

WENCESLAO ARROYO MACHADO

***Big data techniques applied to
the study and characterisation of
scientific activity on social media***



UNIVERSIDAD
DE GRANADA

Editor: Universidad de Granada. Tesis Doctorales
Autor: Wenceslao Arroyo Machado
ISBN: 978-84-1195-035-0
URI: <https://hdl.handle.net/10481/84688>



UNIVERSIDAD DE GRANADA

Big data techniques applied to the study and
characterisation of scientific activity on social media

DOCTORAL DISSERTATION

presented to obtain the

DOCTOR OF PHILOSOPHY DEGREE

in the

Doctoral Programme in Information and Communication Technologies

by

Wenceslao Arroyo Machado

PhD Advisors

Enrique Herrera Viedma & Daniel Torres Salinas

Granada, July 2023

To Mercedes & Enzo

Influ Science
ediciones

Cover image extracted from a frame from the series The Sopranos. The use of this image is non-profit.
This thesis has been funded by the Spanish Ministry of Universities FPU Grant (FPU18/05835).



Agradecimientos (Acknowledgments)

Quiero empezar agradeciendo a Enrique Herrera Viedma y Daniel Torres Salinas, mis directores de tesis y sin los cuales este viaje habría sido imposible. Gracias Enrique por confiar y guiarme. Gracias Daniel por tu dedicación y cuidado. Si algo me llevo de estos años es un mentor y, sobre todo, un amigo. Nunca tendré palabras para agradecerte todo lo que has hecho por mí y todo el tiempo que me has regalado. Has sabido motivarme y conseguido transferirme todo cuanto sabes con mucha pasión e infinita paciencia. Muchas veces has bromeado con que somos como Tony y Christopher o Batman y Damian, pero en verdad somos Alfredo y Totó. Por muchos años más a tu lado.

La siguiente en la lista no puede ser otra que Mercedes. Sin ella los resultados no habrían llegado. En todo momento me has dado un profundo apoyo y me has cuidado sin contemplaciones. Son muchas las horas que esta tesis nos ha robado, pero desde luego que sin ti no habrían tenido sentido. Estos años también han sido muy importantes para nosotros. Hemos tenido altibajos, pero, sobre todo, nos hemos casado y ahora esperamos un niño. Te quiero Mercedes.

Mi madre. Nadie en el mundo creo que me entienda como tú lo haces. Durante este tiempo me has apoyado ciegamente y motivado. Pero, lo que más has hecho, es cuidarme cuando solo tú veías que algo no iba bien y siempre has hecho tuyos mis problemas y preocupaciones. Ojalá podamos seguir yendo a conciertos de los *Rolling Stones* y ferias del libro por muchos años más. Evidentemente, no has estado sola y ahí también se encuentran mi padre y mi hermana. Muchas gracias por cuidarme y preocuparos por mí, no hay nada que me haga más feliz que veros y saber que estáis ahí.

Durante todos estos años de tesis también he tenido la oportunidad de cruzarme con gente maravillosa. Muchas gracias Nicolás, empezaste guiándome en un congreso, me has enseñado a investigar y, después de todo este tiempo, me enorgullezco de poder llamarte amigo. Muchas gracias Rodrigo, no sé cuánto tiempo te habré hecho perder en reuniones virtuales y cuanto me has enseñado desde una pantalla, pero tu paciencia infinita, horas de dedicación y salidas por Leiden son algo que no podré agradecerte suficientemente nunca. Muchas gracias también a Esteban, Juan, Domingo y Adrián, he aprendido muchísimo con vosotros y disfrutado enormemente.

Y un especial agradecimiento a mis amigos. Sandra, eres la persona más increíble del mundo. Diego, sin ti los días serían grises. Joaquín, tu nobleza y dedicación son un ejemplo para todos.

Table of contents

I PhD dissertation.....	12
1 Introduction.....	13
2 Literature review.....	17
2.1 Major Challenges in Collecting and Processing Social Media Data	17
2.2 From Classical Horizons: Adapting Standard Scientometrics Methods in Altmetrics Research	20
2.3 Towards New Horizons: Exploring Original Methods and New Opportunities in Social Media	22
3 Objectives.....	24
4 Methodology.....	25
5 Summary	26
5.1 Facing Heterogeneity of Sources and Data From a Data Science Perspective.....	27
5.2 Importing Classical Methods for Scientific Mapping of Social Media	28
5.3 Developing Novel Approaches for Mapping Social and Semantic Relationships	28
6 Discussion of results	30
6.1 Facing Heterogeneity of Sources and Data From a Data Science Perspective.....	30
6.2 Importing Classical Methods for Scientific Mapping of Social Media	30
6.3 Developing Novel Approaches for Mapping Social and Semantic Relationships	31
7 Concluding remarks	33
References	34
II Publications.....	39
Exploring WorldCat Identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics	40
1. Introduction	41
2. Methodology	43
3. Results	46
4. Discussion & Conclusions.....	53
5. References	56
Wikinformetrics: Construction and description of an open Wikipedia knowledge graph dataset for informetric purposes.....	60

1. Introduction	61
2. Wikipedia from an informetric perspective	66
3. Wikipedia knowledge graph.....	75
4. Case study: informetric analysis of the English Wikipedia	76
5. Discussion.....	80
References.....	81
Mapping the backbone of the Humanities through the eyes of Wikipedia	88
1. Introduction	89
2. Material and methods.....	91
3. Analysis and results	95
4. Conclusions	102
References.....	104
Science through Wikipedia: a novel representation of open knowledge through co-citation networks.....	108
Introduction.....	109
Materials and Methods.....	111
Results.....	113
Discussion	124
References.....	129
Mapping social media attention in Microbiology: Identifying main topics and actors	133
Introduction.....	134
Objectives	135
Materials and Methods.....	135
Results.....	137
Discussion	142
References.....	145
Identifying and characterizing social media communities: a socio-semantic network approach to altmetrics.....	147
1. Introduction	148
2. Background.....	149
3. Data and methods	155

4. Case study: Information Science & Library Science.....	159
5. Case study: Microbiology.....	163
6. Discussion.....	166
7. Concluding remarks.....	167
8. References	168

Resumen

La llegada de los medios sociales ha propiciado todo un ecosistema digital para la comunicación y la gestión de la información. Este cambio ha afectado de lleno a la ciencia y la forma en la que se publican y difunden sus resultados. Twitter, Wikipedia o las noticias son ahora la cabeza visible de un extenso número de canales para la comunicación científica, integrándose y siendo visible este discurso y difusión de resultados científicos para toda la sociedad. Esto ha dado el paso a la exploración de la forma en que se consume la ciencia en dichos entornos y cuál es la atención que captan más allá del reino académico. No obstante, se ha identificado una falta de profundidad y aprovechamiento de los medios estudiados más allá de la contabilización de menciones a trabajos científicos y que ponga en mayor contexto la actividad en torno a la ciencia, así como la existencia de plataformas inexploradas y la adaptación limitada de métodos tradicionales de la ciencimetría para el estudio cuantitativo de la ciencia. Esta tesis tiene por objetivo afrontar estos retos para ahondar en el potencial de los masivos datos de medios sociales y la heterogeneidad de los medios sociales para el estudio de la ciencia, combinando para ello la ciencia de datos y la ciencimetría. Como resultado se han elaborado propuestas de marcos conceptuales y metodológicos para el uso y mapeo de datos de medios sociales. Para ello se han adaptado técnicas clásicas de ciencimetría para el análisis de redes sociales y se han propuesto nuevos métodos para la elaboración de mapas científicos combinando información social y semántica. Esto permite la identificación de las estructuras del conocimiento establecidas a través de la actividad social y la identificación de comunidades cognitivas de actores sociales. Además, las propuestas metodológicas han sido puestas en práctica mediante estudios de caso y a gran escala para validarlas y ofrecer resultados novedosos sobre la discusión y difusión de la ciencia en Twitter y Wikipedia, en especial en comparación con el ámbito académico.

Abstract

The advent of social media has spawned an entire digital ecosystem for communication and information management. This change has had a profound effect on science and the way its results are published and disseminated. Twitter, Wikipedia, and news outlets are now the visible heads of an extensive number of channels for scientific communication, integrating and making the discourse and dissemination of scientific results visible to society as a whole. This has led to the exploration of how science is consumed in such environments and the attention it captures beyond the realm of academia. However, a lack of depth and exploitation of the media studied beyond counting mentions of scholarly outputs has been identified, along with putting the activity around science into greater context. There also exists the unexplored platforms and limited adaptation of traditional methods of scientometrics for the quantitative study of science. This thesis aims to address these challenges to delve into the potential of massive social media data and the heterogeneity of social media for the study of science by combining data science and scientometrics. As a result, proposals for conceptual and methodological frameworks for the use and mapping of social media data have been developed. For this purpose, classic scientometric techniques have been adapted for social network analysis, and new methods have been proposed for the creation of scientific maps that combine social and semantic information. This allows the identification of knowledge structures established through social activity and the identification of cognitive communities of social actors. Furthermore, the methodological proposals have been put into practice through case studies and large-scale studies to validate them and provide novel results on the discussion and dissemination of science on Twitter and Wikipedia, particularly in comparison to academia.

Structure

The thesis is divided into two parts: the PhD dissertation and the associated publications. In the first part, Section 1 provides a primarily historical context to the research problem. Section 2 presents the main topics addressed in the thesis: challenges in data processing of social media (Section 2.1), adaptation of classical scientometric methods to altmetric research (Section 2.2), and exploration of new methodological approaches to altmetrics (Section 2.3). This is followed by a presentation of objectives in Section 3. Section 4 introduces the main research methods employed. Section 5 presents the primary results and publications. Section 6 comprises the thesis' conclusions, and Section 7 concludes with a discussion and future lines of research. The second part includes publications that have been published during the thesis and that are directly linked to the achievement of the objectives.

I

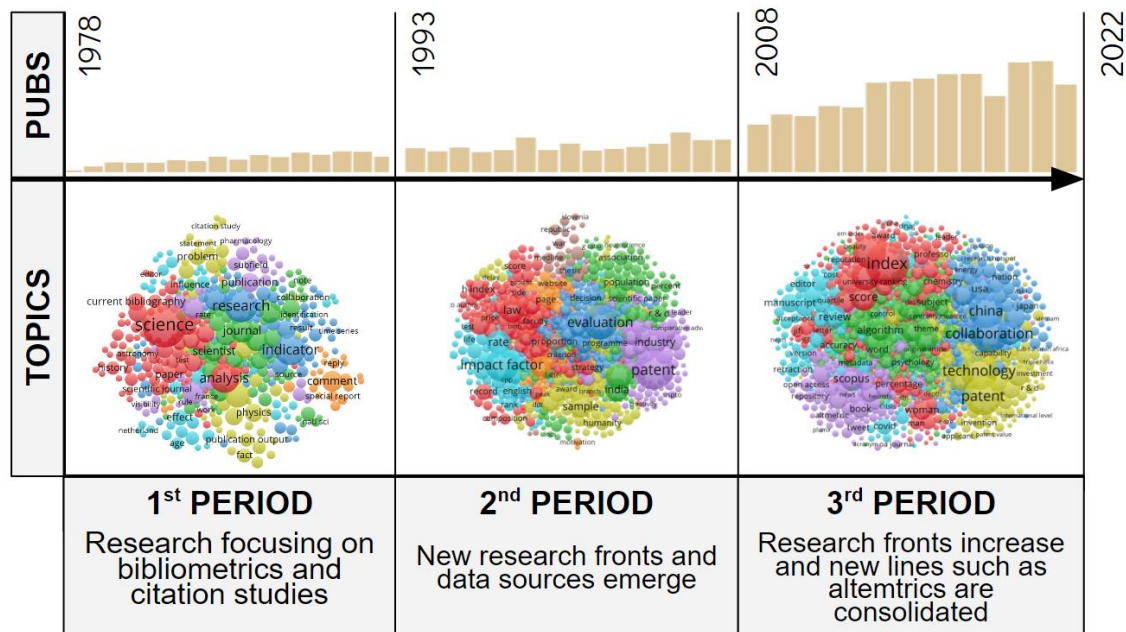
PhD dissertation

1 Introduction

The study of science is not a recent subject for research, particularly when carried out from a historical or philosophical approach. The application of quantitative techniques and the use of bibliographic records as an object of study, however, is a relatively recent development, closely linked to advances in information and communications technology which have made it possible to move from indexes in physical formats to large full-text databases and from the Z39.50 protocol to the APIs. One of the pioneers in this regard was Eugene Garfield, who conceived the Science Citation Index as a revolutionary system for information management (Garfield, 1964). This development can thus be traced back to the technological progress that emerged in the 20th century, most notably between the 1930s and 1950s—a period profoundly shaped by the Second World War. During this time, J.D. Bernal made a seminal contribution through his work *'The Social Function of Science'* (Bernal, 1939), which is now regarded as the precursor, or genesis, of what Price would later term *'Science of Science'* (Price, 1963). This is an interdisciplinary, and re-emerging (Wang & Barabási, 2021), field dedicated to the study of the roots, progression, and advancements of science through the analysis of large-scale data. However, despite the use of big data, it is not a purely quantitative field and brings together methods from disciplines as diverse as history and sociology.

This revolution in the study of science has given birth to scientometrics—a research area that studies how scientific information is created, shared, and used quantitatively, aiding our understanding of scientific research as a social activity (Braun et al., 1985). The term was initially conceived by V.V. Nalimov (Nalimov & Mul'chenko, 1969), who proposed it in 1969 as *'naukometriya'* (in Russian, *наукометрия*). However, it was later adapted by Tibor Braun, the founder of the journal *'Scientometrics'* (Garfield, 2009). Since its inception in 1978, this journal has chronicled significant advancements in the field and also functioned as a conduit between Eastern and Western research (Bensman & Kraft, 2007). Consequently, tracing the evolution of this journal offers a vivid reflection of the progress within this discipline, where technology plays an indispensable role (**Figure 1**). With the advances in information and communication technology, there has been a substantial increase in the accessibility and volume of bibliographic data. This growth has, in turn, led to a proliferation of proposals and evidence relating to academic performance, spanning from productivity and scientific impact to collaboration and mobility.

Figure 1. Evolution of scientometrics research topics through the journal *Scientometrics*

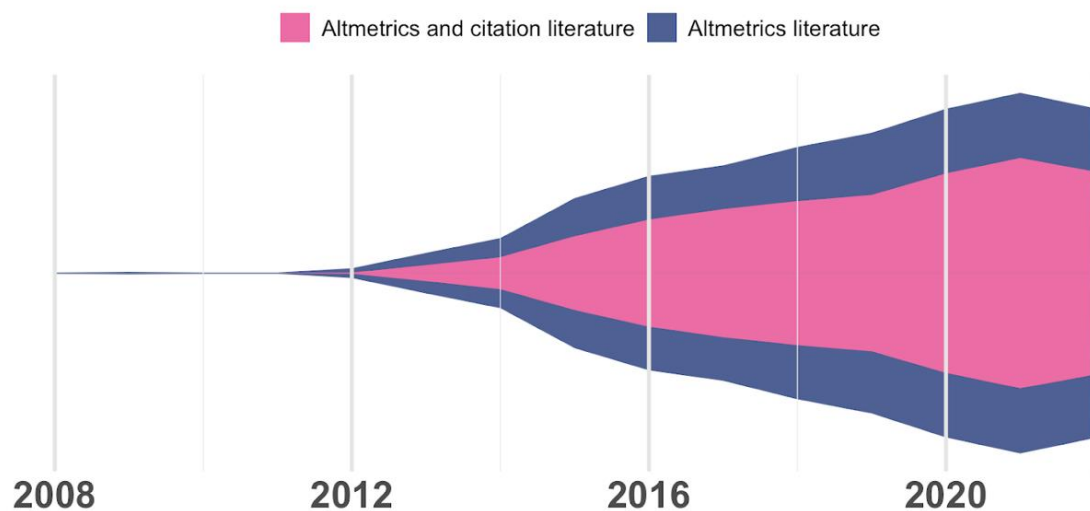


Because of this technological dependence, until well into the 1990s, *Scientometrics* was focused almost exclusively on the study of scholarly outputs, and there was only a single source of data, which is now integrated into Clarivate's Web of Science. However, the advent of the Internet and the development of the web triggered a major revolution in communication. The rise of social media has induced significant shifts in the interaction with science, both in the academic domain and in wider society. With Web 2.0, a sophisticated digital ecosystem of platforms emerged, generating a high volume of data surrounding scientific activity (Torres-Salinas, 2009). Ultimately, this led to an explosion in the diversity of data sources, giving rise to what is now known as '*Scientometrics as Big Data Science*' due to the lack of standards and the major challenge of managing this complex landscape of data (Moed et al., 2014). Consequently, a valuable opportunity presents itself to explore further this intricate interplay of scientific and social data.

In the early altmetric years around 2010, original research addressing social media activity primarily focused on exploring the various platforms through which science is disseminated or has a presence. Through these analyses, the potential usefulness of these media within a scientific context became clear, while also offering practical solutions. These include using Wikipedia references as an indicator of the quality of encyclopaedic articles (Nielsen, 2007), counting library holdings to approximate academic or educational use (Torres-Salinas & Moed, 2009), or using of bookmarks to analyse the use of journals (Haustein & Siebenlist, 2011). The landscape shifted, however, with the formal birth of altmetrics, following the publication of the Altmetric Manifesto by Priem et al. (2010). In this manifesto, not only was the phenomenon of

alternative metrics derived from social media activity surrounding science named, but the label 'impact' was also appended to it. Consequently, the first generation of altmetrics studies emerged, predominantly characterised by a research aiming to identify the relationship between academic impact, as measured through citations, and social impact, attributed to the altmetrics (**Figure 2**). In this phase, mentions from various media were treated as raw metrics, even leading to the proposition of aggregate indicators, such as the Altmetric Attention Score (Gumpenberger et al., 2016).

Figure 2. Distribution of studies that do or do not consider citations in the altmetric literature stream



However, the multitude of studies correlating (early) social media mentions to scholarly outputs with their corresponding (later) citations has primarily served to underscore the lack of relationship between these two realms (Costas et al., 2015). The only altmetric wherein this relationship is positive lies in Mendeley readerships (Thelwall, 2018). Yet, these are not only closely linked to the research itself within a purely scientific context, but they are also discouraged due to their opaque calculation process. Furthermore, it has also become evident that rather than social impact as a whole, there are different social impacts (Thelwall, 2020), or more properly, social attentions (C. Sugimoto, 2015). While the pursuit of a positive correlation between both dimensions remains active, a new generation of altmetrics recognises that their value lies within the context of these mentions and the multiple interactions that arise (Díaz-Faes et al., 2019). Consequently, the advent of this second generation altmetrics provides an ideal framework for the exploration of science-related diversity of entities and social interactions (Costas et al., 2020).

We can thus conclude that scientometrics faces a dual challenge in which data science plays a pivotal role. First, advances in information and communication technologies have

revolutionised the scientometric landscape, providing a multitude of resources for the study of science. Second, altmetric studies demand new methodological efforts and proposals to harness the as-yet unexplored possibilities they present. It is within this context that this thesis is conceived, with the aim of establishing synergy between scientometrics and data science, to provide answers to the posed challenges.

2 Literature review

2.1 Major Challenges in Collecting and Processing Social Media Data

Access to scientific information has been democratised with the advent of the web, which has broken down the barriers to consuming scientific literature and driven the open science movement. In this new context, information shifts into a liquid state, characterised by the instability of the media and its ongoing evolution (Area-Moreira & Ribeiro-Pessoa, 2012). Not only are there different electronic formats, but the scholarly output itself can have multiple versions. Bibliographic records have also undergone a revolution, not only multiplying in number and content, but also expanding in their range of possibilities, much like the databases that gather them. The bibliographic revolution is evidenced by the fact that, within a span of 13 years from 2006, we have moved from comparing two bibliographic databases (Web of Science and Scopus) and the Google Scholar search engine (Bakkalbasi et al., 2006), to 56 databases (Gusenbauer, 2022) and 12 search engines (Gusenbauer, 2019). However, this is far from constituting bibliographic anarchy; Web of Science still remains the contender to surpass. Despite the availability of numerous open alternatives with much broader coverage, it is in the standardisation and quality of its bibliographic records where no other tool can yet match it (Gusenbauer, 2022).

Numerous bibliographic databases strive to offer the broadest possible coverage, with OpenAlex standing out as particularly significant due to the potential and hopes vested in this 'massive open index' (Singh Chawla, 2022). Conversely, there are bibliographic databases that, rather than seeking extensive coverage, concentrate on providing unique functionality, often relying on machine learning-based solutions. Semantic Scholar and scite are noteworthy in this respect, conducting semantic analysis of the context in which citations occur. Access to bibliographic data is no longer the primary challenge. From a multidisciplinary perspective, there are six major bibliographic databases (**Figure 3**). However, here lies a critical distinction: while emerging solutions like Crossref and OpenAlex hold appeal with their extensive coverage and free access, they require considerable data processing and curation. Their metadata, therefore, fall short of the detail and quality provided by traditional options such as Web of Science and Scopus, which have these aspects well under control. This represents a clash between big data, with its vast but raw resources, and smart data, as embodied by traditional, well-curated databases.

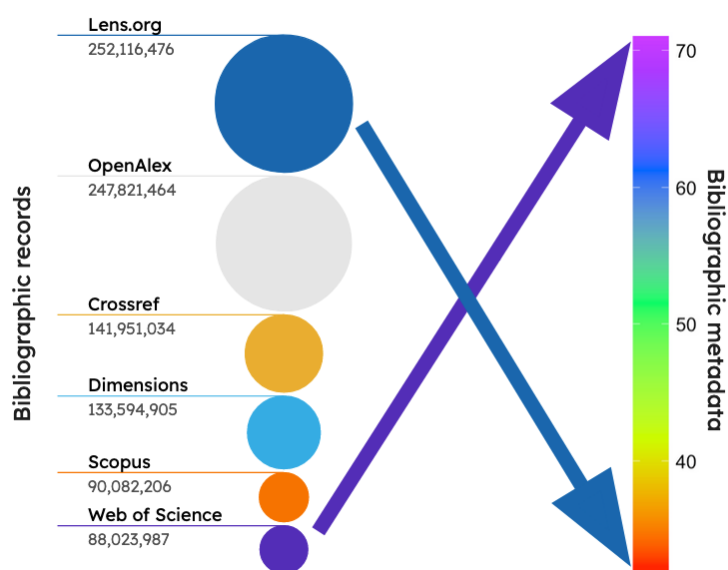


Figure 3. Difference between volume of coverage and metadata of the main multidisciplinary bibliographic databases

The volume of coverage of bibliographic databases is inversely proportional to the metadata it offers about them.

In the realm of altmetrics, data aggregators have rapidly emerged. These tools monitor and collect the mentions of scholarly outputs across the primary social media channels, all within a single interface. The most notable among these tools are Altmetric.com¹ and PlumX², both of which are commonly used for altmetric research. Crossref Event Data³ also warrants attention as an open-source alternative (**Table 1**). Notably, there are other aggregators, some of which have become obsolete, such as Lagotto⁴, which has not been updated since 2015. Furthermore, some have ceased to exist altogether, as in the case of Cobaltmetrics, which terminated its services in 2021. In terms of coverage, PlumX provides the most extensive overview of scientific literature. Altmetric.com distinguishes itself by its extensive inclusion of publications mentioned on Twitter, the main altmetric source in terms of volume of altmetric activity and research. In relation to blogs and news, this aggregator also offers superior coverage. However, disparities are evident in other metrics, particularly in references to Wikipedia and Mendeley. Interestingly, Mendeley offers better coverage on PlumX than on its own platform.

¹ <https://www.altmetric.com/>

² <https://plumanalytics.com/>

³ <https://www.crossref.org/services/event-data/>

⁴ <https://www.lagotto.io/>

Table 1. Differences in publication coverage of the main altmetrics data aggregators by social media and paper

Paper	Data aggr.	Publ.	Twitter	Wikipedia	News	Facebook	Mendeley	Blog
Bar-Ilan et al. (2019)	<i>Altmetric.com</i>	Red	Blue	White	White	White	Red	White
	<i>Mendeley</i>	Blue	White	White	White	White	Blue	White
	<i>PlumX</i>	Red	Red	White	White	White	Red	White
Karmakar et al. (2021)	<i>Altmetric.com</i>	Red	Blue	White	White	Blue	Red	Blue
	<i>PlumX</i>	Blue	Red	White	White	Red	Blue	Red
Zahedi & Costas (2018)	<i>Altmetric.com</i>	Red	Blue	Red	White	Red	Red	White
	<i>CrossRef E.D.</i>	Red	Red	Red	White	White	White	White
	<i>Lagotto</i>	Red	Red	Blue	White	Red	Red	White
	<i>Mendeley</i>	White	White	White	White	White	Red	White
	<i>PlumX</i>	Blue	Red	Red	White	Blue	Blue	White
Ortega (2018)	<i>Altmetric.com</i>	Red	Blue	Red	Blue	White	Red	Blue
	<i>CrossRef E.D.</i>	Red	Red	Red	Red	White	White	Red
	<i>PlumX</i>	Blue	Red	Blue	Red	White	Blue	Red

The colour indicates the overall results of each study, with **blue** indicating the aggregator with the highest source coverage and **red** indicating the sources with the lowest coverage.

The multiplicity of social media is one of the most positive factors considered in the study of altmetrics. This diversity is advantageous, as it enables the capture of scientific attention from various perspectives beyond the scientific domain (Adie, 2014). However, this situation simultaneously introduces new challenges in data processing and analysis (Haustein, 2016). Contrary to bibliographic databases, where all essentially collect a common set of metadata, social media contain data that bear no relation to each other and reflect entirely different phenomena. It is not feasible to compare the number of tweets that mention an article to the number of times the article is cited in a report, or the number of news items mentioning the publication. Indeed, this inconsistent comparison is a major argument against the use of aggregate metrics such as the Altmetric Attention Score (Thelwall, 2020). Therefore, while the challenges of altmetrics data are congruent with those of bibliographic data, they constitute a heightened level of difficulty.

Moreover, despite the wide range of possibilities, the bias in proposals and studies cannot be ignored. Due to its popularity in scientific communication and the consequential wealth of available data, Twitter has become a favoured source for such studies (Haustein, 2019). Wikipedia, Mendeley, news outlets, blogs, and policy reports are other frequently analysed

sources, albeit to a lesser extent. Some researchers have even proposed a meaning-dimension to these sources⁵. This bias is primarily due to the ease of data retrieval and their inclusion in altmetrics data aggregators. However, this situation leads to a neglect of other platforms that may harbour significant potential. Such is the case with Goodreads, whose rating system has been found useful in evaluating humanities books (Zuccala et al., 2015).

Considering this heightened level of complexity, it is clear that despite the facilities offered by altmetric data aggregators, they are only a first step in altmetric research. Indeed, they may even prove ineffective when examining other media that are outside the norm. It becomes essential, therefore, to explore new horizons and examine these mentions and the context in which they occur, necessitating an in-depth exploration of the social media platforms themselves. Similarly, this contextualisation also necessitates accessing and processing bibliographic data, then merging it with the aforementioned datasets. This investigation subsequently demands renewed efforts in capturing, processing, cleansing, and merging large volumes of data, to ultimately derive knowledge from them.

2.2 From Classical Horizons: Adapting Standard Scientometrics Methods in Altmetrics Research

Scientometrics is the informetric discipline that has paid most attention to altmetrics. Consequently, a logical initial step in studying social media mentions involves adapting the traditional scientometric methods to this new context. Scientometrics possesses a robust and rich methodological framework, readily transferable to altmetrics, particularly when these emerging metrics were initially equated directly with citations and scientific impact. Among the foundational scientometric methods are descriptive statistical analyses, commonly utilised to detail and explore research metrics. The use of correlations and regression models is especially prevalent for delineating similarities among them. In the context of altmetrics, the profusion of such studies has already been noted. Similarly, crucial studies concerning data sources and content coverage are frequent also in altmetrics to not only define scientific coverage on one specific social media but also to enable comparison among them (C. R. Sugimoto et al., 2017).

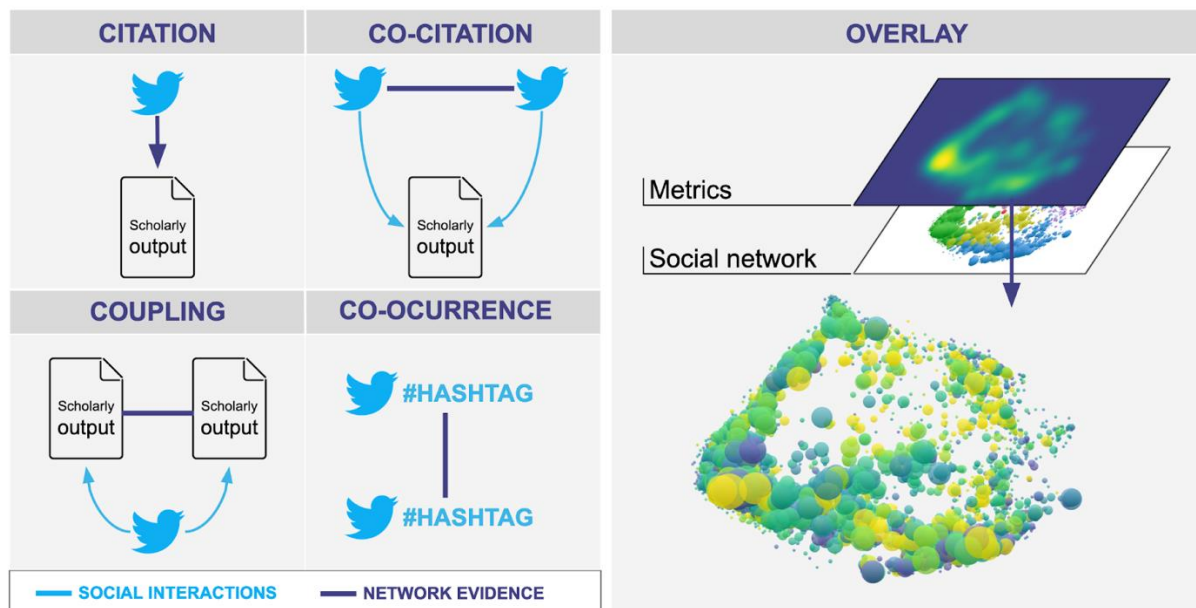
Nevertheless, one method that has generated considerable interest is social network analysis. The basic relationship established through a citation, linking two documents (the citing and the cited), forms the foundation of citation indexes (Garfield, 1955; Gross & Gross, 1927). Within multiple interpretations, this relationship has consistently been perceived as a reflection of the similarity in content between the two documents (Smith, 1958). It is upon this premise that various approaches and abstractions have been constructed to map the social and

⁵ <https://influscience.eu/metodologia/>

cognitive structures of science. Two fundamental techniques, bibliographic coupling (Kessler, 1963) and co-citation (Small, 1973), exploit this similarity relationship extensively. Whereas bibliographic coupling connects documents sharing bibliographic references, co-citation links documents that appear jointly in the bibliographies of third-party sources. Furthermore, this document-based relationship can be aggregated to other levels, such as authors, universities or other types of entities. Alongside these traditional methods, co-occurrence analysis also stands out (Callon et al., 1983). This is predominantly used with words and provides an overview of the thematic domain and has led to the development of what are known as science maps (Noyons, 2005).

The adaptation of traditional methods of social network analysis to the context of social media is straightforward (**Figure 4**). The fundamental link transitions from a citation made by one document to another, to a mention made by a social actor (such as a Twitter user, Wikipedia article, or news outlet) to a scholarly output. From this relationship, it is feasible to construct networks of co-citation and bibliographic coupling, aimed at revealing the similarity between social actors and the scholarly objects that underpin social activity (Costas et al., 2017). Likewise, the co-occurrence of words is also readily adaptable to this field, either to map the terms used in the social discussion or the research topics disseminated. Overlay maps, however, have garnered significant interest, as they facilitate the construction of base maps onto which information can be projected (Rafols et al., 2010). This allows for illumination of aspects such as social attention to a particular topic or the interests of social actors.

Figure 4. Adaptation of the main methods of social network analysis in scientometrics to altmetrics



Although social network analysis approaches have been incorporated into altmetrics research, they predominantly focus on Twitter. There is a plethora of approaches specifically designed to map communities of tweeters who engage in discussions about common topics of interest. Therefore, there is a scarcity of proposals seeking to explore other altmetric sources and provide further evidence of the structures of science in such contexts. More critically, there is a dearth of studies that not only directly transfer these traditional methodologies to the altmetrics field, but also provide a comparative analysis with the scientific domain. Such contributions can be instrumental not only in understanding how scientific knowledge is consumed outside academia but also in discerning whether the image portrayed deviates from reality.

2.3 Towards New Horizons: Exploring Original Methods and New Opportunities in Social Media

The rapid and effective adaptation of scientometric methods to social media has quickly exposed a dearth of novel approaches. This demand for new methodologies has had a transformative influence on altmetric research, forming one of the main reasons for transitioning to the second generation of altmetrics (Wouters et al., 2019). In the first generation of altmetrics, the citation model was applied to social media mentions, perfectly adapting the traditional methods of scientometrics. However, once this phase was complete, research interest pivoted to the nature and richness of interactions taking place on social media, moving beyond mere mentions of scientific publications (Díaz-Faes et al., 2019). As a result of this renewed perspective, the conventional methods of scientometrics have begun to fall short, necessitating fresh efforts to exploit the yet unexplored potential of social media.

The recognition of a lack of innovation in altmetric methods has sparked a surge of new proposals. Although some authors have given up on demonstrating that indicators can be a reflection of social impact, the research focus has shifted from measuring research dissemination through alternative metrics to understanding the patterns of scientific communication in these new channels. Notably, the conceptual framework proposed with 'heterogeneous couplings' stands out (Costas et al., 2020). It advocates for exploiting the relationships that can be found between scientific and non-scientific entities. Beyond considering the broad array of elements involved in social media and their interactions, this approach allows for the aggregation and integration of all these elements, thus enhancing evidence. However, these approaches are predominantly theoretical and require more rigorous scrutiny and implementation, not only to test them but also to generate novel findings.

A central element in approaches based on social network analysis is the detection of communities, a common aspect of scientometric studies. Various algorithms, including the

traditional Louvain algorithm (Blondel et al., 2008) and the more recent Leiden algorithm (Traag et al., 2019), are often employed for this purpose. Fundamentally, these algorithms aim to group entities that are highly interconnected and minimally intra-connected, reflecting a pattern of similarity amongst the grouped entities. In the context of altmetrics, the value of this approach is evident in detecting communities of shared interests (Schalkwyk et al., 2020) or identifying key topics of discussion (Hellsten & Leydesdorff, 2019).

Some innovative proposals have been realised and applied in case studies, aligning with this direction. As such, a variety of social network analyses have emerged, serving to illuminate the diverse relationships between science and society, as well as science communication. Some of these approaches have proven useful, for instance, by comparing research topics with those most frequently discussed by the public (Haunschild et al., 2019). Furthermore, beyond social network analysis, other studies with an exploratory focus have also been introduced. These include research exploring the various types of engagement that occur around scholarly output in social media. For example, such studies may seek to determine the extent of interaction and access to research (Fang et al., 2021, 2022).

It can be concluded that there are still many opportunities and possibilities to be explored in this respect. There is a call for innovation in the development of science maps and methods based on social network analysis. Meanwhile, the rich and volatile social media landscape needs up-to-date research to provide valuable insights to guide the community on the usefulness and potential of these platforms.

3 Objectives

The main objective of this thesis is to delve into the potential that social media harbours for studying science-society interactions, using social data mining as a tool. More precisely, we aim to confront the challenges faced by the altmetric community in terms of data extraction and processing to furnish innovative methods and novel findings. To achieve this, the following specific objectives have been established:

- 1. To determine the main difficulties involved in processing and analysing large amounts of data from bibliographic sources and social media.** Bibliographic records pose only a minor challenge in their processing as they come from highly structured and curated databases, notably Web of Science, with the biggest difficulty being the combination with social media data. However, social media such as Wikipedia are valuable sources of data, but given the volume of data and complexity of their structure, they remain to be fully explored. We therefore aim to delve into these problems in order to ultimately provide open curated datasets for altmetric study.
- 2. To adapt standard methods of scientometrics in the context of social media.** Through the specific use of techniques based on social network analysis, we aim to provide maps of science from Wikipedia. These maps serve as a reflection of social attention and the collaborative, open construction of knowledge. The use of proven and widely used techniques in scientometrics will make it possible to confront the social and academic representations of science.
- 3. To improve scientific mapping methods that combine social and semantic information.** Given the conceptual breadth of altmetrics, which encompasses the various interactions between science and society on social media, there is a need for renewed effort. Our aim is to develop and apply innovative methodologies that integrate such information and heterogeneous sources.

4 Methodology

This thesis follows the structure of the traditional scientific method, requiring an integration of both practical and theoretical methodologies throughout its development. Specifically, the following guidelines are applied to the research work and experiments:

- 1. Observation:** Through the exploration and mapping of the interaction between science and society on social media, with a particular emphasis on the treatment and processing of large data volumes.
- 2. Hypothesis formulation:** adaptation of traditional scientometric methods to fit social media environments, coupled with the design of new methods and techniques for mapping social media activity related to science. These methods are implemented within the data science framework, with big data processing forming an integral part of these techniques.
- 3. Observation gathering:** taking the results acquired from the application of proposed methods to social media, and using various performance measures derived from social network analysis as indicators for validation.
- 4. Contrasting the hypothesis:** the comparison of the acquired results with those published by other state-of-the-art related proposals.
- 5. Hypothesis validation or refusal:** validation of the hypothesis through the conducted experiments and the acquired results. If rejected, the previous steps should be repeated, thus formulating a new hypothesis to ensure the quality of the results.
- 6. Scientific thesis:** extraction, redaction and acceptance of the conclusions based on the research process, and compiling these findings along with the entire process into thesis memory and journal publications.

5 Summary

This section summarises the proposals and studies conducted in the publications associated with this thesis. Following this, Section 6 presents the principal results obtained in each research paper. The associated journal publications are listed below and organized following the objectives presented in Section 3:

OBJECTIVE 1 To study and address the difficulties involved in processing and analysing large amounts of data from bibliographic sources and social media	<p>Torres-Salinas, D., Arroyo-Machado, W., & Thelwall, M. (2021). Exploring WorldCat identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics. <i>Scientometrics</i>, 126, 1725-1743. https://doi.org/10.1007/s11192-020-03814-w</p> <p>JCR (SCIE) – Computer Science, Interdisciplinary Applications JIF Q2 (54/112)</p>
	<p>Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. <i>Quantitative Science Studies</i>, 1-22. https://doi.org/10.1162/qss.a.00226</p> <p>JCR (ESCI 2021) – Information Science & Library Science JCI Q1 (19/164)</p>
OBJECTIVE 2 To map science through the lens of social media adapting standard methods of scientometrics	<p>Torres-Salinas, D., Romero-Frías, E., & Arroyo-Machado, W. (2019). Mapping the backbone of the Humanities through the eyes of Wikipedia. <i>Journal of Informetrics</i>, 13(3), 793-803. https://doi.org/10.1016/j.joi.2019.07.002</p> <p>JCR (SCIE) – Computer Science, Interdisciplinary Applications JIF Q1 (16/109)</p>
	<p>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. <i>PLOS ONE</i>, 15(2), e0228713. https://doi.org/10.1371/journal.pone.0228713</p> <p>JCR (SCIE) – Multidisciplinary Sciences JIF Q2 (26/72)</p>
OBJECTIVE 3 To propose new methods for scientific mapping that merge social and semantic information	<p>Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. <i>FEMS Microbiology Letters</i>, 366(7). https://doi.org/10.1093/femsle/fnz075</p> <p>JCR (SCIE) – Microbiology JIF Q3 (99/136)</p>
	<p>Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. <i>Scientometrics</i>, 126(11), 9267-9289. https://doi.org/10.1007/s11192-021-04167-8</p> <p>JCR (SCIE) – Computer Science, Interdisciplinary Applications JIF Q2 (54/112)</p>

Below is a summary of the research developed in the publications, according to the thesis objectives. Firstly, Section 5.1 presents the efforts and solutions for source and data

heterogeneity from a data science perspective. Then, Section 5.2 encompasses the traditional methods of mapping science, adapted for social media. Finally, Section 5.3 introduces new solutions for mapping semantic and social relationships.

5.1 Facing Heterogeneity of Sources and Data From a Data Science Perspective

Social media data present a double novelty to the scientometric community due to their recent emergence and the traditional focus on bibliographic records as the primary object of study. Our aim is to delve into social media, providing guidance for the community and producing clean datasets for altmetric analyses. Thus, we have explored a new data source and analysed a well-known but largely unexplored media.

Firstly, we focused on WorldCat Identities, a novel tool within the global library catalogue, WorldCat. This tool provides indicators based on author profiles, aiding in the assessment of the impact and spread of academic books. We undertook a comprehensive study involving a sample of Bibliometrics and Scientometrics authors. We analysed and compared the Library Catalog Analysis indicators generated by WorldCat Identities with citations from Google Scholar and Web of Science. This allowed us to highlight the potential of this tool and its value as a source of altmetrics data. Ultimately, this process allowed us to build a Python package from which to access WorldCat Identities data and carry out related studies.

Secondly, we turned our focus towards Wikipedia, a globally visited website and frequent subject of scientific study, yet untapped in its potential for altmetric research. We compared Wikipedia page features to those of scientific publications, revealing the similarities and differences between these document types. This comparative study allowed us to uncover various analytical opportunities within Wikipedia and its diverse data sources, culminating in a methodological framework for Wikipedia analysis. Simultaneously, we constructed and shared a comprehensive Wikipedia knowledge graph dataset dedicated to the English Wikipedia. Lastly, we performed a descriptive case study on the dataset to demonstrate its capabilities.

The journal publications related to this section are:

WorldCat Identities	Torres-Salinas, D., Arroyo-Machado, W., & Thelwall, M. (2021). Exploring WorldCat identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics. <i>Scientometrics</i> , 126, 1725-1743. https://doi.org/10.1007/s11192-020-03814-w
Wikipedia knowledge graph	Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikinformatrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. <i>Quantitative Science Studies</i> , 1-22. https://doi.org/10.1162/qss_a_00226

5.2 Importing Classical Methods for Scientific Mapping of Social Media

Social network analysis is a traditional method in scientometric studies, demonstrating its versatility in adapting to social media environments. One of the most notable techniques is co-citation analysis, as it enables the linkage of scholarly outputs through social interactions. An approach that sheds light on the structure of knowledge underlying social activity.

We developed a methodological proposal to adapt the co-citation method for Wikipedia, utilising references to scholarly outputs within encyclopaedic articles. In this proposal, we also considered the possibility of aggregating these relations by the research areas of the scholarly outputs, thus constructing conceptual co-citation networks. To verify and validate this method, a case study was performed using Humanities literature referenced in the English version of Wikipedia. We subsequently applied this method to the total number of references in the English Wikipedia, conducting extensive mapping of science through this platform based on social and collaborative construction. This facilitated the highlighting of not only the method's relevance but also the provision of valuable results regarding Wikipedia's scientific perspective, and the differences between how scientific knowledge is consumed in the academic and social realms.

We also undertook a descriptive approach to Wikipedia references from a scientometrics perspective. We analysed the extent of coverage of scientific literature, topic biases, and ageing of the scientific literature. This collectively helped illuminate which scholarly aspects attract greater interest on Wikipedia, as well as the potential factors that lead to a scholarly output being referenced on Wikipedia.

The journal publications related to this section are:

Methodological framework proposal	Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. <i>FEMS Microbiology Letters</i> , 366(7). https://doi.org/10.1093/femsle/fnz075
Large-scale analysis	Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. <i>Scientometrics</i> , 126(11), 9267-9289. https://doi.org/10.1007/s11192-021-04167-8

5.3 Developing Novel Approaches for Mapping Social and Semantic Relationships

Some approaches have been used to understand thematic interests using the maps of science popularised in scientometrics, while basic approaches to social network analysis have been employed to map social relations. Both approaches are valuable for identifying communities of social actors based on their interests and social relations. However, the integration of these

insights into a single visualisation, combining both types of relationships, has not yet been explored.

We initially proposed the use of an overlay map to identify the distinct types of social attention reflected by each social media. By creating a base thematic landscape specific to all research topics within a particular area, and then projecting the attention from each media onto this, we managed to develop a new type of visualisation that combines social and semantic dimensions. We showcased this method through a case study that used scholarly outputs on Microbiology, with mentions sourced from Twitter, news outlets, and policy reports.

Moreover, we have also proposed a method that integrates social discussions and semantic interests directly. This approach commences with n-mode networks and socio-semantic networks, but we simplify the diverse network elements and their relationships. We take as our starting point a 2-mode network of social actors who are socially interconnected by discussing identical research topics on social media, and then identify the actor communities. We subsequently project the topic communities obtained from the co-occurrence network of research topics. The result is a 2-mode social network on which a 2-mode semantic network is overlaid. The overlap facilitates the identification of cognitive communities, that is, groups of actors sharing interests, even if they are not necessarily socially connected. This allows the detection of gaps between intellectual and social interaction. We verified this approach through a dual case study in Twitter involving Information Science & Library Science and Microbiology.

The journal publications related to this section are:

Overly map	Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. <i>FEMS Microbiology Letters</i> , 366(7). https://doi.org/10.1093/femsle/fnz075
Socio-semantic network	Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. <i>Scientometrics</i> , 126(11), 9267-9289. https://doi.org/10.1007/s11192-021-04167-8

6 Discussion of results

In the following sections, we delve into the primary findings and engage in detailed discussion inspired by the research conducted in this thesis.

6.1 Facing Heterogeneity of Sources and Data From a Data Science Perspective

The research conducted underscores both the challenges and the potential inherent in the thorough exploration of social data sources. Numerous untapped possibilities and platforms exist, the results of which could hold significant value for altmetric research.

The potential of WorldCat Identities for comparative author analyses based on library holdings has been proved. The case study has facilitated the identification of different publication profiles and the potential to capture scholarly interest beyond what is measured by traditional citations. However, we have also issued a caution about the need for data validation and the risks of misinterpretation due to factors such as the massive electronic offerings and the selection process of library holdings. We have additionally pointed out the potential for geographical and linguistic biases. Moreover, we have highlighted other classic methodological challenges in scientometrics, such as incorrect author disambiguation and improper assignment of works, which are prevalent in WorldCat Identities.

The findings made with Wikipedia are illuminating in the need and efforts to process data. While this digital encyclopaedia is one of the longest-standing social media platforms and one of the first to be leveraged from an altmetric perspective, many of its possibilities remain undiscovered and unexplored. We have not only proposed a robust conceptual framework, but we have also highlighted various metrics that allow us to contextualise activity and interactions with science. The page views of encyclopaedic articles serves as a proxy for social attention, discussions for detecting controversial content, or references to scholarly outputs for identifying science-related topics. This paves the way for new approaches to contextualise not only general encyclopaedic contents, but also those that have a scientific orientation or where science is featured. However, data processing in Wikipedia requires particular efforts to handle the vast volumes of data it provides, both in terms of retrieval and processing, and in terms of merging it with other sources for added value.

6.2 Importing Classical Methods for Scientific Mapping of Social Media

The research conducted has firstly served to propose a practical adaptation of co-citation to the Wikipedia environment, and secondly to apply it on a large scale.

A detailed study of the Humanities has highlighted the importance of platforms such as Wikipedia in understanding how the public consumes scientific information. The results

showed that within the Humanities, History emerged as a main research topic, gathering the highest number of citations for individual journals (531) and scientific articles (11,661), with a total of 15,969 Wikipedia citations. We also observed an increase in annual citations in the Humanities, from an average of 2,500 to 7,500 between 2013 and 2017. However, the Humanities account for only 5% of the citations made by Wikipedia.

In our comprehensive study, we have successfully conducted a large-scale co-citation analysis of all articles referenced on Wikipedia. Notably, our methodology helped us produce a holistic map, capturing the perspectives of Wikipedia editors—who are not necessarily scientists—on scientific activity. Using the Pathfinder algorithm, we managed to create a more efficient data processing system. Our findings show that the most cited research areas on Wikipedia are ‘Medicine’ (32.58% of the referenced scholarly outputs), ‘Biochemistry, Genetics and Molecular Biology’ (31.5%), and ‘Agricultural and Biological Sciences’ (14.91%). Our methodology also revealed some unique citation practices in Wikipedia, highlighting journals that are not frequently cited in other databases or platforms and that only 13.44% of Wikipedia citations are to Open Access journals.

6.3 Developing Novel Approaches for Mapping Social and Semantic Relationships

The proposal based on overlay maps introduces a novel approach to map the specific interests of social networks around Microbiology research using network analysis and mapping techniques. Our findings highlight that the majority of mentions are from Twitter, revealing separate clusters of discussion where publications spark interest either on Twitter, news outlets, or policy reports, but rarely across all platforms. Topics attracting attention on these platforms were distinct. Our analysis showed thematic differences across them, revealing a dependence on the topic for social attention. An analysis of top Twitter accounts revealed the presence of bots, questioning the utility of tweet mentions as raw counts. The advanced visualisation techniques and altmetric data sources used in our study provide valuable insight into social interest topics, shedding light on how such attention is generated and enabling better interpretation of topic and community differences. Our findings confirm the utility of our method, with potential for refinement to better understand public perceptions and use of research outputs.

In the socio-semantic proposal, we identified communities on social media based on shared scientific interests in the fields of Information Science & Library Science, and Microbiology, overcoming limitations such as data loss due to deleted or blocked accounts. Despite these challenges, our unique approach provided us with a detailed and granular insight into scientific literature consumption patterns. Particularly in Microbiology, we found numerous small groups showing interest in multiple areas of the field. By focusing on keywords rather than

publications, we minimised potential relationships derived from social ties, providing a truer representation of common research interests. This methodology, though initially applied to Twitter, holds promise for broader applications across various social media platforms and different content types.

7 Concluding remarks

This thesis presents a pioneering exploration into the profound potential of social media as an untapped reservoir to comprehend the interactions between science and society. Our central methodological tool is social data mining, a technique we believe is pivotal in navigating the multifaceted and intricate challenges associated with data extraction and processing in the realm of altmetrics.

We embark on this exploration by dissecting the complexities of handling vast volumes of data originating from both bibliographic and social media sources. Traditionally, the processing of bibliographic records from structured databases, such as Web of Science, is perceived as a straightforward task due to their well-curated nature. However, when it comes to the amalgamation of this information with social media data, we encounter substantial challenges. Our work, therefore, holds a spotlight on one such under-explored social media platform - Wikipedia. Despite its rich volume of data and complex structure, the full potential of Wikipedia as a data source for altmetrics remains largely untapped. This presents a riveting frontier for our study.

Our investigation evolves as we move towards the mapping of the scientific landscape as perceived through the lens of social media. This endeavour is not merely about employing existing scientometric methods. Instead, it seeks to adapt and repurpose these methods, thus creating comprehensive maps of science that accurately reflect the focus of social attention. These maps also serve as an illustrative testament to the collaborative, and open-ended process of knowledge construction that defines our modern digital era.

In the latter part of the thesis, we venture into uncharted territory, proposing novel methodologies to map the scientific landscape. These methods are unique in their capability to amalgamate social and semantic information. Faced with the broad conceptual scope of altmetrics and the diverse ways in which science and society interact on social media platforms, this progression is critical. The ultimate ambition here is not just the creation, but also the application of these innovative methodologies. Their deployment is geared towards the integration of a wealth of information drawn from an array of diverse and heterogeneous sources.

By accomplishing these objectives, our thesis strives to provide a fresh, nuanced perspective on the dynamic relationship between science and society. It serves as a testament to the transformative power of social data mining and its ability to shed light on the ways in which the digital age shapes and drives scientific discourse.

References

- Adie, E. (2014). Taking the alternative mainstream. *El Profesional de La Informacion*, 23(4), 349–351. <https://doi.org/10.3145/epi.2014.jul.01>
- Area-Moreira, M., & Ribeiro-Pessoa, M. T. (2012). From Solid to Liquid: New Literacies to the Cultural Changes of Web 2.0. *Comunicar*, 19(38), 13–20. <https://doi.org/10.3916/C38-2012-02-01>
- Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3(1), 7. <https://doi.org/10.1186/1742-5581-3-7>
- Bar-Ilan, J., Halevi, G., & Milojević, S. (2019). Differences between Altmetric Data Sources – A Case Study. *Journal of Altmetrics*, 2(1), 1. <https://doi.org/10.29024/joa.4>
- Bensman, S., & Kraft, D. (2007). Tibor Braun, the journal *Scientometrics* and the international development of a new discipline. *The Multidimensional World of Tibor Braun: A Multidisciplinary Encomium for His 75th Birthday: The International Society for Scientometrics and Informetrics E-Zine Newsletter*, 3.
- Bernal, J. D. (1939). The social function of science. *The Social Function of Science*.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Braun, T., Glänzel, W., & Schubert, András. (1985). *Scientometric indicators: A 32-country comparative evaluation of publishing performance and citation impact*. World Scientific.
- Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235. <https://doi.org/10.1177/053901883022002003>
- Costas, R., de Rijcke, S., & Marres, N. (2017). Beyond the dependencies of altmetrics: Conceptualizing ‘heterogeneous couplings’ between social media and science. *The 2017 Altmetrics Workshop*. http://altmetrics.org/wp-content/uploads/2017/09/altmetrics17_paper_4.pdf
- 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>
- Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019. <https://doi.org/10.1002/asi.23309>

- 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>
- Fang, Z., Costas, R., Tian, W., Wang, X., & Wouters, P. (2021). How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *Journal of the Association for Information Science and Technology*, *72*(7), 918–932. <https://doi.org/10.1002/asi.24458>
- Fang, Z., Costas, R., & Wouters, P. (2022). User engagement with scholarly tweets of scientific papers: A large-scale and cross-disciplinary analysis. *Scientometrics*, *127*(8), 4523–4546. <https://doi.org/10.1007/s11192-022-04468-6>
- Garfield, E. (1955). Citation Indexes for Science. *Science*, *122*(3159), 108–111. <https://doi.org/10.1126/science.122.3159.108>
- Garfield, E. (1964). Towards the world brain. *Current Contents*, *6*, 8–9.
- Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics*, *3*(3), 173–179. <https://doi.org/10.1016/j.joi.2009.03.009>
- Gross, P. L. K., & Gross, E. M. (1927). College Libraries and Chemical Education. *Science*, *66*(1713), 385–389. <https://doi.org/10.1126/science.66.1713.385>
- Gumpenberger, C., Glänzel, W., & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, *108*(2), 977–982. <https://doi.org/10.1007/s11192-016-1991-5>
- Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*(1), 177–214. <https://doi.org/10.1007/s11192-018-2958-5>
- Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, *127*(5), 2683–2745. <https://doi.org/10.1007/s11192-022-04289-7>
- Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics*, *13*(2), 695–707. <https://doi.org/10.1016/j.joi.2019.03.008>
- 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>
- Haustein, S. (2019). Scholarly Twitter Metrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 729–760). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_28

- Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3), 446–457. <https://doi.org/10.1016/j.joi.2011.04.002>
- Hellsten, I., & Leydesdorff, L. (2019). Automated analysis of actor–topic networks on twitter: New approaches to the analysis of socio-semantic networks. *Journal of the Association for Information Science and Technology*, 71(1), 3–15. <https://doi.org/10.1002/asi.24207>
- Karmakar, M., Banshal, S. K., & Singh, V. K. (2021). A large-scale comparison of coverage and mentions captured by the two altmetric aggregators: Altmetric.com and PlumX. *Scientometrics*, 126(5), 4465–4489. <https://doi.org/10.1007/s11192-021-03941-y>
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. <https://doi.org/10.1002/asi.5090140103>
- Moed, H. F., Luwel, M., & Daraio, C. (2014, March 25). *Scientometrics as Big Data Science: On integration of data sources and the problem of different types of classification systems*. OECD Workshop, Paris, Paris. <https://www.oecd.org/sti/inno/4.2.%20Henk%20Moed.pdf>
- Nalimov, V. V., & Mul'chenko, Z. M. (1969). *Наукометрия, Изучение развития науки как информационного процесса [Naukometriya, the study of the development of science as an information process] (in Russian)*. Nauka.
- Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*, 12(8). <https://doi.org/10.5210/fm.v12i8.1997>
- Noyons, C. M. (2005). Science Maps Within a Science Policy Context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (pp. 237–255). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_11
- 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>
- Price, D. J. D. S. (1963). *Little Science, Big Science*. Columbia University Press. <https://doi.org/10.7312/pric91844>
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Altmetrics. <http://altmetrics.org/manifesto/>
- Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 1871–1887. <https://doi.org/10.1002/asi.21368>
- Schalkwyk, F., Dudek, J., & Costas, R. (2020). Communities of shared interests and cognitive bridges: The case of the anti-vaccination movement on Twitter. *Scientometrics*, 152(2), 1499–1516. <https://doi.org/10.1007/s11192-020-03551-0>

- Singh Chawla, D. (2022). Massive open index of scholarly papers launches. *Nature*, d41586-022-00138-y. <https://doi.org/10.1038/d41586-022-00138-y>
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Smith, M. (1958). The trend toward multiple authorship in psychology. *American Psychologist*, 13, 596–599. <https://doi.org/10.1037/h0040487>
- Sugimoto, C. (2015). ‘Attention is not Impact’ and Other Challenges for Altmetrics [Wiley]. *The Wiley Network*. <https://www.wiley.com/en-us/network/publishing/research-publishing/promoting-your-article/attention-is-not-impact-and-other-challenges-for-altmetrics>
- 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>
- Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, 115(3), 1231–1240. <https://doi.org/10.1007/s11192-018-2715-9>
- Thelwall, M. (2020). Measuring societal impacts of research with altmetrics? Common problems and mistakes. *Journal of Economic Surveys*, 35(5). <https://doi.org/10.1111/joes.12381>
- Torres-Salinas, D. (2009). *Indicadores 2.0 para la ciencia 2.0*. IX Workshop REBIUN Proyectos Digitales, Salamanca. REBIUN. https://repositoriorebiun.org/bitstream/handle/20.500.11967/770/ws_2009_torressalinas_P.pdf?sequence=1
- Torres-Salinas, D., & Moed, H. F. (2009). Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics*, 3(1), 9–26. <https://doi.org/10.1016/j.joi.2008.10.002>
- Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), Article 1. <https://doi.org/10.1038/s41598-019-41695-z>
- Wang, D., & Barabási, A.-L. (2021). *The Science of Science*. Cambridge University Press. <https://doi.org/10.1017/9781108610834>
- Wouters, P., Zahedi, Z., & Costas, R. (2019). Social Media Metrics for New Research Evaluation. In W. Glänzel, M. Henk F, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 687–713). Springer International Publishing.
- 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>

Zuccala, A. A., Verleysen, F. T., Cornacchia, R., & Engels, T. C. E. (2015). Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Journal of Information Management*, 67(3), 320–336. <https://doi.org/10.1108/AJIM-11-2014-0152>

II

Publications

Exploring WorldCat Identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics



Daniel Torres-Salinas¹, Wenceslao Arroyo-Machado^{1,*} and Mike Thelwall²


¹Department of Information and Communication Sciences, University of Granada, Faculty of Communication and Documentation, Granada, Spain

²Statistical Cybermetrics Research Group (SCRG), University of Wolverhampton, Wulfruna Street, Wolverhampton, WV1 1LY, UK

*Corresponding author: wences@ugr.es

Journal

Scientometrics

0138-9130 

Index

SCIE – Q2

DOI

10.1007/s11192-020-03814-w

Data

https://github.com/Wences91/library_catalog_wi/

Version

Published

References

APA 7th

Funding

Influ Science

Abstract

Assessing the impact of scholarly books is a difficult research evaluation problem. Library Catalog Analysis facilitates the quantitative study, at different levels, of the impact and diffusion of academic books based on data about their availability in libraries. The WorldCat global catalog collates data on library holdings, offering a range of tools including the novel WorldCat Identities. This is based on author profiles and provides indicators relating to the availability of their books in library catalogs. Here, we investigate this new tool to identify its strengths and weaknesses based on a sample of Bibliometrics and Scientometrics authors. We review the problems that this entails and compare Library Catalog Analysis indicators with Google Scholar and Web of Science citations. The results show that WorldCat Identities can be a useful tool for book impact assessment but the value of its data is undermined by the provision of massive collections of ebooks to academic libraries.

Citation

Torres-Salinas, D., Arroyo-Machado, W., & Thelwall, M. (2021). Exploring WorldCat identities as an altmetric information source: A library catalog analysis experiment in the field of Scientometrics. *Scientometrics*, 126, 1725–1743. <https://doi.org/10.1007/s11192-020-03814-w>

1. Introduction

The importance of books and monographs in scientific communication has been recognized for a long time (Archambault, Vignola-Gagné, Côté, Larivière & Gingrasb, 2006; Hicks, 1999; Huang & Chang, 2008). Early bibliometric impact evaluations of books were restricted to the limited data available in citation indexes. A key problem was that citation analysis databases primarily indexed journal articles, with limited coverage of books. This was addressed by new indicators based on library catalogs that became universally accessible through the Z39.50 protocol for internet-based search/retrieval and the launch of WorldCat.org, which used Z39.50 to collate library holdings from all over the world. The WorldCat.org open-access catalog unified millions of libraries in a single search engine enabling users to count how many libraries contained any given book, creating an alternative type of impact evidence (Nilges, 2006).

The library count method was termed Library Catalog Analysis (LCA) (Torres-Salinas & Moed, 2008) or Library Holdings Analysis (Linmans, 2008) and challenged the use of traditional citations two years before the publication of the Altmetric manifesto suggested that social media mentions could be used to track the societal impacts of academic publications (Priem, Taraborelli, Groth & Neylon, 2010). Since then, many other methods of analyzing book diffusion have appeared. These include the number of reviews recorded by the Book Review Index or the number and score available on the Goodreads or Amazon Reviews websites—the latter being related to popularity (Kousha & Thelwall, 2016). Similarly, in recent years, mentions in course syllabi, the Mendeley social reference sharing site, or YouTube comments (Kousha & Thelwall, 2015) have also been used.

In 2009, LCA was defined as “the application of bibliometric techniques to a set of library online catalogs” and in a case study, Torres-Salinas and Moed (2008, 2009) selected the field of Economics and analyzed 121 147 titles included on 417 033 occasions in 42 libraries. Similarly, White et al. (2009) proposed an identical method for what they termed “libcitations” that focused on the micro-level: 148 authors from different departments (Philosophy, History and Political Science) at two Australian universities (New South Wales, Sydney). They used WorldCat as their source of information. These studies showed that LCA was practical and gave plausible results.

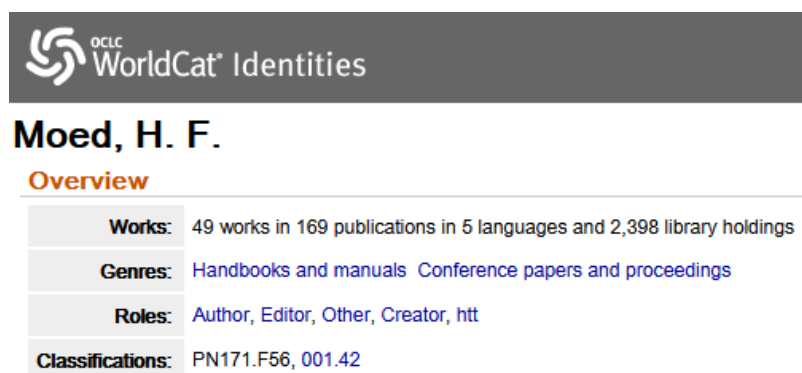
Libcitations offer an alternative vision of the impact of books, with correlations with citations usually being low (Linmans, 2010; Zuccala & Guns, 2013; Zhang, Zhou & Zhang, 2018). Linmans’ data set generated a correlation of 0.29, rising to 0.49 for English language books. Zuccala obtained correlations of 0.24 and 0.20 for History and Literature, respectively. Kousha and Thelwall (2016) compared libcitation correlations with various Amazon indicators. Their

strongest correlation with Amazon Reviews was 0.348 for the humanities. Torres-Salinas, Gumpenberger and Gorraiz (2017) also reported a low correlation between library counts and other altmetrics included in PlumX Analytics.

WorldCat dominates library catalog studies (e.g., Linmans, 2010; Zuccala & White, 2015; Neville & Henry, 2014; Halevi, Nicolas, & Bar-Ilan, 2016) partly because of its uniquely large size. The OCLC directory currently identifies 15 194 libraries, 5804 of which are academic⁶. PlumX, currently owned by Elsevier, includes WorldCat library holdings (Holdings: WorldCat) among its altmetric indicators and facilitates both their calculation through ISBNs and large-scale searches (Torres-Salinas, Robinson-Garcia, Gorraiz, 2017). One of the first studies to use this source was Halevi, Nicolas and Bar-Ilan (2016), which used 71 443 eBook ISBN numbers.

WorldCat offers a wide range of functions in addition to its core index, such as the experimental WorldCat Identities (WI). This tool brings together the “complete works” of any given author, reporting the library diffusion data of their works overall for the author and by individual publication. It also integrates context-based data (e.g., genre, topics, name variants, co-authors) (Fig. 1). WI therefore provides Library Catalog Metrics profiles similar to that of other academic sites (Google Scholar Profiles, ResearchGate or custom current research information systems (CRIS) at research institutions), except WI is based mainly on books and bibliographic records indexed in library catalogs. These profiles open an interesting methodological door because library holdings have previously been studied at the level of record or work rather than aggregated by author.

Fig. 1 Basic information offered by WorldCat Identities for a given author



Despite the potential usefulness of WI, no study of its use has been undertaken from a metric perspective. Consequently, there is a need to test the value of WI as a source of information to

⁶ Information drawn from the Directory of OCLC Members: <https://www.oclc.org/en/contacts/libraries.html>. Note that some OCLC sources put the number of member libraries at 17 983: <https://www.oclc.org/en/about.html>

obtain indicators based on library catalog for authors. Furthermore, WI includes information drawn from other sources and uses unspecified methods—presumably computational—to match works to authors⁷. Although other studies have used library holdings at the author level to a limited extent (White & Zuccala, 2018), no prior study has analyzed the authors in a specific scientific specialty. Here we address this gap for the field of Scientometrics and its most frequently cited authors as a sample, with the following objectives:

- 1) To analyze the strengths and weaknesses of WI as a source of information about the presence of an author's works in library catalogs.
- 2) To assess the value of LCA for describing a scientific field by analyzing WI indicators for the library holdings of scientometricians.
 - 1) To compare citation indicators and indicators based on library holdings at the author level.
 - 2) To identify the most widely distributed library works in a given scientific field from WI author profiles.

This paper is an extension of a book chapter analysing Informetrics authors (Torres-Salinas & Arroyo-Machado, 2020). It is organized as follows: first we describe the methodological process of identifying authors and gathering WI data. Secondly, we present results on the number of library holdings of Informetrics researchers (author level analysis) and compare these with Google Scholar citations. We also apply LCA to determine which books it identifies as being the most important (book level analysis). Finally, we discuss our results, acknowledging their limitations, and stressing the potential use of WI in the analysis of any specific field.

2. Methodology

2.1. Selecting our sample of researchers

To apply WI to a set of researchers who specialize in Scientometrics, we selected a sample of authors included in the “Scholar Mirrors” portal⁸—which gathers profiles of researchers in Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics on platforms like Google Scholar or ResearchGate. The “Scholar Mirrors” was created in 2015 by the EC3 Group at the University of Granada (Spain) and has identified a total of 813 Google profiles of which 398 are classified as core authors⁹. We selected these 398 authors for our study to represent a large sample of active authors in the broad field of Scientometrics.

⁷ <https://www.oclc.org/research/areas/data-science/identities.html>

⁸Scholar Mirrors: [2020-02-11]: <http://www.scholar-mirrors.infoec3.es/>

⁹Scholar Mirrors: Methodology [11/02/2020]: <http://www.scholar-mirrors.infoec3.es/layout.php?id=methodology>

2.2. WorldCat Identities

The WI author profile provides information about an author's bibliographic production. It includes the following four indicators (Fig. 1):

- Works: Number of different works indexed in WorldCat.
- Publications: Total number of works indexed in WorldCat, including separate book editions.
- Languages: Number of different languages in which an author's works, including different editions, have been published.
- Library holdings: Number of different WorldCat member libraries that hold the author's works.

Table 1 shows how WI calculates these four indicators for authors. The example concerns the hypothetical profile of an author who has published a total of 3 works. The different versions of these works (editions, translations, reprints, etc.) have generated five publications that are indexed in 159 library catalogs associated with WorldCat. This example, as other authors have pointed out (Zuccala et. al, 2018), indicates that we are not counting books in a physical sense (i.e. items with a unique ISBN) but are measuring intellectual contributions. These contributions are the sum of different types/versions and formats of works, as exemplified in Table 1. In the present paper, we have used these indicators together with each author's citation record taken from Google Scholar profiles and the Web of Science (WoS) Core Collection using the beta Author Search.

Table 1 Example of the indicators computed for a fictitious author with five books indexed in WorldCat

Work	Publication	Language	Holdings
Work 1	1 ed – country a	Spanish	10
	2 ed - country a	Spanish	2
	1 ed - country b	English	12
Work 2	1 ed - country a	Spanish	13
Work 3	1 ed - country b	English	122
3 works	5 publications	2 languages	159 holdings

2.3. Retrieving Information from WorldCat Identities

WorldCat Identity information can be retrieved directly from the interface or from the API (Applications Programming Interface). First, we used the WI API¹⁰ to automatically search for each author by name, select the most relevant personal identity result, and retrieve all profile information. However, WI has a disambiguation issue that meant we needed to process this search manually in order to check whether the data was correct and then add records from duplicate identities, recording the URLs of all the author's records. Once we had reviewed and

¹⁰ WorldCat Identities API: <https://pypi.org/project/worldcatidentities/>

corrected this, we used the API to automatically obtain each author's data on 26 November 2019. The data retrieved were: I) basic information and WorldCat indicators; II) the author's 20 most widely held WorldCat works and related indicators (WI does not provide access to further works) and III) the works' language distributions. Data were then combined by author and all previously-gathered information added. We also identified each author's professional role, status, affiliation, and Google Scholar citation record on 16 December 2019 - the EC3 Scholar Mirrors portal includes this information but it has not been updated. The Python software used to analyze the data is available in a GitHub repository¹¹.

2.4. Verification process

A library count indicator is difficult to create because complete lists of holdings are needed in order to identify duplicates. The holdings indicator includes duplicate libraries when the same library holds multiple publications and, for any given author, books they have authored or co-authored, edited or contributed to (e.g., as a chapter author). It also includes works dedicated to the author or about the author, including *festschriften* in their honor, and theses they have supervised. The results can include non-book sources, such as newspaper interviews with the author or historical letters sent to the author. Profiles sometimes contain mistakes, such as works written by other authors, but the results for the present sample seemed to be largely correct. To check the validity of the profiles, we verified the results for the authors listed in Tables 2 and 3.

For the checks, through the API, we downloaded the works by the author with the highest number of holdings returned by the API and checked them manually. The API returns a maximum of 20 works per profile, but as some authors have more than one record in some cases the works recovered is higher than that limit. A total of 1125 works were checked, 98 (9%) of which were not correctly assigned. At the author level, 20 profiles included at least one wrongly assigned work. However, four authors had more than 10 incorrectly assigned works (Aparna Basu, Paul Wouters, Henry Small and José Luis Ortega). Details of these errors and how they affected the results are in Table 5 and in electronic supplementary material 1. The aforementioned 98 works appear in 1751 library holdings but the authors' total number of holdings is 103 796 (Tables 2 and 3). In other words, at profile level, the final count of holdings has a 1.65% error rate, although this is much higher for some authors.

¹¹ https://github.com/Wences91/library_catalog_wi/

3. Results

3.1. Author level analysis

Of the 398 authors drawn from EC3's Scholar Mirrors, 129 researchers were not in WorldCat Identities. These authors were not present for three reasons: a) The authors have not published any book, b) the authors do not have works indexed in any library catalogue (For example, they produce grey literature), c) Library catalogues where books are catalogued are not part of WorldCat. Therefore 269 were in WI and 461 records were recovered, including duplicates. In total, 113 authors had more than one record and 156 had only one. We also excluded four authors we considered not to be closely involved in Scientometrics, giving a final sample of 265 authors distributed in 456 records, which we subsequently classified. In July 2020, 232 of this sample were active, 12 emeritus, 11 had died, and 10 retired. Furthermore, 150 were professors, 70 researchers, 42 librarians, and 3 were professionals in the field. A total of 141 105 mentions have been collected from WorldCat. Electronic supplementary material 2 shows the complete list of authors and indicators used in this study.

Table 2 Top 25 historical (i.e. emeritus, retired or deceased) authors according to library holdings and WI information

Author	Main Affiliation	Library Holdings	Works	Holdings / Work	Publications
Blaise Cronin	Indiana University	6785	144	47.12	586
Derek J. de Solla Price	Yale University	6775	179	37.85	484
Jose Maria López Piñero	CSIC	5551	750	7.40	1836
Eugene Garfield	Institute for Scientific Information	3399	148	22.97	448
Péter Jacsó	University of Hawaii	3362	25	134	77
Michael E. D. Koenig	Long Island University	2871	34	84.44	136
Tibor Braun	Loránd Eötvös University	2600	163	15.95	452
Alan Pritchard	National Computing Centre (UK)	2515	78	32.24	199
Vasily V. Nalimov	----	2441	83	29.41	262
Henk F. Moed	Leiden University	2398	49	48.94	169
Michael J. Moravcsik	University of California	2225	59	37.71	192
Peter Ingwersen	University of Copenhagen	1862	113	16.48	270
Howard D. White	Drexel University	1821	18	101.17	63
Ronald Rousseau	KU Leuven	1381	23	60.04	119
Yves-Francois Le Coadic	Cnam – Paris	1302	30	43.40	91
Loet Leydesdorff	University of Amsterdam	1234	65	18.98	192
Sven Hemlin	University of Gothenburg	1124	34	33.06	99
Bertram C. Brookes	University College London	963	47	20.49	203
Samuel C Bradford	Science Museum London	721	48	15.02	140
Anthony F.J. van Raan	Leiden University	439	39	11.26	76
Francis Narin	CHI Research	422	46	9.17	96
Belver C. Griffith	Drexel University	352	24	14.67	54
András Schubert	Hungarian Academy of Sciences	349	21	16.62	62
Aparna Basu*	NISTADS	297	10	29.70	68

*This author's record includes many incorrectly assigned works.

The 265 authors found have an average of 22.4 works and 52.0 publications, in an average of 1.94 languages. The average number of library holdings is 532. In the authors' Google Scholar profiles, the average total number of citations is 3186, of which they received 1651 between 2014 and 2019. There are substantial differences between authors, with the 25 authors with the most works accounting for 49.8% of the total. Although historical authors account for only 12.5% of our sample, they produced 39.5% of the works. To introduce our results, we have divided the authors into two subsets: the historical figures of Scientometrics (Table 2) and currently active figures (Table 3).

Table 3 Top 25 active authors according to library holdings and WI information

Author	Main Affiliation	Library Holdings	Works	Holdings / Work	Publications
Caroline S. Wagner	Ohio State University	7157	32	223.66	147
Chaomei Chen	Drexel University	5879	42	139.98	243
Katy Börner	Indiana University Bloomington	5077	46	110.37	163
Paul Wouters*	Leiden University	3582	60	59.70	123
Nick Tomaiuolo	Connecticut State University	3186	5	637.20	36
Ben R Martin	Prof Ben Martin	3014	60	50.23	192
Peter Van den Besselaar	Vrije Universiteit Amsterdam	2920	52	56.15	179
Lokman Meho	American University of Beirut	2420	11	220.00	46
Cassidy R. Sugimoto	Indiana University Bloomington	2301	11	209.18	88
Andrea Scharnhorst	The DANS KOS Observatory	2234	15	148.93	67
Ian Rowlands	University of Waterloo	2090	23	90.87	101
Fiorenzo Franceschini	Politecnico di Torino	2042	28	72.93	98
Radhamany Sooryamoorthy	University of KwaZulu-Natal	1876	27	69.48	85
Bart Van Looy	KU Leuven	1810	90	20.11	175
Koenraad Debackere	KU Leuven	1637	105	15.59	175
Kim Holmberg	University of Turku	1617	12	134.75	46
Ying Ding	Indiana University Bloomington	1374	14	98.14	83
HD Daniel	ETH Zurich	1351	36	37.53	92
Stefanie Haustein	University of Ottawa	1317	11	119.73	32
Javier Ruiz-Castillo	Universidad Carlos III	1186	171	6.94	301
Wolfgang Glänzel	KU Leuven	1111	54	20.57	114
José Luis Ortega*	CSIC	1095	23	47.61	50
Svein Kyvik	Nordic Institute for Studies in Innovation	1063	65	16.35	140
Mustar Philippe	Ecole des Mines de Paris	1057	39	27.10	112
Mike Thelwall	University of Wolverhampton	998	34	29.35	114

*This author's record includes many incorrectly assigned works.

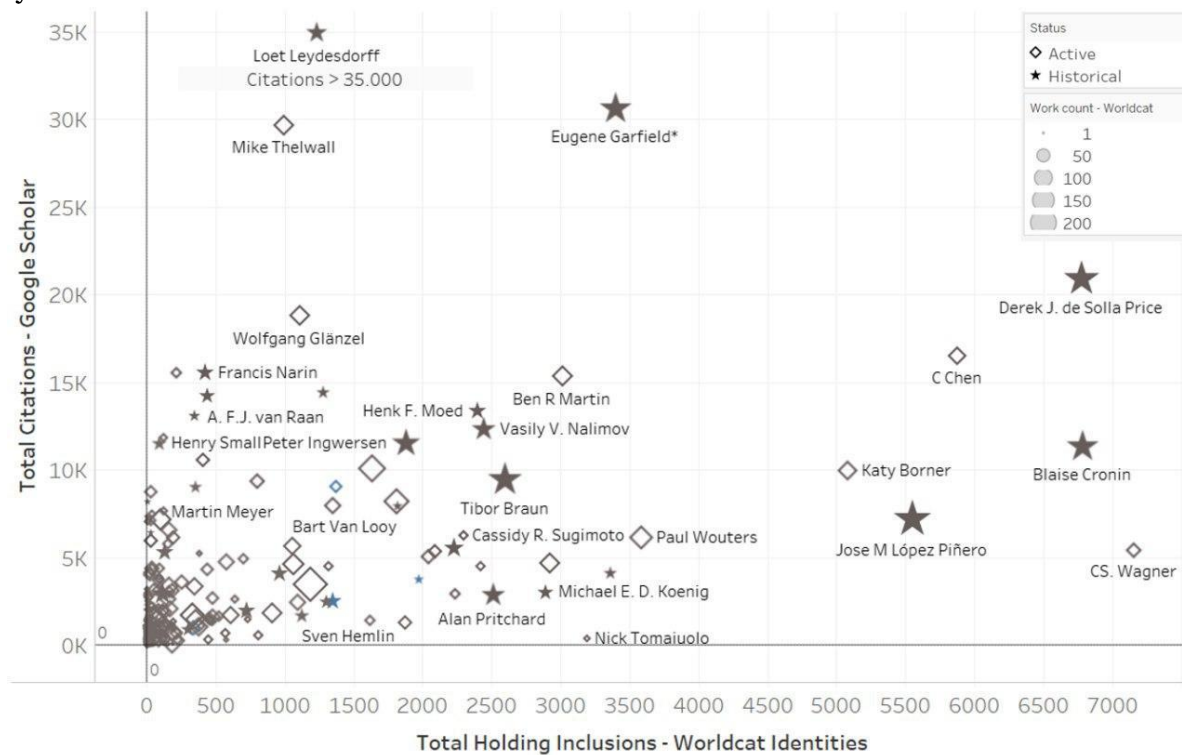
The historical researcher with the highest number of library holdings (6785) is Blaise Cronin, from the University of Indiana and a former editor of JASIST. He is followed by Derek J de Solla Price—one of the fathers of Bibliometrics—and José María López-Piñero—who introduced Bibliometrics into Spain—with 6775 and 5562 holdings, respectively. There seem to be no notable authors absent from the table. It includes early contributors like Alan Pritchard (2515)—one of the first to define Bibliometrics in 1969—and the generation of the 1950s and 1960s with Eugene Garfield (3339) or Nalimov (2441)—the father of Science in the

Soviet Union and author of *Naukometria*. The list also includes the more recent generation which definitively consolidated the field, with figures like Tibor Braun (2706)—who founded *Scientometrics* in 1978—Henk F. Moed (2398)—one of the first members of the Centre for Science and Technology Studies (CWTS) at Leiden University—and Loet Leydesdorff (1234). It excludes figures that were influential in *Scientometrics* but were not primarily scientometricians, such as Robert Merton.

Table 3 shows those researchers who currently remain active and have not retired. It includes the researcher with the largest number of library holdings in the sample: Caroline S Wagner from Ohio State University with 7157 holdings. There are also two researchers with over five thousand library holdings: Chaomei Chen from Drexel University (5879) and Katy Börner from Indiana University at Bloomington (5077). The list also includes two librarians: Nick Tomaioulo at Connecticut State University (3186)—who has published just five books mainly related to library management—and Lockman Meho (2421) from Lebanon. At the university level, KU Leuven has the most researchers on the list (Van Looy, Debackere and Glänzel). If we compare Tables 2 and 3, we see that at author level library holdings favor current researchers as almost all of them have at least 1000 library holdings. However, we have only classified 38 as historical, so their averages are higher. Some 59% of active authors have fewer than 50 library holdings and 30% have three or fewer.

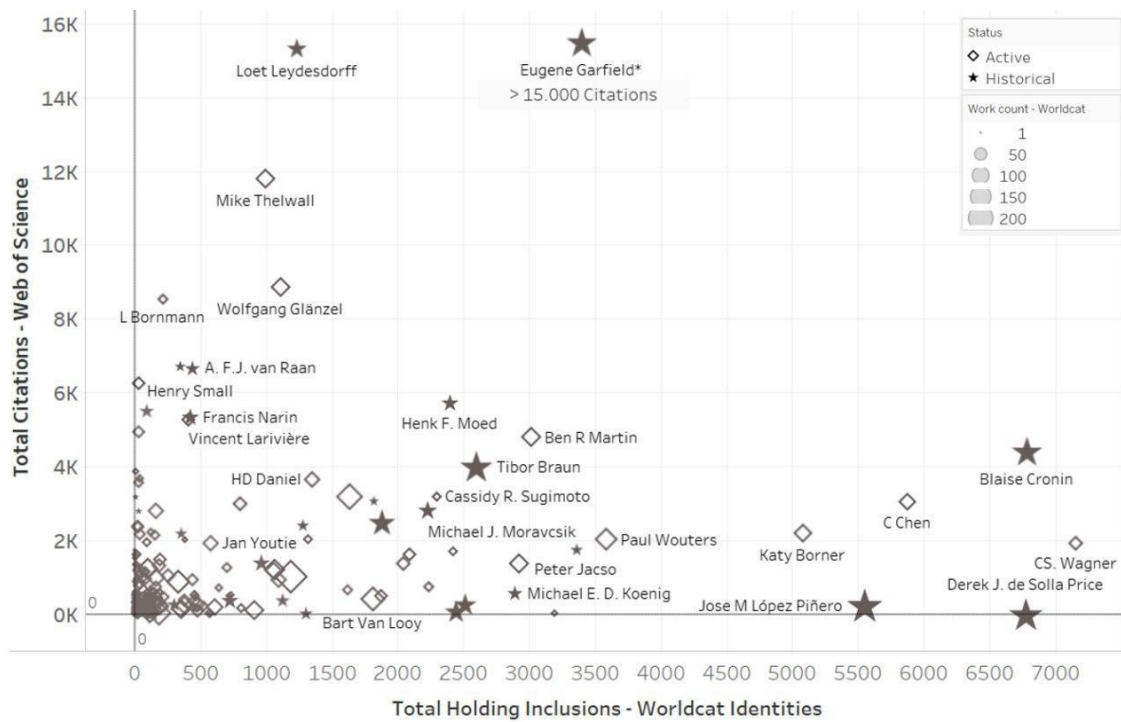
To complement this analysis we compared the library holdings data with total citations in Google Scholar profiles, —we used the Spearman correlation rather than the Pearson correlation—and found a weak positive correlation with the number of citations (0.49). A group of authors with higher values for both indicators stands out, but we also find highly cited authors (for example Leydesdorff, Thelwall or Glänzel) with relatively few library holdings for their citations. These authors may focus more on publishing scientific articles and write comparatively few books, hence their lower library holdings. This suggests that these indicators may help to distinguish between highly visible authors with significant book authorship and those who are highly visible overall.

Fig. 2 Library holding and Google Scholar Citations for main Bibliometrics authors classified by status



Library holdings were also compared with indicators calculated from the WoS Core Collection. If we compare WoS citations with library holdings, the image we obtain is similar to that in Fig. 2, although the Spearman correlation of 0.22 is weaker (Fig. 3). Google Scholar indexes books, potentially bringing its citation indicator closer to the number for library holdings than that for WoS citations. Fig. 3 shows that some of the most frequently cited researchers today (e.g., Bormann, Lariviere) have no impact on library holdings. Again, we have clear evidence of a researcher profile that is oriented towards journals and ignores books as a channel of communication. The two comparative figures therefore reveal different author profiles and demonstrate the value of using multiple indicators.

Fig. 3 Library Holding and Web of Science Core Collection citations for the main Bibliometrics authors classified by status



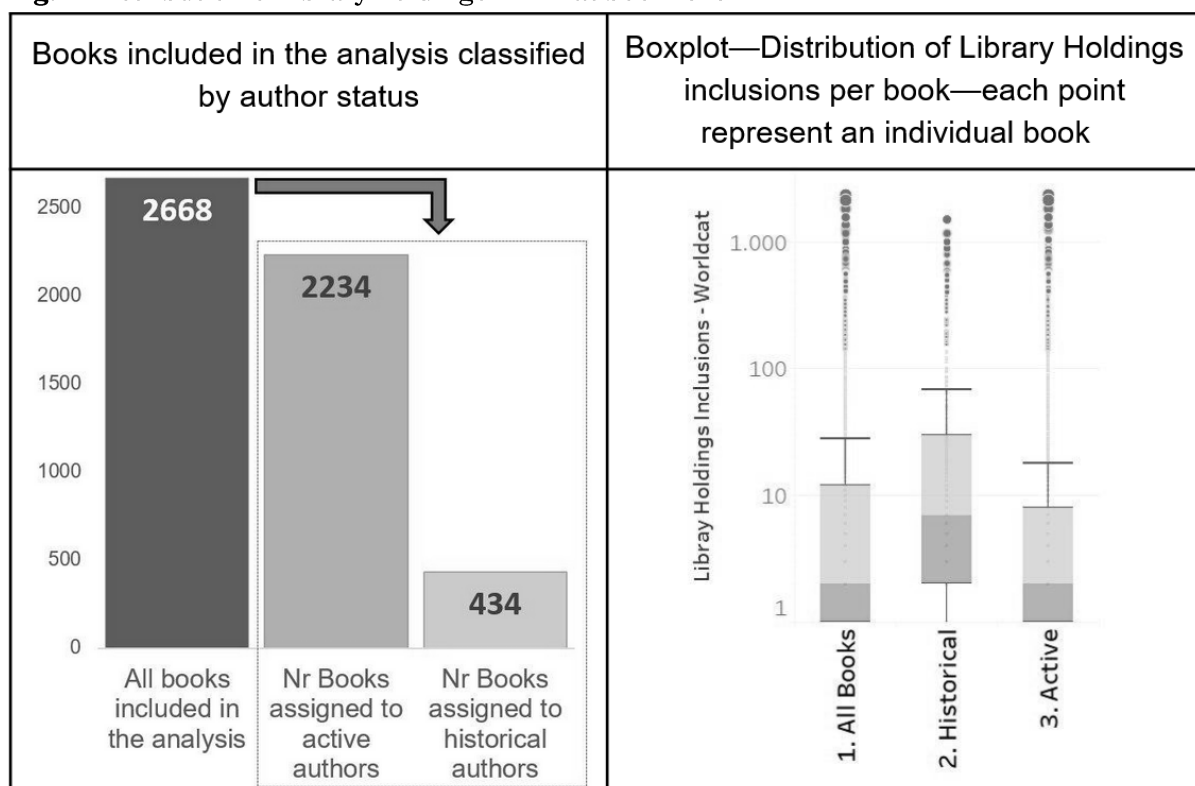
3.2. Book-level analysis

The authors in our sample have 5925 works and 13 786 publications in WorldCat. However, we have been unable to recover all of them, so our final sample is limited to 3134 works (52.9%) and 9484 publications (68.8%). Our total sample of books was 2668 following a cleaning process that involved eliminating duplicate works caused by co-authorship and books wrongly assigned to the authors being analyzed (Electronic supplementary material 1). Some 223 of these authors have at least one publication in English, while 78 have one in Spanish, and 45 have one in German. Of the works recovered, the most common language is English, accounting for 68.3% of the total, with 13.8% in Spanish and 4.2% in German. Most of the books analyzed (83%) correspond to active authors, and these account for 76% of library holdings. In total, 119 264 books are included in WorldCat and 89 959 are cited on Google Scholar, which means an average of 44 and 33 library holdings per book, respectively. The boxplot distributions of library holdings in Fig. 4 shows that books published by historical authors have higher averages even though the books with higher library holdings figures correspond to active authors. Some of the works are monographs and others are edited volumes. The contribution of an editor is presumably less than the contribution of the monograph author since editors should share credit with chapter authors.

Table 4 ranks books published by authors who specialize in Scientometrics, ordered by the number of library holdings. The first three books have two common characteristics: they have

an applied/professional nature and receive few citations. The book that is most visible in catalogs is *Global science & technology information: a new spin on access* by Wagner, with 2378 library holdings and 6 citations. This is followed by *Build your own databases* by Jacsó (2137 libraries, 14 citations) and *The Web library* by Tomaiuolo (1845 libraries and 9 citations). Format seems to have an impact on which books are available. For example, *Global science & technology information: a new spin on access* is listed by WorldCat as available at the University of Wolverhampton (to which one of the present authors is affiliated). However, it is only in electronic format as part of a ProQuest Ebook Central subscription despite being listed in WorldCat as a “Print Book” (which does also exist). This title is part of the UK & Ireland Academic Complete package of 115 000 ebooks provided to UK and Irish universities by ProQuest¹²and, therefore, part of a package selected by ProQuest rather than by a librarian¹³. Similarly, *Compass for intercultural partnerships* and *Theories of informetrics and scholarly communication: a Festschrift in honor of Blaise Cronin* are part of the ProQuest Academic Complete ebook offering for the USA, presumably accounting for their high library holdings.

Fig. 4 Distribution of library holdings in WI at book level



¹² <https://about.proquest.com/blog/pqblog/2015/Brand-New-UKI-Edition-of-Academic-Complete.html>

¹³ See also: <https://about.proquest.com/products-services/Academic-Complete.html>

In terms of content, these books are not necessarily oriented towards bibliometric issues and many focus instead on technological issues (web design, databases, etc.). Books of a combined professional and educational nature dominate the table.

Table 4 Ranking of books with a higher number of Library Holdings

Book. Bibliographic reference	Library holdings	Google Scholar Citations
Caroline Wagner, Allison Yezril. Global science & technology information: a new spin on access. RAND, 1999	2378	6
Péter Jacsó; F Wilfrid Lancaster. Build your own database. American Library Association, 1999.	2137	14
Nicholas G Tomaiuolo; Barbara Quint. The Web library: building a world class personal library with free Web resources. Information Today, 2004.	1845	9
Katy Börner. Atlas of knowledge: anyone can map. The MIT Press, 2015	1565	97
Steven W Popper; Eric V Larson; Caroline S Wagner. New forces at work: industry views critical technologies, RAND, 1998	1557	88
Derek de Solla Price. Science since Babylon. Yale Univ. Press 1962	1515	1381
R Elsen; Ignace Pollet; Patrick Develtere; Koenraad Debackere. Compass for intercultural partnerships. Leuven University Press, 2017	1371	2
Paul Wouters; Anne Beaulieu; Andrea Scharnhorst; Sally Wyatt. Virtual knowledge: experimenting in the humanities and the social sciences. MIT Press, 2013	1342	47
Nicholas G Tomaiuolo. UContent: the information professional's guide to user-generated content. Information Today, 2012	1336	6
Stefanie Haustein Multidimensional journal evaluation: analyzing scientific periodicals beyond the impact factor. De Gruyter/Saur, 2012.	1305	52
Katy Börner; David E Polley. Visual insights: a practical guide to making sense of data. MIT Press, 2014.	1289	107
Caroline S. Wagner. The new invisible college: science for development. Brookings Institution Press, 2008.	1269	547
Chris Steyaert; Bart van Looy. Relational practices, participative organizing. Emerald, 2010.	1255	28
Howard D. White. Brief tests of collection strength: a methodology for all types of libraries. Greenwood Press, 1995.	1179	60
Henk F Moed. Citation analysis in research evaluation. Springer, 2005	1015	1171
Vladimir Geroimenko; Chaomei Chen. Visualizing the semantic Web: XML-based Internet and information visualization. Springer, 2003	948	294
Cassidy R Sugimoto; Blaise Cronin. Theories of informetrics and scholarly communication: a Festschrift in honor of Blaise Cronin. De Gruyter, 2016.	921	12
Derek de Solla Price. Frontiers of science: on the brink of tomorrow. The Society, 1982.	894	--
Péter Jacsó. Content evaluation of textual CD-ROM and Web databases. Libraries Unlimited, 2001.	872	16

This list also includes at three books with both a high number of citations and a high number of library holdings. Unlike those featured earlier, these books were written by the historical authors of the field and their contents are not specialized in Scientometrics in two cases. The first is the classic foundation book *Science since Babylon* which describes the exponential growth of scientific literature. It is fifth in the library holdings ranking (1515) and is the most cited book in our collection (1381). It is followed by a title with a similar profile: Wagner's *New invisible college: science for development* (1269 library and 547 citations) and *Citation analysis in research evaluation* by Henk F. Moed (1015 libraries and 1171 citations). Table 4

therefore reveals two distinct book profiles: those with a Scientometric focus and those focusing more on the sociology of science.

4. Discussion & Conclusions

This paper has investigated WorldCat Identities for comparative analyses of authors through their library holdings. We investigated the viability of applying this source of information to a specific scientific field. The investigation shows that this type of analysis is possible but that WI has a series of methodological limitations (Table 5). Most of the problems described relate to the incorrect disambiguation of author names and the incorrect assigning of works. This means we cannot directly use the indicators given to an author without first verifying and validating the data collected. When using the API directly, we successfully located 221 authors (82.2% of the total collected in WI). In addition, 125 of these authors had no duplicate records (46.6% percent of the total and 80.1% of the total of non-duplicated records). Furthermore, we recovered 96 authors but found duplicate records (85.0% of the authors had duplicates; 31.5% of the records were duplicated).

Table 5 List of main methodological limitations of WI

Limitation	Description
Does not disambiguate well in Spanish	Does not disambiguate homonyms, thus sometimes generating multiple entries.
Does not aggregate authors from different sources	Many authority records are generated from other authority sources, such as VIAF or LCCN, but the authors who are present in several of them are not unified in WI.
Separation of personal and corporate identity	Some authors have separate records for personal identity and corporate identity.
Does not only include books	Includes materials such as theses, articles or conference papers and proceedings.
Incorrect assignment of records	Authors may have other authors' work associated with their records.
Conflict between works written by the author and works about the author	WI differentiates between the author's own work and works about the author—such as biographies—although these are sometimes confused.
Includes catalog entries from large scale ebook subscriptions	Books can have high values if they are included in a version of ProQuest Academic Complete or any other ebook service that integrates with library catalogs.

Furthermore, database use can involve geographic or linguistic bias, both of which are very common in citation indexes. Despite WorldCat's obvious advantages, few studies have critically analyzed its coverage even though it has a clear English language bias (Wakeling, Clough, Connaway, Sen & Tomás, 2017). Table 6 shows that 44.8% of platform users and 43% of academic libraries are from the US—much higher figures than those for any European

country. For example, only 1.5% of users and 0.7% of libraries are in Spain—of 76 Spanish university libraries, only 42 are in WorldCat. Thus, Spanish researchers may need to use complementary sources when analyzing the diffusion of books in Spain. Thus, when conducting an LCA with WorldCat, researchers should consult the OCLC members’ directory to verify the catalogs’ territorial distribution. In addition, Table 6 may allow us to take an objective approach to the significance and the terminological debate. In this paper we use the terms 'impact' as well as 'book diffusion' for library catalog measures but the term ‘cultural presence’ or ‘cultural visibility’ (Zuccala, 2018) may be more precise. The results have shown that the vast platforms and users in WorldCat are from the United States and this could be a signal of this cultural ‘availability’, ‘presence’ or ‘visibility’.

Table 6 User location and number of academic libraries in WorldCat (OCLC members only)

Countries	User Location ^{note1}	Number of academic libraries ^{note2}	Number of public libraries ^{note2}	Number of other types of library ^{note2}	Total number of libraries ^{note2}
United States	44.8	2198 (40.69 %)	3394 (80.79 %)	3264 (76.96 %)	8856 (63.97 %)
China	5.3	13 (0.24 %)	8 (0.19 %)	4 (0.09 %)	25 (0.18 %)
Canada	5.2	117 (2.17 %)	73 (1.74 %)	113 (2.66 %)	303 (2.19 %)
United Kingdom	3.7	72 (1.33 %)	96 (2.29 %)	82 (1.93 %)	250 (1.81 %)
Germany	3.2	274 (5.07 %)	16 (0.38 %)	75 (1.77 %)	365 (2.64 %)
France	2.3	1076 (19.92 %)	8 (0.19 %)	29 (0.68 %)	1113 (8.04 %)
India	1.8	23 (0.43 %)	0 (0 %)	11 (0.26 %)	34 (0.25 %)
Italy	1.7	87 (1.61 %)	112 (2.67 %)	12 (0.28 %)	211 (1.52 %)
Indonesia	1.7	32 (0.59 %)	0 (0 %)	42 (0.99 %)	74 (0.53 %)
Spain	1.5	15 (0.28 %)	4 (0.1 %)	14 (0.33 %)	33 (0.24 %)
Netherlands	1.5	38 (0.7 %)	95 (2.26 %)	24 (0.57 %)	157 (1.13 %)
Mexico	1.3	24 (0.44 %)	0 (0 %)	5 (0.12 %)	29 (0.21 %)
Australia	1.3	156 (2.89 %)	232 (5.52 %)	376 (8.87 %)	764 (5.52 %)
Brazil	1.3	17 (0.31 %)	0 (0 %)	3 (0.07 %)	20 (0.14 %)
Poland	1.2	17 (0.31 %)	4 (0.1 %)	3 (0.07 %)	24 (0.17 %)
Japan	0.9	64 (1.18 %)	1 (0.02 %)	12 (0.28 %)	77 (0.56 %)
Malaysia	0.9	23 (0.43 %)	0 (0 %)	3 (0.07 %)	26 (0.19 %)
Korea, Republic of	0.7	6 (0.11 %)	0 (0 %)	1 (0.02 %)	7 (0.05 %)
Russian Federation	0.7	13 (0.24 %)	0 (0 %)	9 (0.21 %)	22 (0.16 %)
Singapore	0.7	12 (0.22 %)	2 (0.05 %)	11 (0.26 %)	25 (0.18 %)
^{note1} Geographical location of users, results from log (Wakeling, Clough, Connaway, Sen & Tomás, 2017)					
^{note2} Data from the Directory of OCLC members					

As with other scientific databases, WorldCat has common information retrieval and coverage problems that must always be borne in mind when conducting a study. Nonetheless, despite these limitations, WI is a relevant source for studies of authors. In this article, we have analyzed the field of Scientometrics and our results include the most frequently cited

researchers in the field, both historical and current. The results confirm that authors have different publication profiles so that focusing on journal articles alone may disadvantage book-oriented scholars. For example, highly cited authors like Price and Wagner have high library holdings numbers whereas others, like Leydesdorff, have library holdings numbers that do not correspond to them and, finally, other authors, like the librarian Nick Tomaiuolo, have few citations and high library holdings numbers.

Library holdings are most relevant to authors or editors of handbooks, monographs, or textbooks. This classification reflects a different sphere of activity and academic contributions related to the generation of teaching/educational contents (e.g. textbooks) or the author's engagement in their field (e.g. editing conference proceedings). Clearly some authors contribute to a specific field as well as undertaking other activities or publishing materials other than articles. In this context, both Chaomei Chen and Blaise Cronin serve as examples. For example in Chen's profile two of his most relevant contributions are textbooks oriented towards undergraduate and graduate students (*Visualizing information using SVG and X3D* and *Information visualization: beyond the horizon*). In the case of Cronin's profile, some works have a professional and/or educational approach (*The marketing of library and information services*) and even some with a clear humorous entertainment component (*Pulp friction*). We have also detected profiles that are 100% professional, especially those of librarians who have remained outside of scientific circles. Our classification captures the value of an academic activity beyond journal citations.

We have also analyzed the most relevant books and this has allowed us to discuss the previous results and distinguish between two contrasting phenomena. Firstly, we have a set of books with great scientific impact and diffusion in libraries. These include foundational texts in the field like *Science since Babylon* (Price, 1962) and contemporary manuals such as *Citation analysis in research evaluation* (Moed, 2005). These books enjoy widespread scientific recognition and, moreover, are reference manuals or handbooks—a value encapsulated in the library counts indicator. Secondly, we have a set of books that are present in many libraries but which have few citations. They may have a practical, professional profile, are not oriented towards a scientific readership and, thanks to the library counts, can now be analyzed from a different standpoint. One example of this profile is *Build your own database* (Jacsó & Lancaster, 1999). The list of works also shows that many authors in the field of Scientometrics publish non-specialized works that are more of a professional or educational nature and which go unnoticed in the more traditional bibliometric analyses.

The list of books should be interpreted in the context of its limitations. With our methodology we have identified a large number of books with the highest "Library Catalog Counts". This

methodology does not identify all important books, however. For example, if an important book's author is not in EC3's Scholar Mirror, that book would not be in our study. Therefore, the list is likely to be incomplete. Another issue that could question the value and significance of this list is the fact that books found in a catalog do not always respond to a librarian's choice since some are donations (Biagetti, 2018). Book holdings are primarily the result of decisions made by collection librarians, with the exception of big deal packages, which are collated by publishers. Thus, the library holding results are indirect indicators of impact or diffusion in the sense that they rely on the channels of information available to collection librarians to make judgements about the types of books that they believe to be relevant to their audience. However, perhaps the factor that most distorts the value of citations—as a consequence of the selection process—is the purchase of e-book collections, since these integrate books into library collections *en masse* (Lewis & Kennedy, 2019). In our case the holdings counts for some books are substantially boosted by their presence in ProQuest (or other) mass electronic offerings. This substantially undermines the value of library holdings as indicators of academic interest in books because they do not always reflect the decisions of individual librarians and academics when purchasing books, even though they do reflect the availability of books in libraries.

Finally, whilst WI may be useful for indicators based on library holdings, the limitations above should be taken into account when using them or the results may be highly misleading. Like most current metric profiles, the indicators have to be reviewed and corrected. This problem has been reported, for example, for Google Scholar, ResearchGate and Mendeley profiles (Martín-Martín, Orduña-Malea & Delgado López-Cózar, 2016) although the advantage of WI is that, unlike other sources, it cannot be manipulated (Thelwall & Kousha, 2017). In addition, library holdings are influenced by the presence of a book in a package deal, such as that of ProQuest, and there does not seem to be a practical way to detect this (there does not seem to be a public ProQuest list of Academic Complete that could be checked, for example). Nevertheless, WI may be useful tool because at author level it offers quickly-obtained indicators that allow us to present an alternative vision of impact even though it must be used with checks to safeguard against inflation due to books in package deals.

5. References

- Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68(3), 329–342. <https://doi.org/10.1007/s11192-006-0115-z>
- Biagetti, M. T., Iacono, A., & Trombone, A. (2018). Testing library catalog analysis as a bibliometric indicator for research evaluation in Social Sciences and Humanities. In

- Challenges and Opportunities for Knowledge Organization in the Digital Age: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal* (pp. 892–899). Baden-Baden: Ergon-Verlag. <https://doi.org/10.5771/9783956504211-892>
- Halevi, G., Nicolas, B., & Bar-Ilan, J. (2016). The Complexity of Measuring the Impact of Books. *Publishing Research Quarterly*, 32(3), 187–200. <https://doi.org/10.1007/s12109-016-9464-5>
- Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193–215, <https://doi.org/10.1007/BF02457380>
- Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, 59(11), 1819–1828. <https://doi.org/10.1002/asi.20885>
- Jacsó, P., & Lancaster, F. W. (1999). *Build your own database*. American Library Association.
- Kousha, K., & Thelwall, M. (2015). Web indicators for research evaluation: Part 3: books and non standard outputs. *El Profesional de La Información*, 24(6), 724–736. <https://doi.org/10.3145/epi.2015.nov.04>
- Kousha, K., & Thelwall, M. (2016). Can Amazon.com reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*, 67(3), 566–581. <https://doi.org/10.1002/asi.23404>
- Lewis, R. M., & Kennedy, M. R. (2019). The Big Picture: A Holistic View of E-book Acquisitions. *Library Resources & Technical Services*, 63(2), 160. <https://doi.org/10.5860/lrts.63n2.160>
- Linmans, A. J. M. (2008). Een exploratieve studie van de onderzoeksprestaties van de Faculteit Letteren aan de Universiteit Leiden (in Dutch). Internal CWTS Report
- Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link - Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, 83(2), 337–354. <https://doi.org/10.1007/s11192-009-0088-9>
- Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2016). The Role of Ego in Academic Profile Services: Comparing Google Scholar, ResearchGate, Mendeley, and ResearcherID (SSRN Scholarly Paper ID 2745892). Social Science Research Network. <https://doi.org/10.2139/ssrn.2745892>
- Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Springer Netherlands. <https://doi.org/10.1007/1-4020-3714-7>
- Neville, T. M., & Henry, D. B. (2014). Evaluating Scholarly Book Publishers—A Case Study in the Field of Journalism. *The Journal of Academic Librarianship*, 40(3), 379–387. <https://doi.org/10.1016/j.acalib.2014.05.005>
- Nilges, C. (2006). The Online Computer Library Center's Open WorldCat Program. *Library Trends*, 54(3), 430–447. <https://doi.org/10.1353/lib.2006.0027>
- Price, D. J. D. S. (1962). *Science since babylon*. Yale University Press New Haven.

- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. <http://altmetrics.org/manifesto/> Accessed 20 March 2020.
- Thelwall, M., & Kousha, K. (2017). ResearchGate versus Google Scholar: Which finds more early citations?. *Scientometrics*, *112*(2), 1125–1131. <https://doi.org/10.1007/s11192-017-2400-4>
- Torres-Salinas, D., & Arroyo-Machado, W. (2020). Library Catalog Analysis and Library Holdings Counts: Origins, Methodological Issues and Application to the Field of Informetrics. In C. Daraio & W. Glänzel (Eds.), *Evaluative Informetrics: The Art of Metrics-Based Research Assessment: Festschrift in Honour of Henk F. Moed* (pp. 287–308). Springer International Publishing. https://doi.org/10.1007/978-3-030-47665-6_13
- Torres-Salinas, D., Gumpenberger, C., & Gorraiz, J. (2017). PlumX As a Potential Tool to Assess the Macroscopic Multidimensional Impact of Books. *Frontiers in Research Metrics and Analytics*, *2*, 5. <https://doi.org/10.3389/frma.2017.00005>
- Torres-Salinas, D., & Moed, H. F. (2008). *Library catalog analysis is a useful tool in studies of social sciences and humanities*. In A New Challenge for the Combination of Quantitative and Qualitative Approaches. 10th International Conference on Science and Technology Indicators, Viena.
- Torres-Salinas, D., & Moed, H. F. (2009). Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics*, *3*(1), 9–26. <https://doi.org/10.1016/j.joi.2008.10.002>
- Torres-Salinas, D., Robinson-Garcia, N., & Gorraiz, J. (2017). Filling the citation gap: measuring the multidimensional impact of the academic book at institutional level with PlumX. *Scientometrics*, *113*(3), 1371–1384. <https://doi.org/10.1007/s11192-017-2539-z>
- Wakeling, S., Clough, P., Connaway, L. S., Sen, B., & Tomás, D. (2017). Users and uses of a global union catalog: A mixed-methods study of WorldCat.org. *Journal of the Association for Information Science and Technology*, *68*(9), 2166–2181. <https://doi.org/10.1002/asi.23708>
- White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, *60*(6), 1083–1096. <https://doi.org/10.1002/asi.21045>
- White, H. D., & Zuccala, A. (2018). Libcitations, worldcat, cultural impact, and fame. *Journal of the Association for Information Science and Technology*, *69*(12), 1502–1512. <https://doi.org/10.1002/asi.24064>
- Zhang, H., Zhou, Q., & Zhang, C. (2018). Multi-discipline correlation analysis between citations and detailed features of library holdings. *Proceedings of the Association for Information Science and Technology*, *55*(1), 946–947. <https://doi.org/10.1002/pr2.2018.14505501188>

- Zuccala A. (2018). Language, Culture and Traversing the Scholarly Evaluation Landscape. In: A. Bonaccorsi (Eds.), *The Evaluation of Research in Social Sciences and Humanities*. Springer, Cham. https://doi.org/10.1007/978-3-319-68554-0_17
- Zuccala, A., Breum, M., Bruun, K. & Wunsch, B.T. (2018). Metric assessments of books as families of works. *Journal of the Association for Information Science and Technology*, 69(1), 146–157. <https://doi.org/10.1002/asi.23921>
- Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus ‘perceived cultural benefit’. In *14th International Society of Scientometrics and Informetrics Conference, ISSI 2013*, pp. 353–360. Vienna.
- Zuccala, A., & White, H. D. (2015). Correlating lib citations and citations in the humanities with WorldCat and scopus data. In A.A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al. (Eds.), *Proceedings of the 15th International Society of Scientometrics and Informetrics Conference, (ISSI), Istanbul, Turkey, 29th June to 4th July, 2015*. (pp. 305–316). Denmark: Bogazici Universitesi.

Wikiformetrics: Construction and description of an open Wikipedia knowledge graph dataset for informetric purposes

Wenceslao Arroyo-Machado^{1,*}, Daniel Torres-Salinas¹ and Rodrigo Costas^{2,3}



¹Department of Information and Communication Sciences, University of Granada, Granada, Spain

²Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

³DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

*Corresponding author: wences@ugr.es

Journal

Quantitative Science Studies
2641-3337 

Index

ESCI (2021)

DOI

10.1162/qss_a_00226

Data

10.5281/zenodo.6346899

Version

Published

References

APA 7th

Funding

Influ Science

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 (di)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 dataset of the English Wikipedia was built (and shared) following a relational model. Finally, a descriptive case study is carried out on the English Wikipedia dataset to illustrate the analytical potential of the knowledge graph and its metrics.

Citation

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikiformetrics: Construction and description of an open Wikipedia knowledge graph dataset for informetric purposes. *Quantitative Science Studies*, 1-35. https://doi.org/10.1162/qss_a_00226

1. Introduction

On January 15, 2001 Wikipedia was born under the umbrella of Nupedia, an encyclopedia project whose edition 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 two years, experiencing a steady growth. Since then, Wikipedia has become one of the top visited websites of the world (<https://www.semrush.com/website/top/>, consulted on August 4, 2022), having 328 different editions, 285 of them having more than 1000 articles (https://meta.wikimedia.org/wiki/List_of_Wikipedias, consulted on 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 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 so-called Wikipedia talks' pages (Maki et al., 2017; Yasseri 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 since 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 (C. E. Adams et al., 2020; Candelario et al., 2017; Weiner 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, for example racial and gender gaps in its biographies (J. Adams et al., 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 et al., 2020; Zheng 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 et al., 2019).

In spite of all previous 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 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 et al., 2010), which considered citations on Wikipedia to scientific literature as part of its realm¹. Wikipedia citations are one of the most popular sources covered in altmetric aggregators (Ortega, 2020; Zahedi & Costas, 2018) like 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 et al., 2020; Zagorova et al., 2022). Moreover, other proposals like Scholia, enable exploring bibliographic data at different levels through Wikidata (F. Å. Nielsen et al., 2017). In Table 1 a summary of previous studies on Wikipedia bibliographic references are presented.

¹ Although Wikipedia references had been already studied for years before the birth of altmetrics, like the citation analysis by F. A. 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
<i>Mühlhauser and Oser</i> (Mühlhauser & Oser, 2008)	Qualitative	Content and quality analysis	---	Check list	German	Health care
<i>Candelario et al.</i> (Candelario et al., 2017)		Content and quality analysis	33 pages	Scoring system	English	Medication
<i>Kaffee and Elsahar</i> (Kaffee & Elsahar, 2021)		Analyze the editors' citation process	---	Survey and interviews	Multilingual	Multidisciplinary
<i>Nielsen</i> (F. A. Nielsen, 2007)	Quantitative	Analyze citation patterns	30,368 citations	Descriptive statistics	English	Multidisciplinary
<i>Kousha and Thelwall</i> (Kousha & Thelwall, 2017)		Evaluate the impact of references	36,191 citations	Descriptive statistics	Multilingual	Multidisciplinary
<i>Lewoniewski et al.</i> (Lewoniewski et al., 2017)		References coverage across languages	6.8 million pages 41 million citations	Descriptive statistics	Multilingual	Multidisciplinary
<i>Maggio et al.</i> (Maggio et al., 2017)		Analyze citation patterns	229,857 pages 1,049,025 citations	Descriptive statistics	English	Medicine
<i>Pooladian and Borrego</i> (Pooladian & Borrego, 2017)		Evaluate the impact of references	982 citations	Descriptive analysis	Multilingual	Multidisciplinary
<i>Jemielniak et al.</i> (Jemielniak et al., 2019)		Rank journals by citations	11,325 pages 137,889 citations	Citation analysis	English	Medicine
<i>Torres-Salinas et al.</i> (Torres-Salinas et al., 2019)		Mapping of knowledge structure	25,555 pages 41,655 citations	Co-citation analysis	English	Arts & Humanities
<i>Arroyo-Machado et al.</i> (Arroyo-Machado et al., 2020)		Mapping of knowledge structure	193,802 pages 847,512 citations	Co-citation analysis	English	Multidisciplinary
<i>Colavizza</i> (Colavizza, 2020)		Publications coverage	3,083 ref. pub.	Topic modeling and regression analysis	English	COVID-19
<i>Nicholson et al.</i> (Nicholson et al., 2021)		Reviewing citation quality	1,923,575 pages 824,298 ref. pub.	Classification modeling	English	Multidisciplinary
<i>Singh et al.</i> (Singh et al., 2020)		Dataset creation	4 million citations	Text mining	English	Multidisciplinary
<i>Zagorova et al.</i> (Zagorova et al., 2022)		Dataset 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 co-citations patterns, showing fundamental differences in the use of scientific literature in Wikipedia compared to the academic realm. Bould et al. (2014), Li et al. (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 et al., 2016; Mittermeier et al., 2019, 2021; Roll et al., 2016; Vilain et al., 2017), and even demonstrating the potential of Wikipedia pages to monitor the spread of diseases (Generous 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 et al. (2018) suggested the value of using metrics in specific moments of

the life cycles, for example the number of editors in the first three 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 of 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 datasets (e.g. Mittermeier et al. (2021) who studied a large set of features in a dataset of Wikipedia pages of 10,099 bird species across 251 language editions). Thus, the large-scale study of Wikipedia, both from 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, as well as 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 as those developed in scientometric research. Such knowledge graph could work as complement of other Wikipedia knowledge graph like 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, etc.) 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 do 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 dataset 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, and ultimately providing a complete Wikipedia knowledge graph. More specifically three objectives of different nature 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 mapping 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 dataset, 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 et al., 2019; Wouters et al., 2019), contributing to the exploration of different science-society interactions that can be captured on Wikipedia (Costas et al., 2020). Our ambition is to frame Wikipedia as a data source with multiple informetric research possibilities. Furthermore, a dedicated dataset of the English edition of Wikipedia is constructed for informetric purposes and is freely available at Zenodo (doi:[10.5281/zenodo.6346899](https://doi.org/10.5281/zenodo.6346899)). R and Python were used together for its elaboration, with the scripts available on GitHub (doi:[10.5281/zenodo.6959428](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 or data scientists, among others.

2. Wikipedia from an informetric perspective

2.1. Analogy between Wikipedia pages and scientific publications

In Wikipedia the key component 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, talk pages, etc., 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 2 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, being possible to identify the first 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 connecting Wikipedia pages 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, like 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 a static document. The access and editing of the contents also differ between Wikipedia pages and scientific publications, since 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).

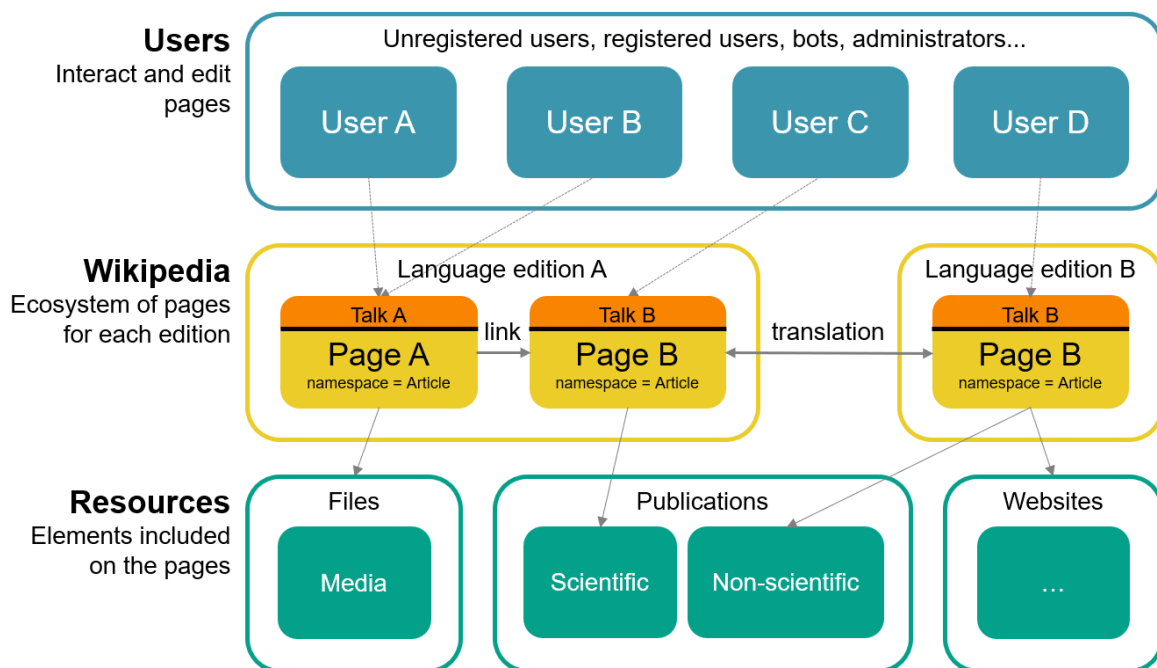
Table 2. Comparison of features between Wikipedia pages and scientific publications.

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

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

The living nature of Wikipedia pages puts them at the center of a complex system (Ladyman et al., 2013), whose main elements are represented in Fig 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 et al., 2016), since 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 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 et al., 2017).

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



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 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, for example, by studying their semantic domain (Aghaebrahimian et al., 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 et al., 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 pages 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 et al., 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 counts with 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 models proposed (Ross-Hellauer, 2017), the three basic principles (open identities, reports, and participations) are clearly recognizable in Wikipedia (S2 Table). 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 post-publication correction system and readers' comments (Xiao & Askin, 2014).

Wikipedia also counts with 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, which 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 to modify them. When there are disagreements among wikipedians in the quality levels of a page, this leads to a discussion and quest of 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 candidature and an evaluation by independent wikipedians according to pre-established criteria. These two levels also have their own badges on the article page.

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

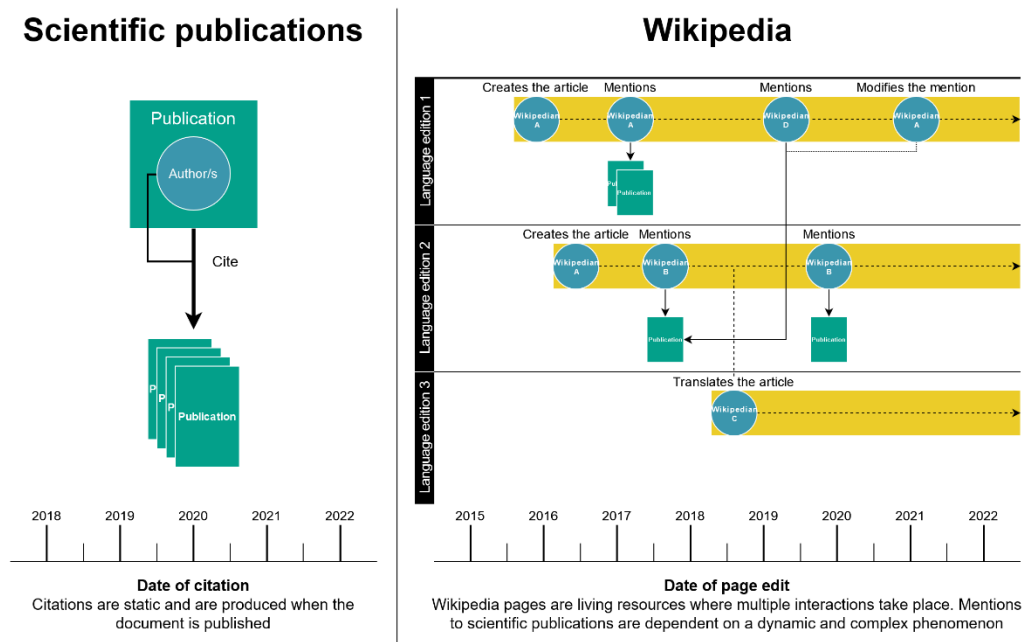
2.4. Sources

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 regards to other encyclopedic systems (Reagle & Koerner, 2020). These internal links have been studied previously, showing both the semantic relationships they can establish and other potential utilities (Consonni et al., 2019; Presutti et al., 2014), as well as the possibility of calculating network indicators like 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, i.e., 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 Fig 2 the main differences between scientific publications references and Wikipedia references are schematized.

Fig 2. Differences between traditional citations and Wikipedia mentions to scientific publications.



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. Although, there are some methodological proposals for this purpose (Zagorova et al., 2022).

- Wikipedia citation counts can be distorted by the translations of articles into different languages, since 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. Such limitation does not occur in scientific publications, since 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 S3 Appendix. 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.

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	Realtime	Small data	General	Data recovery
Wiki Replicas	Metadata, page content, and relationships	Online	SQL	Near-realtime	Small data	General	Data recovery
Event Streams	Real-time logs	Online	SSE, JSON	Realtime	-	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	Realtime	Small data	Specific	Data recovery
Repositories	Dedicated Wikipedia datasets	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.

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 2GB 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. Wikiformetrics

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 while the number of edits and editors reflect the volume of activity, the number 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., contributions, content development, links and interactions, and impact), being also easy to interpret in an informetric framework.

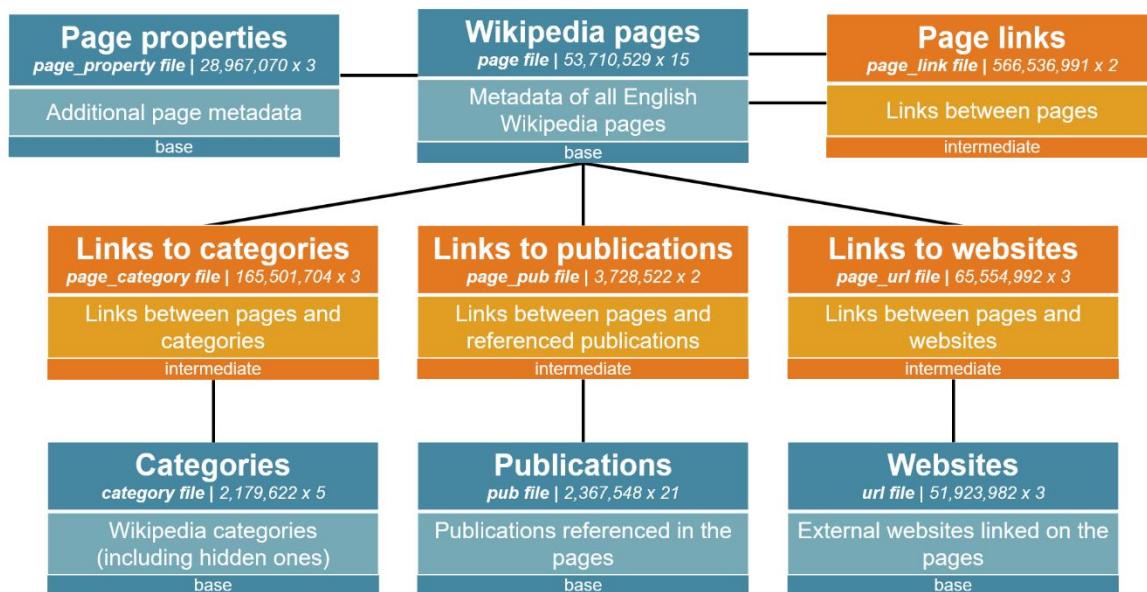
Table 5. Description of the metrics obtained for Wikipedia articles by analytical dimension.

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

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 (doi:[10.5281/zenodo.6346899](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 dataset of Singh et al. (2020) in which they share references made in Wikipedia articles. The data included in this dataset 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 (doi:[10.5281/zenodo.6959428](https://doi.org/10.5281/zenodo.6959428)). The construction of this dataset is described in S1 Appendix. The resulting dataset consists of 9 files connected to each other by a relational structure summarized in Fig 3.

Fig 3. Diagram of files and relationships of the Wikipedia knowledge graph dataset.



Wikipedia Knowledge Graph, dataset and description free at: [10.5281/zenodo.6346899](https://doi.org/10.5281/zenodo.6346899)

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 media (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, to name a few). This is the case of the work of Arroyo-Machado et al. (2022), who found a positive relationship between the research performance of universities and their social attention on Wikipedia, using data from this dataset.

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 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 is also available in other languages. The only limitation in this respect is in the reference data, as they come from a specific dataset (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 (doi:[10.5281/zenodo.6958972](https://doi.org/10.5281/zenodo.6958972)).

4.1. Wikipedia metrics and article's content

There are a total of 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. For all of them, the metrics proposed in Fig 4 have been obtained.

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

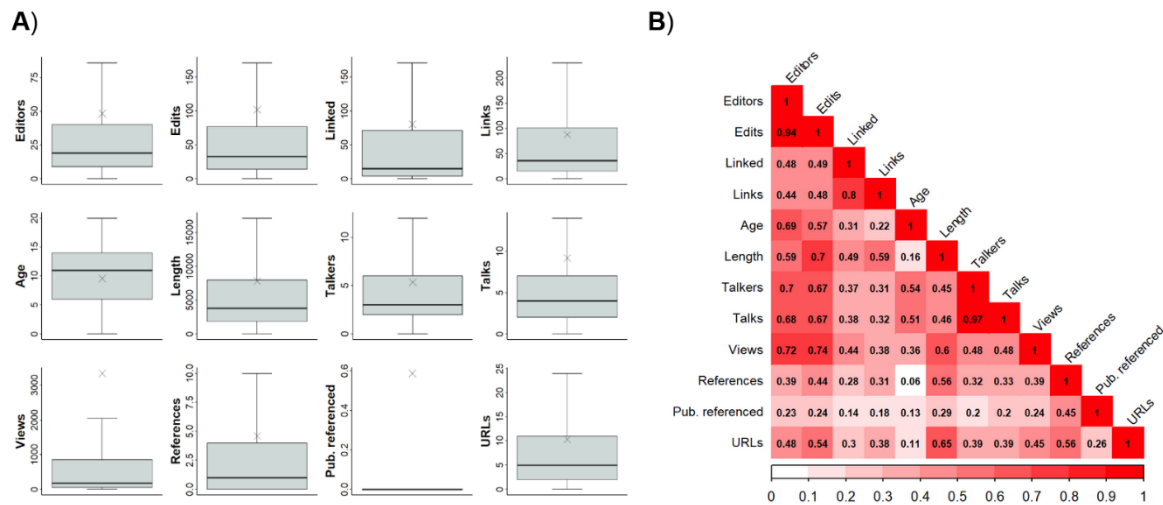
	All articles	Featured articles	Featured lists	A	Good	B	C	List	Start	Stub
<i>N. of articles</i> → <i>Wiki Metrics</i> ↓	6,328,134	5945	3816	958	34,004	109,019	394,065	253,066	1,818,356	3,079,778
Editors	48.38	516.93	179.13	176.80	275.71	297.62	165.36	56.27	63.13	22.85
Edits	101.92	1491.35	593.61	564.91	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

Fig 4 shows the descriptive statistics of the main variables, differentiating between total Wikipedia articles and those classified based on their quality. A total of 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 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 of Wikipedia pages are not of recent creation (Fig 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 3346.59. 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, since the median for the former is 36 and for the latter 15.

Fig 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.



The correlations between these variables are all positive (Fig 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 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 number of views. Other 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 R Shiny 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 (S4 Table). 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 dataset—. Thus, in our dataset 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 twenty most viewed pages are series or movies released in the period analyzed, which also highlights that contents related to entertainment occupy 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 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 stand out political figures, religion topics, and scientific controversies. 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...), it is also possible to identify what are 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 target to more general audiences. The live nature of Wikipedia is probably its main distinctive feature when compared to scientific publications. Such live nature of Wikipedia articles 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 among others. 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 (doi:[10.5281/zenodo.6346899](https://doi.org/10.5281/zenodo.6346899)). The data are gathered under a comprehensive dataset 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 on the one hand to characterize any Wikipedia page, while also allowing 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 dataset shared by Singh et al. (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 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 dataset 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.

References

- Adams, C. E., Montgomery, A. A., Aburrow, T., Bloomfield, S., Briley, P. M., Carew, E., Chatterjee-Woolman, S., Feddah, G., Friedel, J., Gibbard, J., Haynes, E., Hussein, M., Jayaram, M., Naylor, S., Perry, L., Schmidt, L., Siddique, U., Tabakert, A. S., Taylor, D., ... 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>
- 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*, 0165551520977438. <https://doi.org/10.1177/0165551520977438>

- 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., 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>
- 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. *Proceedings of the 17th International Conference on World Wide Web*, 1095–1096. <https://doi.org/10.1145/1367497.1367673>
- Boldi, P., & Monti, C. (2016). Cleansing Wikipedia Categories Using Centrality. *Proceedings of the 25th International Conference Companion on World Wide Web*, 969–974. <https://doi.org/10.1145/2872518.2891111>
- Bould, M. D., Hladkowitz, E. S., Pigford, A.-A. E., Uffholz, L.-A., Postonogova, T., Shin, E., & 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>
- 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.e1. <https://doi.org/10.1016/j.japh.2016.12.063>
- Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies*, 1–32. 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. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 598–607.
- 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>
- Dzogang, F., Lansdall-Welfare, T., & Cristianini, N. (2016). Seasonal Fluctuations in Collective Mood Revealed by Wikipedia Searches and Twitter Posts. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 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. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 777–786.
- Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Friedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, 10(11), e1003892. <https://doi.org/10.1371/journal.pcbi.1003892>
- 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.
- Hill, B. M., & Shaw, A. (2015). Page Protection: Another Missing Dimension of Wikipedia Research. *Proceedings of the 11th International Symposium on Open Collaboration*. <https://doi.org/10.1145/2788993.2789846>
- History of Wikipedia*. (2021, May 28). Wikipedia. https://en.wikipedia.org/wiki/History_of_Wikipedia
- Jang, M., Foley, J., Dori-Hacohen, S., & Allan, J. (2016). Probabilistic Approaches to Controversy Detection. *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 2069–2072. <https://doi.org/10.1145/2983323.2983911>
- Jemielniak, D. (2012). *Wikipedia: An effective anarchy*. Society for Applied Anthropology, Baltimore, United States.
- 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>

- 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–e11429. PubMed. <https://doi.org/10.2196/11429>
- Kaffee, L.-A., & Elsahar, H. (2021). References in Wikipedia: The Editors' Perspective. *Companion Proceedings of the Web Conference 2021*, 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. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 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*, 1–16. <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.
- 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>
- Maki, K., Yoder, M., Jo, Y., & Rosé, C. (2017). Roles and Success in Wikipedia Talk Pages: Identifying Latent Patterns of Behavior. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 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.
<https://www.frontiersin.org/article/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>
- 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>
- 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>
- Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*.
<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.
- Olleros, F. X. (2008). Learning to Trust the Crowd: Some Lessons from Wikipedia. *2008 International MCETECH Conference on E-Technologies (Mctech 2008)*, 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-oneil/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., Bannour, I., & 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.
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Altmetrics. <http://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.
- Roll, U., Mittermeier, J. C., Diaz, G. I., Novosolov, M., Feldman, A., Itescu, Y., Meiri, S., & Grenyer, R. (2016). Using Wikipedia page views to explore the cultural importance of global reptiles. *Advancing Reptile Conservation: Addressing Knowledge Gaps and Mitigating Key Drivers of Extinction Risk*, *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*, 588–588. PubMed. <https://doi.org/10.12688/f1000research.11369.2>
- Singh, H., West, R., & Colavizza, G. (2020). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies*, 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. Mladenčić, S. Auer, & C. Lange (Eds.), *The Semantic Web* (pp. 227–240). Springer International Publishing.
- 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>
- 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. *J Med Internet Res*, *21*(3), e12450. <https://doi.org/10.2196/12450>
- Wilkinson, D. M., & Huberman, B. A. (2007). Assessing the value of cooperation in Wikipedia. *First Monday*. <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, M. Henk F, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 687–713). Springer International Publishing.
- 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>
- 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>
- 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. (Nico), Albano, C. M., Vora, N. M., Mai, F., & Nickerson, J. V. (2019). The Roles Bots Play in Wikipedia. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW). <https://doi.org/10.1145/3359317>

Mapping the backbone of the Humanities through the eyes of Wikipedia



Daniel Torres-Salinas¹, Esteban Romero-Frías¹ and Wenceslao Arroyo-Machado^{1,*}

¹Medialab UGR, University of Granada, Gran Vía de Colón, 48, 18071 Granada, Spain.

*Corresponding author: wences@ugr.es

Journal

Journal of Informetrics
1751-1577

Index

SCIE – Q1

DOI

10.1016/j.joi.2019.07.002

Data

None

Version

Preprint

References

APA 7th

Funding

None

Abstract

The present study aims to establish a valid method by which to apply the co-citation methodology to Wikipedia article references and, subsequently, to map these relationships between scientific papers. This method, originally applied to scientific literature, will be transferred to the digital environment of collective knowledge generation. To this end, a dataset containing Wikipedia references collected from Altmetric and Scopus' Journal Metrics journals has been used. The articles have been categorized according to the disciplines and specialties established in the All Science Journal Classification (ASJC). They have also been grouped by journal of publication. A set of articles in the Humanities, comprising 25 555 Wikipedia articles with 41 655 references to 32 245 resources, has been selected. Finally, a descriptive statistical study has been conducted and co-citations have been mapped using networks of degree centrality and intermediation.

Citation

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>

1. Introduction

When Wikipedia was created in 2001 (DiBona, Cooper & Stone, 2006), few could have imagined that in a short time a voluntary, collective project would become the main encyclopedic work of reference for a large part of Humanity. The birth of Wikipedia, in the middle of the dot-com bubble, occurred during the prelude to the emergence of the Web 2.0 paradigm (O'Reilly, 2005) and was destined to become one of the greatest exponents of the Web's ability to activate the collective intelligence of Internet users (Surowiecki, 2005). In January 2018, 17 years later, the English language version of Wikipedia accounted for 5.5 million of the 47 million articles in the more than 290 editions of Wikipedia; although it had more than 32 million registered users only 123 966 were active editors¹. The Wikipedia in English—its largest edition—represents approximately 11.7% of the whole of Wikipedia, creating more than 600 new articles per day in 2017. According to the Community Engagement Insights 2018 Report², prepared by the Wikimedia Foundation, 85% of contributors to Wikimedia communities have post-secondary education (12% have a doctorate).

According to Alexa,³ at the beginning of 2018, Wikipedia ranked 5th among the most visited websites in the world with a remarkable 66.4% of traffic received coming from user searches. These data refer to organic traffic received by the website and demonstrate that, for a wide variety of terms, Wikipedia is one of the first options that search engines offer as a relevant result on the Web. Hence, it constitutes a much-used reference resource that is of great importance for educational purposes in Science, the Humanities, and other fields. For example, as an encyclopedic digital project, Wikipedia is considered a "*very fertile ground for the creation of innovative projects related to the Digital Humanities*"⁴. It is argued that Wikipedia might be the best and the largest educational platform in history (Tramullas, 2016).

Wikipedia is conceived of as a tool for the dissemination of knowledge through articles generated by its users under Creative Commons licenses (attribution-share alike). Wikipedia has overtaken its competitors by revolutionizing the industry through a profound epistemological transformation that focuses on the social dimension (Fallis, 2008, Fuchs, 2008). Over time, Wikipedia has developed complex rules—generated by the community itself—that are not rigid and remain subject to revision but, at the same time, are strictly observed. Articles should always be verifiable and have reliable sources. Insofar as encyclopedic content is concerned, secondary sources that are "*reliable, independent and published*" prevail. Among these, particular mention is made of specialized publications:

¹ <https://en.wikipedia.org/wiki/Wikipedia> (consulted on January 3, 2018)

² https://meta.wikimedia.org/wiki/Community_Engagement_Insights/2018_Report (consulted on February 5, 2019)

³ <https://www.alexa.com/topsites> (consulted on February 21, 2018)

⁴ <https://blog.wikimedia.org/2016/08/17/wikipedia-largest-digital-humanities-project/> (consulted on February 21, 2018)

"Many Wikipedia articles rely on scholarly material. When available, academic and peer-reviewed publications, scholarly monographs, and textbooks are usually the most reliable sources. However, some scholarly material may be outdated, in competition with alternative theories, or controversial within the relevant field. Try to cite current scholarly consensus when available, recognizing that this is often absent. Reliable non-academic sources may also be used in articles about scholarly issues, particularly material from high-quality mainstream publications. Deciding which sources are appropriate depends on context. Material should be attributed in-text where sources disagree."⁵

At the same time, from the perspective of scientific knowledge evaluation, in recent years digital indicators have been used as an alternative measure of academic impact: the so-called altmetrics indicators (Piwowar, 2013a, 2013b; Priem et al., 2010; Torres-Salinas, Cabezas-Clavijo & Jiménez-Contreras, 2013).

In this context, Wikipedia faces a dual challenge: on the one hand, the call to guarantee rigor in Wikipedia contents by referencing articles published in scientific journals; on the other, the opportunity to use Wikipedia references to scientific articles as a highly valuable altmetric information source to assess the social impact of research. Evidence of the value of references included in Wikipedia is its high weighting in a synthetic indicator such as the Altmetric Attention Score⁶. In this indicator, Wikipedia articles receive a rating of 3, which is higher than those corresponding to mentions on Twitter (1) or Facebook (0.25), but lower than references to news feeds (8) and blogs (5).

The connection between Wikipedia as a social platform and scientific articles has been explored in different ways. For example, through the analysis of reference and citation patterns in a specific scientific area (Serrano-López, Ingwersen & Sanz-Casado, 2017), as a platform for the promotion of open access scientific literature (Teplitskiy, Lu & Duede, 2016), or by exploring its limitations as a source in the evaluation of scientific activity (Kousha & Thelwall, 2016). Knowledge representation has also been formulated through reference maps connecting articles (Silva et al., 2011), or by analyzing differences between the Universal Decimal Classification (UDC) category structure and that generated by Wikipedia itself (Salah et al., 2012).

From a bibliometric perspective, co-citations constitute a classic instrument (Small, 1973) that allows knowledge to be mapped by taking account of common references received from a third

⁵ https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources (consulted on February 21, 2018)

⁶ <https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated->

document. Co-citations can be interpreted as a measure of the similarity between two documents. This approach has been used to observe the connections between words (Leydesdorff & Nerghe, 2017), or between areas of knowledge through scientific articles (Leydesdorff, Carley & Rafols, 2012). More recently, with the development of the Web, this concept has been transferred to this new space by discussing co-link analysis (Thelwall, 2009)—an approach based on sites or web pages that simultaneously link to other sites or web pages. Co-link analysis has proved a useful means of revealing the cognitive or intellectual structure of a field of study (Zuccala, 2006). Moreover, it has allowed investigators to broaden their scope of study beyond scientific production, having been applied to business (Vaughan & Romero-Frías, 2010), politics (Romero-Frías & Vaughan, 2010) or universities (Vaughan, Kipp & Gao, 2007).

In this regard, to our knowledge, no study has used Wikipedia as a reference to map science by extrapolating classical co-citation methodology to this digital platform in order to discover the structure of journals corresponding to different areas of knowledge and different disciplines. With this approach, scientific knowledge could be mapped from a social perspective, thus offering a radically different view to that of the traditional maps constructed from the relationships between the scientific studies themselves. This approach is in line with the proposal made by Costas, Rijcke and Marres (2017) for the study of co-social mediation interaction. Based on this framework, we have focused on the Humanities in order to achieve the following objectives:

1. to establish a methodology to transfer co-citation methodology to a digital environment taking as a reference an altmetric indicator linked to the collective generation of knowledge in Wikipedia; and,
2. to analyze how scientific knowledge is established in the field of the Humanities as this is represented in Wikipedia.

2. Material and methods

2.1 Information sources and data processing



This study uses Altmetric.com as its source of information and the Altmetric Explorer to extract the references to scientific articles that are included in Wikipedia articles. To do this we have used the platform's download functions to obtain a csv file in which each scientific article appears with its basic data and information about the Wikipedia article in which it is referenced. So, all the scientific articles indexed in Altmetric.com and cited in Wikipedia have been downloaded. We have also used the Altmetric API to obtain complementary information (ISSN).

A database with 261 079 Wikipedia entries was generated with a total of 1 214 322 references to 848 079 individual resources dated between 2004 and 2017. It should be noted that in 2004, 2005 and 2006 only 12 citations were counted. Only references for which Altmeter.com provides an associated publication date have been included, leading to an 8.8% (107 008) reduction in the dataset. When several citation dates were associated with the same Wikipedia article, only the most recent date has been taken into account, thus discarding duplications. Given the diversity of existing journals and their varied scientific nature, we decided to filter only those journals indexed in Scopus. The extension of Scopus to include journals in the humanities could provide us with a new opportunity in the absence of a JCR for the A&HCI (Leydesdorff, de Moya-Anegón & Guerrero-Bote, 2010). Thus, we hoped to achieve two objectives: firstly, to guarantee that each reference corresponded to a valid scientific journal and, secondly, to obtain complementary information—such as the scientific category to which each article belonged. To do this, we used the Elsevier journal dataset in Cite Score Metrics,⁷ indexed in 2016, as our source of information. Thus, the references were linked to the entire collection of Scopus journals. The final dataset contained 179 329 Wikipedia articles with 784 209 references to 549 782 individual resources, mainly scientific articles.

The present study focuses on scientific articles belonging to all 3209 journals in Scopus under the All Science Journal Classification (ASJC) code "Arts and Humanities" (discipline). Every journal within this discipline is attached to one or more specialties (subcodes within Scopus). Once our dataset had been merged with the Scopus data, our final sample comprised references to 1717 journals (54% of the total in Scopus), including: 25 555 articles (14.25% of all Wikipedia articles citing articles in Scopus included in all disciplines) with 41 655 references (5.31%) to 32 245 resources (5.86%). The vast majority (99.25%) of articles in the final sample corresponded to Wikipedia in English, the language on which Altmeter.com is based. Only 0.75% of articles corresponded to other languages: Swedish (0.6%) and Finnish (0.15%). Figure 1 summarizes the process of collection and the evolution of sample size, as reported above.

⁷ <https://www.scopus.com/sources>

Figure 1. Process of collection and the evolution of sample size

	1	Download resources cited in Wikipedia as indexed in Altmetric.com	261 079 Wikipedia entries with 1 214 322 references
	2	Download ISSN codes of resources by using the Altmetric.com API	
Scopus [*]	3	Limitation of the sample to cited articles published in Scopus journals	179 329 Wikipedia entries with 784 209 references to 549 782 articles
Scopus [*]	4	Assignment of Scopus thematic category to entries and articles	
	5	Reduction of the sample to articles published in journals in the Humanities	25 555 Wikipedia entries with 41 655 references to 32 245 articles

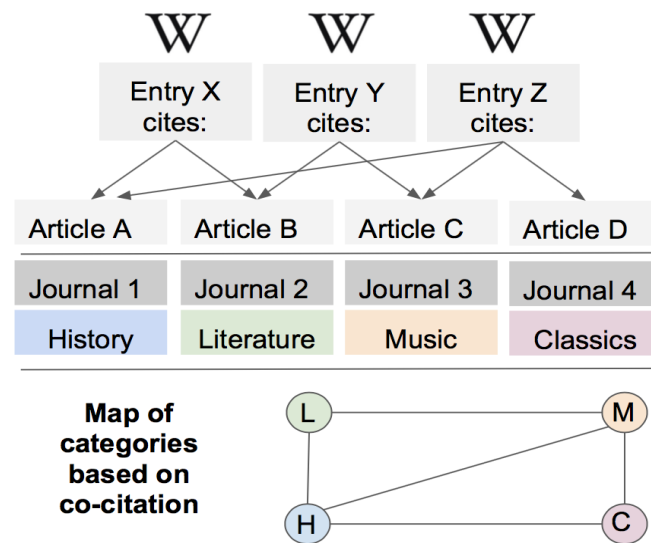
Descriptive indicators have been calculated (mean, median, mode, standard deviation and range), as well as those corresponding to the degree of centrality and intermediation.

2.2 Development of science maps

In classic bibliometrics, the source of information is the scientific article. A co-citation is established when one scientific article cites another two articles, creating a relationship between them that could be interpreted as a measure of similarity between the authors, journals or the categories to which they belong (McCain, 1990). In the present study, we have used the co-citations established by Wikipedia entries (Figure 2) to allow us to draw a map of co-cited journals. Of the 1717 journals represented in the sample, 1408 were co-cited in the 13 specialties in the Humanities included in the Scopus classification.

For our analysis of the journals and specialties, we pruned these data by eliminating relationships with fewer than 6 co-citations in order to facilitate their visualization and interpretation. The nodes isolated in this process were subsequently eliminated. Finally, due to the high co-citation value range, we used Pajek software to normalize them to a 0-to-1 scale, with the min-max normalization technique, which linearly transforms the values from the original range to another between the minimum value (0) and maximum value (1), conserving the relative differences.

Figure 2. Diagram of co-citations from entries in Wikipedia



Next, we divided the data by components in order to extract the largest subset, consisting of 163 nodes. Finally, for both journals and specialties networks obtained, we applied the Pathfinder algorithm (Vargas-Quesada, 2005) to prune them, creating as a result Pathfinder networks (PFNETs). We used the common configuration $r=\infty$ and $q=n-1$ that removes irrelevant links according to triangle inequalities to reduce the networks to their shortest path, that is, their minimum spanning tree. This way, nodes are connected only by their most important links. The parameter r defines the measure to calculate the path between nodes using the Minkowski distance and q the number of intermediate links to consider, where n is all the nodes. In our case, the links preserved between journals or specialties are the strongest co-citations. This technique has previously been used to map thematic domains in science (Moya-Anegón et al., 2004; White, 2003). As a result, two maps show how journals and specialties in the Humanities are linked to each other from the social perspective provided by Wikipedia. The tools used throughout this process were: Notepad++, to correct and prepare the data downloaded from Altmetric.com through regular expressions; Microsoft Access, to store and treat data and for information retrieval; Microsoft Excel, for descriptive statistical analysis; Pajek, to elaborate maps and conduct the centrality study; Gephi, to design the maps; and the programming language R, to download data from the API and for data processing (for example, to combine categories using colors).

Furthermore, we encountered the problem of latent co-citation arising because journals may be assigned to more than one specialty (Vargas-Quesada, 2005). We have resolved this in two different ways: firstly, when an article appeared in more than one specialty, we added an extra copy for each occurrence so that each copy was linked to a single specialty; we then removed the co-citations between the same specialties or articles. Secondly, to visualize journals, the specialties were combined under a single label.

3. Analysis and results

3.1 General data and annual evolution

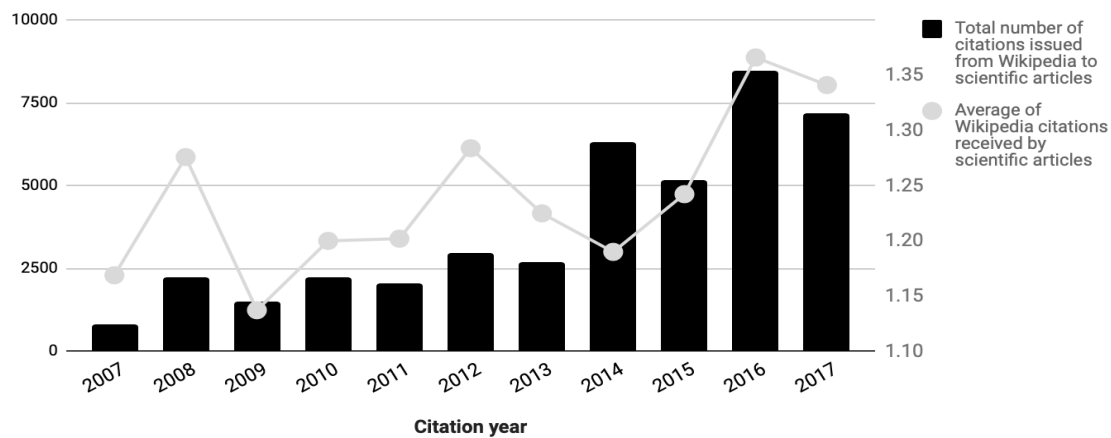
Table 1 shows descriptive statistics of the Wikipedia article references to scientific articles published in Scopus journals, and of the citations received by these scientific articles both for the whole of Wikipedia (global) and for the Humanities discipline. Note that we only take account of Wikipedia articles that include at least one citation to a scientific journal and scientific journals referenced at least once in Wikipedia. Hence, the minimum mean for references is 1. In total, 784 209 citations to scientific articles in all disciplines have been identified; of these 41 655 citations (5%) correspond to works in the Humanities. More specifically, 25 555 individual Wikipedia entries have been compiled, citing 32 245 independent articles. If we focus on the citations of scientific articles found in Wikipedia entries, we find a considerable difference between the global average for all disciplines (4.37) and that for the Humanities (1.63). In addition, there is greater homogeneity in terms of the average number of citations that articles receive: 1.42, globally, versus 1.29, for the Humanities.

Table 1. Descriptive statistical analysis of the distribution of references and citations in Wikipedia articles globally and for the Humanities

References to scientific articles included in Wikipedia entries*	Global	Humanities
Mean	4.37	1.63
Median	2	1
Standard deviation	8.25	1.76
Range	440	54
Total entries with at least 1 reference	179 329	25 555
Total citations in Wikipedia	784 209	41 655
Citations of scientific articles received from Wikipedia*	Global	Humanities
Mean	1.42	1.29
Median	1	1
Standard deviation	10.59	1.23
Range	5067	106
Total articles cited	549 782	32 245

If we depict the annual evolution of the Humanities, the number of citations has been especially dense since 2014: the period 2007-2013 saw some 2500 citations annually; however, since 2014, this has increased to around 7500 citations per year. The most active year was 2016 with 8464 citations. The average number of citations received per scientific article has also shown a positive growth trend, reaching its highest level in 2016 (mean 1.37) and 2017 (mean 1.37).

Figure 3. Annual evolution of the number of citations included in Wikipedia and the average number of citations received per article in the Humanities during the period 2007-2017



3.2 Analysis of specialties in the Humanities

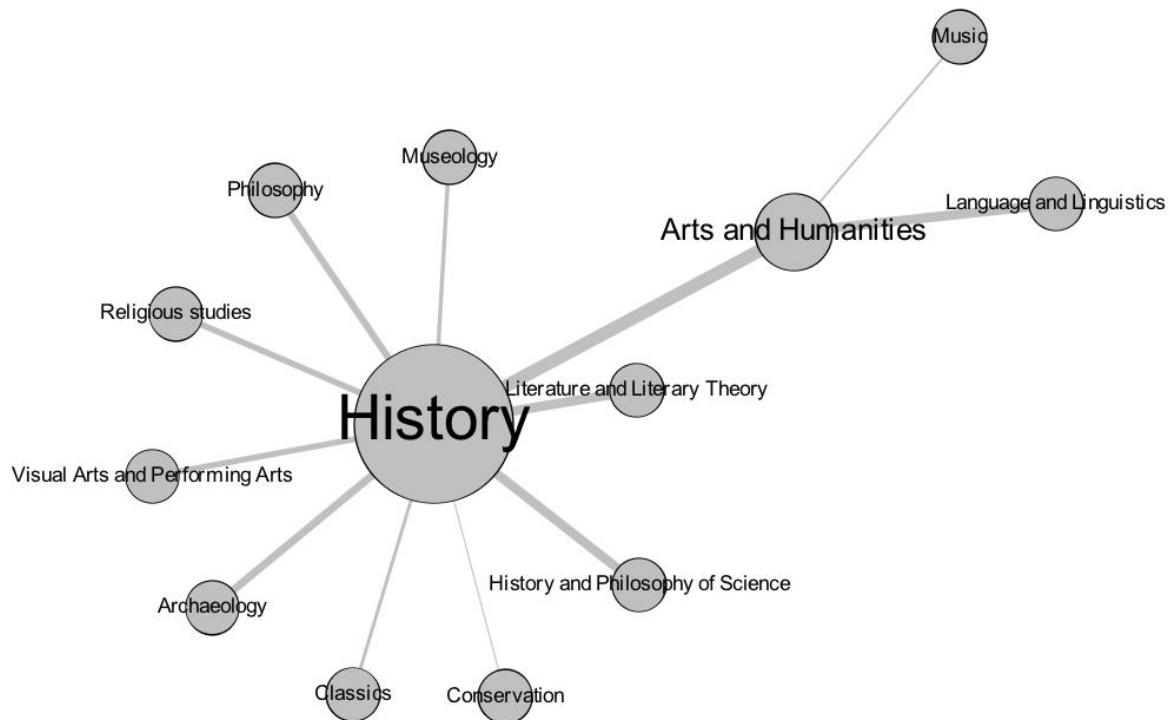
Table 2 shows the Scopus classification specialties in the Humanities, allowing us to identify those that receive most attention in Wikipedia. The most outstanding, at a considerable distance from the rest, is *History*, which concentrates the largest number of single journals cited (531), scientific articles cited (11 661) and total citations (15 969). The specialties *Language and Linguistics* and *History & Philosophy of Science* stand out in terms of the number of citations (without considering the miscellaneous category *Arts & Humanities*). The *Museology* category, despite being smaller, receives higher average citations per article (1.43). However, the average number of citations per article is generally quite homogeneous, ranging between 1.18, corresponding to *Literature and Literary Theory*, and the aforementioned 1.43, corresponding to *Museology*. Among the specialties that receive less attention are *Classics* and *Conservation*, which account for only 1.4% and 0.4%, respectively, of the total number of citations in Wikipedia.

Table 2. Citation indicators of journals and articles referenced in Wikipedia for specialties in the Humanities

	No. of journals cited in Wikipedia indexed in Scopus		No. scientific articles cited in Wikipedia	Total number of citations received in Wikipedia		Average number of citations in Wikipedia received by article
Archeology	111	4.9%	2026	2785	5.2%	1.37 ± 1.03
Arts and Humanities	266	19%	7881	10 034	18.7%	1.27 ± 0.86
Classics	37	1.4%	589	744	1.4%	1.26 ± 0.85
Conservation	18	0.3%	144	188	0.4%	1.30 ± 1.42
History	531	28.1%	11 661	15 969	29.8%	1.36 ± 1.76
History and Philosophy of Science	90	8.5%	3524	4574	8.5%	1.29 ± 0.88
Language and Linguistics	261	9.6%	3990	4796	9%	1.20 ± 0.65
Literature and Literary Theory	282	7.5%	3140	3729	7%	1.18 ± 0.68
Museology	17	1.9%	811	1166	2.2%	1.43 ± 2.03
Music	65	2.8%	1194	1473	2.8%	1.23 ± 0.75
Philosophy	211	6.2%	2588	3147	5.9%	1.21 ± 0.66
Religious studies	170	4.2%	1723	2190	4.1%	1.27 ± 0.87
Visual Arts and Performing Arts	188	5.3%	2197	2719	5.1%	1.23 ± 1.32

Figure 4 shows the co-citation map for specialties in the Humanities after editing the data following the application of the Pathfinder algorithm. In this map, the thickness of the edges indicates the degree of co-citation. The size of the nodes represents the number of articles within the specialty that establish a co-citation. Note that strong connections seldom occur between categories with a highly homogeneous co-citation pattern in which *History* is the common factor. This category stands in a highly important position as it is related to 11 specialties, showing the strongest links with the categories of *Literature and Literary Theory*, *History and Philosophy of Science* and the miscellaneous *Arts and Humanities*. The only two specialties not linked to *History* are *Music* and *Language and Linguistics*, directly connected to *Arts and Humanities*.

Figure 4. Co-citation map of specialties in the Humanities from the co-citations received from Wikipedia entries during the period 2007-2017 using the Pathfinder algorithm



3.3 Analysis of journals in the Humanities

Table 3 lists the first 25 journals ordered according to the number of citations received from Wikipedia. We would conclude that these are among the publications with higher social use on this platform. The journal that receives the highest number of citations (869) is *Annals of the New York Academy of Sciences*. This is a multidisciplinary journal, founded in 1823, that publishes on biomedicine and biology, but also on philosophy and anthropology. The profile for the remaining journals is not homogeneous, including topics like: History (*English Historical Review*, *American Historical Review*), Anthropology (*Current Anthropology*), Linguistics (*International Journal of American Linguistics*) or multidisciplinary topics such as sex (*Archives of Sexual Behavior*). Note that none of these journals is published in open access, in marked contrast to the open nature of the encyclopedia. It is also remarkable that 18 of these publications are high impact journals because they are among the top 10 of those with the greatest impact in their specialty according to the Scopus Journal Metrics. Therefore, we can conclude that Wikipedia editors consider that journals with higher impact on the scientific community are also more reliable sources of information.

Table 3. Most cited journals in the Humanities in Wikipedia during the period 2007-2017

		No. of citations received in Wikipedia	No. of articles cited in Wikipedia	Average number of citations per article	Open Access journal?	Top journal? **
1	Annals of the New York Academy of Sciences	869	698	1.24	No	Yes
2	Journal of the Acoustical Society of America	621	519	1.20	No	No
3	Archives of Sexual Behavior	591	373	1.58	No	No
4	Isis	502	357	1.41	No	Yes
5	English Historical Review	499	320	1.56	No	Yes
6	American Historical Review	444	378	1.17	No	Yes
7	Current Anthropology	416	271	1.54	No	Yes
8	Journal of Archaeological Science	396	270	1.47	No	Yes
9	Quaternary Science Reviews	355	260	1.37	No	Yes
10	Social Science and Medicine	333	267	1.25	No	Yes
11	American Museum Novitates	333	128	2.60	No	Yes
12	Journal of the American Oriental Society	322	215	1.50	No	No
13	Bulletin of the School of Oriental and African Studies	316	219	1.44	No	No
14	Cognition	296	221	1.34	No	Yes
15	Intelligence	291	203	1.43	No	Yes
16	Speculum	287	215	1.33	No	Yes
17	Journal of Asian Studies	275	196	1.40	No	Yes
18	International Journal of American Linguistics	271	218	1.24	No	No
19	Medical History	270	187	1.44	No	Yes
20	Language	255	190	1.34	No	No
21	Economic History Review	243	117	02.08	No	Yes
22	Journal of American History	227	191	1.19	No	No
23	American Antiquity	224	164	1.37	No	Yes
24	Journal of Sex Research	218	144	1.51	No	Yes
25	Journal of African History	215	147	1.46	No	Yes

**** Top is defined as being among the 10% most cited journals in the Scopus/Elsevier Score Metrics categories**

Figure 5 shows the co-citation map between journals. The scientific journals in the Humanities cited in Wikipedia have been grouped into 10 clusters each of which is represented by a color. If we first consider the specialties, not all the clusters are homogeneous as they are composed of journals from different specialties. However we should distinguish between clusters with a lower degree of heterogeneity (0, 1, 3 or 4) and more heterogeneous clusters (5, 8 or 9). Clusters 6 and 9 are identified at the center of the network with a mediating role and connecting specialties. In these two clusters we find multidisciplinary journals belonging mainly to three areas, *History*, *Archeology* and *Linguistics*. Cluster 9 connects with cluster 0, which includes journals from *Language and Linguistics*, and cluster 1, which includes *Philosophy of Science*. Cluster 6 connects with clusters 5 and 2—formed by *History* and *Philosophy of Science*—and clusters 7 and 8—also formed in the main by history journals. To summarize, in this representation of knowledge from Wikipedia, the upper part (Clusters 0, 9, 1 and 3) represents *Language and Linguistics* and *History and Philosophy of Science*; the lower part is dominated by *History* and *Archeology*, although it is closely related to other specialties.

Figure 5. Map of co-citation in Wikipedia of scientific journals in the Humanities grouped according to similarity clusters. Each node corresponds to a journal and the color indicates the cluster to which it belongs



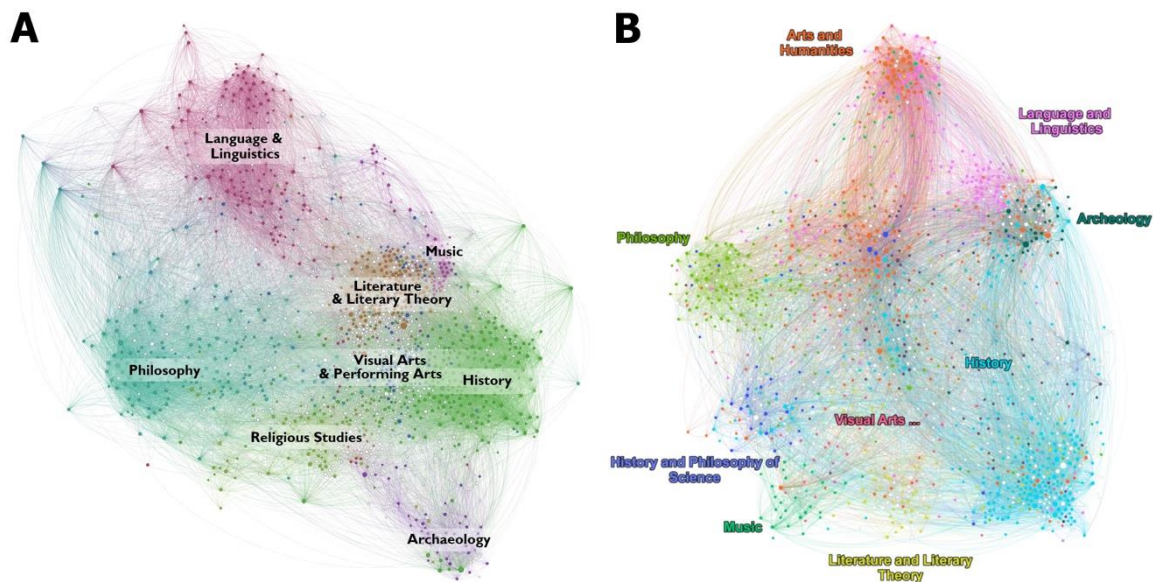
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Arts & Humanities: 59% Language & Linguistics: 36%	History: 46% History & Phil. Science: 46%	Arts & Humanities: 62% History: 12% History & Phil. Science: 12% Philosophy: 12%	Philosophy: 80% History & Phil. Science: 14%	Archaeology: 60% Arts & Humanities: 40%
Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
Arts & Humanities: 73% History & Phil. Science: 13% History: 6.7% Literature & Lit. Theory: 6.7%	Archaeology: 35% History: 32% Arts & Humanities: 31%	History: 75% Arts & Humanities: 9% Literature & Lit. Theory: 7.27	History: 73% Archaeology: 10% Language & Linguistics: 6% Religious studies: 4%	Language & Linguistics 37% History 28% Arts & Humanities 12% Religious studies 12% Philosophy 6%

3.4 Comparison with other studies

Following we compare our results with those from similar studies. For instance, Richardson (2013) used the same database and thematic categorization (Figure 6A). To compare our results with his we have replicated the same methodology, but applying it to the Wikipedia data with the aim of getting a similar map. We apply Richardson’s methodology to our data. The initial network of 1408 co-cited journals in Wikipedia entries has been pruned to extract only the main component, made of 1388 nodes and 12 121 edges. Nodes have been distributed applying the OpenOrd layout algorithm to better identify the communities. As a result, the

network (Figure 6B) shows how the position of *History* is kept with a weight and role more determinant than the rest of specialties. *History* is positioned in the center of the network and is highly connected to other specialties (*Archeology* and *History and Philosophy of Science*). It is important to highlight the secondary role of *Literature and Literary Theory*, with a much smaller size, which is relegated to the periphery of the network and loosely connected. The main role of *History* and the secondary role of *Literature and Literary Theory* are the principal differences found in relation with the studies of Richardson (2013) and Leydesdorff et al. (2011).

Figure 6. Comparison of networks based on: (A) a journal citation map covering 1570 journals from the Arts & Humanities (Richardson, 2013), and (B) the main component of the network of co-citation humanities journals in Wikipedia



Colours correspond to journals specialties, the journals with more of one area are white. In Figure 6B due to the low presence of the journals with the unique area of *Classics*, *Conservation* and *Museology*, they aren't tagged in the network, while *Religious studies* are too disseminated for it

Hence, there are evident differences between the two studies. Following we provide a plausible explanation for them. The maps generated are different due to the different coverage size and proportion of Scopus and Wikipedia. As observed in table 4, *History* accumulates 28% of the total number of Wikipedia articles, while in Scopus, *History* represent only 11% of the database. This fact contrasts, for example, with the case of *Literature and Literary Theory*, which has 7.57% papers of Wikipedia while in Scopus represents 14.18% of the database. This affects the positioning and degree of these specialties in the two networks. Furthermore, it evidences that social interest does not always align with scientific interest.

Table 4. Coverage of humanities specialties in Wikipedia and Scopus in function to the number of articles and cites

	Coverage Wikipedia		Coverage Scopus	
	% of articles	% of citations	% of articles	% of citations**
History	28.12%	29.84%	17.39%	11.70%
Arts and Humanities	19.01%	18.75%	15.41%	35.31%
Language and Linguistics	9.62%	8.96%	11.89%	16.01%
History and Philosophy of Science	8.5%	8.55%	3.82%	8.20%
Literature and Literary Theory	7.57%	6.97%	14.18%	2.92%
Philosophy	6.24%	5.88%	10.76%	7.78%
Visual Arts and Performing Arts	5.30%	5.08%	8.83%	2.43%
Archaeology	4.89%	5.20%	5.15%	10.10%
Religious studies	4.16%	4.09%	7.07%	2.66%
Music	2.88%	2.75%	2.28%	1.17%
Museology	1.96%	2.18%	0.72%	0.45%
Classics	1.42%	1.39%	1.32%	0.26%
Conservation	0.35%	0.35%	1.17%	1.01%
**Scopus citation data from the CiteScore 2016				

4. Conclusions

In the present study, we have extrapolated the methodology for representing science on the basis of co-citation maps to a different context. Traditionally, science maps have been drawn up from scientific articles, using large databases such as the Web of Science or Scopus and demonstrating their validity as a means of establishing relationships between areas and of determining the structure of science from the scientific knowledge itself (Noyons & Van Raan, 1998). In the present study, these co-citation techniques have been extrapolated to a digital, social environment—Wikipedia—illustrating the use of articles as a source of citizen information, and the vision of the structure—from a social point of view—of scientific knowledge. More specifically, a vision of the Humanities has been shown from Wikipedia, the main encyclopedic project, based on collaborative and open principles.

The mapping technique has been successfully extrapolated to create co-citation maps based on categories pruned by applying the Pathfinder algorithm proposed by Moya-Anegón et al. (2004), showing that social platforms can be used to offer an alternative vision of scientific knowledge. However, it should be noted that the methodology used, which combines various sources (Altmetric.com, Wikipedia and Journal Metrics by Elsevier), has some limitations. For example, we have only taken account of scientific articles since they are the only resources in the Journal Metrics dataset provided by Elsevier; books or chapters of special relevance in the Humanities are excluded (Torres-Salinas et al., 2013). This problem is present in other classical approaches that are limited to scientific journals (Leydesdorff, Hammarfelt & Salah, 2011). This explains why the data sources we have used make it difficult to adapt the methodology. In our case, the scope of the results was limited because we were unable to find a category similar to that of scientific journals that would allow us to identify the context in

which the citations appeared. In addition, we would like to emphasize that these results are limited to the English language Wikipedia.

Science maps based on categories pre-assigned by databases always offer a biased view since journals and studies do not always belong to the category assigned by the database (Rafols, Porter & Leydesdorff, 2010). An obvious example in the classification of the Scopus ASJC is the use of insignificant generic categories such as *Arts and Humanities (miscellaneous)* and *General Arts and Humanities*, which we had to unify under the label *Arts and Humanities*. One further problem is the fact that the classification of a journal may not correspond to the classification of the articles therein. Therefore, we should consider the fact that some of the co-citations analyzed are probably not really from the humanities as a cognitive limitation.

Despite its limitations, this study has served to illustrate the use of scientific information in a social context; for example, we have determined that the mean of works in the Humanities cited in Wikipedia is lower than the general mean including all the areas. Also, only 5% of the 784 209 citations in Wikipedia of scientific articles in Scopus correspond to articles in journals in the Humanities. This could suggest the need to strengthen the visibility of work in the Humanities so that it achieves greater social impact. It is well worth noting that since 2013 the annual evolution of citations in the Humanities has risen from an average of 2500 to 7500 per year. Also, despite the open philosophy of Wikipedia—a platform that works thanks to the legal support provided by Creative Commons licenses—the data indicate that of the 25 most cited journals on Wikipedia, none is open access, while more than 70% are among the 10% most cited in their category.

In relation to the maps, if we look at the specific categories within the Humanities, *History* is presented as the main knowledge Domain from a social point of view. It concentrates the largest number of citations of individual journals (531) and scientific articles (11 661), and the highest number of total citations (15 969). Co-citation analysis also places it in a central position, connecting specialties. Important connections between specialties have been determined, such as those between *History* and *Archeology* (Cluster 6) and *History* and *Language and Linguistics* (Cluster 9), around which the other specialties are articulated. *Philosophy* and *Philosophy of Science* are less well represented and occupy more peripheral positions than other specialties (for example Clusters 1, 2 and 3).

If we relate this social vision of science with more traditional bibliometric studies (Richardson, 2013; Leydesdorff, Hammarfelt & Salah, 2011), we encounter interesting differences. Richardson (2013), taking data from Scopus citations in 1570 journals in the *Arts & Humanities*, formulated a map in which the various themes are grouped around *Literature and*

Arts, which occupies a central position. They are closely connected with *History*, that does not occupy as central a position as in our analysis.

On the other hand, Leydesdorff, Hammarfelt and Salah (2011) used Web of Science data to map relationships between 1157 *Arts & Humanities Citation Index* journals in 2008. They observed that *Literature* continued to occupy a central position connecting categories such as *Music, Philosophy, Linguistics, Art* and *History*. *History* in this study was subdivided into three parts: *American History, History and Philosophy of Science* and *History*, properly speaking. Although it was more centrally positioned than in Richardson's study (2013), it was far from the nuclear role it occupies in our research. This is an indicator of how, from a social point of view, *History* is the key specialty that connects with other areas of humanistic knowledge and may reflect how the consumption of information and its relationships can differ in a social context by comparison with a scientific context.

Despite the comparison, there are some limitations that may explain the differences found. For instance, there is a methodological difference between our co-citation analysis and the direct citation analysis used by Richardson (2013). Likewise, Leydesdorff et al. (2011) use a different data source, Web of Science instead of Scopus.

Thus, regardless of the algorithm used, it is clear that Wikipedia offers a different view of science from the traditional maps, which represent the readers and editors interests, but it does not necessarily coincide, nor should it coincide, with the scientists and specialists vision.

To conclude, firstly, a reproducible methodology has been proposed to map scientific knowledge in Wikipedia through bibliometric techniques while, secondly, we have been able to analyze how the "global brain" perceives scientific knowledge and the interrelationship between specialties, offering a new vision of science as a counterpoint to the traditional maps. This methodology, based on the combination of sources such as Altmetric and Scopus, opens the door to other analyses drawing on sources such as Twitter, the News (news feeds) or report (policy feeds) that reflect the social vision of science from different perspectives (social, political, the mass media, among others).

References

- Costas, R., de Rijcke, S., & Marres, N. (2017). Beyond the dependencies of altmetrics: Conceptualizing 'heterogeneous couplings' between social media and science. In *The 2017 Altmetrics Workshop*. Retrieved from http://altmetrics.org/wp-content/uploads/2017/09/altmetrics17_paper_4.pdf

- DiBona, C., Cooper, D., & Stone, M. (2006). *Open Sources 2.0: The Continuing Evolution*. O'Reilly Media.
- Fallis, D. (2008). Toward an Epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(May), 1662–1674. <https://doi.org/10.1002/asi.20870>.
- Fuchs, C. (2008). *Internet and society: Social theory in the information age*. Routledge. New York: Routledge.
- Kousha, K., & Thelwall, M. (2016). 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>.
- Leydesdorff, L., Carley, S., & Rafols, I. (2012). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, 94(2), 589-593. <https://doi.org/10.1007/s11192-012-0784-8>.
- Leydesdorff, L., de Moya-Anegón, F., & Guerrero-Bote, V. P. (2010). Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI. *Journal of the American Society for Information Science and Technology*, 61(2), 352-369.
- Leydesdorff, L., Hammarfelt, B., & Salah, A. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *Journal of the Association for Information Science and Technology*, 62(12), 2414-2426. <https://doi.org/10.1002/asi.21636>.
- Leydesdorff, L., & Nerghe, A. (2017). Co-word maps and topic modeling: A comparison using small and medium-sized corpora (N < 1,000). *Journal of the Association for Information Science and Technology*, 68(4), 1024-1035.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American society for information science*, 41(6), 433-443. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<433::AID-ASI11>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<433::AID-ASI11>3.0.CO;2-Q).
- Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Muñoz-Fernández, F. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129-145. <https://doi.org/10.1023/B:SCIE.0000037368.31217.34>.
- Noyons, E. C., & Van Raan, A. F. (1998). Advanced mapping of science and technology. *Scientometrics*, 41(1-2), 61-67. <https://doi.org/10.1007/BF02457967>.
- O'Reilly, T. (2005). What is web 2.0? Design patterns and business models for the next generation of software. <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html> Accessed 6 March 2018.
- Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493(7431), 159. <http://doi.org/10.1038/493159a>.

- Piwowar, H. (2013). Introduction altmetrics: What, why and where? *Bulletin of the American Society for Information Science and Technology*, 39(4), 8–9. <https://doi.org/10.1002/bult.2013.1720390404>.
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. <http://altmetrics.org/manifesto/> Accessed 22 February 2018.
- Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for information Science and Technology*, 61(9), 1871-1887. <https://doi.org/10.1002/asi.21368>.
- Richardson, M. (2013). Mapping the multidisciplinary of the Arts & Humanities. *Research Trends*, 32, 15-19.
- Romero-Frías, E., & Vaughan, L. (2010). European political trends viewed through patterns of Web linking. *Journal of the American Society for Information Science and Technology*, 61(10), 2109-2121. <https://doi.org/10.1002/asi.21375>.
- Salah, A. A., Gao, C., Suchecki, K., & Scharnhorst, A. (2012). Need to categorize: A comparative look at the categories of universal decimal classification system and Wikipedia. *Leonardo*, 45(1), 84-85. http://doi.org/10.1162/LEON_a_00344.
- Serrano-López, A. E., Ingwersen, P., & Sanz-Casado, E. (2017). Wind power research in Wikipedia: Does Wikipedia demonstrate direct influence of research publications and can it be used as adequate source in research evaluation?. *Scientometrics*, 112(3), 1471-1488. <https://doi.org/10.1007/s11192-017-2447-2>.
- Silva, F. N., Viana, M. P., Travençolo, B. A. N., & Costa, L. D. F. (2011). Investigating relationships within and between category networks in Wikipedia. *Journal of informetrics*, 5(3), 431-438. <https://doi.org/10.1016/j.joi.2011.03.003>.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4), 265-269. <https://doi.org/10.1002/asi.4630240406>.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Teplitskiy, M., Lu, G., & Duede, E. (2016). 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. (2009). Introduction to webometrics: Quantitative Web Research for the Social Sciences. *Synthesis lectures on information concepts, retrieval, and services*, 1(1), 1-116. <https://doi.org/10.2200/S00176ED1V01Y200903ICR004>.
- Torres-Salinas, D., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in web 2.0. *Comunicar*, 21(41), 53–60. <https://doi.org/10.3916/C41-2013-05>.

- Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J., & García, J. A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of informetrics*, 7(2), 412-424. <https://doi.org/10.1016/j.joi.2013.01.004>.
- Tramullas, J. (2016). Competencias informacionales básicas y uso de Wikipedia en entornos educativos. *Gestión de La Innovación En Educación Superior*, 1, 79–95. <http://eprints.rclis.org/29624/1/16-72-1-PB.pdf> Accessed 8 March 2018.
- Vargas-Quesada, B. (2005). *Visualización y análisis de grandes dominios científicos mediante Redes Pathfinder (PFNET)*. Granada: Universidad de Granada.
- Vaughan, L., Kipp, M. E., & Gao, Y. (2007). Why are websites co-linked? The case of Canadian universities. *Scientometrics*, 72(1), 81-92. <https://doi.org/10.1007/s11192-007-1707-y>.
- Vaughan, L., & Romero-Frías, E. (2010). Web hyperlink patterns and the financial variables of the global banking industry. *Journal of Information Science*, 36(4), 530-541. <https://doi.org/10.1177/0165551510373961>.
- White, H. D. (2003). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.
- Zuccala, A. (2006). Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...?. *Journal of the American Society for Information Science and Technology*, 57(11), 1487-1502. <https://doi.org/10.1002/asi.20468>.

Science through Wikipedia: a novel representation of open knowledge through co-citation networks

Wenceslao Arroyo-Machado^{1,2}, Daniel Torres-Salinas^{1,2,3,*},
Enrique Herrera-Viedma⁴ and Esteban Romero-Frías^{1,5}



¹Medialab UGR, University of Granada, Granada, Spain

²Department of Information and Communication, University of Granada, Faculty of Communication and Documentation, Granada, Spain

³EC3metrics spin off, University of Granada, Granada, Spain


⁴Department of Computer Science and Artificial Intelligence, University of Granada, Faculty of Communication and Documentation, Granada, Spain

⁵Department of Accountancy and Finance, University of Granada, Faculty of Economics and Business, Granada, Spain

*Corresponding author: torressalinas@ugr.es

Journal

PLOS ONE

1932-6203 

Index

SCIE – Q2

DOI

10.1371/journal.pone.0228713

Data

None

Version

Published

References

Vancouver

Funding



Abstract

This study provides an overview of science from the Wikipedia perspective. A methodology has been established for the analysis of how Wikipedia editors regard science through their references to scientific papers. The method of co-citation has been adapted to this context in order to generate Pathfinder networks (PFNET) that highlight the most relevant scientific journals and categories, and their interactions in order to find out how scientific literature is consumed through this open encyclopaedia. In addition to this, their obsolescence has been studied through Price index. A total of 1 433 457 references available at Altmetric.com have been initially taken into account. After pre-processing and linking them to the data from Elsevier's CiteScore Metrics the sample was reduced to 847 512 references made by 193 802 Wikipedia articles to 598 746 scientific articles belonging to 14 149 journals indexed in Scopus. As highlighted results we found a significant presence of “Medicine” and “Biochemistry, Genetics and Molecular Biology” papers and that the most important journals are multidisciplinary in nature, suggesting also that high-impact factor journals were more likely to be cited. Furthermore, only 13.44% of Wikipedia citations are to Open Access journals.

Citation

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>

Introduction

Since its creation in 2001, Wikipedia has become the largest encyclopedic work human beings have ever created thanks to the collaborative, connected opportunities offered by the Web. Probably one of the most significant examples of Web 2.0 [1], Wikipedia represents a success story for collective intelligence [2]. With more than 170 editions, the English language version accounted for 5.5 million entries in January 2018 (approximately 11.7% of the entire encyclopedia). Given that worldwide Wikipedia is a top ten website in terms of traffic—according to Alexa (<https://www.alexa.com/topsites>, consulted on July 24, 2019)—and is one of the preferred results provided by search engines, it has become an outstanding tool for the dissemination of knowledge within a model based on openness and collaboration.

Perhaps Wikipedia's most important achievement has been to challenge traditional epistemologies based on authorship and authority and move towards a more social, distributed epistemology [3]. Wikipedia is therefore the result of a negotiation process that provides us with a representation of knowledge in society, offering tremendous research opportunities. For instance, some authors have studied the discursive constructions of concepts such as globalization [4] or historical landmarks like the 9/11 attacks [5]. The process of negotiation behind an article is often driven by the principles of verifiability and reliability in relation to the sources supporting the statements made. Specialized publications are among the preferred sources of reference (https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources, consulted on July 24, 2019), mainly in the form of scholarly material and prioritizing academic and peer-reviewed publications, as well as scholarly monographs and textbooks.

Consequently, the social construction of knowledge on Wikipedia is explicitly and intentionally connected to scholarly research published under the peer-review model. This has offered us the opportunity to investigate how Science and Wikipedia interrelate. Although Wikipedia is not a primary source of information, some studies have examined citations of Wikipedia articles [6,7]. Moreover, numerous studies have analyzed how Wikipedia articles cite scholarly publications because contributors are strongly recommended to do so by the encyclopedia itself. Studies have focused on the analysis of reference and citation patterns in specific areas of knowledge [8], on exploring Wikipedia's value as a source when evaluating scientific activity [9], or on Wikipedia's role as a platform that promotes open access research [10].

Furthermore, some studies undertaken within the last decade could be said to be framed within the Altmetric perspective because they have used indicators extracted from the social media to measure dimensions of academic impact [11,12,13]. Wikipedia references to scientific articles can provide highly valuable altmetric information given that the inclusion of references is not a trivial activity and is usually subject to community scrutiny. For instance,

the Altmetric Attention Score—an indicator created by Altmetric.com—gives this type of citation a high value (3) that is higher than mentions on Facebook (0.25) or Twitter (1), but lower than references to blogs (5) and news feeds (8).

Networks have also been used for knowledge representation in order to visualize differences between the Universal Decimal Classification category structure and that generated by Wikipedia itself [14], to generate automated taxonomies and visualizations of scientific fields [15], and to show connections between articles [16]; furthermore, studies based on the complex networks approach have also been reported [17]. One way to address knowledge representation from a bibliometric perspective is through the use of co-citations [18], an approach that uses references in common received from a third document as a proxy for similarity between two scientific documents. Co-citations have been used to observe similarities between words [19] or areas of knowledge [20].

From an Altmetric perspective, the concept of co-citation was transferred to the online world giving rise to co-link analysis [21], where documents are replaced by webpages or websites, and citations are replaced by links. Co-link analysis has successfully mapped scientific knowledge [22] and analyzed fields such as universities [23], politics [24] or business [25].

These different concepts were recently combined and applied to Wikipedia by Torres-Salinas, Romero-Frías and Arroyo-Machado [26] and tested in the field of the Humanities by mapping specialties and journals. The present study uses Wikipedia to draw a social representation of scientific knowledge and the areas into which it is divided. After collecting all the references in Wikipedia, we concluded that only 5.49% correspond to the Humanities (see Table S1). Therefore, in the present study we take the same approach in investigating Science as a whole, including the Humanities. We seek to achieve the following objectives:

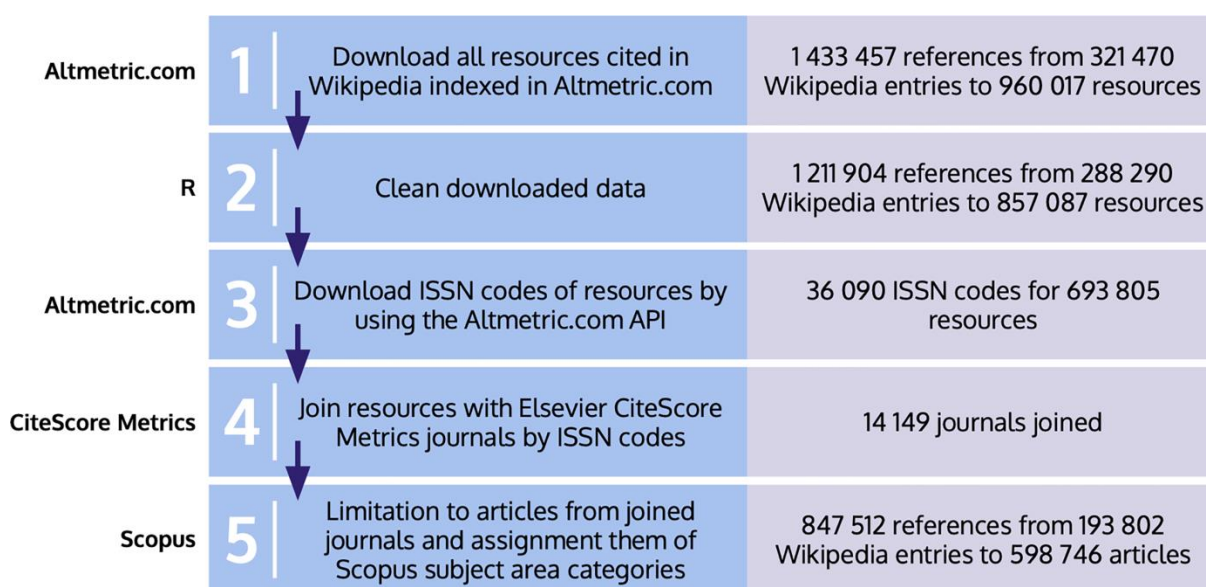
- To apply co-citation analysis to all the articles referenced in Wikipedia in order to validate their usefulness in analyzing open knowledge platforms;
- To offer a general portrait of the use of scientific literature published in journals through the analysis of references and their obsolescence. Thus, we hope to be able to describe the consumption of scientific information by the Wikipedia community and detect possible differences between fields; and lastly, as the nuclear objective of our paper
- To discover the different visions offered by Wikipedia by using co-citation networks at different levels of aggregation: 1) journal co-citation maps 2) main field co-citation maps 3) field co-citation maps. Through these representations we intended to obtain a holistic view of how scientific articles in Wikipedia are used and consumed.

Materials and Methods

Information sources and data pre-processing

The main source of information in this study was Altmetric.com, one of the most important platforms gathering altmetric data about scientific papers. The total volume of references to scientific papers made by Wikipedia articles was downloaded on April 11, 2018. This amounted to 1 433 457 references published between October 15, 2004 and April 10, 2018, citing 960 017 discrete resources. Initially we pre-processed the data with R in order to clean it up. This involved correcting errors to facilitate links with other data sources, eliminating duplicate references, and deleting references lacking the data needed for our study, such as publication dates. As a result, the total number of references fell to 1 211 904 citing 857 087 individual resources. ISSNs corresponding to these resources were collected using the Altmetric API. A total of 36 090 ISSNs corresponding to 693 805 scientific articles were obtained. In addition, we used Elsevier's CiteScore Metrics (with data updated to February 6, 2018) to link each scientific article to its source through journal identifiers and thus obtain additional information. The references were linked to Elsevier's CiteScore Metrics's entire collection. Fig 1 summarizes this process.

Fig 1. Methodological process of collecting the massive dataset of papers referenced in Wikipedia and assigning them to different scientific categories.



The Scopus ASJC (All Science Journal Classification) offered in the CiteScore Metrics collection has been used to attribute areas, main fields, and fields to the scientific articles being studied.

To use Scopus (https://service.elsevier.com/app/answers/detail/a_id/12007/supporthub/scopus/) terminology, we

would say that the ASJC identifies four major areas each of which includes several *Subject Area Classifications* (termed *main fields* in our study). Given that *multidisciplinarity* is a common main field in each of the four areas, we have decided to include this category as a main area as well. As a result, there are 27 main fields (*subject area classifications* in Scopus terminology) and 330 fields (*fields*) within five main areas (*subject areas*): namely "Health Sciences", "Life Sciences", "Physical Sciences", "Social Sciences & Humanities", and "Multidisciplinary". Hence, the final sample consists of 847 512 references included in 193 802 Wikipedia entries, citing 598 746 individual scientific articles from 14 149 journals. This process of attribution enabled us to identify references to scientific articles and, at a more aggregated level, references to journals, fields and main fields, giving rise to three different co-citation networks.

Statistical analysis

As part of the descriptive statistics, the mean, median, standard deviation and interquartile range have been calculated for the number of references made by Wikipedia and the citations received by the scientific articles, as well as for the dates of citation and publication of the papers, at all the different levels under study. We would emphasize the fact that in our dataset all Wikipedia entries include at least one reference to scientific papers and all articles and journals included have been cited at least once by Wikipedia articles. Furthermore, the obsolescence of the scientific references has also been calculated using the Price index [27], which has been applied to intervals of up to 5, 10, 15 and 20 years, the entire dataset, and by scientific fields. The Price index refers to the percentage of publications cited not older than a specific number of years and is a means of showing the level of immediacy of publications cited, which differs according to the scientific area [28]. Similarly, the distributions of citations between Wikipedia and Scopus have been compared using the citation value recorded by Elsevier's CiteScore Metrics—, which corresponds to the sum of citations in 2016 to articles published between 2013 and 2015—, and by Wikipedia, and adjusted to allow for this limitation. Finally, the distribution of journal citations from Wikipedia has been analyzed, to determine whether it fits power law and log-normal distributions using the `powerlaw` package [29].

Analysis of co-citation networks

Co-citation networks, bibliographic coupling and direct citations are some of the most significant bibliometric networks we can use to map citations from Wikipedia entries; of these, co-citation networks are the most popular in research [30,31]. If we take into account other types of network such as co-author and co-word, the aforementioned three methods show a high degree of similarity [32]. Within the field of altmetrics, the concepts of co-citation and coupling have both been adapted [33], but co-citations offer more varied alternatives [34].

Furthermore, they are of special interest as they have been identified as capable of enhancing transdisciplinarity [35]. Hence, we have generated co-citation maps at the level of journal, field and main field.

The Pathfinder algorithm [36] has been applied as a pruning method following a common configuration ($r=\infty$, $q=n-1$) that reduces the networks to a minimum covering tree. This algorithm—successfully applied in the field of Library and Information Sciences [26, 37]—keeps only the strongest co-citation links between all pairs of nodes and offers a diaphanous view of large networks. Given the huge amount of co-citations, especially between journals, we use this technique to prune them in order to make the networks more explanatory. Since it is applied to values in relation to distances, the inverse value of the co-cites has been used in our analysis. Local measures of proximity, betweenness and eigenvector centrality have also been calculated. In the case of journals, the data has undergone a second pruning to eliminate those entries with a co-citation degree lower than 50.

Results

General Description

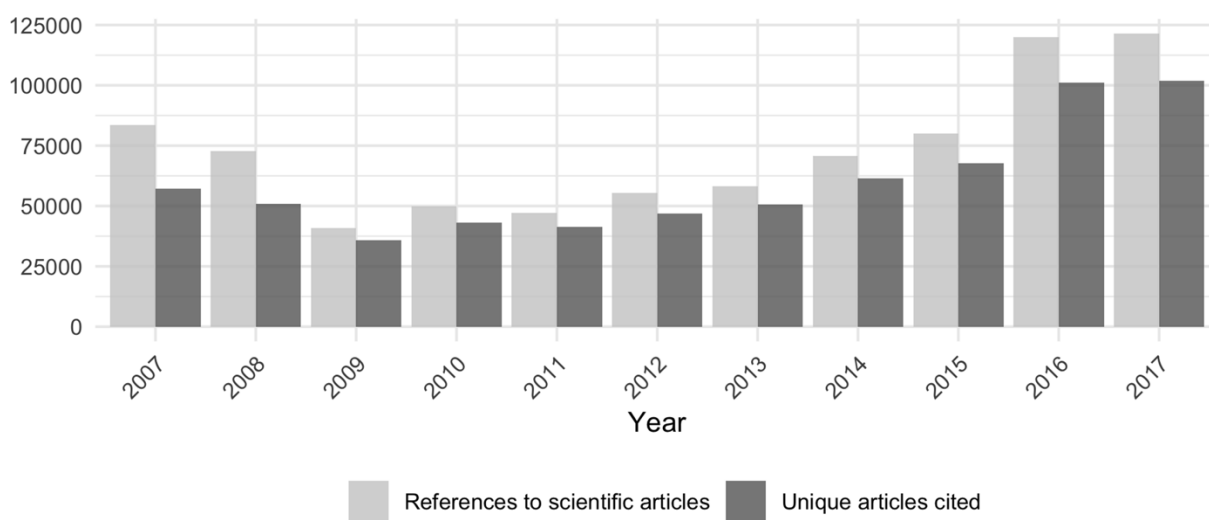
We have analyzed 847512 references to scientific articles distributed across 193802 Wikipedia entries. A total of 598746 scientific articles published in 14149 journals are cited. Each Wikipedia entry includes 4.373 (± 8.351) references to scientific articles, while they receive a mean of 1.415 (± 10.15) citations. Some 81.71% of the total number of scientific articles (489235) receive only one citation and this corresponds to 57.73% of all references in the study sample.

This high standard deviation can be explained by looking at the top 1% of Wikipedia entries with more references, some 60.874 references (± 32.752), representing 13.92% of all references in the study. This top 1% of entries is related to listings—highlights of scientific events in a discipline during a given year—history, genes, common diseases or drugs and medicines. For instance, the highest number of references recorded for a single entry is 550 (https://en.wikipedia.org/wiki/2017_in_paleontology). Furthermore, the level of variation in the standard deviation is not unfamiliar in metrics of this type since the distribution studied here has an especially marked asymmetry because 81% of papers receive only one citation and 97% of the total only receive between one and three. Moreover, 20% of the most cited papers only receive 40% of the total number of citations. This phenomenon occurs in almost all bibliometric indicators [38].

Analyzing the evolution of Wikipedia citations over time, we find that in 2009 the number of individual articles cited and references in Wikipedia fell with respect to 2007 and 2008.

However, since then constant growth has been observed (Fig 2). If we take as a reference the first citation year per Wikipedia entry, in its first year each entry receives 2.793 citations (63.871% of all references), falling to 0.341 (7.795%) and 0.249 (5.682%) in the second and third years, respectively, and further decreasing year after year. Hence, old entries do not accumulate more citations and—except those referenced in 2007 (an average of 6.448) and 2008 (4.022)—these amount to between 2 and 3 per year.

Fig 2. Annual values of total references made by Wikipedia and single articles cited.



The mean publication date of the scientific papers cited is 2001; most were published between 1988 and 2018 (88.51%) as Fig 3 shows. Some 39.43% were published between 2008 and 2018. To analyze the literature on obsolescence, we used the Price index [27], which reflects the percentage of references within a given period. Our results indicate that 36.84% of citations appear within 5 years of publication, twice as many appear within 15 years, and 83.46% appear within 20 years (Fig 4).

Fig 3. Box and violin plots for the years of publication of the scientific articles referenced in Wikipedia (outliers are shown in red).

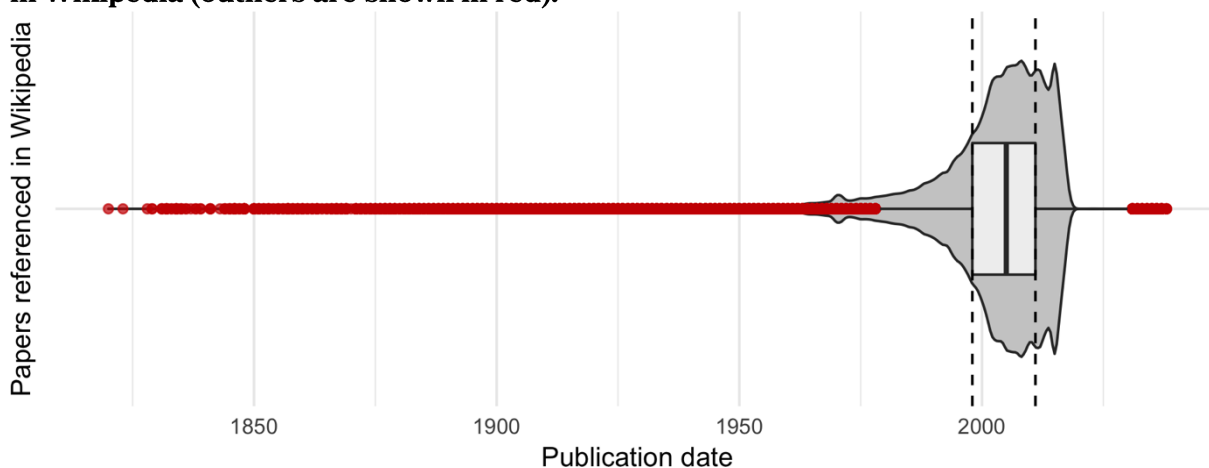
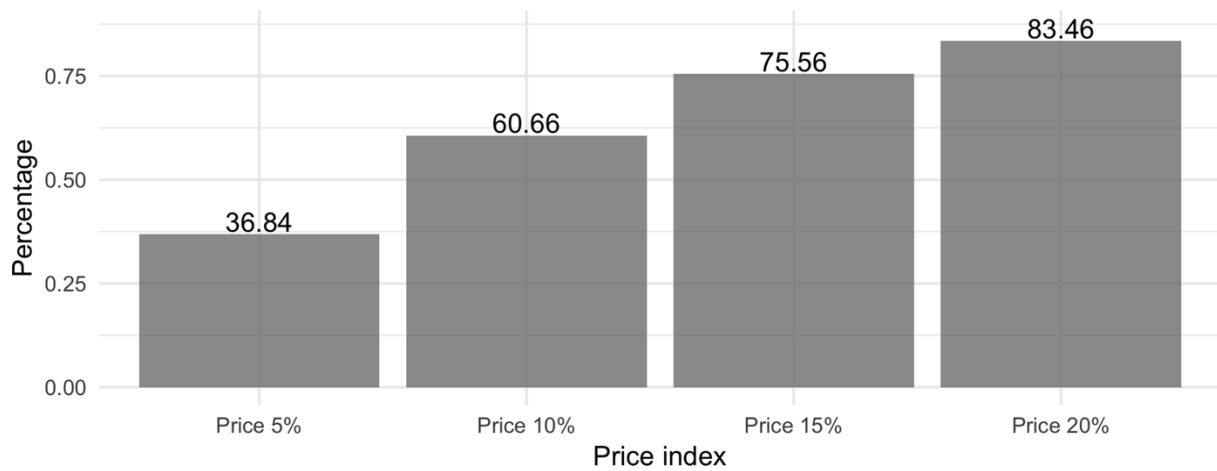


Fig 4. Literature obsolescence of Wikipedia article references using the Price index for 5, 10, 15 and 20 years.

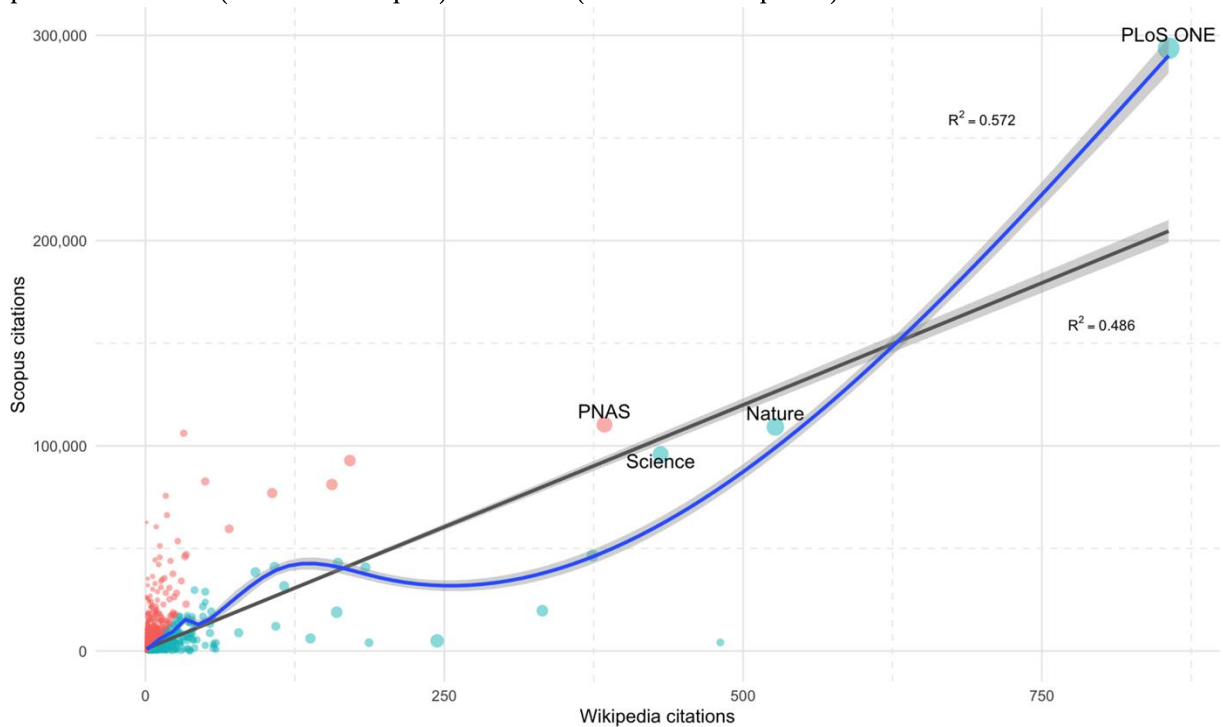


A total of 549 201 papers (91.72%) are co-cited through Wikipedia references and give rise to 7 810 091 co-citations with an average of 28.442 (± 77.088) per paper. To better understand this huge variation, the two most co-cited papers are <https://www.altmetric.com/details/3201729> (4997 mentions) and <https://www.altmetric.com/details/3216022> (3591 mentions) with 3559 co-citations. From the total number of co-citations, 6 110 250 (78.24%) establish connections between papers in different main fields and 1 699 841 (21.76%) do so with papers in the same field. Multidisciplinary co-citations are also slightly more broadly distributed as they have an average of 1.626 (± 3.513) co-citations, compared to non-multidisciplinary co-citations 1.045 (± 1.021).

We have studied the distributions of Wikipedia and Scopus citations at the journal level, considering in both sets only those made in 2016 to articles published between 2013 and 2015. The relationship between the two has been analyzed using linear ($R^2=0.486$) and generalized additive models ($R^2=0.572$)—quantile-quantile (Q-Q) plot shows that both distributions are highly skewed to the right (See Figure S1)—. As can be seen in the scatter plot (Fig 5) and log-log scatter plot (See Figure S2), several journals stand out in both metrics: *PLoS One*, *Nature*, *Science* and *Proceedings of the National Academy of Sciences of the United States of America* (PNAS). In this sense we have obtained the journals' citation percentiles in Wikipedia and Scopus, using only journals with a minimum of three citations in both platforms and two articles cited to avoid noise, and then the ratio between these percentiles have been calculated. While the commented journals have the same attention (ratio=1), the over-cited ones in Wikipedia are *Mammalian Species* (3559), *Art Journal* (192.56), *Northern History* (126.92), *European Journal of Taxonomy* (83.92) and *Art Bulletin* (80.92), and the under-cited ones are *Physical Review A - Atomic, Molecular, and Optical Physics*, *Dalton Transactions* and *Applied Surface Science* (all of them with 0.00027). Furthermore, the distribution of total Wikipedia

citations follows a power law, obtaining a p-value of 0.29 through the goodness-of-fit test, using a bootstrapping procedure. Power law and log-normal distributions offer acceptable fits to the data and do not differ (See Figure S3), giving a p-value of 0.971 via Vuong's test.

Fig 5. Scatter plot of journals by citation collected in Scopus and Wikipedia in 2016 to articles published between 2013 and 2015. The size of the points corresponds to the number of articles published in that period and the color corresponds to the ratio between citation percentiles: red (more on Scopus) and blue (more on Wikipedia).



To illustrate these differences, we have analyzed the 20 most cited scientific articles in Wikipedia (see Table S2). 14 are related to biology (mostly genetics-oriented), while the rest are related to astronomy, physics and computer science, although they also focus on astronomy-related topics. When comparing the Wikipedia citations of these articles (mean 1223.1, ± 1167.19) with the Scopus database (534.1, ± 716.07), we found a mean absolute difference of 1000.4 citations (± 965.42). Only four of these articles received more citations in Scopus than in Wikipedia. The most cited article in Wikipedia is "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences" with 4997 citations (compared to 1228 in Scopus).

Journals by areas

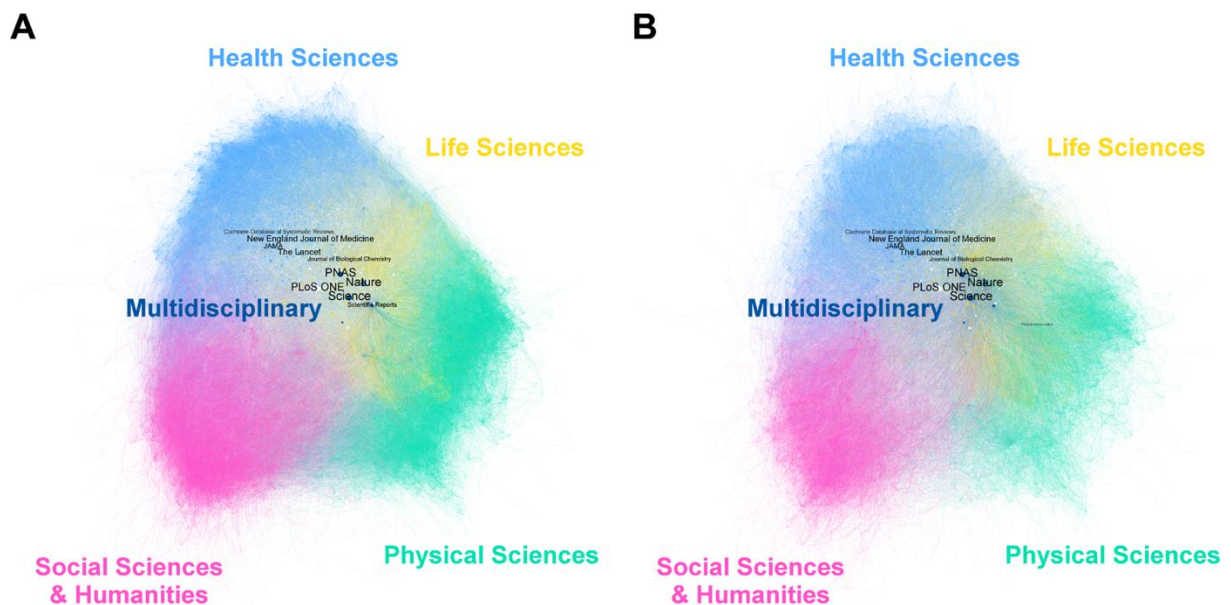
The 14 149 journals in our sample have a mean 42.36 (± 269.22) articles cited in Wikipedia, with each journal receiving a mean 59.9 (± 458.54) citations. Wikipedia entries include a mean 3.25 (± 4.82) references to different scientific journals. So, there are five areas and each journal

belongs to one or more of them with 3279 in "Social Sciences & Humanities", 3077 in "Health Sciences", 2489 in "Physical Sciences", 1298 in "Life Sciences" and 31 in "Multidisciplinary", while the rest belong to more than one area.

The most cited journals are *Nature* (26434 citations); *PNAS* (24104); and the *Journal of Biological Chemistry* (21921), which also has the highest number of individual articles cited (16611). What is remarkable is the fact that only 13.44% of citations are to Open Access journals, when Wikipedia explicitly supports free content. Only two of the 20 most cited journals (see Table S3) are open access resources (*PLoS One* and *Nucleic Acids Research*).

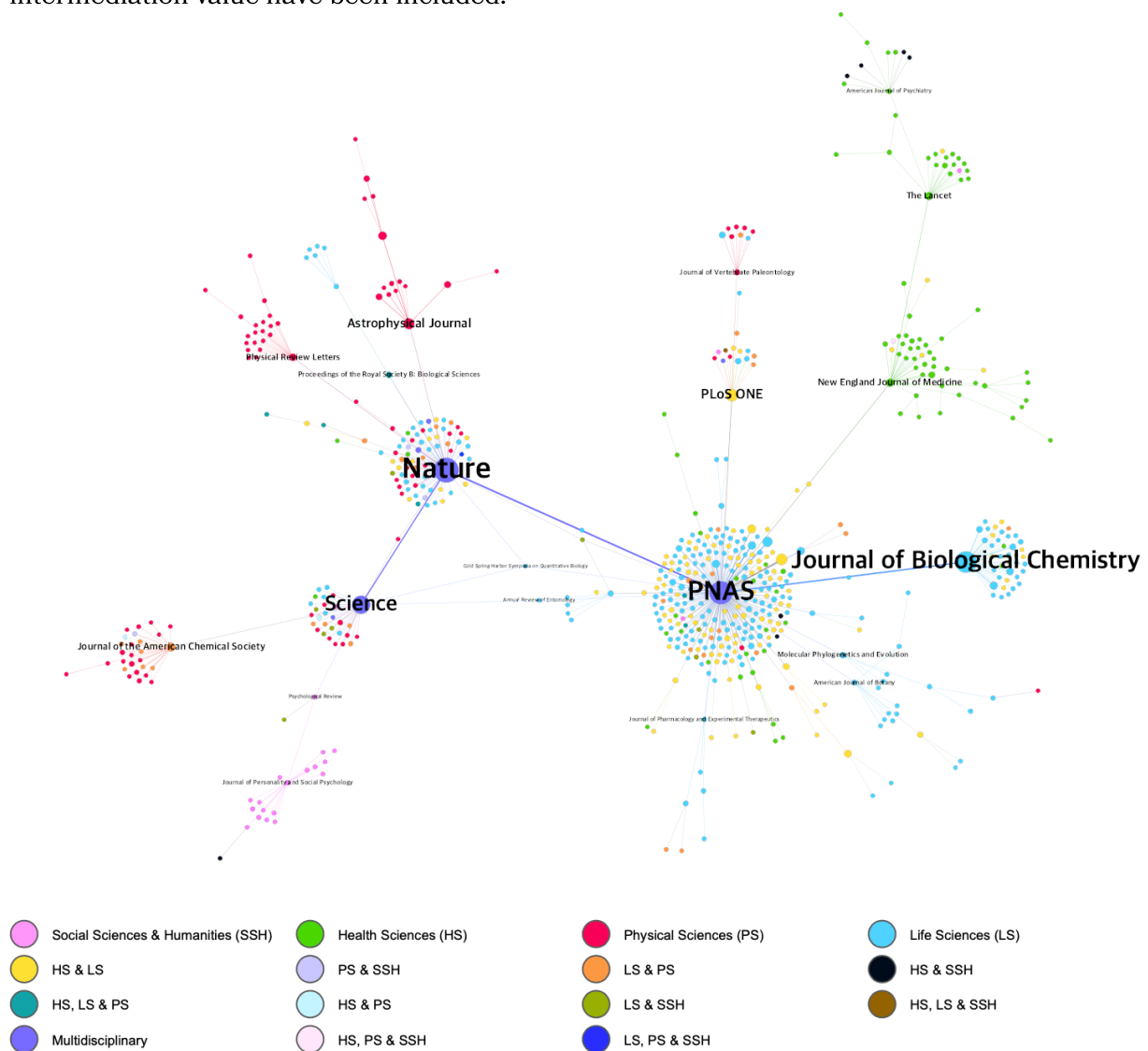
Our map of co-citation networks between journals reveals that 13474 journals (95.2% of the total) are co-cited—each journal has an average of 10.165 co-citations (± 28.292)—, with only 30 (0.22%) having no relationship with the main component (Fig 6). This giant component is made up of 1156668 relationships (Fig 6A), but when we apply the Pathfinder algorithm it is reduced to 684473 (Fig 6B). While the first figure shows that *Science* is the most important journal with the highest number of co-citations (7119) and the highest betweenness, proximity and eigenvector centrality scores (see Table S4), second comes *PNAS* (1604). By analyzing the co-citations between journals by areas in the global network, we find a similar proportion of them are co-cited with others from the same area and from a different one in "Health Sciences" (47.64%, 52.36%), "Physical Sciences" (50.71%, 49.29%) and "Social Sciences & Humanities" (53.93%, 46.07%), but there are significant differences in "Life Sciences" (31.04%, 68.96%). "Multidisciplinary" (0.66%, 99.34%) shows the highest contrast but consists of only a few journals.

Fig 6. Co-citation network of journals based on Wikipedia article references. A) Main component of the full network; B) Pathfinder of the full network. Each node represents one journal and node size corresponds to the total number of citations received; color corresponds to the area but those with more than one are white; the thickness of the edges corresponds to the degree of co-citation between the two. The titles of the 10 journals with the highest intermediation value have been included.



However, after applying the Pathfinder algorithm (Table S4), which eliminates the weakest co-citation links between a journal and co-cited journals, the network obtained is also pruned to display only nodes with a minimum of 50 co-cites. So the score for *Science* falls to 33, below *PNAS* (251), *Nature* (76) and the *Journal of Biological Chemistry* (41). Fig 7 shows the network resulting from applying the Pathfinder algorithm, based on a minimum of 50 co-cites.

Fig 7. Co-citation network of journals based on Wikipedia article references. This network is produced by applying the Pathfinder algorithm—based on a minimum of 50 co-cites—and shows a total of 629 relationships. Each node represents one journal and node size corresponds to the total number of citations received; color corresponds to the area or combination of subject areas to which it belongs; and the thickness of the edges corresponds to the degree of co-citation between the two. The titles of the 20 journals with the highest intermediation value have been included.



If we look at the journals' areas of knowledge we find that Scopus distinguishes between four main subject areas ("Physical Sciences", "Health Sciences", "Social Sciences" and "Life Sciences") and one transversal area called "Multidisciplinary". As Table S5 shows, "Life Sciences" is the most frequently referenced area in Wikipedia (414 400 references and 4.03 mean references per entry) whereas "Multidisciplinary" has the highest average citation (1.88). Given that some journals can be attributed to more than one of the four areas, additional areas have been generated as a result of the possible existing combinations for viewing journals on

the net. The network in Fig 5 shows that most journals belong to "Life Sciences" (36.6% of the total), followed by "Life Sciences & Health Sciences" (19.2%), "Health Sciences" (14.5%) and "Physical Sciences" (14.5%). "Social Sciences & Humanities" is in sixth position (3.5 %) and "Multidisciplinary" is eighth (1.1 %). *PNAS*, *Nature* and *Science* not only act as major intermediaries in the network but also show their multidisciplinary nature by reflecting very strong co-citations with journals from different fields. This is particularly notable in both *Nature* and *Science*. Most connections linked with *PNAS* are to journals in "Life Sciences" and "Health Sciences & Life Sciences". *PLoS ONE* also shows strong co-citation links with journals in many areas despite being cataloged in "Health Sciences & Life Sciences".

Main fields

Wikipedia entries that reference articles within the same main field do so with an average of 1.466 (± 1.504) references, while entries that mix articles from different main fields do so with 5.764 (± 9.799).

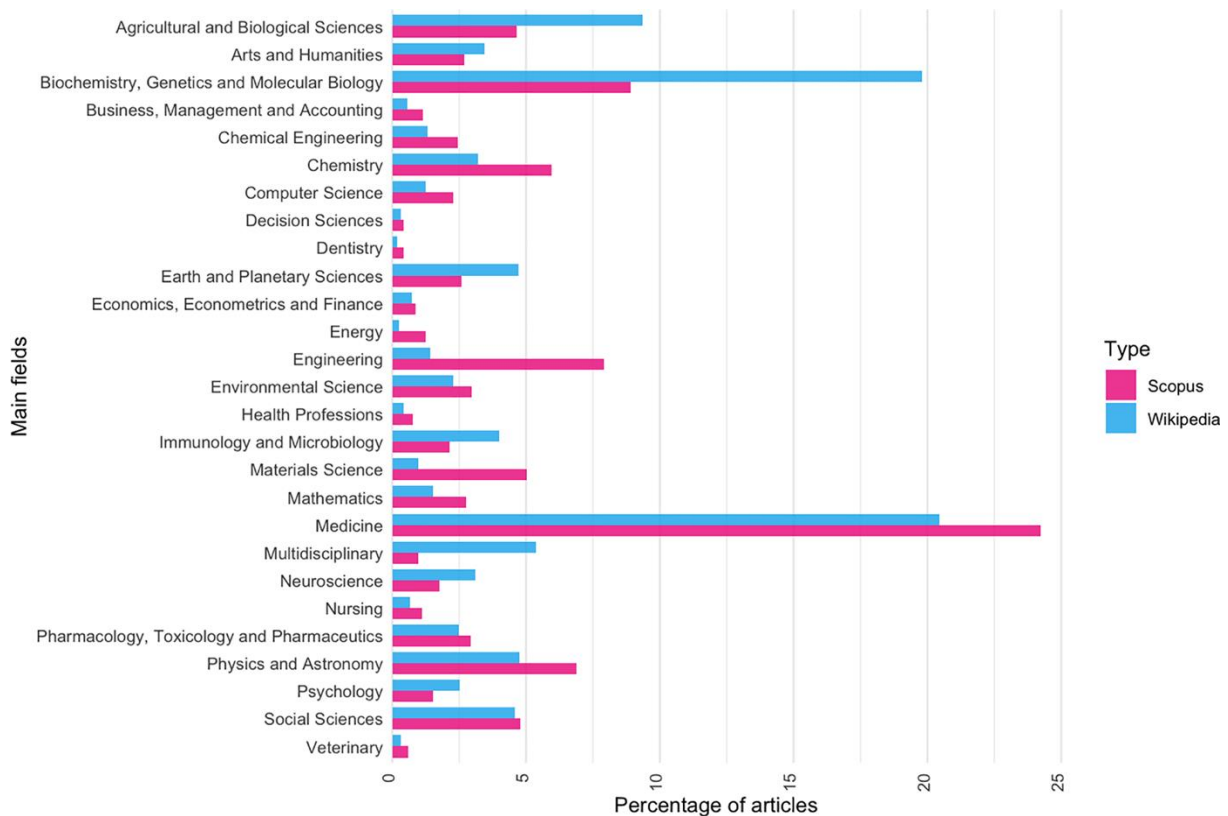
Within the 27 main fields (see Table S1), "Medicine" (referenced in 72384 Wikipedia entries; 3.81 mean references per entry) and "Biochemistry, Genetics and Molecular Biology" (referenced in 64945 Wikipedia entries; 4.11 mean references per entry) are the most significant. In contrast, "Dentistry" has the lowest level of presence in Wikipedia entries (992), although the mean number of references is 2.42. The main fields with the lowest means are: "Arts and Humanities" (1.65) and "Decision Sciences" (1.6).

In relation to citations received by main fields from scientific papers (see Table S1), in absolute terms articles in "Medicine" (206576 citations received) and "Biochemistry, Genetics and Molecular Biology" (181954) stand out. However, on average, the outstanding main fields are "Multidisciplinary" (1.88 citations per article) and "Earth and Planetary Sciences" (1.88). "Dentistry" remains the least frequently cited area and has the lowest mean number of citations (1.14).

Fig 8 shows the distribution by main field of all articles included by Scopus, a total of 62821260 scientific articles indexed in the database, by comparison with the distribution by main field of articles cited in Wikipedia (see Table S6). The main fields attributed to the articles correspond to those of the journals in which they are published. Note that from the Wikipedia perspective, there is a greater presence of articles from "Biochemistry, Genetics and Molecular Biology" (10.86% more), "Agricultural and Biological Sciences" (4.72% more), "Multidisciplinary" (4.37% more), "Earth and Planetary Sciences" (2.11% more), "Immunology and Microbiology" (1.88% more), and "Neuroscience" (1.34% more) than that found in Scopus. In contrast, in Scopus the proportion of articles from "Engineering" (6.49%

more), "Materials Science" (4.05% more), "Medicine" (3.76% more), Chemistry (2.72% more) and "Physics and Astronomy" (2.13% more) is higher than that in Wikipedia. The main fields for which the distribution of articles is similar both in Scopus and Wikipedia are: "Social Sciences"; "Economics, Econometrics and Finance"; "Decision Science" (with differences of less than 0.2% in absolute terms).

Fig 8. Comparison of the percentage of articles by main field in Scopus and Wikipedia.



Analysis of the Price index for each of these main fields shows that "Energy" and "Material Sciences" reflect a rather limited degree of obsolescence compared to the rest (See Fig 9). This phenomenon is more noticeable in the former, with a value of 55% for the first five years, reaching 91.56% when we extend the time interval to 20 years. "Arts and Humanities" and "Decision Sciences" are in a very different situation, with Price indexes for the first five years of 22.76% and 24.67%, respectively—half that of "Energy" for the same period. When we look at Price indexes for 20 years, we also see considerably lower values with 68.44% in "Arts and Humanities" and 60.54% in "Decision Sciences", the latter also having the lowest value of all main fields over 20 years.

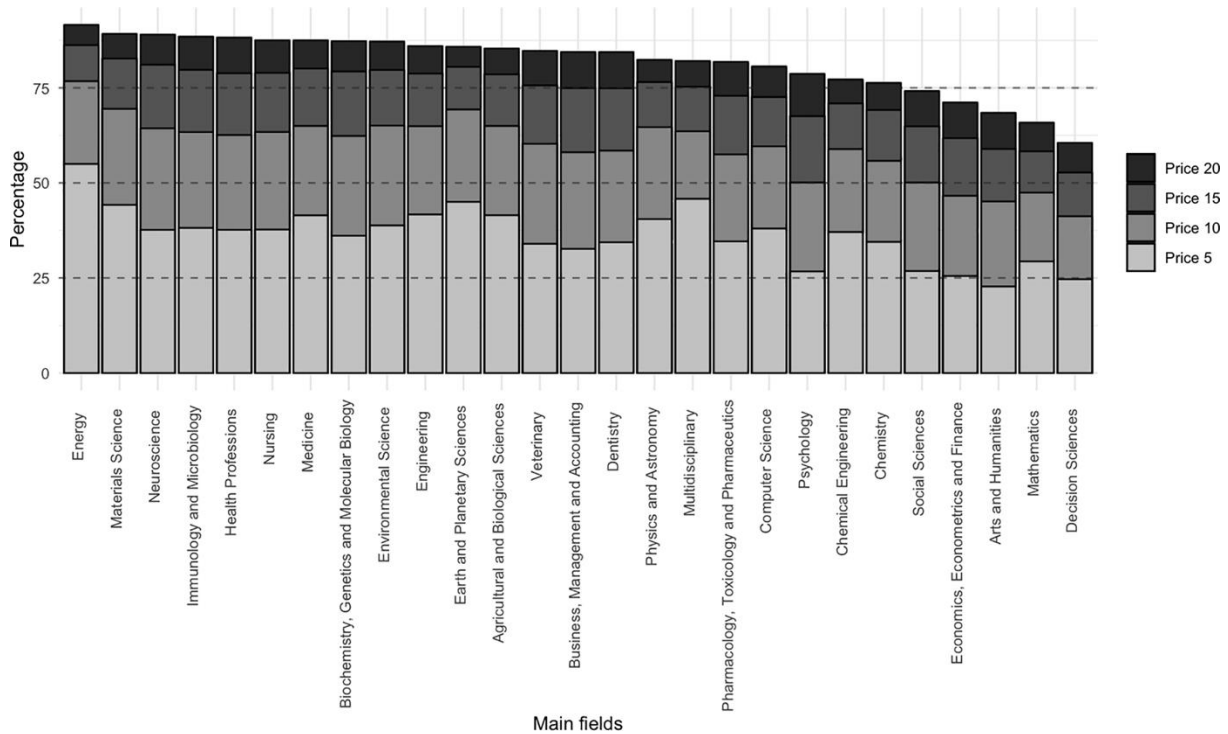
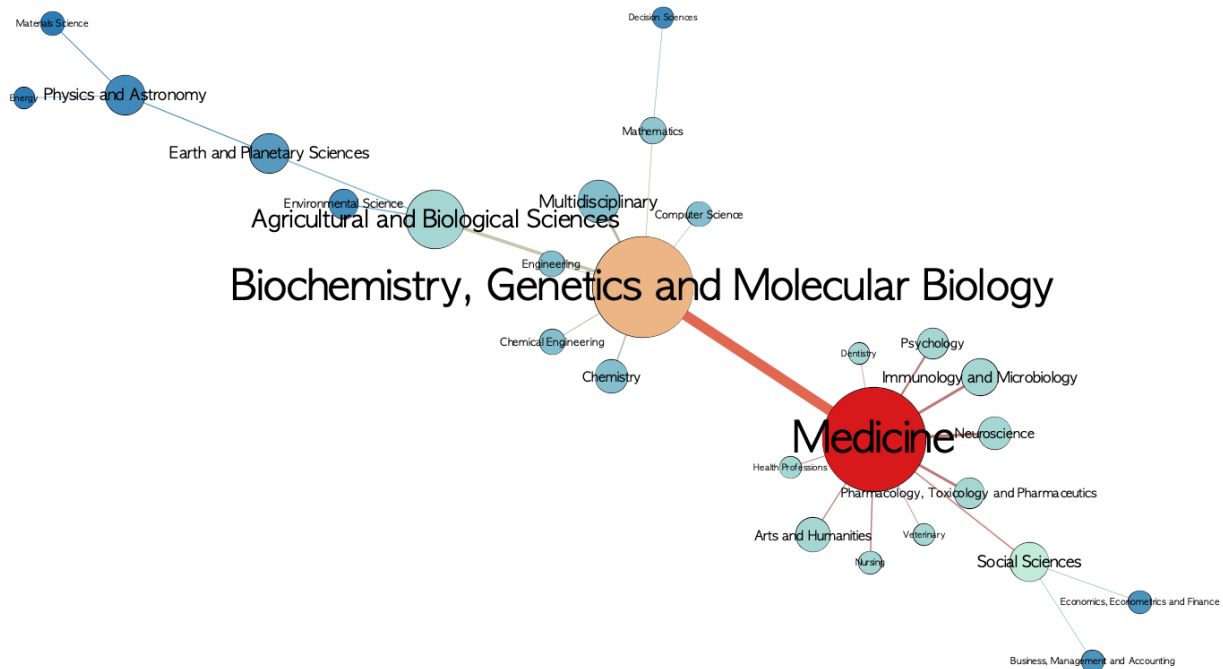
Fig 9. Price index for Wikipedia main fields.

Fig 10 shows the co-citation network of the 27 main fields after applying the Pathfinder algorithm. The two main actors are "Medicine" and "Biochemistry, Genetics and Molecular Biology", which constitute the core of the network and share strong co-citation links. Apart from the connection between "Medicine" and main fields linked to Health, the strong relationship with "Arts and Humanities" and "Social Sciences" (also as a link between "Business, Management and Accounting" and "Economics, Econometrics and Finance") stands out. Furthermore, it is also noteworthy that "Biochemistry, Genetics and Molecular Biology" is closely linked to "Agricultural and Biological Sciences", highlighting the connections with more tangential main fields such as "Computer Science", "Engineering", "Multidisciplinary" and "Mathematics".

Fig 10. Co-citation network of the 27 main fields after applying the Pathfinder algorithm. The nodes represent each main field; node size corresponds to the total number of citations received, color corresponds to own vector centrality; and the thickness of the edges corresponds with degree of co-citation.

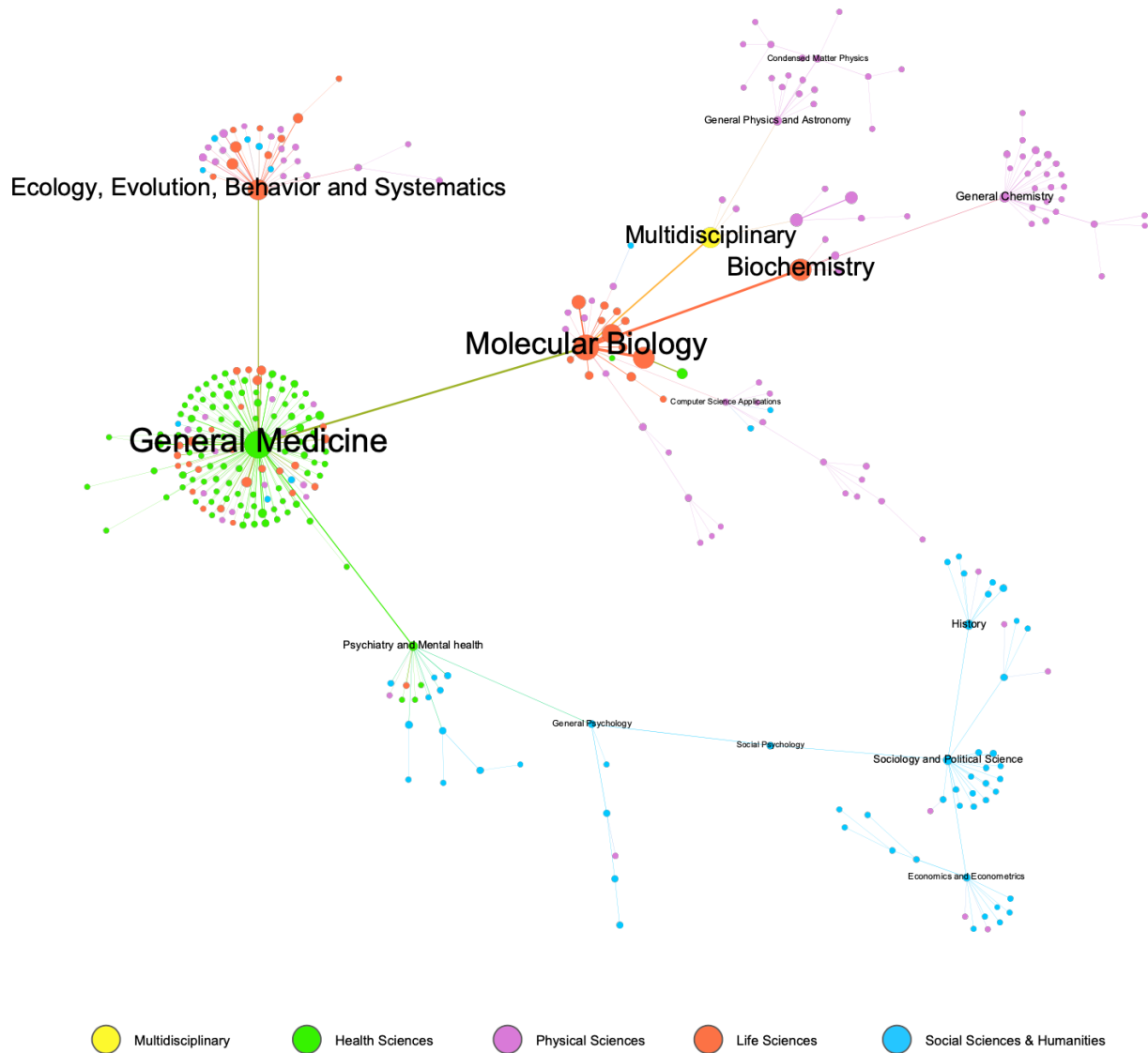


Field

As can be seen in Table S7, within the 330 fields studied, the two most outstanding fields are: "General Medicine" and "Molecular Biology", with 108131 and 98118 total citations, respectively. Both stand at a considerable distance from the next outstanding specialties: "Biochemistry" (78704), "Genetics" (77920) and "Multidisciplinary" (72346).

Fig 11 shows the co-citation network of the 330 after applying the Pathfinder algorithm. This network shows the prominent position of "General Medicine", which has the highest number of relevant co-citations and is central to the majority of fields in "Health Sciences". We should also mention the role of "General and Social Psychology" as a connection between "Health Sciences & Humanities" and "Social Sciences". Despite the link to "Social Psychology", the "Social Sciences & Humanities" appear disconnected and are structured around three fields: "Sociology and Political Science", "Economics" and "Econometrics and History". Finally, the fields related to Physics also appear in peripheral areas of the graph and are linked to "Multidisciplinary"; "General Chemistry", which is linked to "Biochemistry", is in a similar situation.

Fig 11. Co-citation network of the 330 fields after applying the Pathfinder algorithm. The nodes represent each field, indicating size, total number of citations received, color, thematic area or areas, and the thickness of the edges indicates the degree of co-citation. Field titles are given for the 15 fields with the highest levels of intermediation.



Discussion

We have conducted a large-scale application of co-citation analysis to all articles referenced in Wikipedia. Previous research [26] had experimented with this approach in the Humanities alone, presenting promising results in mapping Science from the Wikipedia perspective. However, it was necessary to validate this on a more ambitious scale, that is, the entire open online encyclopedia, in which the Humanities represent only 5.49% of all the references collected. The results presented above are indicative of how this innovative approach allows us to depict a complete picture of Science.

Thus we can produce science maps that complement the traditional co-citation maps focused on scientific articles [39, 40] and provide a representation of knowledge that focuses on the vision and use of information in the scientific community. The methodology presented, which focuses on the co-citation of Wikipedia articles, offers holistic maps of the use of scientific information by Wikipedia users/editors who are not necessarily scientists. Therefore these maps represent the user's vision of scientific activity and in this sense they are close to other mapping methodologies that are not exclusively centered on citations but centered on the user—maps such as those based on Clickstream Data [41], readership network maps using Mendeley [42] or maps based on Co-Tweet [34]. By comparison with earlier research, the main novelties of the present study are that for the first time a source of information as important as Wikipedia has been used, several sources have been combined (Altmetrics, Scopus), and we have used Pathfinder, which is a much more efficient algorithm.

The Wikipedia references, unlike those collected in other social media, offer remarkable quality control and transparency. In relation to the problem of trolling, the encyclopedia is based on a solid quality management system of post-publication peer review in which, in the case of discrepancies, changes are resolved through consensus between editors. Wikipedia also has two manuals: for non-academic experts (https://en.wikipedia.org/wiki/Help:Wikipedia_editing_for_non-academic_experts) and for researchers, scholars, and academics (https://en.wikipedia.org/wiki/Help:Wikipedia_editing_for_researchers,_scholars,_and_academics) both specifying the importance of the use of citations under the principles of verifiability and notability. This substantially minimizes the likelihood that references in entries will be tampered with.

Wikipedia also offers a complete list of its bots (<https://en.wikipedia.org/wiki/Special:ListUsers/bot>), including those such as the Citation bot (https://en.wikipedia.org/wiki/User:Citation_bot), which in addition to adding missing identifiers to references, corrects and completes them, something for which the digital encyclopedia offers several tools (https://en.wikipedia.org/wiki/Help:Citation_tools). However, this does not prevent the appearance of publications with a high, anomalous number of citations [43]. For instance, we found a report (<https://www.altmetric.com/details/3171944>) cited in 1450 lunar crater entries, not attributable to a bot. So, although the use of citations is not compromised, practices of this sort must be taken into account, for example, if their use is in an evaluative context. Given all of the aforementioned, we consider that in this context in which 193802 Wikipedia entries and 847512 article citations have been analyzed, it is very difficult to produce manipulations that could significantly alter the system and, consequently, the results achieved here.

About the results

This study illustrates the use of scientific information from the Wikipedia perspective, which is the most important and largest encyclopedia available nowadays. We have been able to determine the main fields that receive citations in Wikipedia entries. The most relevant fields are “Medicine” (32.58%), “Biochemistry, Genetics and Molecular Biology” (31.5%) and “Agricultural and Biological Sciences” (14.91%). In contrast, “Dentistry” (0.28%), “Energy” (0.43%), “Decision Sciences” (0.49%) and “Veterinary” (0.52%) are the main fields that globally receive fewer references. We would emphasize the fact that these areas need to strengthen the visibility of their work. In general, we find it remarkable that Science disciplines should dominate the Humanities and Social Sciences.

If we look at the maps at journal level, we find that the most important publications are multidisciplinary in nature and the main journals in terms of centrality are *Science*, *Nature*, *PNAS*, *PLoS ONE*, and *The Lancet*. However, after applying the Pathfinder algorithm to discard less relevant relationships, we note that *PNAS* rises to first position, limiting the centrality of journals like *Science* and *PLoS ONE*, which have more but weaker co-citation relations. Without a doubt, this places *Science* and *PLoS ONE* in an interesting centrality space and turns them into nodes that connect with a curiously wide variety of journals. Likewise, our proposed methodology has enabled us to detect the strongest links between the main journals and their scientific uniqueness; in this sense, it is worth highlighting *Nature's* strong relationship with “Physics”, *Science's* relationship with “Chemistry” and that of *PNAS* with “Life Sciences”.

Like other platforms, multidisciplinary journals are hubs and articulate the Wikipedia network. However, despite being a common global phenomenon, Wikipedia does have, and contains unique citation practices mentioning journals that are not cited or mentioned in a relevant way in other databases or platforms. This is evidenced in the scatter plot of Wikipedia and Scopus citations (see Fig 5).

This difference is also illustrated by a comparison of the journals most mentioned in Altmetric.com with those most mentioned in Wikipedia. As we can see, Wikipedia has the major multidisciplinary journals in common, as does Scopus. However, some of the most frequently mentioned journals in Wikipedia are the *Journal of Biological Chemistry* or *Zookeys* which are located in JCR's Q2. Nonetheless, the *Journal of Biological Chemistry*, for example, is one of the most widely cited journals in the field of genetics receiving a large number of citations in various entries such as "Androgen receptor" (45 citations) or "Epidermal growth factor receptor" (25 citations). Therefore, Wikipedia points to another type of specialized

journal in different fields (See Table S8) that are not identified in other rankings. In addition, as we have indicated, this can hardly be due to trolling or a bot.

If we observe the map of main fields, "Medicine" and "Biochemistry, Genetics and Molecular Biology", are the two main nodes. From this perspective, "General Medicine" is the most relevant node, accounting for the highest number of citations received. It acts as a highly important connector in the network. Moreover, this underlines the role of "Psychology" in connecting "Health Sciences" with "Social Sciences and the Humanities".

Given the open nature of Wikipedia, the analysis of references to open access journals is particularly relevant. Firstly, it is remarkable that only 13.44% of citations are to Open Access journals, when Wikipedia explicitly supports free content. Furthermore, only two of the 20 most cited journals are open access resources (*PLoS ONE* and *Nucleic Acids Research*). Teplitskiy et al. [10] determined that the odds in favor of an Open Access journal being referenced in the encyclopedia were about 47% higher than that of closed access journals. They also suggested that high-impact factor journals were more likely to be cited, as we have also observed in our results. In relation to open access resources, the fact that many articles in closed journals can be accessed through their authors or third parties [44] may distort some of these considerations.

PLoS ONE is the most relevant open access journal cited in Wikipedia. And it is the fourth in terms of centrality, just behind *Science*, *Nature* and *PNAS*, all three of which operate under a non-open access model. When applying Pathfinder, *PLoS ONE*'s centrality is reduced. This is due to the fact that this method eliminates the weakest co-citation links, which are highly relevant to the journal because although is cataloged in "Health Sciences & Life Sciences", it occupies a central position in the network in relation to journals from vastly different areas.

Our study has also shown that certain fields have a stronger relative presence in Wikipedia references than in Scopus. This is the case of "Biochemistry, Genetics and Molecular Biology" (10.86% more), "Agricultural and Biological Sciences" (4.72% more) and "Multidisciplinary" (4.37% more), among others. This could indicate that from the Wikipedia perspective some fields receive more attention than from the scientific community as a whole. Finally, in relation to obsolescence we have observed significant differences between main fields. For instance, for the first five years, "Energy" has 55% of references, whereas the "Arts and Humanities" receives only 22.76%.

Comparison of Wikipedia and Scopus

Wikipedia's view of science differs from that of Scopus. While linear regression and generalized additive models have a correlation statistically significant, we do not establish

causality due to the high presence of outliers that do not obey these patterns. Also, the focus of thematic attention provided by Wikipedia editors shows striking differences, an aspect that is clearly evident in Scopus and Wikipedia's differences of coverage and the presence in the latter of journals in very prominent positions that do not coincide with the views of other altmetrics sources. At the level of article citations the strong asymmetry in the distribution curve of Wikipedia citations is due to the fact that most receive only between one and three citations, showing a much more extreme phenomenon than Pareto's law. However, at the journal level, we can confirm the existence of a power-law distribution that shows a phenomenon similar to that observed in citations in Scopus [45]. This is why these differences allow us to appreciate that we are dealing with two phenomena that are not the same.

Limitations

As in our previous study [26], some limitations derive from the attribution of categories to journals since journals do not always belong to the category assigned by the database [46]. Latent co-citation can also arise [37] because some journals may be assigned to more than one field. We have resolved this issue by combining all of them under the same label.

It should be noted that the methodology used, which combines various sources (Altmetric.com, Wikipedia and Journal Metrics by Elsevier), generates certain limitations. For example, we have only taken account of scientific articles since only resources with an ISSN and indexed by Scopus have been used; this excludes books or chapters of special relevance in the Humanities [47]. This is an issue present in the classical approaches that are limited to scientific journals [48].

Other problems derive from the sometimes inaccurate Altmetric description of their records. The dataset frequently presents duplicate records; errors in the year of publication, DOI and identifiers assigned to records; or records with many fields containing null values. For instance, the field presenting most problems has been that of the ISSN, which is sometimes incorrect in both Altmetric and CiteScore Metrics.

One of the most striking limitations detected with regard to the use of Wikipedia references as a measure of activity impact lies in their volatility because many references can be created or eliminated very quickly, making data collection and subsequent use difficult. In this respect Altmetric.com's Altmetric Attention Score can also mislead because given that it is a static measure, it only takes account of presence and makes no allowance for frequency. However, none of these limitations affects the overall results of our study because the large number of references and processed articles in our sample minimizes their impact.

Finally, we must point out that these types of map are an interesting complement to quantitative information offered by platforms such as Altmetric.com or PlumX. Thus, thanks to these contextual methodologies [49] it is possible to elucidate more clearly the social impact (societal impact) of scientific articles in particular and of Science in general of platforms such as Wikipedia. In the future we will extrapolate co-citation studies to other documentary typologies and platforms included in Altmetric.com such as news or policy reports in order to clearly establish the different representations of knowledge generated by different users and consumers of scientific information.

References

1. O'Reilly T. What is web 2.0? Design patterns and business models for the next generation of software [Internet]. O'Reilly Media, Inc.; 2005. Available from: <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>
2. Surowiecki J. The wisdom of crowds. Anchor; 2005.
3. Fallis D. Toward an epistemology of Wikipedia. *J Assoc Inf Sci Technol*. 2008;59(10):1662–74.
4. Rubira R, Gil-Egui G. Wikipedia as a space for discursive constructions of globalization. *Int Commun Gaz* [Internet]. 2019 Oct 30;81(1):3–19. Available from: <https://doi.org/10.1177/1748048517736415>
5. König R. Wikipedia: Between lay participation and elite knowledge representation. *Information, Commun Soc* [Internet]. 2013 Mar 1;16(2):160–77. Available from: <https://doi.org/10.1080/1369118X.2012.734319>
6. Bould MD, Hladkowicz ES, Pigford A-AE, Ufholz L-A, Postonogova T, Shin E, et al. References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature. *BMJ Br Med J* [Internet]. 2014 Mar 6;348:g1585. Available from: <http://www.bmj.com/content/348/bmj.g1585.abstract>
7. Brazzeal B. Citations to Wikipedia in chemistry journals: A preliminary study. *Issues Sci Technol Librariansh*. 2011;67.
8. Serrano-López AE, Ingwersen P, Sanz-Casado E. Wind power research in Wikipedia: Does Wikipedia demonstrate direct influence of research publications and can it be used as adequate source in research evaluation? *Scientometrics*. 2017;112(3):1471–88.
9. Kousha K, Thelwall M. Are wikipedia citations important evidence of the impact of scholarly articles and books? *J Assoc Inf Sci Technol*. 2017;68(3):762–79.
10. Teplitzkiy M, Lu G, Duede E. Amplifying the impact of open access: Wikipedia and the diffusion of science. *J Assoc Inf Sci Technol* [Internet]. 2017 Sep;68(9):2116–27. Available from: <https://doi.org/10.1002/asi.23687>
11. Piwowar H. Value all research products. *Nature*. 2013;493(7431):159.

12. Priem J, Taraborelli D, Groth P, Neylon C. Altmetrics: A manifesto [Internet]. 2010. Available from: <http://altmetrics.org/manifesto/>
13. Torres-Salinas D, Cabezas-Clavijo Á, Jiménez-Contreras E. Altmetrics: New indicators for scientific communication in web 2.0. *Comunicar*. 2013;21(41):53–60.
14. Salah AA, Gao C, Suchecki K, Scharnhorst A. Need to categorize: A comparative look at the categories of universal decimal classification system and Wikipedia. *Leonardo*. 2012;45(1):84–5.
15. Silva FN, Amancio DR, Bardosova M, Costa L da F, Oliveira ON. Using network science and text analytics to produce surveys in a scientific topic. *J Informetr* [Internet]. 2016 May;10(2):487–502. Available from: <http://www.sciencedirect.com/science/article/pii/S1751157715301966>
16. Silva FN, Viana MP, Travençolo BAN, Costa L da F. Investigating relationships within and between category networks in Wikipedia. *J Informetr*. 2011;5(3):431–8.
17. de Arruda HF, Silva FN, Costa L da F, Amancio DR. Knowledge acquisition: A Complex networks approach. *Inf Sci (Ny)* [Internet]. 2017;421:154–66. Available from: <http://www.sciencedirect.com/science/article/pii/S0020025517309295>
18. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J Am Soc Inf Sci* [Internet]. 1973 Jul 1 [cited 2017 Nov 26];24(4):265–9. Available from: <http://doi.wiley.com/10.1002/asi.4630240406>
19. Leydesdorff L, Nerghees A. Co-word maps and topic modeling: A comparison using small and medium-sized corpora ($N < 1,000$). *J Assoc Inf Sci Technol* [Internet]. 2017 Apr 1;68(4):1024–35. Available from: <https://doi.org/10.1002/asi.23740>
20. Leydesdorff L, Carley S, Rafols I. Global maps of science based on the new Web-of-Science categories. *Scientometrics*. 2013;94(2):589–93.
21. Thelwall M. Introduction to webometrics: Quantitative web research for the social sciences. *Synth Lect Inf concepts, retrieval, Serv*. 2009;1(1):1–116.
22. Zuccala A. Author cocitation analysis is to intellectual structure as web colink analysis is to...? *J Assoc Inf Sci Technol*. 2006;57(11):1487–502.
23. Vaughan L, Kipp ME, Gao Y. Why are websites co-linked? The case of Canadian universities. *Scientometrics*. 2007;72(1):81–92.
24. Romero-Frías E, Vaughan L. European political trends viewed through patterns of Web linking. *J Assoc Inf Sci Technol*. 2010;61(10):2109–21.
25. Vaughan L, Romero-Frías E. Web hyperlink patterns and the financial variables of the global banking industry. *J Inf Sci*. 2010;36(4):530–41.
26. Torres-Salinas D, Romero-Frías E, Arroyo-Machado W. Mapping the backbone of the Humanities through the eyes of Wikipedia. *J Informetr* [Internet]. 2019;13(3):793–803. Available from: <http://www.sciencedirect.com/science/article/pii/S1751157718302955>

27. de Solla Price DJ. Networks of Scientific Papers. *Science* (80-) [Internet]. 1965 Jul 30;149(3683):510 LP – 515. Available from: <http://science.sciencemag.org/content/149/3683/510.abstract>
28. De Bellis N. *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Lanham, Maryland, United States: Scarecrow Press; 2009.
29. Gillespie CS. Fitting Heavy Tailed Distributions: The *powerLaw* Package. *J Stat Softw* [Internet]. 2015 Mar 20;64(2). Available from: <http://doi.org/10.18637/jss.v064.i02>
30. Chang Y-W, Huang M-H, Lin C-W. Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics* [Internet]. 2015;105(3):2071–87. Available from: <https://doi.org/10.1007/s11192-015-1762-8>
31. van Eck NJ, Waltman L. Visualizing Bibliometric Networks BT - *Measuring Scholarly Impact: Methods and Practice*. In: Ding Y, Rousseau R, Wolfram D, editors. Cham: Springer International Publishing; 2014. p. 285–320. Available from: https://doi.org/10.1007/978-3-319-10377-8_13
32. Yan E, Ding Y. Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *J Am Soc Inf Sci Technol* [Internet]. 2012 Jul 1;63(7):1313–26. Available from: <https://doi.org/10.1002/asi.22680>
33. Costas R, de Rijcke S, Marres N. Beyond the dependencies of altmetrics: Conceptualizing ‘heterogeneous couplings’ between social media and science. In: *The 2017 Altmetrics Workshop* [Internet]. 2017. Available from: http://altmetrics.org/wp-content/uploads/2017/09/altmetrics17_paper_4.pdf
34. Didegah F, Thelwall M. Co-saved, co-tweeted, and co-cited networks. *J Assoc Inf Sci Technol* [Internet]. 2018 Aug 1;69(8):959–73. Available from: <https://doi.org/10.1002/asi.24028>
35. Trujillo CM, Long TM. Document co-citation analysis to enhance transdisciplinary research. *Sci Adv* [Internet]. 2018 Jan 1;4(1):e1701130. Available from: <http://advances.sciencemag.org/content/4/1/e1701130.abstract>
36. Vargas-Quesada B. *Visualización y análisis de grandes dominios científicos mediante redes pathfinder (PFNET)*. Universidad de Granada; 2005.
37. Moya-Anegón F, Vargas-Quesada B, Herrero-Solana V, Chinchilla-Rodríguez Z, Corera-Álvarez E, Muñoz-Fernández FJ. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics* [Internet]. 2004;61(1):129–45. Available from: <https://doi.org/10.1023/B:SCIE.0000037368.31217.34>
38. Seglen PO. Why the impact factor of journals should not be used for evaluating research. *BMJ* [Internet]. 1997 Feb 15;314(7079):498–502. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/9056804>

39. Bornmann L, Leydesdorff L. Topical connections between the institutions within an organisation (institutional co-authorships, direct citation links and co-citations). *Scientometrics* [Internet]. 2015;102(1):455–63. Available from: <https://doi.org/10.1007/s11192-014-1425-1>
40. Leydesdorff L, Rafols I. A global map of science based on the ISI subject categories. *J Am Soc Inf Sci Technol* [Internet]. 2009 Feb 1 [cited 2017 Nov 7];60(2):348–62. Available from: <http://doi.wiley.com/10.1002/asi.20967>
41. Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, Rodriguez MA, et al. Clickstream Data Yields High-Resolution Maps of Science. *PLoS One* [Internet]. 2009 Mar 11;4(3):e4803. Available from: <https://doi.org/10.1371/journal.pone.0004803>
42. Bornmann L, Haunschild R. Overlay maps based on Mendeley data: The use of altmetrics for readership networks. *J Assoc Inf Sci Technol* [Internet]. 2016 Dec 1;67(12):3064–72. Available from: <https://doi.org/10.1002/asi.23569>
43. Torres-Salinas D, Gorraiz J, Robinson-Garcia N. The insoluble problems of books: what does Altmetric.com have to offer? *Aslib J Inf Manag*. 2018;70(6):691–707.
44. Björk B-C, Laakso M, Welling P, Paetau P. Anatomy of green open access. *J Assoc Inf Sci Technol* [Internet]. 2014 Feb 1;65(2):237–50. Available from: <https://doi.org/10.1002/asi.22963>
45. Brzezinski M. Power laws in citation distributions: evidence from Scopus. *Scientometrics* [Internet]. 2015;103(1):213–28. Available from: <https://doi.org/10.1007/s11192-014-1524-z>
46. Rafols I, Porter AL, Leydesdorff L. Science overlay maps: A new tool for research policy and library management. *J Am Soc Inf Sci Technol* [Internet]. 2010 Sep 1;61(9):1871–87. Available from: <https://doi.org/10.1002/asi.21368>
47. Torres-Salinas D, Rodríguez-Sánchez R, Robinson-García N, Fdez-Valdivia J, García JA. Mapping citation patterns of book chapters in the Book Citation Index. *J Informetr* [Internet]. 2013;7(2):412–24. Available from: <http://www.sciencedirect.com/science/article/pii/S1751157713000060>
48. Leydesdorff L, Hammarfelt B, Salah A. The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1,157 journals. *J Am Soc Inf Sci Technol* [Internet]. 2011 Dec 1;62(12):2414–26. Available from: <https://doi.org/10.1002/asi.21636>
49. Robinson-Garcia N, van Leeuwen TN, Rafols I. Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Sci Public Policy* [Internet]. 2018 Dec 1;45(6):815–26. Available from: <http://dx.doi.org/10.1093/scipol/scy024>

Mapping social media attention in Microbiology: Identifying main topics and actors



Nicolas Robinson-Garcia^{1,2,*}, Wenceslao Arroyo-Machado³
and Daniel Torres-Salinas^{3,4}

¹INGENIO (CSIC-UPV), Universitat Politècnica de València, Valencia, Spain

²School of Public Policy, Georgia Institute of Technology, Atlanta, United States

³Vicerrectorado de Investigación y Transferencia, Universidad de Granada, Granada, Spain

⁴EC3metrics spin off, Granada, Spain

*Corresponding author: wences@ugr.es

Journal

FEMS Microbiology Letter
1574-6968

Index

SCIE – Q3

DOI

10.1093/femsle/fnz075

Data

None

Version

Preprint

References

Oxford

Funding

None

Abstract

This paper aims to map and identify topics of interest within the field of Microbiology and identify the main sources driving such attention. We combine data from Web of Science and Altmetric.com, a platform which retrieves mentions to scientific literature from social media and other non-academic communication outlets. We focus on the dissemination of microbial publications in Twitter, news media and policy briefs. A two-mode network of social accounts shows distinctive areas of activity. We identify a cluster of papers mentioned solely by regional news media. A central area of the network is formed by papers discussed by the three outlets. A large portion of the network is driven by Twitter activity. When analyzing top actors contributing to such network, we observe that more than half of the Twitter accounts are bots, mentioning 32% of the documents in our dataset. Within news media outlets, there is a predominance of popular science outlets. With regard to policy briefs, both international and national bodies are represented. Finally, our topic analysis shows that the thematic focus of papers mentioned varies by outlet. While news media cover the wider range of topics, policy briefs are focused on translational medicine, and bacterial outbreaks.

Citation

Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. *FEMS Microbiology Letters*, 366(7). <https://doi.org/10.1093/femsle/fnz075>

Introduction

In a rapidly changing scholarly communication system, the number of publications grows exponentially (Van Noorden, Maher and Nuzzo 2014), increasing researchers' difficulties to tap into the relevant literature and identify topics of interest and research fronts (Redfern, Cobo and Herrera-Viedma 2018). In this context, science mapping solutions can become key tools for easing researchers' burden. In this study we aim at identifying topics of social interest within the field of Microbiology and exploring the mechanisms which might explain such social interest. We do so by using data extracted from mentions from news media, policy documents and Twitter to scientific publications, instead of citation data, as it has been traditionally been conducted. Despite the expansion of the use of bibliometric techniques and methods to analyze specific scientific fields and areas, they have been rarely applied to the field microbiology (Nai 2017). The ones that have been conducted have either focused on a particular topic or aspect (Brandt *et al.* 2014) or have focused on the main actors and regions active in the field and their evolution over time (Vergidis *et al.* 2005). But, to our knowledge, no study has tried to fill the science-society gap, by aiming at connecting research topics with societal interest.

Science mapping has been extensively used in the context of research evaluation for identifying research priorities (Cassi *et al.* 2017), to aid governance of specific areas (Rotolo *et al.* 2014) or to profile institutions' research portfolio (Rafols, Porter and Leydesdorff 2010). The expansion of science mapping applications is largely derived to the free availability of academic software and tools that are constantly maintained and updated (Cobo *et al.* 2011). These maps usually combine publication data with citation data, although there are notable exceptions (Klavans and Boyack 2014). In this paper, we use altmetric data in combination with publications.

Altmetrics have become a promising research front in the field of research evaluation and scholarly communication. They are based on the notion that non-formal channels of scholarly communication are shifting to the Internet (Priem 2014). Therefore, by tracking these alternative channels, it is possible to identify and access literature which might not only be relevant to scientists, but also to lay people. Although most interest on altmetrics has focused on applying it for research assessment (Robinson-Garcia, van Leeuwen and Rafols 2018), they can also be used as tools for discovery. In this sense, altmetric data has not been used for science mapping until quite recently (Didegah and Thelwall 2018) pointing out towards it interest as a descriptive tool to showcase thematic landscapes (Wouters, Zahedi and Costas 2018).

Objectives

The main goal of this paper is to visualize the main topics of social interest as identified via altmetric data. Moreover, we will explore how topics are captured by social media and which are the main drivers of such attention. For this, we focus on three specific altmetric sources: Twitter, news media and policy documents. The selection of these sources is due to the following reasons. Twitter is the main general platform feeding altmetric data both in coverage of publications as well as intensity (Robinson-García *et al.* 2014) and is the most researched of the social media platforms conforming altmetrics (Thelwall *et al.* 2013; Robinson-García *et al.* 2017). Policy mentions and news media are the most robust sources in the sense that they are theoretically easier to interpret, and hence, to understand the underlying meaning behind them.

Materials and Methods

Here we analyze Twitter, news media and policy briefs' mentions to publications in the field of Microbiology. In 22 October 2018, we retrieved a total of 382,998 records of publications indexed in the subject categories of Microbiology and Biotechnology & Applied Microbiology from Web of Science in the 2012-2018 period. Altmetric data was obtained from Altmetric.com, one of the main secondary providers of altmetric data (Robinson-García *et al.* 2014). Prior altmetric data is scarce and incomplete as this database started to systematically retrieve altmetric data in 2011. To query Altmetric.com we need to use publications' Digital Object Identifiers (DOI), this means that all publications without DOI will be lost from our final dataset. 88.2% of the dataset included DOI numbers. After querying Altmetric.com we identified a total of 174,799 distinct publications which are at least mentioned once by any of the sources covered by Altmetric.com. Furthermore, we downloaded all mentions retrieved from Altmetric.com, that is, a total of 1,594,856 records. These records do not only indicate the publication being mentioned, but more importantly, the author of such mention. Table 1 includes some descriptive of the total number of mentions by platform, and papers mentioned. As observed, around 90% of mentioned papers were mentioned by Twitter users, being the most prominent source of altmetrics. News stories cover around 10% of mentioned publications, while policy documents barely cite 2% of mentioned publications.

Two mapping approaches were followed in this study. First, based on the dataset of mentions, we conducted a two-mode network analysis to identify the most influential actors and the papers being mentioned by them. That is, papers are connected to each other either through the actors which mention them. These actors can be Twitter users, news media or organizations publishing policy briefs. Two-mode networks consist on two types of actors (i.e., publications and altmetric sources: Twitter accounts, organizations publishing policy briefs and news

media) connected by a direct relation between each other (i.e., altmetric source mentions publication).

Table 1. Descriptive of mentions and papers mentioned in Altmetric.com by platform from all publications indexed in Web of Science subject categories Microbiology and Biotechnology & Applied Microbiology for the 2012-2018 period. In bold platforms used in this paper.

Platforms	Mentions	Share of mentions	Number of mentioned papers	Share of mentioned papers
Tweet	1345909	84.4	156912	89.8
News story	80485	5.1	16529	9.5
Facebook post	73189	4.6	28394	16.2
Blog post	29622	1.9	18090	10.4
Patent	24001	1.5	10100	5.8
Google+ post	14834	0.9	6716	3.8
Wikipedia page	10243	0.6	7623	4.4
Policy document	4414	0.3	3295	1.9
F1000 post	4175	0.3	3589	2.1
Reddit post	3769	0.2	3120	1.8
Peer review	1767	0.1	751	0.4
Weibo post	1083	0.1	298	0.2
Video	1040	0.1	799	0.5
Q&A post	224	0.0	209	0.1
Pin	53	0.0	46	0.0
LinkedIn post	48	0.0	48	0.0
<i>Total</i>	<i>1594856</i>	<i>100.0</i>	<i>174799</i>	<i>100.0</i>

In a second step, we aim at identifying the most discussed topics by each media analyzed. In this case, we create a thematic landscape based on terms contained in the titles of papers with mentions in Altmetric.com. This landscape offers an overview of what is being discussed in social media but does not provide any information about the intensity of such discussions. For this, we overlay the number of mentions from each of the selected platform on the thematic landscape. This term map uses binary counting, that is, we do not consider how many times terms occur in a single title but the number of times they occur in different titles from other publications.

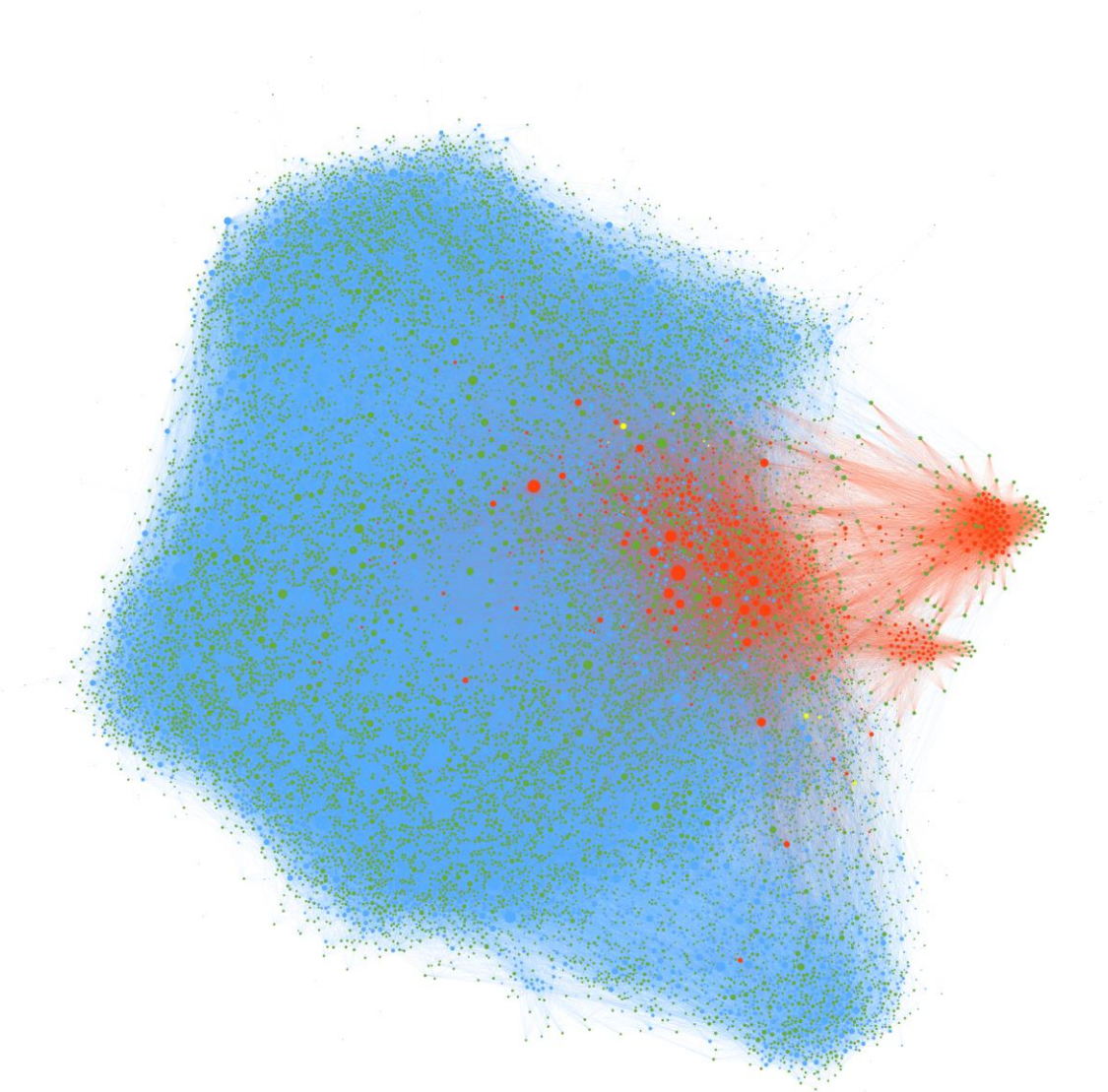


Figure 1. Two-mode network of Microbiology publications and altmetric actors mentioning them. We only show 6% of the network for displaying reasons. Publications are represented in green and they correspond to all publications included in Altmetric.com from the Web of Science subject categories of Microbiology and Biotechnology & Applied Microbiology. Nodes in blue are Twitter accounts, in red are news media and in yellow, organizations producing policy briefs. Map created with Gephi v. 0.9.1

Results

Figure 1 shows a two-mode network of publications and the actor mentioning such publications. The largest portion of mentions to literature in the field of Microbiology come from Twitter discussions. Twitter users represent around 50% of the nodes of the network and involve 94% of the links observed. While some of the discussion generated from Twitter revolves around stories published by news media, there is a large portion of literature which is

only discussed in Twitter. Literature discussed solely in Twitter are literature reviews, publications related with crystallography or news stories affecting either academia (e.g., open science) or specifically with regard to the field of microbiology (e.g., calls for best practices).

There is a smaller cluster of news media mentions to literature which are not discussed in Twitter. These news stories come from local and regional US news media such as Mississippi News Now or KSLA News (from Los Angeles), as well as agencies (e.g., EuropaPress). In the case of news media mentions to publications also discussed in Twitter, we find internationally renowned media such as The Economist, EurekaAlert! or Scientific American. Most discussed publications in this cluster have to do with chemotherapy solutions and topics related with oncology or advancements on the development of vaccines to prevent viral diseases. On the contrary, publications discussed also in Twitter are related with topics which seem to be less applied and more appealing to the curious mind. Here we find papers on the identification of new viruses, calls for the preservation of microbial diversity or new brewing techniques for beer production. Also, news media outlets vary, although there is a high degree of US regional and local media, also some national news media are present such as PBR. However, in all cases we must note high predominance of US media. Policy briefs are scarce and tend to connect publications which are both, discussed in Twitter and by news media. These tend to cite publications discussing specific health issues such as outbreaks in animals or humans in different places of the world. Also, some of the studies cited target specific human groups such as pregnant women.

There is a total of 216735 Twitter accounts that mention at least one publication. 66.05 percent of Twitter accounts mention only one publication, while 12.37 percent do so with 2, 5.42 percent with 3 and from 7 mentions the percentage is below 1 percent. The mean of unique mentions is 5.44 (\pm 53.02). In figure 2 we focus of the top 25 accounts driving the conversation in Twitter. We distinguish between the number of tweets mentioning publications and the actual number of papers which are mentioned. We manually assign an account type to these 25 cases. In all of them, although numbers are similar, there are always papers which are mentioned in more than one occasion. Even more, there is one account (@FarmFairyCrafts) which has only mentioned 5 publications, but these have been mentioned in more than 3500 tweets. While in this case, the account belongs to a firm in Texas, we observe that 12 of these top 25 accounts are bots, followed by 7 accounts from academics, related to scientific journals, one to a news media and one to a physician. These 12 bots are responsible for 4% of the tweets which are directed to 32% of the papers in our sample.

Account name	ALTMETRIC DATA		ACCOUNT DATA			
	Nr Papers tweeted	Nr Mentions	Account type	Nr Tweets	Following	Followers
@AntibioticResis	6497	7699	Bot	27300	931	9024
@yeast_papers	6147	6342	Bot	33300	3	1349
@rnomics	3863	6218	Bot	16200	119	2295
@jcamthrash	5441	5689	Academic	19900	489	7008
@FrontMicrobiol	5007	5290	Journal	7222	816	5913
@EvolvedBiofilm	3898	4940	Academic	29300	1156	4055
@msmjetten	2948	4929	Academic	35800	507	2719
@micro_papers	4448	4587	Bot	12200	11	210
@MicrobiomePaper	4218	4465	Bot	19800	53	3154
@biofilmPapers	4305	4416	Bot	14000	64	920
@pseudo_papers	4030	4048	Bot	16400	35	797
@ndm1bacteria	3221	3981	Press	15500	49	581
@PLOSPathogens	2610	3978	Journal	6601	2530	21700
@Immunol_papers	2920	3969	Bot	61700	0	771
@animesh1977	3129	3962	Professional	954	1059	1592
@phy_papers	3906	3946	Bot	20800	1	2252
@custom_ms	3777	3865	Bot	11100	2	27
@BIOCIENCIA2013	3192	3717	Academic	67200	565	967
@MicrobiomDigest	3356	3597	Academic	34000	15100	20400
@bmgphd	2636	3589	Academic	8375	228	496
@FarmFairyCrafts	5	3578	Company	755	20700	27500
@ASM microbiology	2895	3251	Academic	18900	218	37400
@BioinformaticsP	3052	3062	Bot	4865	27	318
@NatureRevMicro	2686	3059	Journal	10300	1340	38800
@transcriptomes	2638	2804	Bot	21300	4	897

Figure 2. Top 25 Twitter accounts mentioning publications included in Altmetric.com from the Web of Science subject categories of Microbiology and Biotechnology & Applied Microbiology.

Mentions to publications from news media seem to be ridden by news agencies and media specialized in scientific literature. In figure 3A we observe the top 15 news media mentioning microbial literature. EurekaAlert!, a service that provides news releases to journalists, stands out as the main news media. The rest of the list is mostly populated by media focused on medical literature (e.g., Health Medicinet, MedicalXpress).

In the case of policy briefs (Figure 3B), the World Health Organization is the most prominent institution citing microbial literature. Along with this organization, we find other international bodies such as the Food and Agriculture Organization of the United Nations (FAO), but most of the top institutions citing microbial literature in their policy briefs are of a national or regional scope. Here we highlight the presence of the UK Government, the US Centers for Disease Control and Prevention, the European Union or the Netherlands Environmental Agency (PBL).

Next, we expand and focus on the topics discussed by each social source. For this, figure 4 maps terms included in titles of all microbial publications indexed in Altmetric.com and

overlays the focus of discussion for each source. It shows the base map where nodes represent words and noun phrases from titles and colors represent clusters of topics. Seven large topics are identified. The red clusters relates with molecular and cell biology. The yellow cluster represents papers closer to clinical and translational medicine related to virus biology and immunity. The blue cluster in the bottom right, refers to bacterial infections and hubs. A light blue cluster is spread in the middle of the map on the left side of the red cluster and on the left side of the blue one. Such spread is due to the fact that it refers to methodological approaches and techniques for discovery. A separate orange cluster can be observed on the bottom left. Here we find terms related with taxonomies of bacterial species. The large green cluster on the left side represents bioengineering research. Lastly, we observe a purple cluster in the middle of the map just beneath the red cluster. Here we observe terms such as progress, current status or future, which point at future prospects and state of the art papers.

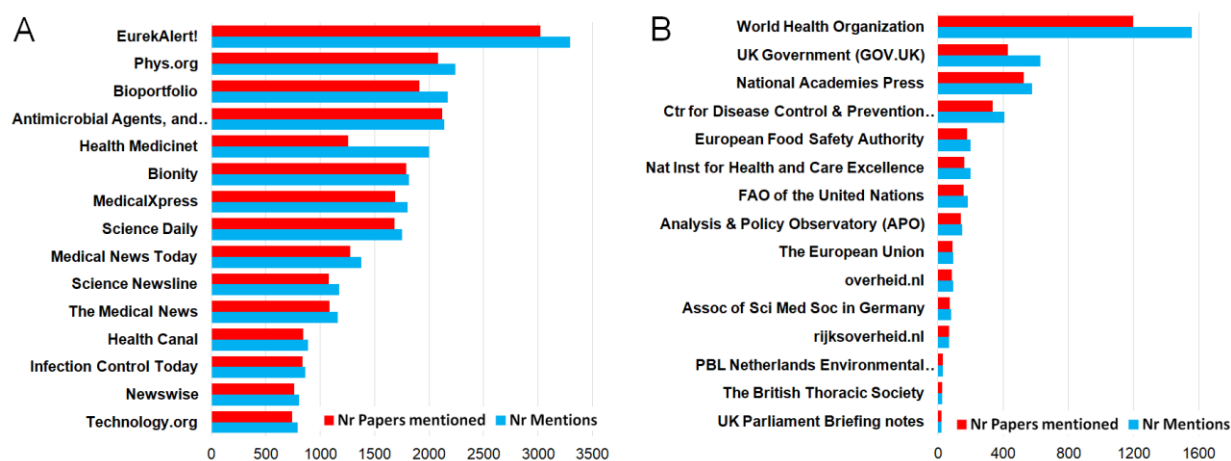


Figure 3. Top 15 **A** news media and **B** international organizations mentioning publications included in Altmetric.com from the Web of Science subject categories of Microbiology and Biotechnology & Applied Microbiology.

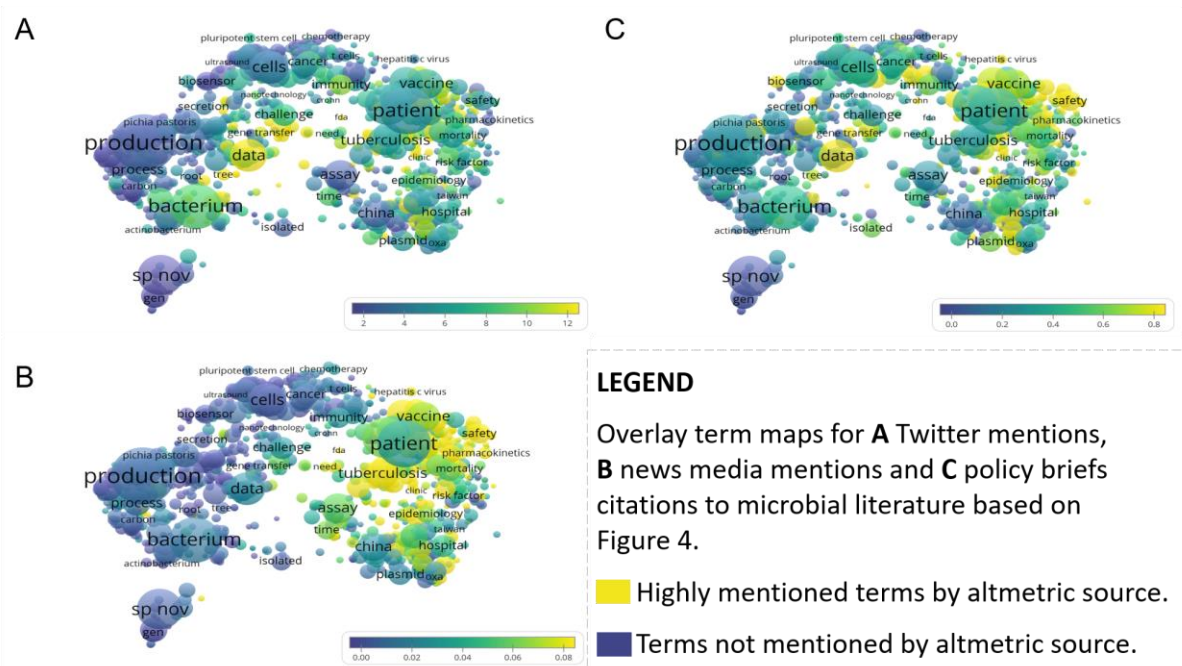


Figure 5. Overlays maps based on figure 4. **A** overlays tweet mentions to terms, **B** overlays news media mentions and **C** overlays mentions from policy briefs. Map created using VOSviewer v. 1.6.10

Discussion

This paper analyzes and identifies topics of social discussion on microbial literature by analyzing mentions to scientific publications in Twitter, news media and policy briefs. We identify and describe which are the actors or channels riding such discussion. For this we make use of mapping techniques and network analysis. Not surprisingly, most of the mentions identified come from Twitter activity. Interestingly, we do find separate clusters of discussion (figure 1): a large cluster of tweets with half of it closely related to news media and two isolated clusters of news media mentioning papers. This signifies that there are publications which drive news media attention but are not discussed socially via Twitter, while there are many other papers which generate a large amount of Twitter attention but are not of interest neither for news media or policy briefs. Policy briefs tend to cite publications which also drive news media and Twitter attention. Differences on news media attention seem to revolve around the locality or globality of topics. Altmetric.com's selection of news media is severely biased towards English language (Robinson-García *et al.* 2014), which explains why news media in these isolated clusters are mainly local and regional US media. Regarding thematic differences, we note that most of the papers mentioned solely by Twitter seem to revolve around academic discussion, confirming the role of Twitter as a non-formal channel of communication for academics, rather than for lay people (Sugimoto *et al.* 2017). On the other hand, we observe that papers mentioned by news media, policy briefs and Twitter combined,

are related with social issues such as advancements on therapies and vaccines and the identification of viral or bacterial diseases.

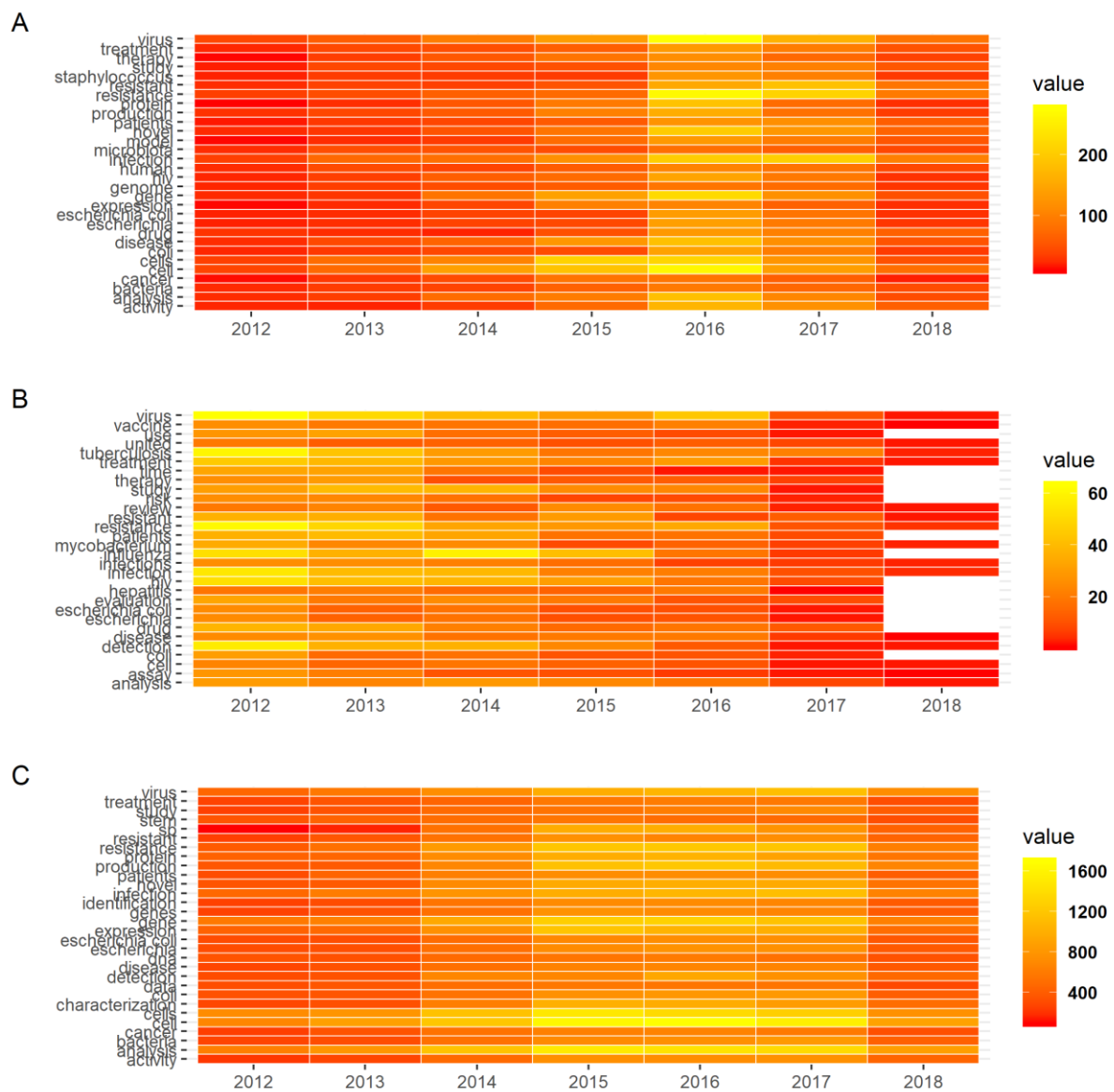


Figure 6 Heatmap of the top 30 most occurring noun phrases in papers mentioned by **A** news media, **B** policy briefs and **C** Twitter accounts.

Our analysis on the top Twitter accounts mentioning microbial literature (Figure 2) confirm that the existence of bots compromises to a great extent any use of tweet mentions as a means to assess the societal relevance of specific papers (Haustein *et al.* 2016; Robinson-Garcia *et al.* 2017). In the case of news media (Figure 3A), there is a preponderance of science specialized media within the media citing the most microbial literature, as well as the above mentioned biased towards English language which obscures local interest from non-English speaking countries, limiting to a great extent the identification of socially interesting topics in peripheral

regions (Alperin 2013). Organization publishing policy briefs which cite microbial literature are both of an international and national scope (Figure 3B). Furthermore, while some of them are field specific (e.g., the British Thoracic Society), others are of a much broader breadth and social influence (e.g., Food and Agriculture Organization of the United Nations or the European Union).

Finally, we observe differences on the topics discussed by each of these sources (Figure 4). While news media seem to cover a wider range of topics, Twitter mentions seem to be more related to future prospects of the field as well as translational research and virology. Policy briefs, on the other hand, are thematically focused on bacterial infections and hubs, and viral diseases and translational medicine. This reveals a great thematic dependency on what drives more social attention (Noyons and Rafols 2018) which compromises general statements and suggestions to push all scientific efforts towards socially relevant topics (e.g., UK's Research Excellence Framework), as it would work on the benefit of some areas and on the detriment of others.

In this paper we have shown that the combination of advanced visualization techniques, network analysis and different altmetric data sources, provides valuable information not only to identify topics of social interest, but also to better assess how such attention is generated and better interpret such differences on topics and communities. While these analyses are still rare with most of the efforts analyzing altmetrics focused on research assessment (Bornmann 2014), recent calls for the use of altmetrics for contextualizing how social attention of research is generated and identifying areas of social engagement (Robinson-Garcia, van Leeuwen and Rafols 2018; Wouters, Zahedi and Costas 2018) will hopefully help to develop advanced methods which can better inform academics and research managers to spot and understand how social attention is generated.

With regard to the topic clusters identified generating more attention by altmetric source, while the results are somehow expected (e.g., viral diseases being of higher interest in news media than bioengineering), they allow to validate the combination of methods. These methods can be of greater interest if more closely refined and directed at specific targets (e.g., social interest of Zika in Latin America) to better understand how research outputs are perceived and used by the public. For instance, by targeting specific terms or noun phrases and analyzing frequency of occurring in publications mentioned by an altmetric data sources and monitoring peaks of attention, similarly to what we show in Figure 6.

References

- Alperin JP. Ask not what altmetrics can do for you, but what altmetrics can do for developing countries. *Bull Am Soc Inf Sci Technol* 2013;**39**:18–21.
- Bornmann L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *J Informetr* 2014;**8**:895–903.
- Cassi L, Lahatte A, Rafols I *et al.* Improving fitness: Mapping research priorities against societal needs on obesity. *J Informetr* 2017;**11**:1095–113.
- Cobo M j., López-Herrera A g., Herrera-Viedma E *et al.* Science mapping software tools: Review, analysis, and cooperative study among tools. *J Am Soc Inf Sci Technol* 2011;**62**:1382-402.
- Didegah F, Thelwall M. Co-saved, co-tweeted, and co-cited networks. *J Assoc Inf Sci Technol* 2018;**69**:959–73.
- Haustein S, Bowman TD, Holmberg K *et al.* Tweets as impact indicators: Examining the implications of automated “bot” accounts on Twitter. *J Assoc Inf Sci Technol* 2016;**67**:232–238.
- Klavans R, Boyack KW. Mapping altruism. *J Informetr* 2014;**8**:431–47.
- Noyons E, Rafols I. Can bibliometrics help in assessing societal contributions of agricultural research? Exploring societal interactions across research areas. *STI 2018 Conference Proceedings*. Leiden (The Netherlands): Centre for Science and Technology Studies (CWTS), 2018, 1049–57.
- Priem J. Altmetrics. In: Cronin B, Sugimoto CR (eds.). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press, 2014, 263–88.
- Rafols I, Porter AL, Leydesdorff L. Science overlay maps: A new tool for research policy and library management. *J Am Soc Inf Sci Technol* 2010;**61**:1871–1887.
- Redfern J, Cobo MJ, Herrera-Viedma E. Editorial: Mapping microbiology with scientometrics - help provide a clearer vision of microbiology research around the globe. *FEMS Microbiol Lett* 2018;**365**, DOI: 10.1093/femsle/fny061.
- Robinson-Garcia N, Costas R, Isett K *et al.* The unbearable emptiness of tweeting—About journal articles. *PLOS ONE* 2017;**12**:e0183551.
- Robinson-Garcia N, van Leeuwen TN, Rafols I. Using Almetrics for Contextualised Mapping of Societal Impact: From Hits to Networks. *Sci Public Policy* 2018;**45**:815–26.
- Robinson-García N, Torres-Salinas D, Zahedi Z *et al.* New data, new possibilities: exploring the insides of Altmetric. com. *El Prof Inf* 2014;**23**:359–366.
- Rotolo D, Rafols I, Hopkins MM *et al.* Scientometric mapping as a strategic intelligence tool for the governance of emerging technologies. *Available SSRN 2239835* 2014.
- Sugimoto CR, Work S, Larivière V *et al.* Scholarly use of social media and altmetrics: A review of the literature. *J Assoc Inf Sci Technol* 2017;**68**:2037–62.

- Thelwall M, Haustein S, Larivière V *et al.* Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE* 2013;**8**:e64841.
- Van Noorden R, Maher B, Nuzzo R. The top 100 papers. *Nature* 2014;**514**:550–3.
- Waltman L, Eck NJ van. A smart local moving algorithm for large-scale modularity-based community detection. *Eur Phys J B* 2013;**86**:471.
- Wouters P, Zahedi Z, Costas R. Social media metrics for new research evaluation. In: Glänzel W, Moed HF, Schmoch U, *et al.* (eds.). *Handbook of Quantitative Science and Technology Research*. Springer, 2018.

Identifying and characterizing social media communities: a socio-semantic network approach to altmetrics




Wenceslao Arroyo-Machado^{1,*}, Daniel Torres-Salinas¹ and Nicolas Robinson-Garcia¹

¹EC3 Research Group, Department of Information and Communication Sciences, University of Granada, Faculty of Communication and Documentation, Granada, Spain

*Corresponding author: wences@ugr.es

Journal

Scientometrics

0138-9130 

Index

SCIE – Q2

DOI

10.1007/s11192-021-04167-8

Data

10.5281/zenodo.4148941

Version

Published

References

APA 7th

Funding



Abstract

Altmetric indicators allow exploring and profiling individuals who discuss and share scientific literature in social media. But it is still a challenge to identify and characterize communities based on the research topics in which they are interested as social and geographic proximity also influence interactions. This paper proposes a new method which profiles social media users based on their interest on research topics using altmetric data. Social media users are clustered based on the topics related to the research publications they share in social media. This allows removing linkages which respond to social or personal proximity and identifying disconnected users who may have similar research interests. We test this method for users tweeting publications from the fields of Information Science & Library Science, and Microbiology. We conclude by discussing the potential application of this method and how it can assist information professionals, policy managers and academics to understand and identify the main actors discussing research literature in social media.

Citation

Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. *Scientometrics*, 126(11), 9267–9289. <https://doi.org/10.1007/s11192-021-04167-8>

1. Introduction

Research literature is increasingly mentioned, shared and discussed on social media. This represents a substantial challenge as well as an opportunity to anyone trying to study the interactions that take place in the digital environment (Stieglitz et al., 2018). It provides researchers with major opportunities to develop novel methodological solutions by which to inform policy managers, journalists and information professionals on the way by which scientific literature is consumed. In vastly differing fields, many ad hoc solutions exemplify the growing interest in social media. In the field of science communication, for example, research has been conducted into the anti-vaccine movement on Twitter (van Schalkwyk et al., 2020), the dissemination of fake medical news (Waszak et al., 2018), or political communication and the influence of Twitter (Davis et al., 2017). In marketing, a substantial, growing number of social media metrics and analytics have been applied (Misirlis & Vlachopoulou, 2018). In disaster management, information propagated by social media such as Facebook and Twitter has formed the basis for new proposals (Kim & Hastak, 2018); and the digital humanities' community on Twitter has been identified and analyzed (Grandjean, 2016).

In scientometrics, these studies have led to the emerging sub-field of altmetrics (Priem et al., 2010), in which mentions to scientific literature on social media are tracked to explore the social reception of research findings. However, this line of research has not been free of controversy. Initial high expectations of the potential value of tracking aspects of social or broader impact on research (Bornmann et al., 2019; Haustein, 2016) were soon rejected in the face of hard evidence (Robinson-Garcia et al., 2017; Sugimoto et al., 2017). Nonetheless, the relevance of social media in scholarly communication remains unquestioned (Robinson-Garcia et al., 2018; Wouters et al., 2019), leading to a new scenario in which novel metrics are being developed to understand and describe aspects of science communication that transcend traditional academic channels.

The rich variety of social platforms (Wikipedia, Mendeley, Twitter, and so on) has given rise to the development of altmetric data aggregators that provide data on a variety of social media sources. These include Altmetric.com, CrossRef Event Data, or Plum Analytics, among others. Despite the evident advantage of offering unique data access points, they do have limitations. Zahedi and Costas (2018) systemically compared altmetric data providers' coverage, metrics and sources. They found differences in data collection, the identification and merging of different versions of a single publication, and data update periodicity. These can be added to other limitations directly related to the nature of social media and the concept of altmetrics, namely heterogeneity, quality and dependencies (Haustein, 2016).

For a variety of reasons, Twitter is the social media platform that has received most attention since the earliest days of altmetric studies. In part, this is because it is the public forum with the second-highest figures for coverage of scientific literature mentions after Mendeley (Robinson-Garcia et al., 2014). Nonetheless, while it is widely used by the general public, it has a relatively low level of acceptance among scientists. Most studies report that around 15% of academics have a Twitter account (Haustein, 2019), although the annual growth rate is constant (Joubert & Costas, 2019).

After initially promising results (Eysenbach, 2011), studies report that Twitter mentions to scientific papers poorly reflect citation impact (Haunschild & Bornmann, 2018). Furthermore, the inclusion of automated bots (Haustein et al., 2016) and the un-informative way in which scientific papers are tweeted (Robinson-Garcia et al., 2017) question the extent to which simple counts of tweets mentioning papers can be informative. Many studies have focused on characterizing the Twitter profiles of individuals who tweet scientific literature to better understand who they are (Díaz-Faes et al., 2019; Ke et al., 2017). The present study adds to this growing trend in the literature by proposing a methodological approach through which communities of actors can be identified on the basis of their scientific preferences. Our goal is to develop tools that can inform on targeted groups interested in specific topics which can later be characterized by other methods, as mentioned earlier. To achieve this, we build on previous studies that investigated differences in topics of interest across social media platforms (Arroyo-Machado et al., 2019; Robinson-Garcia et al., 2019).

The paper is organized as follows: first, we briefly review the literature and focus on three specific topics, Altmetric studies, studies specifically about Twitter, and studies relating to mapping and visualization techniques. Secondly, we formulate our objectives. We then describe our data retrieval and data processing and present our methodological proposal. We apply this in the field of Information Science & Library Science and in the field of Microbiology. We conclude by discussing our findings.

2. Background

2.1 Altmetric studies

Altmetrics were formally proposed in 2010 with the publication of the Altmetrics Manifesto (Priem et al., 2010), although similar proposals had appeared previously (Neylon & Wu, 2009; Nielsen, 2007; Taraborelli, 2008). The emergence of altmetrics led to a fundamental transformation of the field of scientometrics. This occurred at a time when different metrics, sources and indicators co-occurred, moving the field from an almost universal dependence on certain bibliometric databases to a heterogeneous range of data sources. Although scientometricians acknowledged the technical limitations of altmetrics from the very

beginning (Torres-Salinas et al., 2013), an overall optimism led many to consider them an alternative to citation metrics and compared and analyzed their relationship with traditional metrics (Costas et al., 2015; Thelwall, 2018). But, apart from Mendeley (Thelwall, 2018), evidence only suggests the existence of a weak positive correlation.

This led to a change in the discourse and altmetrics began to be presented as a complement to citations (Haustein et al., 2015), rather than an alternative. While acknowledging their potential to inform on other indicators of scientific information consumption, there seems to be a consensus that they cannot be interpreted uniformly and that context plays an important role in their interpretation. This has led many to refer to altmetric indicators as metrics that capture an ‘unknown impact’ of scientific outputs (Bornmann et al., 2019; Kassab et al., 2020).

Since then, effort has been directed at studying the context in which this unknown impact is produced, identifying new channels of scholarly communication that go beyond the traditional (Holmberg et al., 2019). This shift has led some authors to refer to these new studies as studies on social media metrics (Wouters et al., 2019) and define them as ‘second generation metrics’ (Díaz-Faes et al., 2019). While the previous one transferred the citation model to social media, here the focus is on the activity and interactions that take place on social media. This leads to a new scenario in which the altmetric research is focused on the relational attributes of the social media activity rather than focusing on features (i.e, impact) related to scientific publications. To do so, the methodological framing has also changed, focusing now on techniques which help discover and analyze different kinds of social interactions (Costas et al., 2020) that allow a better understanding of science-society relations. However, these new approaches focus mainly on researchers discovering and topic visualizations in social media. But how can communities of social actors with the same interests be identified? Can communities of social actors who consume scientific literature outside the scientific realm be identified?

Numerous examples of these novel approaches to the use of altmetrics can be found in the literature. Table 1 summarizes 14 such methodological proposals. Essentially, these fall into three categories of application or approach: identification and characterization of researchers; visualization of topics discussed; and knowledge maps, which center on descriptive analyses and co-citation and co-word network analyses. Also, most of these studies revolve around the use of Twitter and Wikipedia. Colavizza (2020) estimated how well Wikipedia, as a tool communicating scientific knowledge to the general public, reflects current scientific progress on COVID-19. Similarly, science mapping techniques haven been used to analyze how Wikipedia structured science in comparison with global science maps based on bibliometric databases (Arroyo-Machado et al., 2020); and the humanities (Torres-Salinas et al., 2019).

Table 1 Main altmetric studies and methodological proposals by source of literature.

	Application	Methodology	Source	Scope
Zahedi and van Eck (2018)	Profiling of Mendeley readers	Descriptive analysis and overlay visualizations	Mendeley	Multidisciplinary
Costas et al. (2017)	Identify Researchers on Twitter	Rule-based methods	Twitter	Multidisciplinary
Ke et al. (2017)	Profiling of Twitter users	Descriptive and network analysis	Twitter	Multidisciplinary
Alperin et al. (2018)	Effectiveness of Twitter dissemination on outreach	Descriptive and network analysis	Twitter	Biomedicine
Robinson-Garcia et al. (2018)	Profiling researchers on Twitter	Social network analysis	Twitter	Multidisciplinary
Díaz-Faes et al. (2019)	Characterize Twitter communities interacting with science	Descriptive analysis and overlay visualizations	Twitter	Multidisciplinary
Haunschild et al. (2019)	Topic visualizations based on Twitter discussions	Co-word analysis	Twitter	Climate change
Hellsten and Leydesdorff (2019)	Topic and actor visualizations based on Twitter discussions	Co-occurrence of hashtags and mentions	Twitter	Biomedicine
Haunschild et al. (2020)	Topic visualizations based on Twitter discussions	Co-word analysis	Twitter	Library and Information Science
Robinson-Garcia et al. (2019)	Topic visualizations based on Twitter discussions	Social network analysis and overlay visualizations	Twitter	Microbiology
Torres-Salinas et al. (2019)	Mapping knowledge relationships in Wikipedia	Co-citation analysis	Wikipedia	Humanities
Arroyo-Machado et al. (2020)	Mapping knowledge relationships in Wikipedia	Co-citation analysis	Wikipedia	Multidisciplinary
Colavizza (2020)	Coverage of research topics in Wikipedia	Topic modeling and regression analysis	Wikipedia	COVID-19
Piccardi et al. (2020)	Measuring interactions with Wikipedia references	Engagement metrics	Wikipedia	Multidisciplinary

In addition to Wikipedia, other social media sources have also been used to study the dissemination of scientific activity. For instance, Mendeley has been studied to identify its user types' interests in and their patterns of use of scientific publications (Zahedi & van Eck, 2018). However, in this respect, Twitter is the platform that has most frequently been studied.

2.2 Twitter

Regarding the use of Twitter data, we find a first stream of studies that focus on identifying researchers or users who mention scientific publications and contextualize their activity. Among these we refer to studies like Ke et al. (2017), which identifies scientists from different disciplines; Robinson-Garcia et al. (2018), which proposes the use of mapping techniques to contextualize academics' engagement in social media; or Díaz-Faes et al. (2019), which characterizes Twitter profiles mentioning scientific publications and identifies four dimensions of social media communication patterns.

Secondly, we find studies that focus on using Twitter activity to identify topics of interest. These studies attempt to explain differences between the way scientists communicate research and how research is perceived or characterized by Twitter users. They compare differences between Twitter hashtags and author keywords in tweeted publications (Haunschild et al., 2019, 2020); compare topics of interest by social media platform (Noyons, 2019; Robinson-Garcia et al., 2019); or associate instances of interaction and topic by comparing hashtags co-tweeted by the same profiles (Hellsten & Leydesdorff, 2019).

A third line of research is related to the diffusion of scientific publications. These studies aim to determine the social outreach attained by publications disseminated through Twitter (Alperin et al., 2018).

2.3 Mapping and visualization techniques

One feature common to most of the aforementioned studies is their extensive use of mapping and visualization techniques. Based on network analysis, these techniques seek to construct n-dimensional spatial representations of science (Small, 1999). Most such representations are based on the co-occurrence of given events and are easily interpreted. From a bibliometric point of view, science maps are constructed from three elements: actors, resources and contents (Noyons, 2005), each of which offers a different level of analysis. In recent years, interest in mapping has grown as computational and methodological advances have extended their use. Furthermore, the number of visualization tools has increased considerably (cf. Cobo et al., 2011).

Originally, two types of co-occurrence links between similar publications were proposed: co-citation (Small, 1973) and bibliographic coupling (Kessler, 1963). Both were applied at

different levels of aggregation (i.e., co-citation networks of authors [White & Griffith, 1981] or bibliographic coupling for journals [Small & Koenig, 1977]). But the number of co-occurrence types has grown to include co-author networks (Glänzel, 2001) or co-word maps (Callon et al., 1983), among others. Co-word maps facilitate the exploration of structures across the scientific landscape (Waltman & van Eck, 2012) as an alternative to citation networks (Boyack et al., 2005; Leydesdorff et al., 2013).

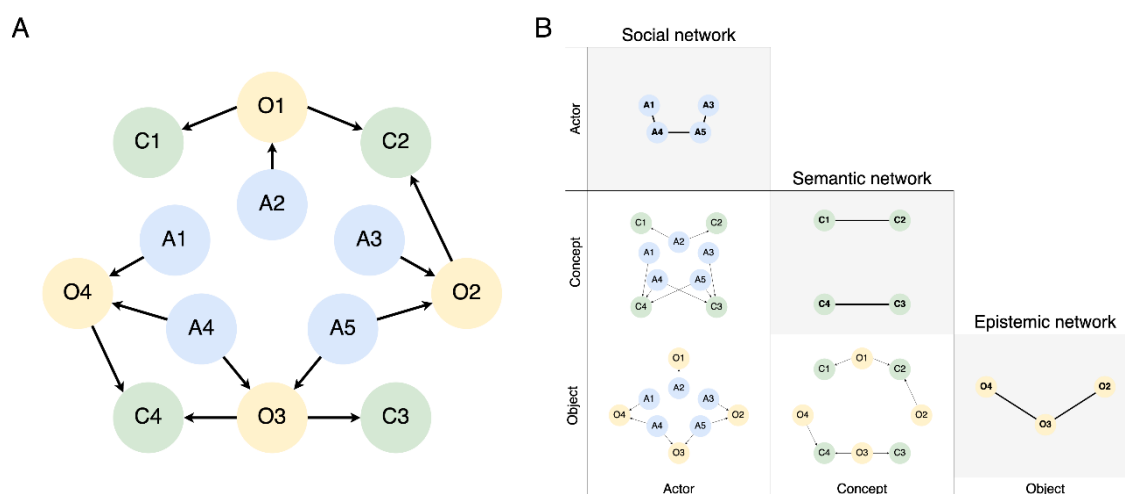
The emergence of new data sources and indicators, including but not exclusively from altmetrics, has led scientometricians to adapt these mapping techniques to the new metrics.

Hence, we find proposals to map scientific literature on the basis of the co-occurrence of publications downloaded by users (Torres-Salinas et al., 2014); to adapt the concepts of co-citation and bibliographic coupling to meet the context of the social media (Costas et al., 2020); and to create thematic landscapes by geographical region (Wouters et al., 2019). These methods can all be used in different contexts. For instance, Arroyo-Machado et al. (2020) created different levels of co-citation networks from Wikipedia entry references. Similarly, Haunschild et al. (2019) built thematic landscapes from co-tweets to visualize public discussion of specific research topics, while Díaz-Faes et al. (2019) used them to characterize the profiles of Twitter users who participate in scientific discussions on the social network. The co-use of hashtags in tweets mentioning scientific literature has also been proposed (Hellsten & Leydesdorff, 2019), as have follower-following networks of scientists who use Twitter (Robinson-Garcia et al., 2018). Clearly, scientific mapping techniques are being adapted to new environments and gaining complexity.

These techniques are based on the social network analysis of actors, relationships and structures (Wasserman & Faust, 1994). They represent any type of entity through nodes and establish relationships between entities that respond to co-occurrences, mentions, or any other type of interaction. Consequently, we can represent science-centered debates on social media at different levels and from different perspectives (Costas et al., 2020).

The rationale behind social network analysis is that by combining co-occurring events, actors can be linked in a 2-mode (bipartite) network. Any such network is based on an asymmetrical matrix in which rows and columns are composed of different entities. Recently, Hellsten et al. (2019) suggested that by aggregating bipartite matrices different combinations could produce additional matrices. Figure 1A shows a 3-mode network that reflects differing but inter-related entities (actors, objects and concepts). Figure 1B shows how these matrices are constructed. Furthermore, the sub-matrices that appear in diagonal, show how entities of the same category are related through the combination of interactions between the other entities.

Fig. 1 n-dimensional matrix constructed by combining the 2-mode matrices of objects, concepts and actors. This representation is based on the conceptual framework proposed by Hellsten et al. (2019)



2.4 Objectives

In the present paper we build on our literature review to better refine methods by which communities with common scientific interests can be identified on social media. We test our methodological proposal using Twitter mentions to scientific papers in two research fields: Information Science & Library Science and Microbiology²². Our main objective is to present a methodological proposal based on social network analysis that allows us to identify cognitive communities by grouping actors who may not necessarily be socially connected but, rather, who are connected through their interests. A proposal that aims to contribute to the new generation of social media metrics (Wouters et al., 2019) as it allows to discover the implicit social and semantic relationships between actors based on the discussion around scientific publications through social media. To this end, we seek to achieve the following objectives:

1. To introduce a novel methodological proposal by which actors in a given network can be grouped on the basis of their cognitive interests thus, to some extent, removing social relationships that could potentially blur the boundaries between communities.
2. To test our methodical approach in a specific case study: Twitter mentions of scientific literature in the field of Information Science & Library Science.
3. To replicate this approach in a different field—Microbiology—to observe potential inconsistencies in the methodology and discuss differences between the two case studies.

²² We selected two categories as distant as possible from each other. Information Science & Library Science and Microbiology belong to very different scientific areas (Social Sciences and Health Sciences) and have significant differences, both, in terms of volume of publications, and communication and collaboration patterns.

Our study closely follows recent work in which a genuine effort has been made to conceptually define and then build a framework in which methodological solutions in the field of altmetrics can be expanded. For instance, Costas et al. (2020) recently proposed the concept of heterogeneous coupling in a study in which, from a theoretical perspective, they explored the potential of social network analysis to reveal links between the social media and science communication. Similarly, Hellsten et al. (2019) present their heterogeneous n-mode method which explores different combinations of interaction between actors. Our proposal could fit well into either of these two except for one noteworthy issue. The goal of our paper is to provide a practical application, showcasing a methodological innovation by which communities can be identified on the basis of common interests.

The present study builds on previous work which analyzed differences in interests of topic by social media platform (Robinson-Garcia et al., 2019) and by clusters (Arroyo-Machado et al., 2019). These earlier studies detected communities of actors who specifically mentioned the same publications and identified the topics that interested them.

3. Data and methods

3.1 Software

The data needed to reproduce our analyses are available at <http://doi.org/10.5281/zenodo.4148941>. We have included supplementary materials at <https://doi.org/10.5281/zenodo.4332921>. Network manipulation of co-word maps (semantic maps) was conducted using Gephi 0.9.2 visualization software (Bastian et al., 2009). As we want an easily replicable methodology fully based on social network analysis, the popular Louvain algorithm is used for community detection (Blondel et al., 2008). Social networks and the overlapping social and semantic networks were constructed using the igraph R package (Csárdi, 2020), and the Louvain algorithm was again used to detect social communities. Both social and semantic networks were tested with the Leiden algorithm (Traag et al., 2019) in Gephi and igraph. In both case studies, the results showed no significant improvements with respect to those derived from applying the Louvain algorithm, so we opted for the original version. Visualizations of intersection sets were constructed using UpSet R software (Lex et al., 2014), a visualization technique that defines the characteristics of the entities studied in order to group them. A detailed description of the data processing and the application of the entire process is available in an R Notebook at https://github.com/Wences91/social_media_communities. All methods have been automated and gathered under the R package 'altanalysis' (<https://github.com/Wences91/altanalysis>).

3.2 Data gathering

We downloaded publication data for two research fields: Information Science & Library Science and Microbiology. We used the former as a case study to test our methodological approach. We then replicated the method in the latter field to compare results and analyze discrepancies in different contexts.

On 17 July 2019 we retrieved all records indexed in the Web of Science (WoS) InCites database (excluding the Emerging Sources Citation Index) published between 2012 and 2018 in the WoS categories of Information Science & Library Science (84 568 publications); and in Biotechnology and Applied Biochemistry (250 577 publications) and Microbiology (187 013 publications)—these two represent a combined total of 413 910 publications, henceforth referred to as ‘Microbiology’. From Altmetric.com’s Altmetric Explorer portal, we extracted all social media mentions of these records by using their DOIs as our search item. Information Science & Library Science has 35 695 publications with DOI (42.21%), and Microbiology has 366 449 (88.53%). Table 2 summarizes the processing tasks undertaken prior to data analysis. We obtained the following datasets:

- Information Science & Library Science: 14 475 publications were mentioned by at least one altmetric source, giving a total of 167 110 mentions from Altmetric.com. Some 151 505 of these (90.66%) were Twitter mentions of 13 458 (92.97%) publications.
- Microbiology: 192 836 publications were mentioned by at least one altmetric source, giving 1 876 599 mentions from Altmetric.com. Some 1 599 315 of these (85.22%) were Twitter mentions of 173 406 (89.92%) publications.

Table 2 Summary of data processing of publications mentioned on social media in Information Science & Library Science and Microbiology

	Information Science & Library Science				Microbiology			
	Publications	%	Twitter mentions	%	Publications	%	Twitter mentions	%
1. Download Web of Science’s InCites records from 2012 to 2018	84,568	100	-	-	413,910	100	-	-
2. Recover all Altmetric.com mentions to InCites publications	14,475	17.12	150,806	100	192,836	46.59	1,585,313	100
3. Data cleaning and filter mentions to only made from Twitter	13,446	15.9	150,723	99.94	173,306	41.87	1,579,896	99.66
4. Remove retweets	13,227	15.64	65,933	43.72	171,085	41.33	695,429	43.87
5. Retrieve Web of Science author keywords and data cleaning	8452	9.99	35,336	23.43	101,206	24.45	327,449	20.66

Our purpose here is to map only those actors who are genuinely involved in Twitter discussions. Retweets have been excluded as they could potentially distort results: they correspond to the platform's social function and do not necessarily indicate participation in scientific debate (Kassab et al., 2020). Twitter mentions retrieved via Altmetric Explorer do not distinguish between tweets and retweets. To identify retweets we searched the Twitter API between 26 December 2019 and 13 January 2020 and removed all retweets from our datasets. This cut the number of Twitter mentions in Information Science & Library Science to 65 933 (43.72% of the original dataset were individual tweets), and in Microbiology to 695 429 (43.87%).

Data processing enabled us to overcome specific limitations. Publications and mentions with no DOI or with a duplicate DOI, were excluded. We also extracted those user names that were missing from the original Altmetric.com dataset from the Twitter API. Thus, in Information Science & Library Science our dataset was further cut to 66 231 mentions (43.72% on Twitter) and in Microbiology to 699 507 (43.74%).

Simultaneously, we extracted author keywords of publications mentioned using terms included in the WoS Author Keywords. These are widely used in bibliometrics and have been previously applied in altmetrics (Haunschild et al., 2019, 2020). Furthermore, we conducted the following processing tasks. All records drawn from the Qualitative Health Research Journal (743 papers) were excluded since it would seem to have been misclassified because most citing journals belong to different categories (Supplementary material, Table C1). Including this journal distorts the semantic map (Supplementary material, Figure C1). Not all publications include author keywords and some journals are left out of the analysis. In Information Science & Library Science there are a total of 239 publication sources, and only 7 journals in the area with more than 10 publications do not include author keywords. From the 747 publication sources of Microbiology there are 18 journals in the area with more than 10 publications not including them either.

Our final Information Science & Library Science dataset constituted 8452 publications (63.9% of the total) with 44 421 keywords, of which 20 027 are unique, and 35 411 Twitter mentions (53.47% of the total); and in Microbiology, our final dataset constituted 101 206 publications (59.16%) with 540 227 keywords, of which 163 674 are unique, and 328 110 Twitter mentions (49.91%).

3.3 Methodological proposal

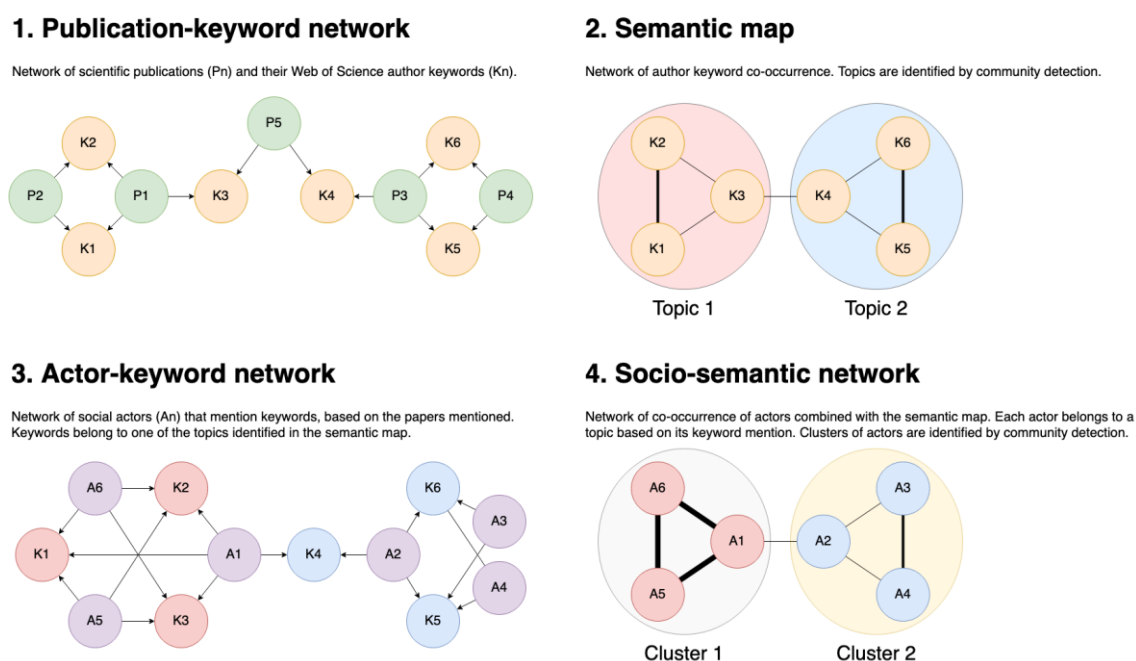
We now describe our methodological proposal to identify communities of interest. This approach can be divided into three distinct phases.

Firstly, we construct a co-word network (semantic map) from the author keywords of publications tweeted in the field. The network is constructed regardless of the number of mentions received and is solely based on the co-occurrence of keywords in scientific publications. It is pruned to remove the weakest co-occurrences, less frequent keywords, and isolated components. Due to the different network sizes and edges' weights (number of times than two keywords co-occur) in the two areas, the established minimums are not the same for both. This map enables us to identify research areas in the field. To do so, we use a social network community detection method. The chosen is the Louvain community detection algorithm (Blondel et al., 2008), where the quality function is the modularity value (Q). We seek a balance between the number and relevance of communities detected and the resulting modularity by applying different resolution values, a parameter which affects the size and number of detected communities. The minimum modularity value set to validate these communities is 0.3 (Newman, 2004). Then the detected communities are tagged taking into account an expert opinion.

Secondly, we assign social actors to topics identified in the map on the basis of the keywords in the papers they discuss. Mentions are combined with the keywords and clusters associated with the papers mentioned. This means that all mentions are divided into as many keyword groups as each paper contains.

Finally, we generate a network of social actors who are linked by the number of tweeted keywords they share (social network). This network is also pruned to remove the weakest relations also following a heuristic strategy, which means that there is no a standard value, but different tests are carried out for this purpose, and reduced to its main component. A community detection is applied to it, using the Louvain community detection algorithm and following the same criteria as in the semantic map. The resulting communities are reflected by areas. To generate the socio-semantic network, each social actor is assigned to its topic, generating a second grouping of social actors, whose quality is calculated by the modularity value. Figure 2 summarizes our approach.

Fig. 2 Overview of our methodological approach to identifying socio-semantic networks of Twitter users on the basis of commonly cited publications



4. Case study: Information Science & Library Science

We identified a total of 13 243 Twitter users mentioning 8452 scientific publications of which 92.65% were articles and 3.42% reviews. Twitter users mention a mean 2.23 publications (SD ±8.79) and 10.59 keywords (SD ±32.32).

The author keywords co-occurrence network is composed of 20 025 nodes and 100 604 edges. It is reduced to 659 nodes and 1315 edges by removing edges with less than 3 co-occurrences and getting its main component. Figure 3 shows the resulting co-word map. We identified four clusters or topics by using a resolution value of 2.5 (Q = 0.62). These were tagged manually on the basis of expert opinions. We found these topics centered on social media (34.14% of nodes in the network), bibliometrics (26.56%), libraries (21.4%), and information retrieval (17.9%). The contents of the clusters were:

- **Social Media:** a community consisting of 5511 Twitter accounts, disseminating 2870 publications in 11 684 tweets and sharing 225 keywords. It includes publications related to social media use, the ethics of their use, their use by young people, and the application of big data techniques in social media analysis.
- **Bibliometrics:** a community consisting of 4989 Twitter accounts, disseminating 2229 publications in 11 984 tweets and sharing 175 keywords. This community includes publications related to bibliometrics and altmetrics analysis and covers issues relating to open science and science policy.

- Libraries: a community consisting of 2854 Twitter accounts, disseminating 1658 publications in 6297 tweets and sharing 141 keywords. This community includes publications relating to general, academic or specialized libraries, their evaluation, and the analysis and training of users.
- Information retrieval: a community consisting of 3522 Twitter accounts, disseminating 1486 publications in 7651 tweets and sharing 118 keywords. This community includes publications relating to information storage and retrieval, its application in electronic health records, the use of ontologies and classification systems and their interoperability.

Fig. 3 Information Science & Library Science thematic landscape. This map shows the main components of the network and those terms that co-occur 3 times or more. It contains 659 WoS author keywords

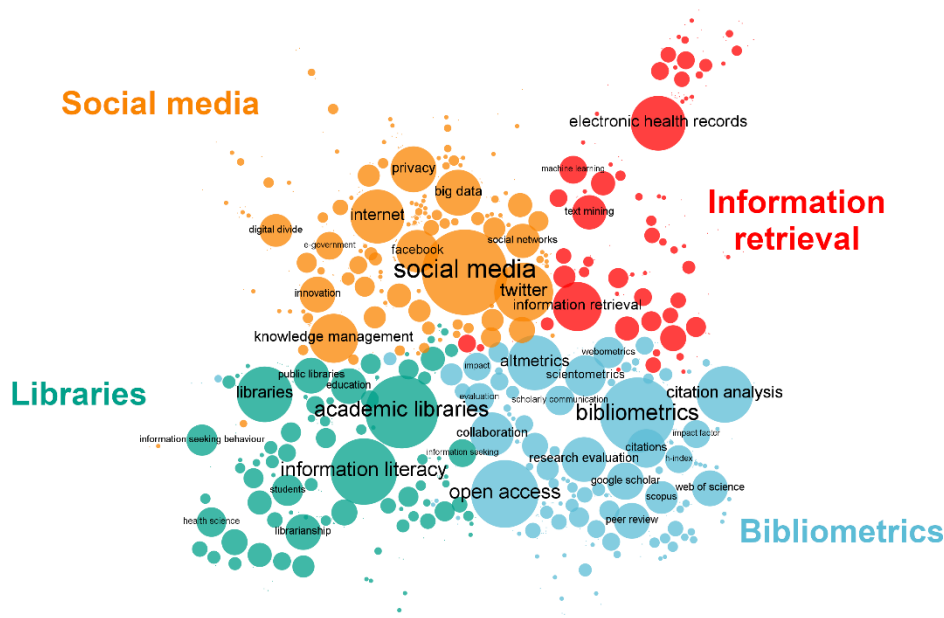
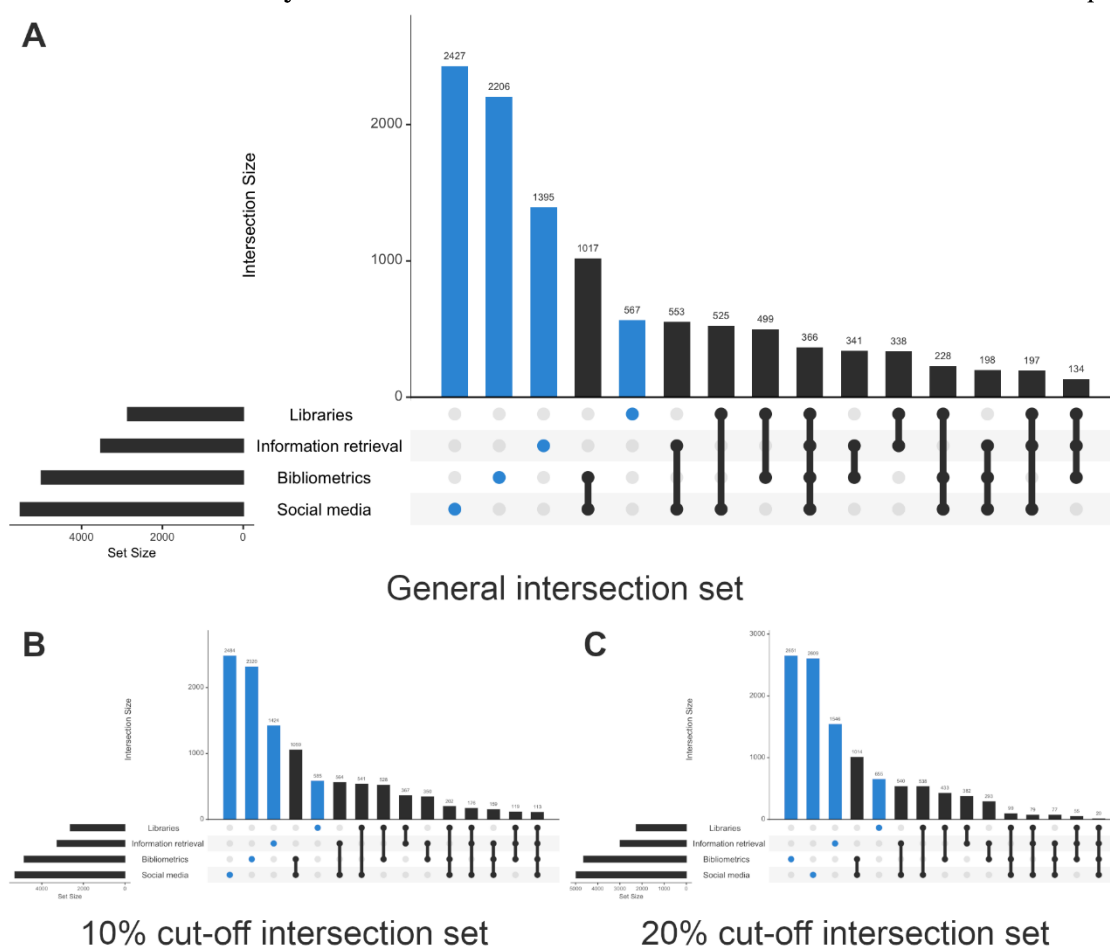


Figure 4 shows the number of Twitter users associated with each topic. As we said earlier, while the largest groups constitute users who discuss topics related to a single area, we found many users who discuss topics related to different areas within the field. We identified 15 communities of interest: four consist of users clearly interested in a single area, whereas the rest combine interests from different areas. In our sample, 10 991 Twitter users (83%) mention one or more of the keywords from the four clusters detected in the semantic network. Those who mention keywords from a single community stand out: 2427 Twitter users discuss topics relating to commercial media (22.08%), 2206 bibliometrics (20.07%) and 1395 information retrieval (12.69%). Among those who refer to topics related to libraries, only 567 Twitter users (5.16%) exclusively mention keywords from this area.

Fig. 4 Intersecting sets for Information Science & Library Science. **A** corresponds to all combinations of actors and topics. **B** shows intersections after introducing a 10% cut-off for the number of times a keyword is mentioned. **C** shows intersections with a 20% cut-off point



Some 1107 Twitter users combine mentions to topics related with social media and bibliometrics (9.25%). In fact, 44.22% of those who discuss topics related to bibliometrics also discuss topics related to social media. This figure falls slightly when combined with information retrieval (39.61%) and drops further when combined with libraries (19.87%). Finally, one singular cluster is that consisting of 366 actors (3.35%) who mention all four topics.

Figure 5 compares communities defined by co-tweeted keywords with those defined by co-occurring keywords in papers. Nodes represent Twitter users. They are colored-coded to reflect communities constructed on the basis of the co-occurring keywords ($Q = 0.27$). Areas are colored-coded to identify Twitter user communities constructed on the basis of co-tweeted keywords ($Q = 0.32$). As we have said, 96.69% of Twitter users tweeting keywords related to bibliometrics, form clearly-defined groups within this community regardless of the cut-off point applied (Figures 4B and 4C). Similarly, 86.96% of users discussing keywords related to social

media are grouped together regardless of the cut-off point applied. This percentage is lower in the case of users discussing topics related to information retrieval (64.29%) or libraries (61.54%). These results corroborate those of the profiles, in which users mentioning retrieval information and, especially, libraries who tend to show interest in a range of topics.

Fig. 5 Information Science & Library Science socio-semantic network. Nodes are color-coded to identify the topics that have greater incidence. Edges are established on the basis of co-tweeted keywords. These have been filtered to a minimum of 12, and the corresponding communities are represented by overlapping areas



Figure 6 details the users belonging to each community and lists those with the highest percentage of terms in each area. We manually assign an account type to these 20 cases. While most of these users only focus on the area to which they have been assigned, we have found some broader profiles. We have also noted that, on the basis of the number of times keywords appear and the percentage of keywords mentioned, the most frequent users in the information retrieval and bibliometrics clusters are more active and engage more intensely with the topics related to their cluster. Finally, most of these users are academics although in the libraries cluster two accounts belong to librarians and three are bots.

Fig. 6 5 Twitter accounts with the highest percentage of terms mentioned for each topic

Cluster	Twitter account	% terms mentioned	Altmetric data		Twitter data			
			Cluster terms mentioned	Most mentioned term	Type	Tweets	Friends	Followers
Libraries	@HealthLib_cccu	93,93	14	information literacy information seeking behaviour libraries	Librarian	1173	223	401
	@PbBrenda	93,93	14	access to information information literacy libraries	Librarian	36	299	61
	@readaloudbooks	91,89	34	information literacy libraries education	Bot	12500	2284	1802
	@megasimplebooks	90,2	46	information literacy library and information professionals lifelong learning	Bot	11200	1840	1673
	@readaloudbook	87,84	65	information literacy information seeking behaviour information skills	Bot	28200	1922	1956
Information retrieval	@MedProProtector	100	29	electronic health records safety clinical decision support	Company	15900	3276	3266
	@soniaebeitezok	100	21	electronic health records electronic health record patient safety	Academic	17900	1877	708
	@VirginiaWalley	100	87	patient safety medication errors electronic health records	Professional	6863	807	1528
	@curtlanglotz	95,83	23	electronic health record natural language processing text mining	Academic	2607	438	6446
	@HardeepSinghMD	95,12	78	electronic health records health information technology patient safety	Academic	3044	328	3233
Bibliometrics	@SciPubLab	100	73	open access publishing science policy	Academic	18700	3049	4261
	@AWHarzing	100	57	scopus google scholar citations	Academic	3502	681	2205
	@FWFOpenAccess	100	26	open access scientific publishing scholarly publishing	Institution	12100	943	3936
	@rpalacioshub	100	26	quality quantity bibliometrics	Academic	7071	8982	17200
	@monogragh	100	23	peer review social sciences journal	Academic	8930	212	1530
Social media	@nathanjurgenson	100	14	social media facebook social networks	Professional	53000	2367	19900
	@SebaValenz	97,83	45	social media facebook political participation	Academic	7764	1477	3834
	@pverdegem	96,77	30	social media twitter privacy	Academic	6200	748	1306
	@CDNJobar	96	24	internet social media innovation	Academic	1125	918	694
	@meta_d	95,24	20	internet ict knowledge	Academic	3439	500	780

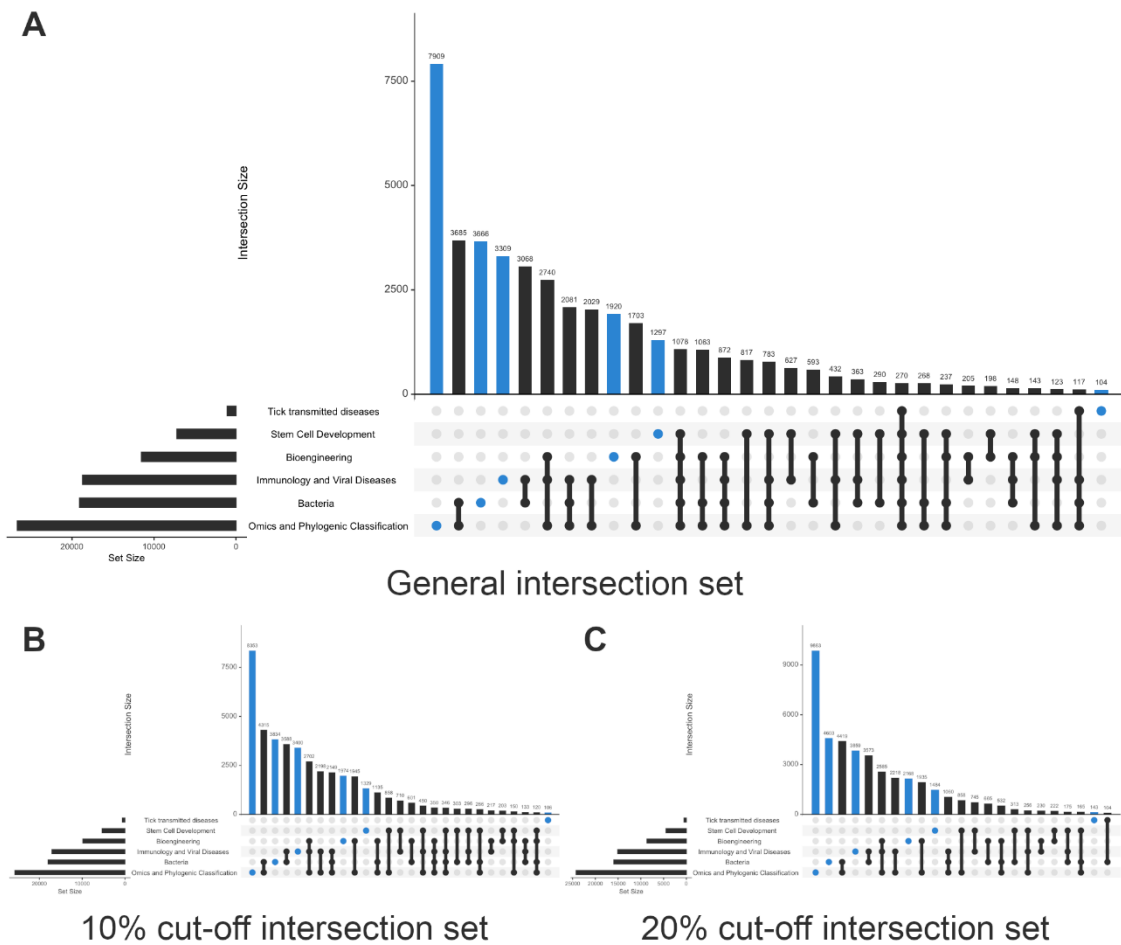
5. Case study: Microbiology

We replicated our approach in a larger field—Microbiology—to see how it would work in a different context. We identified 48 109 Twitter users mentioning 101 206 scientific publications of which 86.52% were articles, 11.03% reviews, and 1.88% editorial material. Twitter users mentioned a mean 5.93 publications (SD±63.65) and 25.27 keywords (SD±197.84).

The author keywords co-occurrence network is composed of 163 650 nodes and 1 173 938 edges. It is reduced to 2309 nodes and 7559 edges by removing keywords with less than 50 occurrences, edges with less than 5 co-occurrences and getting its main component. Figure 7 shows the corresponding co-word map. The community detection algorithm identified 6 clusters or topics using a resolution value of 2.0 (Q = 0.591). We labeled these: bacteria

When assigning Twitter users to each of these six topic groups (Figure 8), we found a much more complex and varied picture than in the previous case study. We identified 58 communities of interest. Although Twitter user groups relating to a single topic still stand out (38.84% of all users), most groups show an interest in more than one topic. Some 7909 Twitter users only mentioned keywords relating to omics and phylogenetic classifications (16.44%); 3666 mentioned keywords relating to bacteria (7.62%); 3309 immunology and viral diseases (6.88%); 1920 bioengineering (3.99%); 1297 stem cell development (2.7%); and 104 tick transmitted diseases (0.22%). The presence of ‘mixed’ profiles was much more common than in Information Science & Library Science. For instance, only 29.67% of Twitter users who mentioned keywords related to omics and phylogenetic classifications solely discussed this topic. This fell to 19.22% in the case of bacteria, 18% for stem cell development; 17.7% for immunology and viral diseases; 16.66% for bioengineering; and 9.92% for tick transmitted diseases.

Fig. 8 Intersecting sets with more than 100 actors in Microbiology. **A** corresponds to all combinations of actors and topics. **B** shows intersections after introducing a 10% cut-off for the number of times a keyword is mentioned. **C** shows intersections with a 20% cut-off point



6. Discussion

In the present study we propose a methodological approach to the identification of social media communities on the basis of common scientific interests. It enables us to link social media users on the basis of the keywords of the publications they mention and then group users by topic. We first applied this to Twitter users who mention publications in the fields of Information Science & Library Science. We then tested its feasibility by replicating the study in the field of Microbiology. Our proposal responds to the need for new efforts in social network analysis (Fu & Li, 2020), is based on recently-published conceptual frameworks, especially the so-called heterogeneous couplings defined by Costas et al. (2020) and n-mode networks proposed by Hellsten et al. (2019), and previous studies in which we looked into differences in topics of interest on social platforms (Robinson-Garcia et al., 2019). This method is in line with the second generation of social media metrics (Díaz-Faes et al., 2019). Twitter mentions are not used here in a quantitative way, not even to filter keywords or actors. The focus of the paper is on social media-objects (Twitter users and tweets) and the papers are treated abstractly as keywords.

The resulting socio-semantic network of this proposal has significant differences with respect to other kinds of networks. 2-mode networks can reflect direct and explicit relationships, such as social actors mentioning publications, as well as implicit ones, such as social actors that are connected by co-mention of the same publications. All of them are easily readable, but when an n-mode network is constructed combining 2-mode networks it becomes complex to interpret. Not only do the nodes represent different kinds of entities, but the relationships that exist between them can be of a different nature. This hinders the analysis, especially when network pruning or community detection methods are applied. Our proposal is to overlap instead of adding 2-mode networks. In this way, communities are detected independently, and then joined. While the n-mode network communities are composed of different types of elements, for example social actors and keywords, in ours the social actors have two types of groupings, one based on their social relationships and the other on keywords mentioned by them. The overlap between the two allows determining if their social relations and interests are in line or differ.

Our study has not been free from limitations. Firstly, some tweets or accounts in our data sample were subsequently removed from Twitter or blocked. Consequently, they were excluded from our study. Second, to create the semantic maps, we initially extracted terms from publication titles. However, these proved too generic and included many distractors, generating widely varying communities. We resolved this by using WoS author keywords even though this limited the publications included to those present in the WoS database and having associated author keywords. Although actors were correctly assigned to the topic mentioned

in most publications and people profiles prevail, bots are also present. In our Microbiology case study, given the complexity of the socio-semantic network, due to the variety of topics and social communities, this was not included.

Altmetrics has a number of well-known limitations—for example, the fact that data aggregators only retrieve tweets that include identifiers such as a DOI. The present study represents a step forward in the creation of applied solutions that use altmetrics beyond mere counting. Elsewhere, studies have already identified researchers (Costas et al., 2017; Ke et al., 2017) and communities on Twitter (Robinson-Garcia et al., 2018) or visualized the topics discussed on social media by using WoS author keywords and hashtags (Haunschild et al., 2019, 2020). Indeed, the thematic landscapes in this study seem more granular and more detailed than those generated elsewhere (Robinson-Garcia et al., 2019) due to our use of WoS author keywords instead of title noun phrases. Our study used both methods but integrates them into a single visualization. In this context, Hellsten et al. (2019) and Hellsten & Leydesdorff, (2019) proposed heterogeneous networks and applied these, respectively, to scientific journals and their attributes and Twitter and user mentions and hashtags. These proposals were based on networks produced by aggregating bipartite matrices that combine actors and objects in the same network. Our proposal also combines co-occurrence relationships of actors, publications and author keywords but we do not directly integrate them all into a network. Instead, we take the co-occurring keyword network and the co-tweeted keyword network and overlap these. Thus, the network is only formed of actors linked by social relations and their social communities are delimited through overlapping areas.

7. Concluding remarks

Our proposed methodology allows us to identify communities of users in an inclusive way, reflecting a complex reality in which actors may be interested in different aspects of a research field. This is especially evident in the case of Microbiology, where there are many groups consisting of only a few individuals assigned to more than one area. This study furthers our understanding on the use of social media to inform on scientific literature consumption by the general public. By isolating communities of common interest as well as finding those with overlapping interest we can narrow the target audience who is discussing scientific literature in social media. This is potentially useful to assess on the effectiveness of social outreach of scientific research, identify social stakeholders or analyze communication strategies. Further research should consider combining methods such as the one proposed with those strictly focused on characterizing user types (cf. Díaz-Faes et al., 2019).

By focusing on concepts (i.e. keywords) rather than objects (i.e. publications), we minimize potential relationships derived from social relations between actors rather than from common research interests (e.g. colleagues from the same institution).

This methodology has the potential of being applied in other scenarios from the ones proposed here. Other social media platforms could be considered, as well as other types of contents shared through social media. Some of the many and varied contexts in which it can be applied are political participation and political engagement (Halpern et al., 2017), trolling interactions in the online gaming sphere (Cook et al., 2019), experiences of mental disorders shared in forums (Yoo et al., 2019), or social communities discussing eating disorders (Wang et al., 2017). Moreover, it is possible to use other social objects and links to construct the social network and other kinds of semantic maps, for example Reddit posts as social object, co-mentioned hashtags for social network, and topic modelling for semantic map. In the specific case of altmetrics, a future line of study is the application of this methodology to different social media and the use of other terms to create the semantic maps. This is an initial approach only using Twitter mentions due to their enormous coverage and the extension of altmetrics studies. However, we would hope to study its applicability further by using altmetric sources other than Twitter, to study source-related differences in the type of users who discuss scientific literature.

8. References

- Alperin, J. P., Gomez, C. J., & Haustein, S. (2018). Identifying diffusion patterns of research articles on Twitter: A case study of online engagement with open access articles. *Public Understanding of Science*, 28(1), 2–18. <https://doi.org/10.1177/0963662518761733>
- 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>
- Arroyo-Machado, W., Torres-Salinas, D., & Robinson-García, N. (2019). Identifying communities of interest in social media: Microbiology as a case study. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th International Conference on Scientometrics and Informetrics, ISSI 2019* (pp. 1201–1209). http://issi-society.org/proceedings/issi_2019/ISSI%202019%20-%20Proceedings%20VOLUME%20I.pdf
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International AAAI Conference on Weblogs and Social Media*. Third International AAAI Conference on Weblogs and Social Media. <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>

- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF). *Journal of Informetrics*, 13(1), 325–340. <https://doi.org/10.1016/j.joi.2019.01.008>
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351–374.
- Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235. <https://doi.org/10.1177/053901883022002003>
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402. <https://doi.org/10.1002/asi.21525>
- Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies*, 1–32. https://doi.org/10.1162/qss_a_00080
- Cook, C., Conijn, R., Schaafsma, J., & Antheunis, M. (2019). For Whom the Gamer Trolls: A Study of Trolling Interactions in the Online Gaming Context. *Journal of Computer-Mediated Communication*, 24(6), 293–318. <https://doi.org/10.1093/jcmc/zmz014>
- 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>
- Costas, R., van Honk, J., & Franssen, T. (2017). *Scholars on Twitter: Who and how many are they?*
- Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019. <https://doi.org/10.1002/asi.23309>
- Csárdi, G. (2020). *igraph: Network Analysis and Visualization*. <https://CRAN.R-project.org/package=igraph>
- Davis, R., Bacha, C. H., & Just, M. R. (2017). *Twitter and elections around the world: Campaigning in 140 Characters or Less*. Routledge.
- 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>

- Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4), e123.
- Fu, J. S., & Lai, C.-H. (2020). Are We Moving Towards Convergence or Divergence? Mapping the Intellectual Structure and Roots of Online Social Network Research 1997–2017. *Journal of Computer-Mediated Communication*, 25(1), 111–128. <https://doi.org/10.1093/jcmc/zmz020>
- Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69–115. <https://doi.org/10.1023/A:1010512628145>
- Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1), 1171458. <https://doi.org/10.1080/23311983.2016.1171458>
- Halpern, D., Valenzuela, S., & Katz, J. E. (2017). We Face, I Tweet: How Different Social Media Influence Political Participation through Collective and Internal Efficacy. *Journal of Computer-Mediated Communication*, 22(6), 320–336. <https://doi.org/10.1111/jcc4.12198>
- Haunschild, R., & Bornmann, L. (2018). Field- and time-normalization of data with many zeros: An empirical analysis using citation and Twitter data. *Scientometrics*, 116(2), 997–1012. <https://doi.org/10.1007/s11192-018-2771-1>
- Haunschild, R., Leydesdorff, L., & Bornmann, L. (2020). Library and Information Science Papers Discussed on Twitter: A new Network-based Approach for Measuring Public Attention. *Journal of Data and Information Science*, 5(3), 5–17. <https://doi.org/10.2478/jdis-2020-0017>
- Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics*, 13(2), 695–707. <https://doi.org/10.1016/j.joi.2019.03.008>
- 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>
- Haustein, S. (2019). Scholarly Twitter Metrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 729–760). Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_28
- Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated “bot” accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1), 232–238. <https://doi.org/10.1002/asi.23456>

- Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *PLOS ONE*, *10*(3), e0120495. <https://doi.org/10.1371/journal.pone.0120495>
- Hellsten, I., & Leydesdorff, L. (2019). Automated analysis of actor–topic networks on twitter: New approaches to the analysis of socio-semantic networks. *Journal of the Association for Information Science and Technology*, *71*(1), 3–15. <https://doi.org/10.1002/asi.24207>
- Hellsten, I., Opthof, T., & Leydesdorff, L. (2019). N-mode network approach for socio-semantic analysis of scientific publications. *Poetics*, 101427. <https://doi.org/10.1016/j.poetic.2019.101427>
- Holmberg, K., Bowman, S., Bowman, T., Didegah, F., & Kortelainen, T. (2019). What Is Societal Impact and Where Do Altmetrics Fit into the Equation? *Journal of Altmetrics*, *2*(1), 6. <https://doi.org/10.29024/joa.21>
- Joubert, M., & Costas, R. (2019). Getting to Know Science Tweeters: A Pilot Analysis of South African Twitter Users Tweeting about Research Articles. *Journal of Altmetrics*, *2*(1), 2. <https://doi.org/10.29024/joa.8>
- Kassab, O., Bornmann, L., & Haunschild, R. (2020). Can altmetrics reflect societal impact considerations?: Exploring the potential of altmetrics in the context of a sustainability science research center. *Quantitative Science Studies*, *1*(2), 792–809. https://doi.org/10.1162/qss_a_00032
- Ke, Q., Ahn, Y.-Y., & Sugimoto, C. R. (2017). A systematic identification and analysis of scientists on Twitter. *PLOS ONE*, *12*(4), e0175368. <https://doi.org/10.1371/journal.pone.0175368>
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. <https://doi.org/10.1002/asi.5090140103>
- Kim, J., & Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, *38*(1), 86–96. <https://doi.org/10.1016/j.ijinfomgt.2017.08.003>
- Lex, A., Gehlenborg, N., Strobelt, H., Vuilleumot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992. <https://doi.org/10.1109/TVCG.2014.2346248>
- Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589–593.
- Misirlis, N., & Vlachopoulou, M. (2018). Social media metrics and analytics in marketing – S3M: A mapping literature review. *International Journal of Information Management*, *38*(1), 270–276. <https://doi.org/10.1016/j.ijinfomgt.2017.10.005>
- Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133. <https://doi.org/10.1103/PhysRevE.69.066133>

- Neylon, C., & Wu, S. (2009). Article-Level Metrics and the Evolution of Scientific Impact. *PLOS Biology*, 7(11), e1000242. <https://doi.org/10.1371/journal.pbio.1000242>
- Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*. <https://doi.org/10.5210/fm.v12i8.1997>
- Noyons, C. M. (2005). Science Maps Within a Science Policy Context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (pp. 237–255). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_11
- Noyons, E. (2019). Measuring societal impact is as complex as ABC. *Journal of Data and Information Science*, 4(3), 6–21.
- Piccardi, T., Redi, M., Colavizza, G., & West, R. (2020). Quantifying Engagement with Citations on Wikipedia. *Proceedings of The Web Conference 2020*, 2365–2376. <https://doi.org/10.1145/3366423.3380300>
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. In *Altmetrics*. <http://altmetrics.org/manifesto/>
- Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. *FEMS Microbiology Letters*, 366(7). <https://doi.org/10.1093/femsle/fnz075>
- Robinson-García, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PloS One*, 12(8), e0183551.
- Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: Exploring the insides of Altmetric.com. *El Profesional de La Información*, 23(4), 359–366. <https://doi.org/10.3145/epi.2014.jul.03>
- Robinson-Garcia, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, 45(6), 815–826. <https://doi.org/10.1093/scipol/scy024>
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799–813. [https://doi.org/10.1002/\(SICI\)1097-4571\(1999\)50:9<799::AID-ASI9>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G)
- Small, H. G., & Koenig, M. E. D. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management*, 13(5), 277–288. [https://doi.org/10.1016/0306-4573\(77\)90017-6](https://doi.org/10.1016/0306-4573(77)90017-6)
- Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International*

- Journal of Information Management*, 39, 156–168.
<https://doi.org/10.1016/j.ijinfomgt.2017.12.002>
- 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>
- Taraborelli, D. (2008). Soft peer review: Social software and distributed scientific evaluation.
- Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. *Scientometrics*, 115(3), 1231–1240. <https://doi.org/10.1007/s11192-018-2715-9>
- Torres-Salinas, D., Clavijo, Á. C., & Contreras, E. J. (2013). Altmetrics: New Indicators for Scientific Communication in Web 2.0. *Revista Comunicar*, 21(41), 53–60. <https://doi.org/10.3916/C41-2013-05>
- Torres-Salinas, D., Jiménez-Contreras, E., & Robinson-García, N. (2014). Tendencias en mapas de la ciencia: Co-uso de información científica como reflejo de los intereses de los investigadores.
- 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>
- Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z>
- van Schalkwyk, F., Dudek, J., & Costas, R. (2020). Communities of shared interests and cognitive bridges: The case of the anti-vaccination movement on Twitter. *Scientometrics*. <https://doi.org/10.1007/s11192-020-03551-0>
- Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392. <https://doi.org/10.1002/asi.22748>
- Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2017). Detecting and Characterizing Eating-Disorder Communities on Social Media. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 91–100. <https://doi.org/10.1145/3018661.3018706>
- Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511815478>
- Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media – The pilot quantitative study. *Health Policy and Technology*, 7(2), 115–118. <https://doi.org/10.1016/j.hlpt.2018.03.002>
- White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163–171. <https://doi.org/10.1002/asi.4630320302>

- Wouters, P., Zahedi, Z., & Costas, R. (2019). Social Media Metrics for New Research Evaluation. In W. Glänzel, M. Henk F, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 687–713). Springer International Publishing.
- Yoo, M., Lee, S., & Ha, T. (2019). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. *Information Processing & Management*, 56(4), 1565–1575. <https://doi.org/10.1016/j.ipm.2018.10.001>
- 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>
- Zahedi, Z., & van Eck, N. J. (2018). Exploring Topics of Interest of Mendeley Users. *Journal of Altmetrics*, 1(1), 5. <https://doi.org/10.29024/joa.7>

