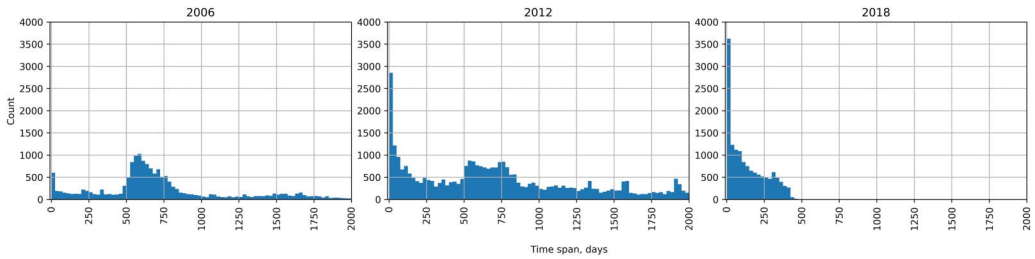


Figure 7. Distribution of the time spans between the creation of the references and the introduction of their DID for the years 2006, 2012, and 2018. The *x*-axis shows the time span in days between reference creation and the introduction of their DIDs—*only including the references created in each of the years in the titles of the plot*. The *y*-axis shows the frequency for each of the time spans. See Appendix G in the Supplementary material for distributions of all other years.

suggests that the editor community and infrastructure have been getting more effective at identifying and adding missing DIDs for references.

As we have already mentioned, DID References correspond to 7% of all the references in our 2019 data set. Had we collected the data set in other years, the percentages would have been slightly different (solid line, Figure H1 of Appendix H in the Supplementary material), especially before 2010. For example, there would have been around 6.6% DID References at the beginning of 2007. After 2010, the number of DID References has stabilized around 7%, with a small increase in the last four years.

Hypothetically, one could collect the histories of references using only DIDs (see Appendix I in the Supplementary material). In that case, one would observe ~4.4% DID References in 2007 (dashed line, Figure H1 of Appendix H in the Supplementary material) while the true number should have been at least 6.6% references; the alternative method would have missed 37.5% (~2.2% out of ~6.6%) of references that got their corresponding DID after the hypothetical data collection. These differences are discussed in more detail in Section 6.

5.2. The Editors of Wikipedia References

In the context of altmetrics, the focus is often placed on which scholarly works receive mentions or interactions from social media or other alternative platforms, while relatively little is known about who is behind these mentions and interactions. In collaborative platforms such as Wikipedia, it is relevant to understand the actors who participate in the inclusion of scholarly publications, as this has a direct impact on visibility. In contrast to traditional publications, where the decision about which material should be cited is attributed to the authors of each publication, in a collaborative environment, the decision is not straightforward but may have to be negotiated over different article revisions. In this section, we investigate whether contributions come from registered editors, bots, or nonregistered sessions (IP addresses) (see Table 4), and explore the behavior of these actors within Wikipedia. We are interested in whether those who edit Wikipedia references differ from the overall Wikipedia editor community, and we inquire if there exist subcommunities of editors that specialize in different types of editing activities.

We found 1,910,667[30] editors, 1,172 bots, and 23,459,838 edits by 4,286,160 IP addresses that worked with Wikipedia references (Table 4). Figure 8 presents the distributions of actions per user type. Registered editors are responsible for most actions: more than 122 million (74%

[30] For comparison, the English Wikipedia had 35.7 million registered editors as of July 2019.

Table 4. Types of Wikipedia editors. The first column lists the types of editors, the number of reference-editing actors of each type, and their actions that we encounter in our data set. The second column elaborates on each

Type of editors	Description
Registered editors Actors: 1,910,667 Actions: 121,681,174	These correspond to individual users who have registered their profile on Wikipedia and edited at least one reference.
Bots Actors: 1,172 Actions: 19,386,851	Bots were identified from bot lists of Wikimedia plus an additional list of bots' names that we created. These sources were combined into a final list consisting of 10,262 unique account names (see Appendix J in the Supplementary material for the sources), out of which 1,172 were associated with at least one action in Wikipedia references.
Nonregistered editors Actors: N/A Actions: 23,459,838	Edits coming from nonregistered IP addresses cannot be attributed to specific anonymous editors. Several persons can share the same IP address (e.g., university addresses or libraries), and one editor can connect via several IPs.

of all actions in our data set). Registered editors focused on the creation of new references (40% of their actions) and modification of existing ones (28.2% of their actions). Bots, in comparison, with 19.4 million (13.7%) of all actions, were focused on modifications (71% of their actions). Nonregistered editors are responsible for only 14.3% of the actions in our data set. And although registered editors made most deletions (around 24.5 million; left plot in Figure 8), nonregistered editors appear to specialize in them (right plot): Nonregistered sessions have proportionally more deletions (53.3% of all their actions) than either registered editors (20.2%) or bots (5.6%), not unlikely due to large amounts of vandalism, especially blanket deletions of large chunks of text. The nonregistered editors generally comprise a diverse and occasional set of editors, and 89.8% of IPs have fewer than 10 actions. Some IPs might be associated with several editors (e.g., school IPs), while a user might also use several IPs. Given the comparably low figures of actions for nonregistered editors but mostly the difficulties of attributing actions to specific actors, we will exclude them from the rest of the analysis in this section.

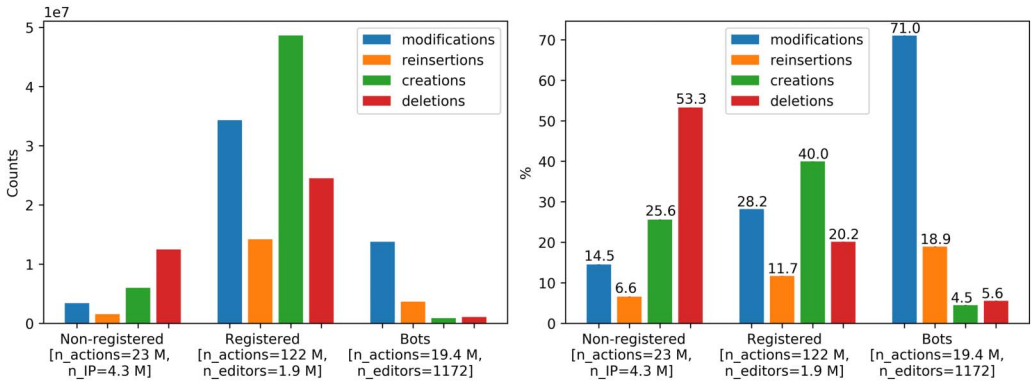


Figure 8. Distribution of actions performed by each type of editors. The left plot shows the total actions (*y*-axis) per type of account (*x*-axis), and type of action (legend). The right plot shows the percentage (*y*-axis) of the type of actions (legend) within the account type (*x*-axis). The *x*-axis also presents the total number of actions (n_actions) and editors or IP addresses (n_editors or n_IP) for each account type.

Table 5. Classification of registered editors according to the type of activity. The first column presents the cluster id and the percentage of registered editors in parentheses. The second column describes the group of registered editors in terms of the actions they perform (Figure L1a of Appendix L in the Supplementary material)

Cluster (% of all editors)	Type of activity
0 (39.56%)	Only create new references
1 (20.55%)	Only delete references
2 (10.3%)	Modify references in 90% of the cases, create new ones in 6% of the cases
3 (11.17%)	Mostly delete and create references (42% of cases for each action), modify in 10% cases and do a few reinsertions
4 (4.06%)	Mostly (70%) reinsert deleted references and do a few deletions, creations and modifications
5 (14.36%)	Mostly create (55%) and modify (35%), and do a few deletions and reinsertions

Most registered editors have performed only a few actions on references in Wikipedia articles, whereas the top contributors have contributed millions of actions (Figure K1, Appendix K in the Supplementary material). We also studied the number of different articles in which each editor has performed actions on references. We see a similar trend as with the number of actions (e.g., the top user has edited references in 226,334 articles). This seems to suggest that some editors are specifically focusing on reference editing beyond a specific topical area of interest.

Using a manually curated list of Wikipedia bots (10,262 unique bot account names; see Table 4), we found that 1,172 bots (0.1% of editors) have taken part in the editing of references. On a per user basis, bots performed more actions on references than registered users (Mann-Whitney U test, $p < 0.001$; Figure K2, Appendix K in the Supplementary material).

Bots and registered editors display very different behavior that is evident by directly looking at the types and quantity of actions (Figure 8). Within the group of registered editors, we were interested in identifying subgroups of users who behave similarly (and distinct from other subgroups), as measured by the types of actions that they usually perform. We use the K-means clustering algorithm with Euclidian distance to find such groups. Each registered editor is represented by four features, one per type of action, that contain the distribution (in percentages) of actions of that editor. We applied the algorithm on a sample of 10,000 random editors.

Table 6. Similarity of action groups with most active Wikipedians. The first column displays the criteria used for our ranking. The second to seventh columns show the Jaccard similarity of top x editors of both rankings. The last two columns represent the RBO scores for two values of the parameter p of RBO

	Jaccard similarity of top x						RBO scores for $p = x$	
Actions	**10**	**100**	**500**	**1,000**	**5,000**	**10,000**	**0.95**	**0.9999995**
Total	0.05	0.25	0.32	0.38	0.48	0.52	0.001	0.635
Modifications	0.11	0.18	0.25	0.29	0.42	0.46	0.002	0.428
Creations	0.00	0.12	0.20	0.26	0.37	0.41	0.003	0.449
Deletions	0.11	0.17	0.25	0.31	0.41	0.42	0.001	0.363
Reinsertions	0.05	0.09	0.16	0.20	0.34	0.37	0.001	0.297

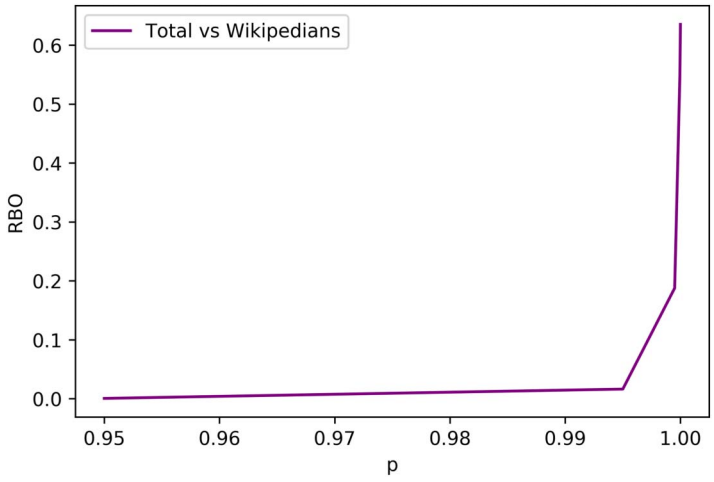


Figure 9. RBO scores between ranked lists of the 10,000 active Wikipedians and the 10,000 most active editors of references. The *x*-axis shows the decrease of the RBO with increasing weight on the top of the ranked lists. The sharp decrease even at low levels of top weight points to the dissimilarity of the editors at the very top.

To determine the optimal number of clusters the following analyses were performed: silhouette coefficients (Rousseeuw, 1987), presented in Appendix L in the Supplementary material, and clustering tree algorithm (Zappia & Oshlack, 2018) with Sugiyama layout (Sugiyama, Tagawa, & Toda, 1981) for tree depiction (Appendix M in the Supplementary material). According to this analysis, we choose to divide the editors into six clusters ($k = 6$, mean silhouette score of 0.69), which are summarized in Table 5.

A clear behavioral pattern can be observed for each of the six clusters. We observe that clusters 0 and 1 are perfectly defined (i.e., all their members dedicate themselves exclusively to creating references [cluster 0] or deleting them [cluster 1]), clusters 2 and 4 focus on modifications and reinsertions respectively, and clusters 3 and 5 are slightly more mixed, focusing on two actions each.

Additionally, we investigate whether editors of references are different from the general Wikipedia editor community (Wikipedians). We therefore compared the 10,000 most active reference editors in our data set with the most active Wikipedians according to the Wikimedia Foundation.[31] The ranking of the most active Wikipedians is based on the total number of revisions they have created, whereas we have used five different rankings of reference editors based on the total counts of actions, modifications, creations, deletions, and reinsertions.

To see if highly active reference editors correspond to highly active Wikipedians, we look at the general overlap of the two lists via Jaccard similarity (Table 6) of different top k groups of editors. We find that the very elites of the reference editors and general editors differ; 52% of editors are in both lists of 10,000 most active users. The lack of overlap is more notable for groups specialized in certain types of actions. For example, the most active reinserters have a Jaccard similarity of 0.37 (for the top 10,000) with active Wikipedians (last row of Table 6).

We also consider the positions of each editor in the two main rankings (most active Wikipedians and the 10,000 most active editors of references) by looking at the rank-biased overlap (RBO) by Webber, Moffat, and Zobel (2010). The RBO similarity scores are relatively high (Figure 9) only when the RBO top weight parameter (called p, $0 \leq p \leq 1$) is over 0.9999;

[31] https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits

such a high value places very low importance on the top of the two lists. For values lower than 0.9999 (placing more importance on the top of the lists), the similarity of both lists is very low (0.001 for $p = 0.95$). In other words, the very elite editors for general edits are substantially different from those working on references, while for the complete lists of 10,000, a moderate rank correlation exists that differs for lists regarding different types of changes.

6. DISCUSSION

This section discusses our data set and the results from Section 5.

6.1. Quality and Applications of the Data Set

To the best of our knowledge, we have created the most comprehensive data set of English Wikipedia references to date, preserving the traceability of each reference across all revisions with very high accuracy. We also contribute a gold standard set based on judgments by 523 crowdworkers as part of this work. According to our evaluation against the cosine similarity (Section 4), the gold standard meets the highest quality standards and could be used to evaluate other bibliometric matching algorithms, such as those provided by the Centre for Science and Technology Studies (CWTS), the Institute for Research Information and Quality Assurance (iFQ), or Web of Science (WoS)[32] (Olensky, Schmidt, & van Eck, 2016). While those methods might not perform well against our gold standard data set because they depend on bibliographical fields, which are often missing in Wikipedia references (Pooladian & Borrego, 2017), the data set could be used to tune such algorithms or develop new algorithms (e.g., based on machine learning) that are able to handle the more unstructured data we have collected and annotated here.

Our full data set is a contribution to the altmetrics community with several application areas. For example, it can be used to compare different data collection approaches (e.g., used by altmetrics aggregators), or to retrospectively analyze previous data sets evaluating the historical evolution of their collected references or the types of editors responsible for their creation. It offers the opportunity to investigate additional research questions related to coverage of specific types of publications over time, background information for evolutions of highly cited publications, topical distributions of references, and the surrounding editors dynamics. Here, we provided insights into the evolution of references based on edit types (actions), DID coverage, and editor characterization, but we believe that a host of further research questions can be answered based on this data.

6.2. Evolution of Wikipedia References

For our first research question (RQ1) we investigated how Wikipedia references evolve over time. Our data clearly highlights that references in Wikipedia are by no means static entities but are subject to amendments or "retractions" by the community in various ways. These insights imply that the point of data collection is crucial for observations: Citation counts for publications based on Wikipedia data will not only increase but may as well decrease over time; for instance, between 19.4% and 31.8% of total references (between 10.8% and 25.7% of DID-Rs) were deleted every year (from 2007 to 2019) and never reinserted again. These full deletions could cause erroneous assumptions drawn from statistics and imply an instability of

[32] https://apps.webofknowledge.com

Wikipedia references as a measurement instrument. In classical bibliometric approaches, comparable issues are negligible, as changes to the reference lists in papers are almost impossible, and retractions of papers (together with their referencing lists) are very rare events (Shema, Hahn et al., 2019). But for altmetrics, the phenomenon of citation data volatility needs to be discussed in the community.

We further find evidence that there is a continuous effort to increase the quality of Wikipedia references, expressed in the constant rise of references added to Wikipedia and the increase of the ratio of modifications to creations, with the peak in the last three years, where there were 20–40% more modifications than creations. Additionally, assuming that the presence of DIDs is an indicator of quality, we can include two further indicators: the increase of new references that include a DID (DBorn) and the presence of modifications directly targeting the absence of DIDs, at an increasing pace (see Figure 7).

We also find evidence that that DID-Rs are treated very differently by the Wikipedia community than general references, even if they are not yet marked as such via a DID. First, we detected periods of time with peaks of modifications only targeting DID-Rs representing efforts to add missing DIDs, as the peaks are often followed by low values (troughs)—probably because there is a decrease in the amount of missing DID-R that can be detected by the normal editor community. Second, the Wikipedia editor community seems to also perceive the reference with DID as more credible, given that DID-Rs are deleted with lower rates than all references.

6.3. The Role of Identifiers and Potential Effects on Altmetrics

Not only because of the discovered differences in the ways Wikipedia editors treat DID-Rs, but also because of the general importance of document identifiers (e.g., for tracking publications that were cited by Wikipedia articles), we placed an additional focus on the evolution of document identifiers as elements within Wikipedia references (RQ2).

Full deletions of references clearly disrupt the measurement of impact based on Wikipedia references, and it affects references with or without DIDs in the same way. We note that the only modifications that have a direct impact are those that change the reference in such a way that either they make it point to a new resource (i.e., the equivalent of removing and adding a reference), or they make the reference detectable (by adding a DID) or invisible (by removing the DID, depending on the mining method).

Assuming that some of the altmetrics aggregators take advantage of the presence of DIDs for identifying and counting references from Wikipedia, we have looked at the specific modifications that introduced a DID to an existing reference in more detail. We analyzed the DID-Lagged References that did not include a reference upon their first introduction but received it through later edits. Those references would potentially have been ignored during their initial lifespans before getting their DID. We were able to show that they correspond to a considerable fraction of DID-Rs (10% corresponding to 12.1% actions before the introduction of the DID). We found important periods regarding the evolution of the DID-Lagged References (dashed line, Figure H1 of Appendix H in the Supplementary material). Before 2010, a method that relied only on DIDs would have missed up to 37.5% (2007) of references for which we know that they should have had a DID (as we see that their DIDs were added by June 2019). The situation quickly improved between 2009 and 2010 (11.3%), and then continued doing so until our data collection. Our findings show that mining methods that rely on DIDs are vulnerable to coverage errors (Sen et al., 2021) that can misrepresent the importance of academic works in the altmetrics community.

6.4. Towards Understanding Who Edits Wikipedia References

For our last research question, we investigated the editor community that creates and maintains Wikipedia references (RQ3). These contributors play a crucial role within Wikipedia, as they judge the relevance of references and shape what Wikipedia readers consume, also influencing whether an article is perceived as relevant or trustworthy based on the presence of references.

We found that most of the references (87.6%) are created by registered editors, whereas bots were only responsible for 1.6% of new references. The concern that the presence of bots (Nielsen, 2008) was dominating reference creation cannot be confirmed. For comparison, this has been the case in Twitter, where Robinson-Garcia, Costas et al. (2017) found that bots (and thoughtless bot-like retweets of user accounts) were responsible for most of the activity containing scholarly articles. These findings support the idea that Wikipedia references are curated by humans and thus involve deliberate selection of sources and materials.

We further demonstrated that according to our similarity metrics (Jaccard and RBO), registered editors shaping Wikipedia references are considerably different from the rest of Wikipedia's most active users in terms of general edits: Only about half of the top Wikipedia general editors are among the top 10,000 reference editors. We were able to identify clusters of editors; two of these clusters are fully specialized in creations and deletions and together add up to ~61% of the editors. We also found single editors that edited references in many different Wikipedia articles (e.g., one editor has edited references in more than 226,000 articles), and thus appeared to be highly specialized on reference editing, independent of topical domains. Bots, on the other hand, have been to the largest extent only used to maintain (modify) references, throughout Wikipedia history.

These observations deserve additional attention in the future, as they remind us of our introductory example (Figure 1). Despite Wikipedia being a community effort, individuals can have substantial influence over certain areas. A single editor has the potential to largely affect the representation of a specific reference.

6.5. Limitations

The collection method covers references indicated as inline citations via ref tags, following Wikipedia's recommendation for how references should be added to articles and implying some quality control for inline citations based on Wikipedia's standards. However, we do not include other forms of references, such as parenthetical references[33] or wikilinks to full references using templates[34]. These forms are not uniform, and we could not guarantee that their extraction would be accurate. We are not sure to what extend this strategy is used among altmetrics aggregators, but, at least, we found that altmetric.com also only considers ref tags for identifying references[35].

The data set was created based on the English Wikipedia as of June 2019, and more DIDs have been (or will be) added after that date. We also worked with a selection of common types of document identifiers: DOI, PubMedID, PMC, ISBN, ISSN, and arXiv ID. The list corresponds to that used by the Wikimedia Foundation project (Halfaker et al., 2019), as it is supposed to

[33] https://en.wikipedia.org/wiki/Wikipedia:Parenthetical_referencing

[34] For example, shortened footnote template (e.g., {{sfn}}), Harvard style templates (e.g., {{harvnb}}), or freehand anchors (e.g., [[#anchor_id]]) https://en.wikipedia.org/wiki/Wikipedia:Citing_sources/Further_considerations#Wikilinks_to_full_references.

[35] https://help.altmetric.com/support/solutions/articles/6000235982-wikipedia

capture most academic citations. We use the presence of identifiers as a weak indicator of the quality of the referenced publications (see Section 6.2 for a discussion), and not as a way to identify types of publications (e.g., scientific vs. nonscientific).

7. CONCLUSIONS

In this paper, we have introduced an overview of the evolution of Wikipedia references, analyzed the historical coverage of reference-mining methods that are based on DIDs, and offered a characterization of the Wikipedia editors. In the scope of our research questions, we conclude that the quality of Wikipedia references has been slowly but persistently increasing. Although our findings do highlight limitations, we believe that the historical registry of Wikipedia contains information that can be leveraged to create more robust methods of mining and assigning importance to references in Wikipedia. We recommend that such methods use this record to reduce manipulations and biases that blur the visibility of references, to increase the overall coverage of references (by looking at all revisions), and to assign impact based on historical activity and the community of (e.g., reputable) editors that surround the references.

These recommendations only open a different path for the creation of altmetrics based on Wikipedia, and there is certainly more to be done. The high-quality data set that accompanies this paper offers the opportunity to extend the research in this direction, for example the following:

- analyzing the longevity and activity of references distinguishing between academic and nonacademic (see Singh et al. (2021) for a classification approach),
- exploring the dynamics of references according to different knowledge fields,
- further investigating the editors by mining (with natural language processing techniques) their profile pages and extract demographics,
- modeling the co-editors network to find important actors and communities, and
- predicting which references are still missing a document identifier, as our data set already provides this information for existing references.

ACKNOWLEDGMENTS

We would like to thank all the *metrics project members, as well as Prof Dr Isabella Peters and Prof Dr Claudia Wagner for their supervision and feedback, student assistants Tara Morovatdar and Alexandra Stankevich for their help with the data curation, and Kenan Erdogan for the insights about the WikiWho service.

AUTHOR CONTRIBUTIONS

Olga Zagovora: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Roberto Ulloa: Conceptualization, Formal analysis, Software, Visualization, Writing—original draft, Writing—review & editing. Katrin Weller: Conceptualization, Funding acquisition, Project administration, Supervision, Writing—original draft, Writing—review & editing. Fabian Flöck: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Validation, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research was supported by the Deutsche Forschungsgemeinschaft, DFG, project number 314727790. Fabian Flöck acknowledges support from the Volkswagen Foundation (grant 92136). The publication of this article was funded by the Open Access Fund of the Leibniz Association.

DATA AVAILABILITY

The data set is made available on Zenodo (Zagovora et al., 2020). We also provide a Python notebook with examples on how to process the data, and the code can be directly executed on the GESIS Notebooks server.

REFERENCES

Bayliss, G. (2013). Exploring the cautionary attitude toward Wikipedia in higher education: Implications for higher education institutions. *New Review of Academic Librarianship*, *19*(1), 36–57. https://doi.org/10.1080/13614533.2012.740439

Bould, M. D., Hladkowicz, E. S., Pigford, A.-A. E., Ufholz, L.-A., Postonogova, T., ... Boet, S. (2014). References that anyone can edit: Review of Wikipedia citations in peer reviewed health science literature. *British Medical Journal*, *348*, g1585. https://doi.org/10.1136/bmj.g1585, PubMed: 24603564

Chen, C.-C., & Roth, C. (2012). {{Citation needed}}: The dynamics of referencing in Wikipedia. *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (pp. 1–4). https://doi.org/10.1145/2462932.2462943

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia risks. *Communications of the ACM*, *48*(12), 152. https://doi.org/10.1145/1101779.1101804

Eijkman, H. (2010). Academics and Wikipedia: Reframing Web 2.0+ as a disruptor of traditional academic power-knowledge arrangements. *Campus-Wide Information Systems*, *27*(3), 173–185. https://doi.org/10.1108/10650741011054474

Flöck, F., & Acosta, M. (2014). WikiWho: Precise and efficient attribution of authorship of revisioned content. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 843–854). https://doi.org/10.1145/2566486.2568026

Flöck, F., Erdogan, K., & Acosta, M. (2017) TokTrack: A complete token provenance and change tracking dataset for the English Wikipedia. *Eleventh International AAAI Conference on Web and Social Media*. https://arxiv.org/abs/1703.08244

Grathwohl, C. (2011). Wikipedia comes of age. *Chronicle of Higher Education*, *57*. https://www.chronicle.com/article/Wikipedia-Comes-of-Age/125899

Halfaker, A., Mansurov, B., Redi, M., & Taraborelli, D. (2019). Citations with identifiers in Wikipedia. *figshare*. https://doi.org/10.6084/m9.figshare.1299540

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*, *108*(1), 413–423. https://doi.org/10.1007/s11192-016-1910-9

Holman Rector, L. (2008). Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review*, *36*(1), 7–22. https://doi.org/10.1108/00907320810851998

Holmberg, K. J. (2015). *Altmetrics for information professionals: Past, present and future*. Chandos Publishing. https://www.sciencedirect.com/science/book/9780081002735

Huvila, I. (2010). Where does the information come from? Information source use patterns in Wikipedia. *Information Research*, *15*(3). https://www.informationr.net/ir/15-3/paper433.html

Imran, M., Akhtar, A., Said, A., Iqra, S., Hassan, S.-U., & Aljohani, N. R. (2018). Exploiting social networks of Twitter in altmetrics big data. *STI 2018 Conference Proceedings* (pp. 1339–1344). https://hdl.handle.net/1887/65219

Kaffee, L.-A., & Elsahar, H. (2021). References in Wikipedia: The editors' perspective. *8th Wiki Workshop at The Web Conference*. https://arxiv.org/abs/2102.12511. https://doi.org/10.1145/3442442.3452337

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in Wikipedia. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–462). https://doi.org/10.1145/1240624.1240698

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. https://doi.org/10.1002/asi.23694

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Analysis of references across Wikipedia languages. In R. Damaševičius & V. Mikašytė (Eds.), *Information and Software Technologies* (pp. 561–573). Springer International Publishing. https://doi.org/10.1007/978-3-319-67642-5_47

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2020). Modeling popularity and reliability of sources in multilingual Wikipedia. *Information*, *11*(5), 263. https://doi.org/10.3390/info11050263

Lin, J., & Fenner, M. (2013). Altmetrics in evolution: Defining and redefining the ontology of article-level metrics. *Information Standards Quarterly*, *25*(2), 20. https://doi.org/10.3789/isqv25no2.2013.04

Lin, J., & Fenner, M. (2014). An analysis of Wikipedia references across *PLOS* publications. *Expanding Impacts and Metrics, An ACM Web Science Conference 2014 Workshop* (pp. 23–26). https://doi.org/10.6084/m9.figshare.1048991.v3

Luyt, B., & Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology*, *61*(4), 715–722. https://doi.org/10.1002/asi.21304

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. https://doi.org/10.1002/asi.23172

Murić, G., Abeliuk, A., Lerman, K., & Ferrara, E. (2019). Collaboration drives individual productivity. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW) (pp. 74:1–74:24). https://doi.org/10.1145/3359176

Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *First Monday*, *12*(8). https://doi.org/10.5210/fm.v12i8.1997

Nielsen, F. Å. (2008). Clustering of scientific citations in Wikipedia. *Wikimania 2008*. https://arxiv.org/abs/0805.1154

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, *65*(12), 2381–2403. https://doi.org/10.1002/asi.23162

Olensky, M., Schmidt, M., & van Eck, N. J. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *Journal of the Association for Information Science and Technology*, *67*(10), 2550–2564. https://doi.org/10.1002/asi.23590

Ortega, J. L. (2018). Reliability and accuracy of altmetric providers: A comparison among Altmetric.com, PlumX, and Crossref Event Data. *Scientometrics*, *116*(3), 2123–2138. https://doi.org/10.1007/s11192-018-2838-z

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: A study of power editors on Wikipedia. *Proceedings of the ACM 2009 International Conference on Supporting Group Work – GROUP '09* (pp. 51–60). https://doi.org/10.1145/1531674.1531682

Piccardi, T., Redi, M., Colavizza, G., & West, R. (2020). Quantifying engagement with citations on Wikipedia. *Proceedings of The Web Conference 2020 (WWW '20)* (pp. 2365–2376). https://doi.org/10.1145/3366423.3380300

Pooladian, A., Borrego, Á. (2017). Methodological issues in measuring citations in Wikipedia: A case study in library and information science. *Scientometrics*, *113*, 455–464. https://doi.org/10.1007/s11192-017-2474-z

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. https://altmetrics.org/manifesto/

Redi, M. (2018). *Research: Characterizing Wikipedia citation usage*. https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Citation_Usage

Redi, M., & Taraborelli, D. (2018). *Accessibility and topics of citations with identifiers in Wikipedia*. https://doi.org/10.6084/m9.figshare.6819710.v1

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PLOS ONE*, *12*(8), e0183551. https://doi.org/10.1371/journal.pone.0183551, PubMed: 28837664

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, *85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Shema, H., Hahn, O., Mazarakis, A., & Peters, I. (2019). Retractions from altmetric and bibliometric perspectives. *Information – Wissenschaft & Praxis*, *70*(2–3), 98–110. https://doi.org/10.1515/iwp-2019-2006

Shuai, X., Jiang, Z., Liu, X., & Bollen, J. (2013). A comparative study of academic and Wikipedia ranking. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 25–28). https://doi.org/10.1145/2467696.2467746

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

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. https://doi.org/10.1002/asi.23833

Sugiyama, K., Tagawa, S., & Toda, M. (1981). Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics*, *11*(2), 109–125. https://doi.org/10.1109/TSMC.1981.4308636

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. https://doi.org/10.1002/asi.23687

Thelwall, M. (2016). Does astronomy research become too dated for the public? Wikipedia citations to astronomy and astrophysics journal articles 1996–2014. *El Profesional de La Información*, *25*(6), 893–900. https://doi.org/10.3145/epi.2016.nov.06

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, *28*(4), 20:1–20:38. https://doi.org/10.1145/1852102.1852106

Zagovora, O., Ulloa, R., Weller, K., & Flöck, F. (2020). *Individual edit histories of all references in the English Wikipedia* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3964990

Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLOS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197326, PubMed: 29772003

Zaldivar, M. S. S., B. Tomlinson, R. LaPlante, J. Ross, L., & Irani, A. (2018). Responsible research with crowds: Pay crowdworkers at least minimum wage. *Communications of the ACM*, *61*(3), 39–41. https://doi.org/10.1145/3180492

Zappia, L., & Oshlack, A. (2018). Clustering trees: A visualization for evaluating clusterings at multiple resolutions. *GigaScience*, *7*(7). https://doi.org/10.1093/gigascience/giy083, PubMed: 30010766