Wikinformetrics: construction and description of an open Wikipedia knowledge graph data set for informetric purposes

Arroyo-Machado, W.; Torres-Salinas, D.; Costas Comesana, R.

RESEARCH ARTICLE

Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes

Wenceslao Arroyo-Machado[1], Daniel Torres-Salinas[1], and Rodrigo Costas[2,3]

[1]Department of Information and Communication Sciences, University of Granada, Granada, Spain
[2]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands
[3]DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

Keywords: altmetrics, data, informetrics, knowledge graph, metrics, Wikipedia

ABSTRACT

Wikipedia is one of the most visited websites in the world and is also a frequent subject of scientific research. However, the analytical possibilities of Wikipedia information have not yet been analyzed considering at the same time both a large volume of pages and attributes. The main objective of this work is to offer a methodological framework and an open knowledge graph for the informetric large-scale study of Wikipedia. Features of Wikipedia pages are compared with those of scientific publications to highlight the (dis)similarities between the two types of documents. Based on this comparison, different analytical possibilities that Wikipedia and its various data sources offer are explored, ultimately offering a set of metrics meant to study Wikipedia from different analytical dimensions. In parallel, a complete dedicated data set of the English Wikipedia was built (and shared) following a relational model. Finally, a descriptive case study is carried out on the English Wikipedia data set to illustrate the analytical potential of the knowledge graph and its metrics.

1. INTRODUCTION

On January 15, 2001, Wikipedia was born under the umbrella of Nupedia, an encyclopedia project that was based on a peer review system. Due to the lack of agility in publishing articles, Wikipedia was created as a feeder project, as its objective was to make the creation of new articles easier before they were reviewed (*History of Wikipedia*, 2021). Wikipedia combined in a single project different elements that were new on the web and that made possible for the first time a universal encyclopedia (Reagle, 2009). It was successful enough to make Nupedia disappear in 2 years, experiencing steady growth. Since then, Wikipedia has become one of the most visited websites in the world (https://www.semrush.com/website/top/, accessed August 4, 2022), having 328 different editions, 285 of them having more than 1,000 articles (https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed August 4, 2022). Although this is the most successful project of Wikimedia Foundation, there are also other well-known knowledge projects using wikis as a basis (e.g., the Wiktionary dictionary or the Wikidata knowledge base).

Wikipedia has been a disruptive innovation, finding in its open nature and decentralized knowledge development one of its key elements (Olleros, 2008). Not only can everyone access

an open access journal

Citation: Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. *Quantitative Science Studies*, *3*(4), 931–952. https://doi.org/10.1162/qss_a_00226

DOI: https://doi.org/10.1162/qss_a_00226

Supporting Information: https://doi.org/10.1162/qss_a_00226

Received: 10 August 2022
Accepted: 28 October 2022

Corresponding Author: Wenceslao Arroyo-Machado
wences@ugr.es

Handling Editor: Vincent Larivière

its contents free of charge, but they can also participate in its construction, in a fully transparent process. This social construction of the knowledge can be seen in the differences found among language editions of the same Wikipedia pages (Hara & Doney, 2015). Wikipedia contents are also the result of consensus among editors or Wikipedians. This consensus is built in open discussions in the Wikipedia talk pages (Maki, Yoder et al., 2017; Yasseri, Sumi et al., 2012), open to anyone and capturing transnational debates around Wikipedia contents (Kopf, 2020). Some of these talks and debates have sometimes transcended Wikipedia itself (O'Neil, 2017).

As an online encyclopedia, Wikipedia is not exempt from problems. The reliability of its content has been much debated, as it is based on contributions from anonymous individuals (Olleros, 2008). The quality of Wikipedia pages' content has been studied numerous times from different perspectives, especially with regard to medical content pages, pointing out limitations, such as occasional incomplete or imprecise information (Adams, Montgomery et al., 2020; Candelario, Vazquez et al., 2017; Weiner, Horbacewicz et al., 2019). The importance of integrating Wikipedia into academia, both in its use and in its development, has been highlighted (Jemielniak, 2019). Social and cultural inequalities have also been pointed out, such as racial and gender gaps in its biographies (Adams, Brückner, & Naslund, 2019; Tripodi, 2021).

Wikipedia is not free of bots and vandalism, although they do not constitute a serious threat to its contents and reliability and Wikipedia's policy does not allow detrimental use of the activity of bots or automated accounts. Most of the bots on Wikipedia are publicly identified (https://en.wikipedia.org/wiki/Special:ListUsers/bot), and they contribute to improving the content and structure of Wikipedia articles (Arroyo-Machado, Torres-Salinas et al., 2020; Zheng, Albano et al., 2019). Bots also help to control and reduce problems of vandalism and trolls, as they eliminate their harmful edits of articles in advance of human editors. There is also no shortage of proposals for methods based on machine learning to prevent this type of harmful activity (Martinez-Rico, Martinez-Romo, & Araujo, 2019).

In spite of all of these issues, the general idea is that Wikipedia is a transparent and reliable source of encyclopedic information (Lageard & Paternotte, 2021), with value of its own to be the subject of scientific research.

1.1. Wikipedia as Source for Informetric Research

Wikipedia has been researched from different scientific perspectives. One of them is informetrics, quantitatively studying the contents and activity generated on Wikipedia. Thus, Wikipedia has been studied from the points of view of scientometrics, bibliometrics, and webometrics, which are discussed in detail below.

Bibliographic references made in Wikipedia have been studied, particularly since the emergence of the notion of "altmetrics" (Priem, Taraborelli et al., 2010), which considered citations on Wikipedia to scientific literature as part of its realm[1]. Wikipedia citations are one of the most popular sources covered in altmetric aggregators (Ortega, 2020; Zahedi & Costas, 2018) such as Altmetric.com, PlumX, or Crossref Event Data. In addition to altmetric data providers, there are also several other open data sources providing extensive metadata on Wikipedia citations (Singh, West, & Colavizza, 2020; Zagorova, Ulloa et al., 2022). Moreover, other proposals, such as Scholia, enable the exploration of bibliographic data at different levels through Wikidata (Nielsen, Mietchen, & Willighagen, 2017). In Table 1 a summary of previous studies on Wikipedia bibliographic references are presented.

[1] Wikipedia references had already been studied for years before the birth of altmetrics, such as in the citation analysis by Nielsen (2007) or, in a more qualitative way, that of Mühlhauser and Oser (2008).

Table 1. Main studies on the bibliographic references included in Wikipedia pages

Reference	Type	Application	Data	Methodological approach	Language edition	Topic analyzed
Mühlhauser and Oser (Mühlhauser & Oser, 2008)	Qualitative	Content and quality analysis	–	Check list	German	Health care
Candelario et al. (Candelario et al., 2017)		Content and quality analysis	33 pages	Scoring system	English	Medication
Kaffee and Elsahar (Kaffee & Elsahar, 2021)		Analyze the editors' citation process	–	Survey and interviews	Multilingual	Multidisciplinary
Nielsen (Nielsen, 2007)	Quantitative	Analyze citation patterns	30,368 citations	Descriptive statistics	English	Multidisciplinary
Kousha and Thelwall (Kousha & Thelwall, 2017)		Evaluate the impact of references	36,191 citations	Descriptive statistics	Multilingual	Multidisciplinary
Lewoniewski et al. (Lewoniewski, Węcel, & Abramowicz, 2017)		References coverage across languages	6.8 million pages 41 million citations	Descriptive statistics	Multilingual	Multidisciplinary
Maggio et al. (Maggio, Willinsky et al., 2017)		Analyze citation patterns	229,857 pages 1,049,025 citations	Descriptive statistics	English	Medicine
Pooladian and Borrego (Pooladian & Borrego, 2017)		Evaluate the impact of references	982 citations	Descriptive analysis	Multilingual	Multidisciplinary
Jemielniak et al. (Jemielniak, Masukume, & Wilamowski, 2019)		Rank journals by citations	11,325 pages 137,889 citations	Citation analysis	English	Medicine
Torres-Salinas et al. (Torres-Salinas, Romero-Frías, & Arroyo-Machado, 2019)		Mapping of knowledge structure	25,555 pages 41,655 citations	Cocitation analysis	English	Arts & Humanities
Arroyo-Machado et al. (Arroyo-Machado et al., 2020)		Mapping of knowledge structure	193,802 pages 847,512 citations	Cocitation analysis	English	Multidisciplinary

Table 1. (*continued*)

Reference	Type	Application	Data	Methodological approach	Language edition	Topic analyzed
Colavizza (Colavizza, 2020)		Publications coverage	3,083 ref. pub.	Topic modeling and regression analysis	English	COVID-19
Nicholson et al. (Nicholson, Uppala et al., 2021)		Reviewing citation quality	1,923,575 pages 824,298 ref. pub.	Classification modeling	English	Multidisciplinary
Singh et al. (Singh et al., 2020)		Data set creation	4 million citations	Text mining	English	Multidisciplinary
Zagorova et al. (Zagorova et al., 2022)		Data set creation	6,073,708 pages 55 million citations	Text mining	English	Multidisciplinary

Kaffee and Elsahar (2021) explored the flow that Wikipedians follow to include references in Wikipedia articles. Kousha and Thelwall (2017), and Pooladian and Borrego (2017) described the problems of Wikipedia citations in performance evaluation. Nicholson et al. (2021) studied the quality of cited references in Wikipedia. Lewoniewski et al. (2017) showed that the different language editions of the same Wikipedia page tended to cite common sources, with the largest overlap between English and German and some differences depending on the topics. Colavizza (2020) studied the coverage of the scientific literature on COVID-19 on Wikipedia, showing that although there was only a small percentage of scientific literature on COVID-19 in Wikipedia, it was sufficiently representative of its various topics. Arroyo-Machado et al. (2020) and Torres-Salinas et al. (2019) mapped Wikipedia cocitations patterns, showing fundamental differences in the use of scientific literature in Wikipedia compared to the academic realm. Bould, Hladkowicz et al. (2014), Li, Thelwall, and Mohammadi (2021), and Tomaszewski and MacDonald (2016) studied academic citations in scientific publications to Wikipedia articles, proving that scientific publications also use Wikipedia content in their citations, as well as other digital encyclopedias, especially in areas such as chemistry, physics, or mathematics.

Wikipedia has also been the subject of webometric studies. For example, "Wikiometrics" were proposed as a rating system to rank universities or journals based on the features of their Wikipedia pages, also finding positive correlations with existing academic rankings (Katz & Rokach, 2017). The estimation of the importance of Wikipedia pages based on the PageRank algorithm was also studied, correlating positively with other page-view-based rankings (Thalhammer & Rettinger, 2016). Miquel-Ribé and Laniado (2018) showed that the different language editions of Wikipedia pages reflect cultural differences, as the contents cover local topics corresponding to different linguistic regions. Other studies focused on metrics about the attention generated around Wikipedia articles (e.g., likes or page view counts), showing how they reflect current topics of interest at a particular time/region (Dzogang, Lansdall-Welfare, & Cristianini, 2016; Mittermeier, Roll et al., 2019; Mittermeier, Correia et al., 2021; Roll, Mittermeier et al., 2016; Vilain, Larrieu et al., 2017), and even demonstrating the potential of Wikipedia pages to monitor the spread of diseases (Generous, Fairchild et al., 2014).

There are also numerous studies around Wikipedia's informetric features. Wilkinson and Huberman (2007) found a correlation between the quality of Wikipedia articles and their number of edits. The relationship between the length of Wikipedia articles and their quality has been highlighted by Blumenstock (2008). Beyond quality, relationships between Wikipedia metrics have also been explored. Previous studies found positive correlations between views and the number of edits and editors (Mittermeier et al., 2021), and weak correlations between the length of Wikipedia pages and the length of their talk pages (Yasseri et al., 2012). Zhang, Ren, and Kraut (2018) suggested the value of using metrics in specific moments of the life cycles, for example the number of editors in the first 3 months of an article's life was not when it was most strongly related to its future quality.

Although, as shown above, there is abundant scientific literature on Wikipedia and its informetric applications, most previous studies tended to focus on either limited sets of metrics (e.g., Nicholson et al. (2021), who were focused on the level of quality of scientific publications referenced in Wikipedia articles), or limited data sets (e.g., Mittermeier et al. (2021), who studied a large set of features in a data set of Wikipedia pages of 10,099 bird species across 251 language editions). Thus, large-scale study of Wikipedia, from both a large volume of pages and attributes, is still missing in the literature. Arguably, a potential reason for this lack of large-scale studies on Wikipedia is the lack of a conceptual framework that highlights both

the large-scale data available from Wikipedia and the multiple informetric metrics that Wikipedia offers. Such absence has hindered the development of broader research perspectives, especially regarding the relationship of Wikipedia with science, where a contextualization of the relationships between the two is still needed.

In this study, we propose such a framework by means of developing an informetric-inspired knowledge graph, with the aim of enabling similar analytical approaches to those developed in scientometric research. Such a knowledge graph could work as a complement of other Wikipedia knowledge graphs such as Wikidata (https://www.wikidata.org/) or DBpedia (https://www.dbpedia.org/). Wikidata and DBpedia provide exhaustive Wikipedia knowledge graphs but they are more focused on content and semantic relationships, transforming Wikipedia pages into entities (e.g., people, places, music bands) and establishing different computer-understandable relationships between them. Our proposed knowledge graph aims at characterizing the attention and usage of Wikipedia pages using a relational model and incorporating activity metadata that are not present in the semantic graphs of Wikidata and DBpedia, capturing the attention and social engagement, such as views or edits, as well as the presence of scientific literature in Wikipedia pages.

The paper is structured as follows: First, we describe our main objectives and our alignment with recent developments in the field of altmetrics. Second, we describe the informetric features of Wikipedia pages and their similarities with scientific publications, together with the existing data sources for data collection. Several informetric-inspired metrics (Wikinformetrics) are proposed for Wikipedia. Third, a Wikipedia knowledge graph, based on the combination of different Wikipedia data sources, is constructed and presented. Fourth, the data set is explored in a descriptive way to show the analytical possibilities of the knowledge graph and the proposed metrics. Finally, we conclude by discussing our findings and proposing future research venues.

1.2. Objectives

The main objective of this work is to explore the research value of Wikipedia from an informetric perspective, ultimately providing a complete Wikipedia knowledge graph. More specifically, three different objectives are targeted:

1. Theoretical objective: To establish a framework for Wikipedia analytics, by exploring the informetric features of Wikipedia pages (composition, categories, sources, data gathering, etc..) and proposing a set of informetric-inspired metrics (Wikinformetrics) for their quantitative study. This objective will help us to map the analytical possibilities of Wikipedia as a scientific object.
2. Instrumental objective: To create a large open Wikipedia knowledge graph. Once we are familiar with the main features of Wikipedia, we will construct a dedicated knowledge graph focused on the English-language edition of Wikipedia with the main information and data relationships coming from combining different data sources.
3. Applied objective: To conduct a descriptive quantitative study of Wikipedia metrics based on the knowledge graph data set, and to explore the proposed metrics and the different types of attention they capture.

This work and its objects align with novel developments on social media metrics (Díaz-Faes, Bowman, & Costas, 2019; Wouters, Zahedi, & Costas, 2019), contributing to the exploration of different science-society interactions that can be captured on Wikipedia (Costas, de

Rijcke, & Marres, 2020). Our ambition is to frame Wikipedia as a data source with multiple informetric research possibilities. Furthermore, a dedicated data set of the English edition of Wikipedia is constructed for informetric purposes and is freely available at Zenodo (https://doi.org/10.5281/zenodo.6346899). R and Python were used together for its elaboration, with the scripts available on GitHub (https://doi.org/10.5281/zenodo.6959428). Many of the results presented here are novel, as to the best of our knowledge there is no previous literature that has explored the same large set of Wikipedia features and with the same large-scale perspective as in this study. This work is intended to be useful for a wide range of researchers, such as librarians, informetricians, sociologists, and data scientists.

2. WIKIPEDIA FROM AN INFORMETRIC PERSPECTIVE

2.1. Analogy Between Wikipedia Pages and Scientific Publications

In Wikipedia, the key components are the individual pages. Wikipedia pages are not only used for the publication of encyclopedia articles but also other numerous typologies of pages, such as categories, users, and talk pages, as well as relationships among them. The different types of pages are given by a pre-established namespace (a type of page with special features identifiable through a prefix included in the title). Wikipedia currently has 12 namespaces in use (*article*, *user*, *Wikipedia*, *file*, *mediawiki*, *template*, *help*, *category*, *portal*, *draft*, *timedtext*, and *module*), each with an associated "talk namespace" (or "talk page") in which discussions are held around the contents and edits of the page, and two virtual namespaces (special and media).

There are several features of Wikipedia pages, in particular namespace article pages, for which it is possible to establish an equivalence with that of a scientific publication. First, they have a title and an associated page identifier (Wikipedia page ID). They may have one or more authors, it being possible to identify the first person who created it, and when, and those who have made a greater contribution or whose edition has been revoked. The contents may include multimedia files, links to external resources, and bibliographic references, among others. There are also internal links that enable Wikipedia pages to connect to each other, just like citations among scientific publications. Finally, Wikipedia pages can be classified with categories according to their contents to carry out its thematic classification, such as keywords and classifications applied to scientific publications. Most of these elements can be seen as metadata to be treated in the study of Wikipedia pages. However, there are several differences between Wikipedia pages and scientific publications that cannot be ignored (Table 2). The most important is that Wikipedia pages are a living resource and not static documents. The access and editing of the contents also differ between Wikipedia pages and scientific publications because Wikipedia pages do not focus on a specific audience (e.g., scientific publications mostly focus on academic audiences), but anyone can take an active part in editing them. It should be also noted that some pages may be temporarily limited or protected for editing (Hill & Shaw, 2015).

The living nature of Wikipedia pages puts them at the center of a complex system (Ladyman, Lambert, & Wiesner, 2013), whose main elements are represented in Figure 1. Many of the elements of the pages are static or unalterable, such as the creation date or page ID, while others are in constant evolution, especially the contents themselves. This makes it difficult to study certain elements in Wikipedia (Détienne, Baker et al., 2016), as Wikipedia content is volatile and authorship and contribution roles can be diluted in contrast to the higher stability of scientific publications. In addition, the same page, especially encyclopedic articles, may have parallel versions in different language editions of Wikipedia, which may

Table 2. Comparison of features between Wikipedia pages and scientific publications

Wikipedia element description		Wikipedia pages vs. Scientific publications	
		Wikipedia page	**Scientific publication**
State	*Document state condition*	Living	Static
ID	*Document identification number*	Page ID	DOI, ISBN, URI ...
Name	*Title of the document*	Title	Title
Type	*Document typologies*	Namespace (12 + 12 types)	Paper, proceeding, letter ...
Creation	*Date from which it is available*	First edition date	Publication date
Authorship	*Responsible for the work*	Wikipedians	Authors
Content	*Type of content*	Structured text	Structured text
Language	*Language of the resource*	Edition dependent	Document dependent
Discussion	*Comments on the contents*	Talk	Peer review
Description	*Work summary*	Short description	Abstract
Tags	*Terms describing the content*	Categories	Keywords
Media	*Audiovisual resources includable*	Images, audios, and videos	Images, audios, and videos
Internal links	*Links to the related resources*	Internal links	Citations
Format	*Standardized structure and content*	Manual of style*	Format guidelines
Bibliography	*References of cited resources*	References	References
Access	*Access model*	Open	Closed/Open
Audience	*Document target audience*	General	Specialized

* The English Wikipedia has its own manual of style https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

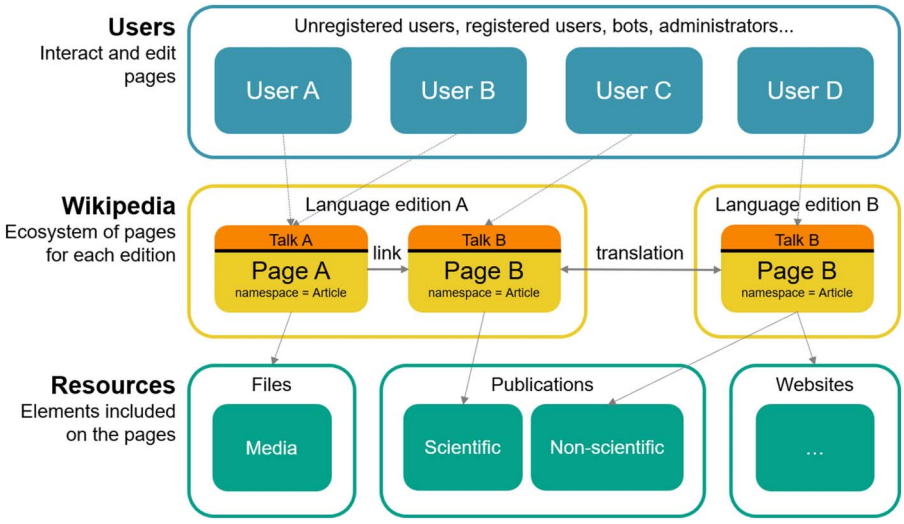

Figure 1. Diagram of the main elements involved in creating and editing Wikipedia articles.

vary in content. This scenario becomes even more complex when taking into account that not only human users are involved in the development of Wikipedia pages but also bots, thus making the interactions that can occur more complex to analyze (Tsvetkova, García-Gavilanes et al., 2017).

2.2. Categorization

Wikipedia pages are not thematically organized according to a controlled language-based classification, such as *Britannica*'s subject organization system. Instead, Wikipedia pages have a category system that works like a folksonomy (Minguillón, Lerga et al., 2017). Wikipedians are free to tag each page under one or more existing categories or to create new ones. Numerous studies have approached them, such as by studying their semantic domain (Aghaebrahimian, Stauder, & Ustaszewski, 2020; Heist & Paulheim, 2019). However, the main problem of this folksonomy is the large number of individual categories and their unstructured (i.e., without a clear hierarchical system) relations at different levels, introducing a lot of noise and making it difficult to have a general thematic view of Wikipedia (Boldi & Monti, 2016; Kittur, Chi, & Suh, 2009). In addition, there are also hidden categories, related to the maintenance or management of the page.

Besides the categories, Wikipedia has other options for accessing and browsing its contents by topics (https://en.wikipedia.org/wiki/Wikipedia:Contents). On the one hand, it offers different curated content lists (e.g., the "list of articles every Wikipedia should have" or the list of "vital articles"). There are other lists that offer collections of articles that respond to the same topic, and even "lists of lists." Similarly, there are "portals," which imitate the classic web portals and are organized in sections that group the main contents of a topic, not only the articles (e.g., the "Science" portal or the "History of science" subportal). WikiProjects, communities of Wikipedians aimed at improving Wikipedia content on a specific topic and which have their own page from which they coordinate their activities, can also work as a classification approach due to their thematic orientation (e.g., "Anthropology" or "The Beatles"). There are also third-party classification systems, such as the "Library of Congress Classification" or the "Universal Decimal Classification." Finally, external to Wikipedia, but within the Wikimedia ecosystem, there are other types of classification solutions, such as Wikidata taxonomies (https://www.wikidata.org/wiki/Wikidata:WikiProject_Taxonomy) or ORES (https://www.mediawiki.org/wiki/ORES), that can be used to identify Wikipedia topics using machine learning techniques. The main limitation with all of the above is that there is no central classification system that covers all Wikipedia pages, and that at the same time it is concise and easy to manage, particularly in terms of the number of subjects and the hierarchical relationships among them. The lack of such central classification in Wikipedia is a major hindrance for the large-scale epistemic study of Wikipedia.

2.3. Content Control

Each Wikipedia page has a discussion space called "talk pages," where Wikipedians discuss with other Wikipedians. Talk pages aim at improving the quality and reliability of the articles. Discussions in talk pages are public (Ferschke, Gurevych, & Chebotar, 2012), resembling the model of open peer review of scientific publications (Black, 2008), and representing a form of public review in contrast to the traditional academic blind peer review system (Cummings, 2020). Wikipedia also includes formal peer review approaches in which Wikipedians request assistance from experts on given topics (https://en.wikipedia.org/wiki/Wikipedia:Peer_review). Despite discrepancies and differences about what open peer review means and the different

Table 3. General quality grading scheme of WikiProject articles

Class	Description	Assignment	Badge
Featured article	The best possible content on Wikipedia, no need for improvement	Review	Yes
Featured list	The best possible list on Wikipedia, no need for improvement	Review	Yes
A	Fully addresses the subject and requires only minor improvements	Review	No
Good article	It satisfies Wikipedia's main criteria and is close to a professional article	Review	Yes
B	The content is almost complete and has no major problems	Free	No
C	The content is considerable, but has significant problems	Free	No
Start	It includes significant content, but is still in development	Free	No
Stub	The content is very short and requires substantial work	Free	No
List	Content displayed in a list linking to Wikipedia articles on a specific topic	Free	No

models proposed (Ross-Hellauer, 2017), the three basic principles (open identities, reports, and participation) are clearly recognizable in Wikipedia (Table S2 in the Supplementary material). Wikipedians are both authors and reviewers of content and their reports are available as comments on the talk pages, all of which are always open and identifiable. Interestingly, Wikipedia-inspired reviewing approaches have even been proposed for scholarly publishing, such as the postpublication correction system and readers' comments (Xiao & Askin, 2014).

Wikipedia also includes a quality control system of the content of the different articles that comes from WikiProjects. It is grounded on an evaluation system to classify pages in higher or lower levels of content quality, with standard grades that are listed on the respective talk page. Although there is a general scheme (Table 3), it is possible that some WikiProjects do not include all grades or that there may be differences in their application. Similarly, the pages are also classified according to their importance within the topic (Top, High, Mid, and Low). Wikipedians can set any level of quality and importance on a given page, as well as modifying them. When there are disagreements among Wikipedians about the quality level of a page, this leads to a discussion and a search for consensus around the quality level of the page. However, at the highest levels of quality (Featured Articles and Good Articles) this assignment requires a stricter review process, including the presentation of a candidacy and an evaluation by independent Wikipedians according to pre-established criteria. These two levels also have their own badges on the article page.

2.4. Sources

A fundamental aspect of Wikipedia lies in the system of links that allows its pages to be connected among them, making Wikipedia unique in this sense with regard to other encyclopedic systems (Reagle & Koerner, 2020). These internal links have been studied, showing both the semantic relationships they can establish and other potential utilities (Consonni, Laniado, & Montresor, 2019; Presutti, Consoli et al., 2014), as well as the possibility of calculating network indicators such as PageRank based on them (Thalhammer & Rettinger, 2016). There are, however, important issues to consider when working with Wikipedia pages links:

1. The links may be redirects; that is, old page versions that automatically redirect to the new versions when accessing them.

2. There are lists of links to other Wikipedia pages. Most of the lists include pages that are conceptually related to each other and share a clear subject matter. However, there are specific lists such as disambiguation pages, which are aimed at reducing the ambiguity of some terms (e.g., "citation" or "granada"), and therefore the links in these lists are not necessarily thematically related.

Another fundamental source for Wikipedia is its bibliographic references. Wikipedia recommends the use of bibliographic references to support its contents and it is an essential requirement for a page to achieve the best quality status (Featured article). These references are the same as those made in scientific publications, in both cases serving as a support for an idea. However, it is necessary to consider that citations in Wikipedia and citations in scientific publications are governed by different norms and dynamics. In Figure 2 the main differences between scientific publications references and Wikipedia references are schematized.

Other relevant particularities of Wikipedia references include

- Unlike scientific publications in which the identity of the citers (i.e., those including the references in the scientific publication) is clear and invariable, in Wikipedia this is more complex (given the live nature of Wikipedia articles) and not always possible. However, there are some methodological proposals for this purpose (Zagorova et al., 2022).

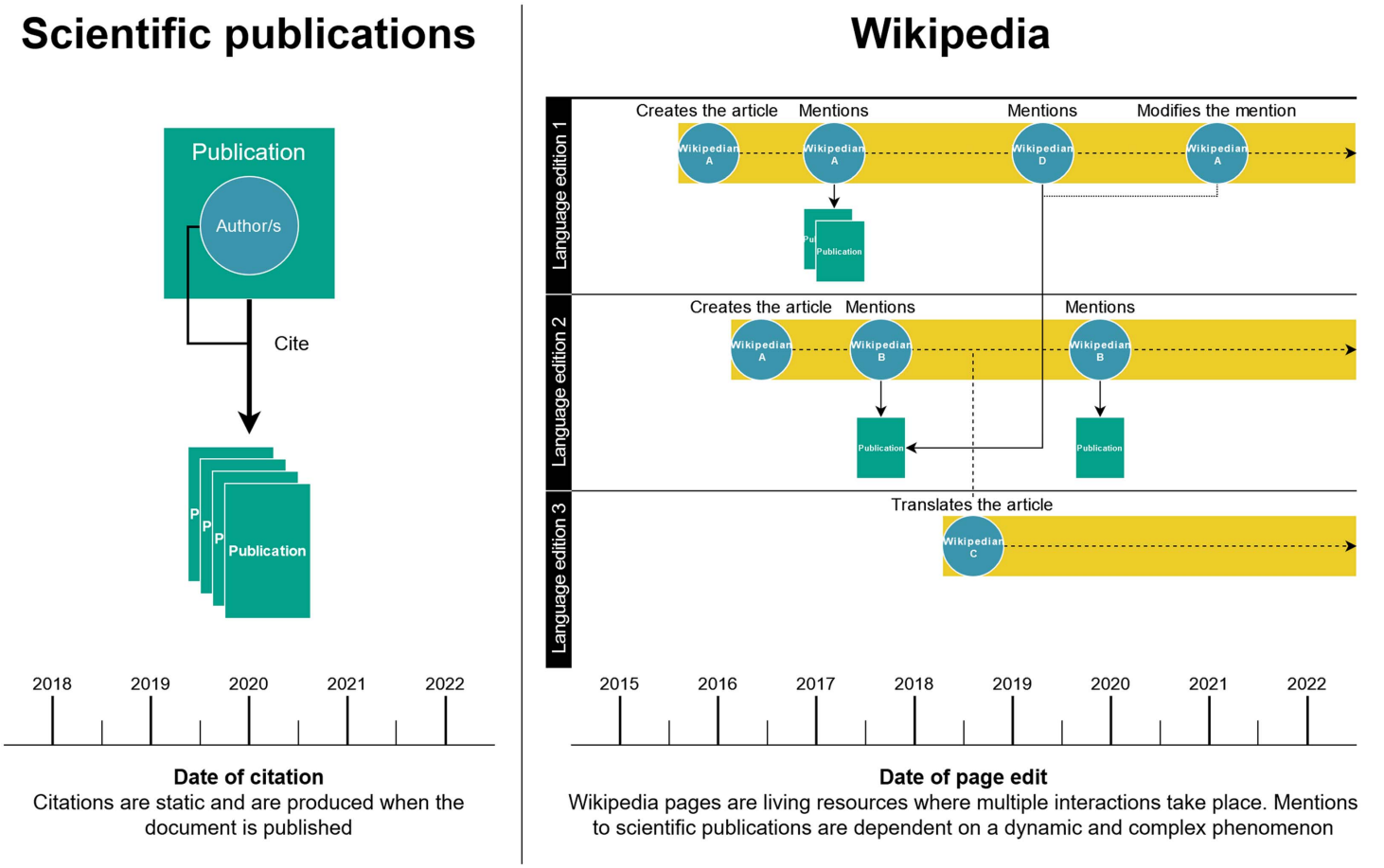

Figure 2. Differences between traditional citations and Wikipedia mentions of scientific publications.

- Wikipedia citation counts can be distorted by the translations of articles into different languages, because it is possible to easily transfer the references across the different language versions of the same article, thus distorting the meaning and value of Wikipedia citation counts. This limitation does not occur in scientific publications, as only one language version of a given publication is usually considered in the counting of citations.
- There are certain Wikipedia pages that function as large bibliographic indexes, bringing together the most relevant literature on a specific topic (e.g., research annuals or bibliographies).
- There are also templates (special Wikipedia pages that are embedded within other pages to facilitate the repetition of information), which are sometimes used to generate pre-established lists of references that are quickly inserted and replicated into numerous Wikipedia pages that are strongly related. This happened, for example, with the listing of lunar crater references (https://en.wikipedia.org/wiki/Wikipedia:Templates_for_discussion/Log/2014_June_8#Template:Lunar_crater_references).

2.5. Data Gathering

There are numerous data sources, and the choice of one or the other depends mostly on the type and volume of data required. In some cases, there are even multiple ways of accessing the same data. These have been summarized in Table 4, but can be found in detail in Section S3 in the Supplementary material. In fact, Wikimedia has a Research community (https://meta.wikimedia.org/wiki/Research) that gathers different resources to help and guide all those people who want to access the data of the Wikimedia projects and that lists the different projects related to it.

The two main sources are dumps and APIs. One of the main problems when working with Wikipedia data dumps is their size, especially when dealing with the more complete editions (e.g., the metadata of the revision of the English Wikipedia pages as of June 2022 is formed by 27 files of more than 2 Gbyte each), so accessing a subset of data requires a lot of time and effort. In the case of using Wikipedia APIs, metadata can be accessed on demand, but the retrieval process is very laborious, especially when large volumes of data are required. Other sources are characterized by offering already preprocessed data, such as the total number of edits or page views, which can be consulted from XTool.

In this paper, we extracted and developed a full Wikipedia knowledge graph with the ambition of facilitating the future of the English Wikipedia, reducing the time and effort that researchers may need in collecting and connecting all the different data sources.

2.6. Wikinformetrics

Finally, there are multiple metrics that can be extracted from the sources presented before and that enable the informetric study of Wikipedia pages. Based on previous studies and the above exploration of the informetric characteristics of Wikipedia, several metrics have been selected (Table 5). Each of them is of interest for measuring a particular dimension of the pages. For example, the number of views can be seen as a measure of the impact and outreach of a particular page, and although the numbers of edits and editors reflect the volume of activity, the numbers of talks and talkers are representative of the discussions that take place around these pages. These are not the only metrics that can be obtained from Wikipedia, but they can be considered to capture some of the most important analytical aspects of Wikipedia pages (e.g.,

Table 4. Summary of Wikipedia data sources by format, update frequency, data quantity, type, and challenges

	Content	Access	Format	Update frequency	Data quantity*	Type**	Main challenge***
Wikimedia Dumps	Metadata, page content, and relationships	Offline	XML, SQL	Once/twice a month	Big data	General	Data processing
MediaWiki and Wikimedia APIs	Metadata, page content, relationships, and statistics	Online	JSON, WDDX, XML, YAML, PHP	Real time	Small data	General	Data recovery
Wiki Replicas	Metadata, page content, and relationships	Online	SQL	Near-real time	Small data	General	Data recovery
Event Streams	Real-time logs	Online	SSE, JSON	Real time	–	Specific	Data recovery
Analytics dumps	Statistics on page views and activity	Offline	TSV	Monthly	Big data	Specific	Data processing
WikiStats	Statistics on page views, content, and activity	Online	JSON/CSV	Monthly	Small data	Specific	Data recovery
Dbpedia	Contents and semantic relationships	Both	RDF/XML, Turtle, N-Triplets, SPARQL endpoint	Live/monthly	–	General	Data recovery
XTools	Statistics on page views, content, and activity	Online	JSON	Real time	Small data	Specific	Data recovery
Repositories	Dedicated Wikipedia data sets	Offline	–	–	–	–	–
Altmetric aggregators	Wikipedia References to publications	Online	CSV/JSON	Daily	–	Specific	Data processing

* Volume of data to be retrieved and processed.

** Data from Wikipedia are included to address different problems or are of a specific nature.

*** Task that will require more effort when using the data source.

Table 5. Description of the metrics obtained for Wikipedia articles by analytical dimension

Metric	Analytical dimension	Description
Editors	Activity	Number of unique editors that have edited a Wikipedia article
Edits	Activity	Number of total edits that have a Wikipedia article
Linked	Connectivity	Number of Wikipedia articles in which the article is linked to
Links	Connectivity	Number of internal links that include a Wikipedia article to others
Age	Description	Years that have passed since the creation of the page to the date of data collection
Length	Description	Length in bytes of the page
Talkers	Discussion	Number of unique editors that have edited a Wikipedia article's talk page
Talks	Discussion	Number of total edits that the talk page of a Wikipedia article has
Views	Outreach	Number of daily views of a Wikipedia page
References	Support	Number of elements listed in the references
Pub. referenced	Support	Number of publications referenced
URLs	Support	Number of external links that include a Wikipedia article

contributions, content development, links and interactions, and impact), being also easy to interpret in an informetric framework.

3. WIKIPEDIA KNOWLEDGE GRAPH

Using the different data sources described above, a knowledge graph of the English edition of Wikipedia has been constructed for informetric purposes and freely shared on Zenodo (https://doi.org/10.5281/zenodo.6346899). The English edition of Wikipedia has been chosen because it is the largest one and has an international scope. For its construction, data from Wikimedia and analytic dumps were used, as well as data shared in repositories, specifically the data set of Singh et al. (2020) in which they share references made in Wikipedia articles. The data included in this data set covers all English Wikipedia activity until July 2021, except page views, which are from April 1, 2021 to June 30, 2021, and bibliographic reference data, until May 2020. R and Python have been used together, with the scripts available on GitHub (https://doi.org/10.5281/zenodo.6959428). The construction of this data set is described in Section S1 in the Supplementary material. The resulting data set consists of nine files connected to each other by a relational structure summarized in Figure 3.

This knowledge graph offers numerous possibilities for the informetric study of Wikipedia, making it possible to study new relationships (and interactions) between science and this social medium (e.g., the attention on Wikipedia to academic topics, the presence of scientific literature on popular Wikipedia pages, or the use of scientific literature in Wikipedia pages with large discussions in their Talk pages). This is the case of the work of Arroyo-Machado, Díaz-Faes, and Costas (2022), who found a positive relationship between the research performance of universities and their social attention on Wikipedia, using data from this data set.

Although the generation of new versions of the knowledge graph cannot be guaranteed by the authors of this paper, the way in which its creation is detailed and the shared scripts ensure

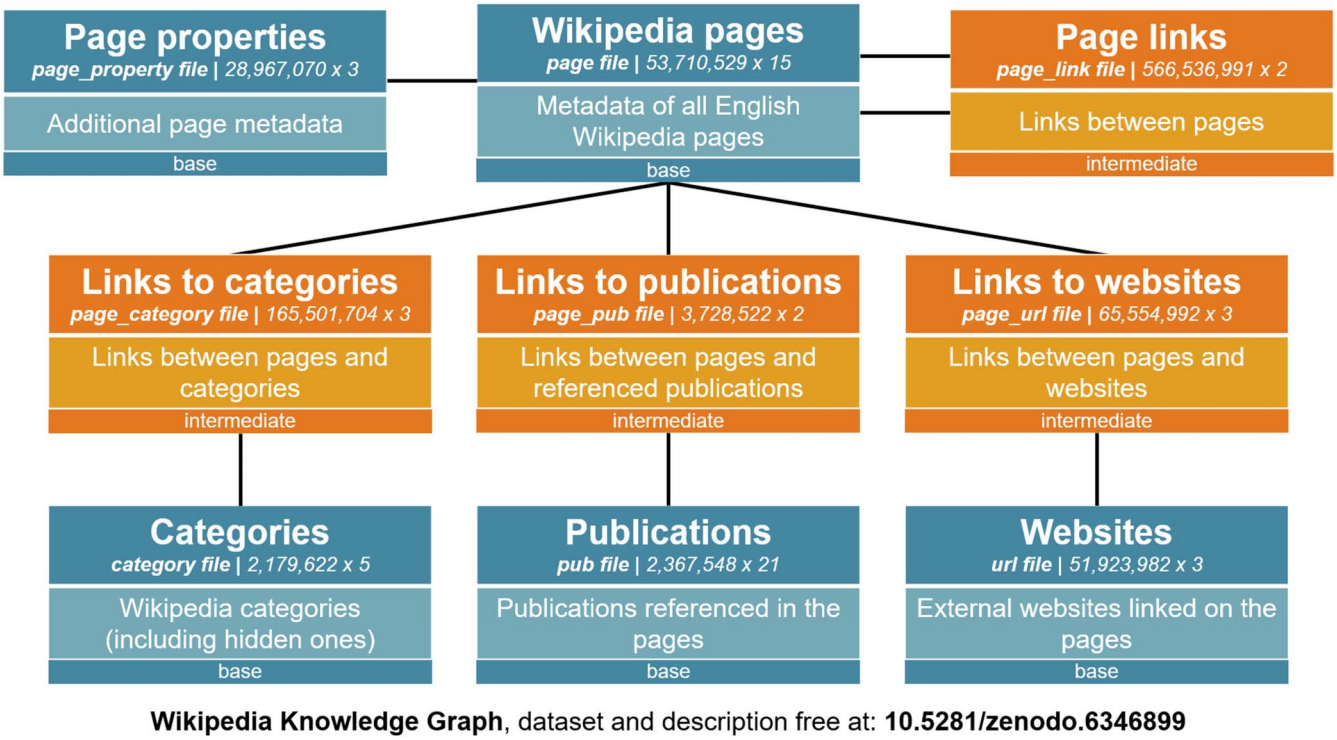

Figure 3. Diagram of files and relationships of the Wikipedia knowledge graph data set.

that new versions can be generated. This is also of importance for the generation of new knowledge graphs in other language editions of Wikipedia, as the data used as a basis are also available in other languages. The only limitation in this respect is in the reference data, as they come from a specific data set (Singh et al., 2020). However, those responsible have also shared the tools used to obtain the references and there are other alternatives such as Zagorova et al. (2022) or altmetric data aggregators.

4. CASE STUDY: INFORMETRIC ANALYSIS OF THE ENGLISH WIKIPEDIA

As a case study, the knowledge graph of the English Wikipedia is used to calculate and study the proposed metrics in a broad manner. The analysis was performed in Python and the code is available at GitHub (https://doi.org/10.5281/zenodo.6958972).

4.1. Wikipedia Metrics and Articles' Content

There are 53,710,529 pages in the English Wikipedia, considering all namespaces as well as pages that are redirects; however, this number is reduced to 6,328,134 pages when the focus is on articles that are not redirects. These represent just 11.79% of the overall English Wikipedia. The metrics proposed in Figure 4 have been obtained for all of them.

Figure 4 shows the descriptive statistics of the main variables, differentiating between total Wikipedia articles and those classified based on their quality; 5,522,676 articles (87.27% of the total) are associated with a WikiProject and with some quality level. Articles with different quality levels have been considered in all of them. It is noticeable that in all metrics, Featured articles have the highest values. The case of class B articles is noteworthy, as they not only show few differences with respect to the Good and A-Class articles, being

N. of articles → / Wiki Metrics ↓	All articles	Featured articles	Featured lists	A	Good	B	C	List	Start	Stub
	6,328,134	5945	3816	958	34,004	109,019	394,065	253,066	1,818,356	3,079,778
Editors	48.38	516.93	179.13	176.80	275.71	297.62	165.36	56.27	63.13	22.85
Edits	101.92	1491.35	593.61	564.91	724.13	705.41	369.89	159.80	129.52	40.23
Linked	80.53	725.25	175.84	202.01	330.18	417.00	234.08	107.34	93.03	55.70
Links	87.77	329.68	270.16	236.56	224.88	233.87	164.23	174.78	101.28	69.90
Age	9.59	14.33	11.52	12.74	12.06	12.47	10.92	9.13	10.45	9.20
Length	7844.68	61,248	51,549	43,329	39,444	35,009	21,676	18,202	10,033	3748
Talkers	5.38	66.17	16.62	27.90	29.64	28.16	15.03	4.98	6.56	3.64
Talks	9.19	258.40	42.36	92.21	88.56	88.35	35.32	9.07	9.69	4.32
Views	3345.07	64,801	26,685	16,011	29,229	30,359	15,829	3777	4094	710
References	4.6	53.95	55.49	31.76	38.87	26.51	15.40	9.20	5.79	1.84
Pub. Ref.	0.59	14.27	2.34	8.51	5.83	4.77	2.37	0.53	0.69	0.22
URLs	10.33	58.03	67.32	33.32	46.10	40.31	25.95	22.82	12.90	6.09

Figure 4. Average of Wikipedia article metrics differentiating by the quality assigned from a project.

also greater in number of articles than both, but in aspects such as views they are positioned above them.

There are important differences in the number of referenced publications, going from an average of 14.27 publications in Featured articles to 8.52 in A and 5.84 in Good articles, while the Start and Stub articles cite on average less than one publication. This reflects compliance with English Wikipedia's criteria for establishing the quality level of articles. The general criteria do not make explicit the need for a greater number of references to increase the level of quality, among others, but they do require an increase in "reliable sources," so that citations to publications can serve as a proxy for this. Likewise, it also corroborates previous findings of a relationship between the level of quality and the number of edits (Wilkinson & Huberman, 2007), and the length of articles (Blumenstock, 2008).

Most Wikipedia pages are not of recent creation (Figure 5A), with a median of 11 years. In some of the metrics, such as edits and talks, extreme outliers are found. This can be seen in the fact that their average values are 102 and 9.19, respectively, above the median and third quartile values. This situation is much more pronounced in the case of views, with an average of 3,346.59. Furthermore, the number of referenced elements has a median of 1 and an average of 4.6. When comparing the links with the linked ones, we find that Wikipedia pages link more than they are linked, because the median for the former is 36 and for the latter 15.

The correlations between these variables are all positive (Figure 5B). The strongest correlation is between talkers and talks (r_s = 0.97), followed by another analogous relationship such as that between editors and edits (r_s = 0.94). When considering pairs of metrics of different nature, the strongest correlation is between edits and views (r_s = 0.74), followed by that of

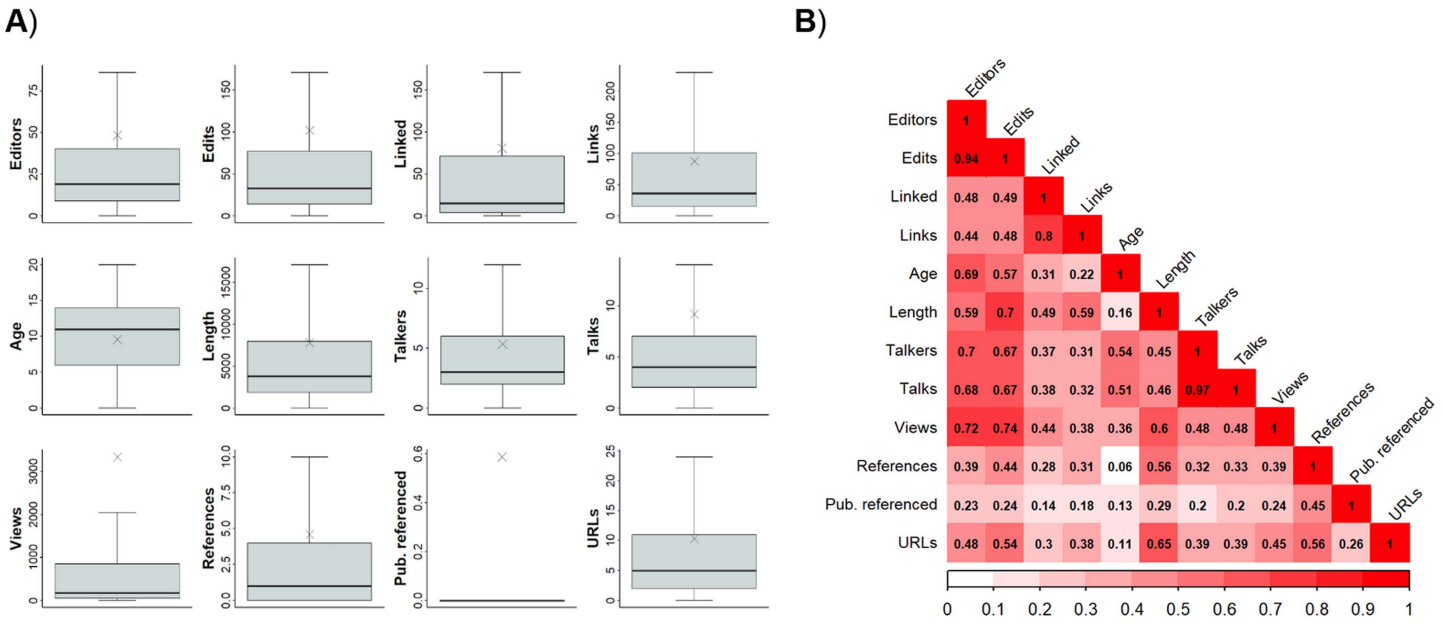

Figure 5. A: Boxplots of the main metrics for Wikipedia articles excluding outliers from the figures and marking the mean with a cross symbol. B: Spearman's rho correlations between the main metrics for Wikipedia English articles.

editors and views (r_s = 0.72), which suggests a relationship between the popularity of Wikipedia pages in terms of visits and their number of edits. Interestingly, a lower correlation was found between views, and both talks and talkers (r_s = 0.48), suggesting that discussions around Wikipedia pages are not necessarily related to higher numbers of views. Another moderate correlation can be found between the length of an article and its views (r_s = 0.6), which may indicate that the larger the article, the more attention it receives or that the more attention it receives, the more it grows in length. There are other moderate correlations, such as between the length and the number of references (r_s = 0.56) and URLs (r_s = 0.65), but which are to be expected as the two elements directly interfere with each other. The number of referenced publications is the metric most weakly correlated, there being for example a weak correlation between this and views (r_s = 0.24) or talks (r_s = 0.2). Our results confirm the same type of relationships reported in previous research (Mittermeier et al., 2021), albeit this time considering the entire population of English language Wikipedia articles.

4.2. Different Types of Attention Captured on Wikipedia

The results of this analysis can also be accessed interactively and in greater detail via the R Shinny app: https://wenceslao-arroyo-machado.shinyapps.io/wikinformetrics/.

A review of Wikipedia's main pages based on different metrics reveals its potential to capture content that responds to different types of attention (Table S4 in the Supplementary material). The page views make it possible to identify those topics that capture the most attention of society in a given period—page views are limited to a period of 3 months in our data set. Thus, in our data set the pages of *Prince Philip, Duke of Edinburgh* (10,860,553 views) and *Elizabeth II* (9,900,275), or *Mare of Easttown* (5,995,513) rank among the most visited in the English-language Wikipedia. Also, five of the 20 most viewed pages are series or movies released in the period analyzed, which also highlights that content related to entertainment

occupies a relevant position in Wikipedia. Sports also receive many views and reflect current events, as evidenced by the *UEFA Euro 2020* page (12,100,455 views), the second most viewed, just after the *Main Page* (554,030,839). There is a clear presence of articles that respond to general interests, such as the *Bible* (11,048,609) or *Cleopatra* (9,516,340) pages. This may indicate that some topics raise general interest and may not be time related.

The number of talks of Wikipedia articles is often used in conjunction with other variables in the construction of models for controversy detection (Jang, Foley et al., 2016). This suggests that this metric may be useful for detecting such controversial content in a simple way. Among the 20 pages with the highest number of talks, those of political figures, religion topics, and scientific controversies stand out. The strong talk that takes place in some of them, as in *Donald Trump* (62,944), and the vandalism and presence of trolls, as in *Gamergate controversy* (27,185), have caused the editing of these pages to be restricted. In fact, there are some articles clearly related to controversial or sensitive issues, such as *Climate change* (40,837) and *Homeopathy* (25,898). In this regard, Wikipedia itself offers a page with a curated list of controversial articles (https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues), with 13 of the 20 pages listed as of 4 July 2021.

Finally, based on the volume of referenced publications, that is, all materials with an associated identifier (DOI, ISBN, arXiv ID, etc.), it is also possible to identify the Wikipedia pages that cite more scientific publications. However, in this case there are many research annuals and bibliographic pages present among the 20 articles, for example *2018 in paleontology* with 569 referenced publications. These lists have been eliminated to select the top 20 articles with encyclopedic content. In these articles there is a clear presence of scientific content, especially in medicine, such as *Feminizing hormone therapy* (329) and *Alzheimer's disease* (277). However, there are also articles related to history, such as *History of Lisbon* (313) or *World War II* (264). This may suggest that the metric of the number of publications cited can be used as a proxy to identify Wikipedia articles that are more scholarly oriented.

5. DISCUSSION

In this study we describe how Wikipedia is a complex system, involving numerous actors and elements, and whose rules and governance depend on the community itself (Jemielniak, 2012). It is not only one of the first and clearest examples of Web 2.0 but also one of the few that remains among the most visited websites and has not deviated from its initial objective. Far from that, over the years it has gained the acceptance and trust of many of those who initially looked at it with skepticism.

We describe many similarities between scientific publications and Wikipedia pages. Both have different typologies of documents, structured content, evaluation of content, and use of links and bibliographic references. There are also notable differences. While scientific publications may have limited access and a more specialized audiences, Wikipedia's content and scope is more open and targeted to more general audiences. The live nature of Wikipedia is probably its main distinctive feature when compared to scientific publications. This must be considered when conducting informetric research on Wikipedia. To help in this endeavor, we propose an informetric-inspired conceptual framework, proposing different metrics that pay attention to the different analytical dimensions of Wikipedia, such as article characteristics, outreach, or citations to scientific publications. Some of these metrics have been already explored in the literature, such as page views (Mittermeier et al., 2019, 2021), but never in a comprehensive conceptual framework. The informetric-inspired conceptual framework presented here is expected to be useful for any Wikipedia study involving informetric,

scientometric, bibliometric, or webometric perspectives. Similarly, different Wikipedia data sources have been identified and described, finding in their differences in coverage, volume, access, or data processing crucial aspects for their selection.

Alongside the conceptual analytical framework proposed, a knowledge graph of the English edition of Wikipedia has been built and shared openly (https://doi.org/10.5281/zenodo.6346899). The data are gathered under a comprehensive data set that follows a relational model and can be used by anyone interested in the study of this encyclopedia from an informetric point of view. It combines different data sources that allow users on the one hand to characterize any Wikipedia page, while also allowing them to establish relationships between each other (e.g., between two articles, an article and a category or an article and a linked website or a scientific publication referenced in it). Together with the metadata and relations of Wikipedia pages, the data of their bibliographic references are also incorporated, which come from the data set shared by Singh et al. (2020). It is precisely in Wikipedia's bibliographic reference data where greater efforts are needed so that they can be efficiently accessed through its official sources, such as dumps or the API.

The case study provides a descriptive overview of Wikipedia articles in its English edition, suggesting interesting valuable analytical possibilities and highlighting the relationships and usefulness of the metrics described. Our results suggest that the low correlations among most of the metrics point to the fact that the analytical dimensions measured through them are rather distinct. The potential analytical usefulness of some of the metrics has been highlighted. For example, the number of Wikipedia page views can be seen as a metric of social attention; the number of talks of Wikipedia pages can be seen as a proxy of controversial topics; and the number of scientific references in Wikipedia pages can help identify scholarly-related content. The use of the quality levels derived from WikiProjects has proved to be useful, showing clear differences between the different levels, but has also provided an overview of the Wikipedia articles.

Finally, it is important to also mention some of the limitations of this work. First, not all possible Wikipedia metrics and their relationships have been explored (e.g., the relationship between pages and users, or the number of users who follow the pages (the so-called *watchers*), or the number of editions in other languages of a given article). The use of large amounts of data and some specific sources leads to a loss of consistency. For example, the Wikipedia dump process takes several days without blocking the edits during that time, so they are not really a snapshot. This loss of consistency also occurs when using different sources, especially when combining 2021 Wikipedia data with references from a third-party data set published in 2020. The knowledge graph and the case study are based on the English Wikipedia; however, future research should study whether the same relationships found in this study also hold for other languages as well as the existing relationships between language editions.

ACKNOWLEDGMENTS

We thank Mercedes and María for their intellectual advice in the early stages.

AUTHOR CONTRIBUTIONS

Wenceslao Arroyo-Machado: Data curation, Formal analysis, Investigation, Software, Visualization, Writing—original draft. Daniel Torres-Salinas: Funding acquisition, Resources, Validation, Writing—review & editing. Rodrigo Costas: Conceptualization, Methodology, Project administration, Supervision, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work was funded by the Spanish Ministry of Science and Innovation with grant number PID2019-109127RB-I00/SRA/10.13039/501100011033. Wenceslao Arroyo-Machado received an FPU Grant (FPU18/05835) from the Spanish Ministry of Universities. Daniel Torres-Salinas received support under the Reincorporation Programme for Young Researchers of the University of Granada. Rodrigo Costas is partially funded by the South African DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

DATA AVAILABILITY

The Wikipedia knowledge graph data set is available in Zenodo (Arroyo-Machado et al., 2022).

The source code for constructing the Wikipedia knowledge graph data set is available in Zenodo (Arroyo-Machado, 2022a).

The case study code is available in Zenodo (Arroyo-Machado, 2022b).

REFERENCES

Adams, C. E., Montgomery, A. A., Aburrow, T., Bloomfield, S., Briley, P. M., ... Xia, J. (2020). Adding evidence of the effects of treatments into relevant Wikipedia pages: A randomised trial. *BMJ Open*, *10*(2), e033655. https://doi.org/10.1136/bmjopen-2019-033655, PubMed: 32086355

Adams, J., Brückner, H., & Naslund, C. (2019). Who counts as a notable sociologist on Wikipedia? Gender, race, and the "Professor Test." *Socius*, *5*, 2378023118823946. https://doi.org/10.1177/2378023118823946

Aghaebrahimian, A., Stauder, A., & Ustaszewski, M. (2020). Testing the validity of Wikipedia categories for subject matter labelling of open-domain corpus data. *Journal of Information Science*, *48*(5), 686–700. https://doi.org/10.1177/0165551520977438

Arroyo-Machado, W. (2022a). Wences91/wikipedia_knowledge_graph [Source code]. https://doi.org/10.5281/zenodo.6959428

Arroyo-Machado, W. (2022b). Wences91/wikinformetrics [Source code]. https://doi.org/10.5281/zenodo.6958972

Arroyo-Machado, W., Díaz-Faes, A. A., & Costas, R. (2022). New insights on social media metrics: Examining the relationship between universities' academic reputation and Wikipedia attention. *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)*, Granada, Spain. https://doi.org/10.5281/zenodo.6962442

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikipedia knowledge graph dataset [Data set]. https://doi.org/10.5281/zenodo.6346899

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

Black, E. W. (2008). Wikipedia and academic peer review. *Online Information Review*, *32*(1), 73–88. https://doi.org/10.1108/14684520810865994

Blumenstock, J. E. (2008). Size matters: Word count as a measure of quality on Wikipedia. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 1095–1096). https://doi.org/10.1145/1367497.1367673

Boldi, P., & Monti, C. (2016). Cleansing Wikipedia categories using centrality. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 969–974). https://doi.org/10.1145/2872518.2891111

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. *BMJ: British Medical Journal*, *348*, g1585. https://doi.org/10.1136/bmj.g1585, PubMed: 24603564

Candelario, D. M., Vazquez, V., Jackson, W., & Reilly, T. (2017). Completeness, accuracy, and readability of Wikipedia as a reference for patient medication information. *Journal of the American Pharmacists Association: JAPhA*, *57*(2), 197–200., https://doi.org/10.1016/j.japh.2016.12.063, PubMed: 28139458

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

Consonni, C., Laniado, D., & Montresor, A. (2019). WikiLinkGraphs: A complete, longitudinal and multi-language dataset of the Wikipedia link networks. In *Proceedings of the 13th International AAAI Conference on Web and Social Media* (pp. 598–607). https://doi.org/10.1609/icwsm.v13i01.3257

Costas, R., de Rijcke, S., & Marres, N. (2020). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, *72*(5), 595–610. https://doi.org/10.1002/asi.24427

Cummings, R. E. (2020). Writing knowledge: Wikipedia, public review, and peer review. *Studies in Higher Education*, *45*(5), 950–962. https://doi.org/10.1080/03075079.2020.1749791

Détienne, F., Baker, M., Fréard, D., Barcellini, F., Denis, A., & Quignard, M. (2016). The descent of Pluto: Interactive dynamics, specialisation and reciprocity of roles in a Wikipedia debate. *International Journal of Human-Computer Studies*, *86*, 11–31. https://doi.org/10.1016/j.ijhcs.2015.09.002

Díaz-Faes, A. A., Bowman, T. D., & Costas, R. (2019). Towards a second generation of "social media metrics": Characterizing Twitter communities of attention around science. *PLOS ONE*, *14*(5), e0216408. https://doi.org/10.1371/journal.pone.0216408, PubMed: 31116783

Dzogang, F., Lansdall-Welfare, T., & Cristianini, N. (2016). Seasonal fluctuations in collective mood revealed by Wikipedia searches and Twitter posts. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 931–937). https://doi.org/10.1109/ICDMW.2016.0136

Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 777–786).

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLOS Computational Biology*, *10*(11), e1003892. https://doi.org/10.1371/journal.pcbi.1003892, PubMed: 25392913

Hara, N., & Doney, J. (2015). Social construction of knowledge in Wikipedia. *First Monday*, *20*(6). https://doi.org/10.5210/fm.v20i6.5869

Heist, N., & Paulheim, H. (2019). Uncovering the semantics of Wikipedia categories. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (Eds.), *The Semantic Web – ISWC 2019* (pp. 219–236). Springer International Publishing. https://doi.org/10.1007/978-3-030-30793-6_13

Hill, B. M., & Shaw, A. (2015). Page protection: Another missing dimension of Wikipedia research. In *Proceedings of the 11th International Symposium on Open Collaboration*. https://doi.org/10.1145/2788993.2789846

History of Wikipedia. (2021). *Wikipedia*. 28 May. https://en.wikipedia.org/wiki/History_of_Wikipedia

Jang, M., Foley, J., Dori-Hacohen, S., & Allan, J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 2069–2072). https://doi.org/10.1145/2983323.2983911

Jemielniak, D. (2012). *Wikipedia: An effective anarchy*. Baltimore, MD: Society for Applied Anthropology.

Jemielniak, D. (2019). Wikipedia: Why is the common knowledge resource still neglected by academics? *GigaScience*, *8*(12), giz139. https://doi.org/10.1093/gigascience/giz139, PubMed: 31794014

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The most influential medical journals according to Wikipedia: Quantitative analysis. *Journal of Medical Internet Research*, *21*(1), e11429. https://doi.org/10.2196/11429, PubMed: 30664451

Kaffee, L.-A., & Elsahar, H. (2021). References in Wikipedia: The editors' perspective. In *Companion Proceedings of the Web Conference 2021* (pp. 535–538). https://doi.org/10.1145/3442442.3452337

Katz, G., & Rokach, L. (2017). Wikiometrics: A Wikipedia based ranking system. *World Wide Web*, *20*(6), 1153–1177. https://doi.org/10.1007/s11280-016-0427-8

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1509–1512). https://doi.org/10.1145/1518701.1518930

Kopf, S. (2020). Participation and deliberative discourse on social media—Wikipedia talk pages as transnational public spheres? *Critical Discourse Studies*, *19*(2), 196–211. https://doi.org/10.1080/17405904.2020.1822896

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

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67. https://doi.org/10.1007/s13194-012-0056-8

Lageard, V., & Paternotte, C. (2021). Trolls, bans and reverts: Simulating Wikipedia. *Synthese*, *198*(1), 451–470. https://doi.org/10.1007/s11229-018-02029-0

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

Li, X., Thelwall, M., & Mohammadi, E. (2021). How are encyclopedias cited in academic research? Wikipedia, Britannica, Baidu Baike, and Scholarpedia. *Profesional de La Información*, *30*(5). https://doi.org/10.3145/epi.2021.sep.08

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

Maki, K., Yoder, M., Jo, Y., & Rosé, C. (2017). Roles and success in Wikipedia talk pages: Identifying latent patterns of behavior. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1026–1035). https://aclanthology.org/I17-1103

Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2019). Can deep learning techniques improve classification performance of vandalism detection in Wikipedia? *Engineering Applications of Artificial Intelligence*, *78*, 248–259. https://doi.org/10.1016/j.engappai.2018.11.012

Minguillón, J., Lerga, M., Aibar, E., Lladós-Masllorens, J., & Meseguer-Artola, A. (2017). Semi-automatic generation of a corpus of Wikipedia articles on science and technology. *Profesional de La Información*, *26*(5), 995–1005. https://doi.org/10.3145/epi.2017.sep.20

Miquel-Ribé, M., & Laniado, D. (2018). Wikipedia culture gap: Quantifying content imbalances across 40 language editions. *Frontiers in Physics*, *6*, 54. https://doi.org/10.3389/fphy.2018.00054

Mittermeier, J. C., Correia, R., Grenyer, R., Toivonen, T., & Roll, U. (2021). Using Wikipedia to measure public interest in biodiversity and conservation. *Conservation Biology*, *35*(2), 412–423. https://doi.org/10.1111/cobi.13702, PubMed: 33749051

Mittermeier, J. C., Roll, U., Matthews, T. J., & Grenyer, R. (2019). A season for all things: Phenological imprints in Wikipedia usage and their relevance to conservation. *PLOS Biology*, *17*(3), e3000146. https://doi.org/10.1371/journal.pbio.3000146, PubMed: 30835729

Mühlhauser, I., & Oser, F. (2008). Does WIKIPEDIA provide evidence based health care information? A content analysis. *Shared Decision-Making in Health Care*, *102*(7), e1–e7. https://doi.org/10.1016/j.zefq.2008.06.020

Nicholson, J. M., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M., & Rife, S. C. (2021). Measuring the quality of scientific references in Wikipedia: An analysis of more than 115M citations to over 800 000 scientific articles. *The FEBS Journal*, *288*(14), 4242–4248. https://doi.org/10.1111/febs.15608, PubMed: 33089957

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

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

Olleros, F. X. (2008). Learning to trust the crowd: Some lessons from Wikipedia. In *2008 International MCETECH Conference on E-Technologies (Mcetech 2008)* (pp. 212–216). https://doi.org/10.1109/MCETECH.2008.17

O'Neil, T. (2017). *Wikipedia erases record of accomplished scientist—'Censored' for his intelligent design position*. PJ Media. https://pjmedia.com/faith/tyler-o-neil/2017/11/21/wikipedia-erases-record-of-accomplished-scientist-censored-for-his-intelligent-design-position-n101002

Ortega, J.-L. (2020). Altmetrics data providers: A meta-analysis review of the coverage of metrics and publication. *Profesional de La Información*, *29*(1). https://doi.org/10.3145/epi.2020.ene.07

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

Presutti, V., Consoli, S., Nuzzolese, A. G., Recupero, D. R., Gangemi, A., ... Zargayouna, H. (2014). Uncovering the semantics of Wikipedia pagelinks. In K. Janowicz, S. Schlobach, P. Lambrix, & E. Hyvönen (Eds.), *Knowledge engineering and knowledge management* (pp. 413–428). Springer International Publishing. https://doi.org/10.1007/978-3-319-13704-9_32

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

Reagle, J. (2009). Wikipedia: The happy accident. *Interactions*, *16*(3), 42–45. https://doi.org/10.1145/1516016.1516026

Reagle, J., & Koerner, J. (Eds.). (2020). *Wikipedia @ 20: Stories of an incomplete revolution*. MIT Press. https://doi.org/10.7551/mitpress/12366.001.0001

Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., ... Grenyer, R. (2016). Using Wikipedia page views to explore the cultural importance of global reptiles. *Biological Conservation*, *204*, 42–50. https://doi.org/10.1016/j.biocon.2016.03.037

Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research*, *6*, 588. https://doi.org/10.12688/f1000research.11369.2, PubMed: 28580134

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

Thalhammer, A., & Rettinger, A. (2016). PageRank on Wikipedia: Towards general importance scores for entities. In H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, & C. Lange (Eds.), *The semantic web* (pp. 227–240). Springer International Publishing. https://doi.org/10.1007/978-3-319-47602-5_41

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

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

Tripodi, F. (2021). Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*, 14614448211023772. https://doi.org/10.1177/14614448211023772

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. *PLOS ONE*, *12*(2), e0171774. https://doi.org/10.1371/journal.pone.0171774, PubMed: 28231323

Vilain, P., Larrieu, S., Cossin, S., Caserio-Schönemann, C., & Filleul, L. (2017). Wikipedia: A tool to monitor seasonal diseases trends? *Online Journal of Public Health Informatics*, *9*(1). https://doi.org/10.5210/ojphi.v9i1.7630

Weiner, S. S., Horbacewicz, J., Rasberry, L., & Bensinger-Brody, Y. (2019). Improving the quality of consumer health information on Wikipedia: Case series. *Journal of Medical Internet Research*, *21*(3), e12450. https://doi.org/10.2196/12450, PubMed: 30882357

Wilkinson, D. M., & Huberman, B. A. (2007). Assessing the value of cooperation in Wikipedia. *First Monday*, *12*(4). https://doi.org/10.5210/fm.v12i4.1763

Wouters, P., Zahedi, Z., & Costas, R. (2019). Social media metrics for new research evaluation. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 687–713). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_26

Xiao, L., & Askin, N. (2014). Academic opinions of Wikipedia and Open Access publishing. *Online Information Review*, *38*(3), 332–347. https://doi.org/10.1108/OIR-04-2013-0062

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PLOS ONE*, *7*(6), e38869. https://doi.org/10.1371/journal.pone.0038869, PubMed: 22745683

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

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), e0197326. https://doi.org/10.1371/journal.pone.0197326, PubMed: 29772003

Zhang, H., Ren, Y., & Kraut, R. E. (2018). Mining and predicting temporal patterns in the quality evolution of Wikipedia articles. *Academy of Management Proceedings*, *2018*(1), 13746. https://doi.org/10.5465/AMBPP.2018.13746abstract

Zheng, L., Albano, C. M., Vora, N. M., Mai, F., & Nickerson, J. V. (2019). The roles bots play in Wikipedia. In *Proceedings of the ACM Conference on Human-Computer Interactions*, *3*(CSCW), 1–20. https://doi.org/10.1145/3359317