Link-based approach to study scientific software usage: the case of VOSviewer

Orduna-Malea, E.; Costas Comesana, R.

Scientometrics (2021) 126:8153–8186
https://doi.org/10.1007/s11192-021-04082-y

Link-based approach to study scientific software usage: the case of VOSviewer

Enrique Orduña-Malea[1] · **Rodrigo Costas**[2,3]

Received: 24 February 2021 / Accepted: 14 June 2021 / Published online: 10 July 2021

Abstract
Scientific software is a fundamental player in modern science, participating in all stages of scientific knowledge production. Software occasionally supports the development of trivial tasks, while at other instances it determines procedures, methods, protocols, results, or conclusions related with the scientific work. The growing relevance of scientific software as a research product with value of its own has triggered the development of quantitative science studies of scientific software. The main objective of this study is to illustrate a link-based webometric approach to characterize the online mentions to scientific software across different analytical frameworks. To do this, the bibliometric software VOSviewer is used as a case study. Considering VOSviewer's official website as a baseline, online mentions to this website were counted in three different analytical frameworks: academic literature via Google Scholar (988 mentioning publications), webpages via Majestic (1,330 mentioning websites), and tweets via Twitter (267 mentioning tweets). Google scholar mentions shows how VOSviewer is used as a research resource, whilst mentions in webpages and tweets show the interest on VOSviewer's website from an informational and a conversational point of view. Results evidence that URL mentions can be used to gather all sorts of online impacts related to non-traditional research objects, like software, thus expanding the analytical scientometric toolset by incorporating a novel digital dimension.

Keywords Scientific software · Link analysis · Informetrics · Webometrics · Scholarly communication · Social media metrics; VOSviewer

✉ Enrique Orduña-Malea
enorma@upv.es

Rodrigo Costas
rcostas@cwts.leidenuniv.nl

[1] Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain

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

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

Introduction

Scientific-purpose software as a non-traditional academic output

Software is an essential component in the ecosystem of modern Science, particularly in those disciplines that follow a data-driven paradigm, guided by the ongoing generation, availability, and consumption of high volumes of scientific data (Hey, Tansley & Tolle 2009; Li & Yan, 2018). Today, software is used in all stages of academic work (Howison et al., 2015), from annotating preliminary ideas to processing large volumes of data or disseminating research results. Among the vast amount of software available in the scientific endeavor, we can distinguish between *general-purpose*[1] and *scientific-purpose* software.

Scientific-purpose applications consist of software explicitly designed to assist on non-trivial scientific tasks (for example, *VOSviewer*,[2] *CitNet Explorer*[3] or *Gephi*,[4] to name a few). Scientific software can play important roles in processes related to data collection, management, formatting, analysis, modelling, simulation, prediction, visualization, and dissemination (Howison et al., 2015; Pan et al., 2017), becoming essential in the scientific discovery process (Pradal et al., 2013). Thus, scientific software has a direct effect on the validity of scientific results, since replacing the software could in turn lead to replacing an underlying procedure or logic assumption (Hannay et al., 2009; Howison & Herbsleb, 2011; Li et al., 2017; Yang et al., 2018).

This rising importance of software in the scientific process prompted the perception of scientific-purpose software as a research product of its own. Research funding agencies are increasingly funding the development of scientific-purpose software (Howison et al., 2015), as well as accepting software creation as an accepted outcome in some grant applications (Piwowar, 2013), like the *U.S. National Science Foundation* (NSF)[5] and the *U.K. Research Excellence Framework* (REF)[6] (Pan et al., 2018).

Tracking citations to scientific-purpose software

While it is commonly accepted that there is no need to mention/cite general-purpose software (Pan et al., 2019), it is recommended that scientific publications using scientific-purpose software should mention it (Niemeyer et al., 2016). The main reasons to encourage the citation of scientific-purpose software include credit allocation, reproducibility, transparency, and discovery (Smith, Katz, & Niemeyer 2016).

Previous literature has provided a significant body of knowledge about the lack of formal mentions of software in scientific publications. For example, Howison and Bullard (2016) found that between 31 and 43% of software textual mentions involved also formal

[1] General-purpose applications are those originally developed for a general usage, which can also be applied to assist and support some trivial scientific tasks, such as writing documents, sending e-mails, videocalls or presentations (Soito & Hwang. 2016). These applications (for example, *Microsoft Word*) have no effect on the validity of scientific results and can be easily replaced by other similar solutions (Pan et al., 2019).

[2] https://www.vosviewer.com.

[3] https://www.citnetexplorer.nl.

[4] https://gephi.org.

[5] https://www.nsf.gov/pubs/policydocs/pappg20_1/nsf20_1.pdf.

[6] https://www.ref.ac.uk/media/1092/ref-2019_01-guidance-on-submissions.pdf.

citations. Pan et al. (2016) discovered that more than 30% of the software mentions in 2014 in articles published in *PLoS ONE* received no formal citations. Park and Wolfram (2019) found that research software was rarely cited in the Clarivate Analytics' Data Citation Index (DCI). This under-citedness of software varies both by discipline (Pan et al., 2016) and the nature (commercial or freeware) of the software (Howison & Bullard, 2016; Pan et al., 2019). Moreover, the mentions of software often lack sufficient information related to the software employed (e.g., version, access, crediting information, etc.) (Howison & Bullard, 2016).

Despite the academic community initiated diverse actions like proposing best software citation practices (Hafer & Kirkpatrick, 2009; Howison & Bullard, 2016; Niemeyer et al., 2016), working groups (e.g., *FORCE11 Software Citation Working Group*,[7] *FORCE11 Software Citation Implementation Working Group*[8] and *WSSSPE Software Credit Working Group*—Katz et al., 2016), publisher guidelines (e.g., the *American Astronomical Society Policy Statement on Software*[9]) or informal statements, such as the *Science Code Manifesto*[10] and *The Research Software Impact Manifesto*,[11] the diversity of ways to referring to software and the still pending proper standardized citation guidelines (e.g., standardized citation styles and publishers sometimes contradict each other make citation counts a limited metric for the proper traceability of scientific-purpose software (Pan et al., 2019).

Textual approaches to track academic software usage

Since citation metrics have shown only a limited applicability to measure software usage in academic settings, it becomes necessary to establish alternative methods to measure the usage of scientific-purpose software, and to obtain evidence about its influence and impact (Hannay et al., 2009; Pan et al., 2018). Thus, the identification of textual mentions of software in the text of scientific papers has been a quite common approach to capture the impact of software on science (Pan et al., 2016).

Different efforts have been made to measure scientific software text-mention patterns in publications at different levels: a) disciplines, such as Biology (Howison & Bullard, 2016; Yang et al., 2018) and Library and information sciences (Pan et al., 2019); b) multidisciplinary journals, such as *PLoS ONE* (Pan et al., 2015, 2016); c) programs stored in software repositories (Thelwall & Kousha, 2016); and d) specific software applications, such as *Geant4 toolkit* (Pia et al., 2009), *R packages* (Li & Yan, 2018; Li et al., 2017, 2019), and bibliometric mapping software (*Citespace*, *VOSviewer* and *Histcite*) (Pan et al., 2017, 2018).

Tracking the use of software via text-mentions introduces some methodological challenges, which might limit the identification of software names in large texts (Du et al., 2021). *First*, there may be different ways to invoke the same software, a software project name (e.g., in *GitHub*), the URL of the software's official website, the URL to the repository where it is hosted, mentions to unpublished manuscripts about the software, users' manuals, etc. In addition, we can find synonyms or even translations to other languages. Consequently, the polymorphous nature of textual mentions is huge (Cronin et al., 1988). *Second*, common words

7 https://www.force11.org/group/software-citation-working-group.

8 https://www.force11.org/group/software-citation-implementation-working-group.

9 http://journals.aas.org/policy/software.html.

10 http://sciencecodemanifesto.org/.

11 https://www.software.ac.uk/blog/2016-10-06-publish-or-be-damned-alternative-impact-manifesto-research-software.

used as software names may also represent other objects, due to polysemy of textual mentions. *Third*, complex software applications might have parts (specific packages and modules, etc.) that make their clear identification through text-mentions complex.

Using software website URLs as traceable objects

A possibility to limit this complexity and facilitate the operationalization of the tracking of software is to consider the URL of the software website as the *traceable object*. The mentioning of URLs has been extensively studied in the webometrics field as an established technique to measure the online importance and impact of websites (Orduna-Malea & Alonso-Arroyo, 2017; Park & Thelwall, 2003; Thelwall, 2004). Likewise, the URL also stands out as a central piece for Altmetric studies, as such universal identifier is often used as a digital object (mainly via DOI URLs) representing research publications mentioned on online social media platforms (Wouters et al., 2019).

Tracking URLs presents fundamental advantages over the tracking of texts: *First*, a URL provides a unique and unequivocal element to identify the software. For example, the text 'vosviewer.com' can only refer to the software VOSviewer, thus reducing polymorphism and polysemy. *Second*, the URL is an actionable element that allows users to navigate from the source document (document where the software is mentioned) to the target (document published/hosted in the software's website), establishing and making explicit a relation between these documents. *Third*, considering that mentioning a URL is time-consuming –more than simply mentioning the software– it can be argued that this action might be related to a more conscious informational purpose (e.g., sources transparency, facilitating resources for readers, etc.), where links are oriented to navigational issues (Halavais, 2008). *Fourth*, the URL not only represents univocally the digital *object* (in this case, the software) but also represents the whole website (and all contents hosted inside) where the software is available. As a website, a wide range of metrics (e.g., traffic, visibility, size, etc.) are available, which can report information related to the consumption and interest on the software.

VOSviewer as a case study

The free bibliometric software *VOSviewer*[12] is analyzed as a case study. *VOSviewer* was developed by Nees Jan van Eck and Ludo Waltman at *Leiden University's Centre for Science and Technology Studies* (CWTS). The application was launched in 2010, and formally introduced through a software paper –a scientific publication describing and analyzing the software (Smith, Katz & Niemeyer 2016)– published in *Scientometrics* (Van Eck & Waltman, 2010). This publication is the most cited article in the journal, according to both *Scopus* (1,621 citations) and *Web of Science*-all databases (1,431), as of August 2020.

Its ease of use and multiple features (including specific clustering and natural language processing techniques) made *VOSviewer* popular not only in the Scientometrics community but also in other disciplines where science maps are used. Given its simplicity as a software product (code and related material is all centralized and available on a website) and the broad interest and diverse audience to this software, *VOSviewer* constitutes an excellent case study to test the proposed approach. Moreover, *VOSviewer* is relevant and well known software for the Scientometric research community. All these features make *VOSviewer* an ideal case study for an illustrative discussion like the one presented in this paper.

[12] https://www.vosviewer.com.

Aim of the study and analytical framework

The objective of this study is to illustrate a webometric and altmetric method to determine the use and interest on a case study of scientific-purpose software: *VOSviewer*. The main purpose of this work is to design an *analytical framework* aimed at studying scientific software's impact metrics by collecting large amounts of data from multiple online data sources. This analytical framework will be made explicit by means of an evaluation technical sheet, which will include a wide list of URL-based metrics specifically and formally defined to measure the use of scientific software. This analytical framework is based on the mentioning of software's main URL across different *scenarios* (academic publications, web at-large, and social media), each of which is operationalized by an online data source (*Google Scholar*, *Majestic*, and *Twitter*, respectively).

In Appendix 1 we describe more specifically the three data sources chosen for this study. The rest of the paper is structured as follows, in Sect. 2 the methodological approach is described, in Sect. 3 we present the main results, discussion in Sect. 4, and finally the main conclusions in Sect. 5.

Methodological approach

We analyze the mentioning of the official URL of *VOSviewer* (www.vosviewer.com) in three *scenarios*: academic publications, web at-large, and social media. Each *scenario* is characterized by the following five elements (Fig. 1):

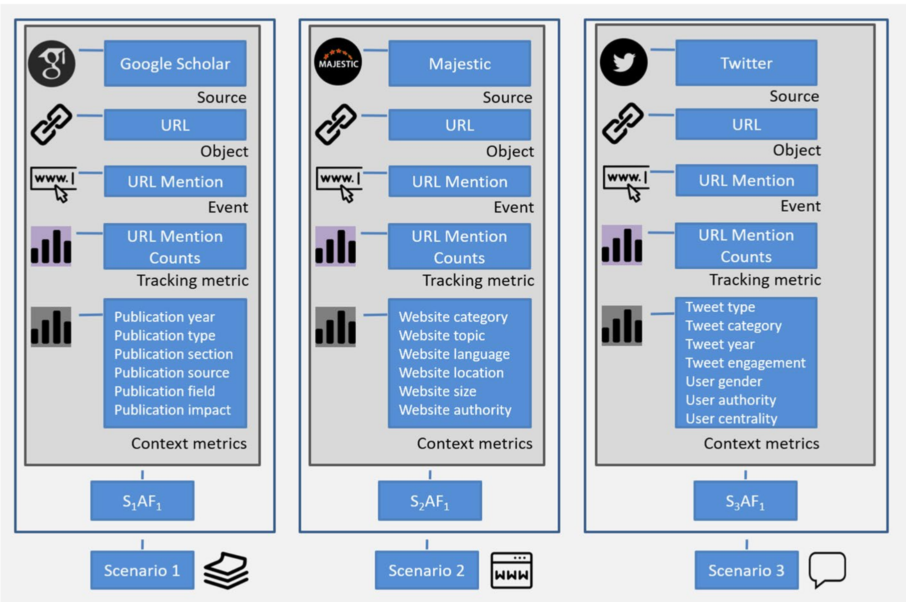

Fig. 1 Scenarios for software use measurement: academic literature (*Google Scholar*), websites (*Majestic*) and social media (*Twitter*). AF = Analytical Framework; S_1AF_1 = Analytical Framework 1 belonging to Scenario 1

- *Source*: the database where the use of the software is tracked (e.g., *Majestic*, *Twitter*, etc.).
- *Object*: an element that represents the software (e.g., the software name, the software website).
- *Event*: an action carried out by users or applications that trigger the appearance or invocation of the object (e.g., a software name's text-mention, a link to the software website).
- *Tracking metrics*: a measure which captures the use of the *object* through a certain *event* (e.g., the number of software name's text mentions).
- *Context metrics*: a measure which captures characteristics of the users that generate the *events* (e.g., users' gender, location, language, etc.).

URL mentions in academic literature via *Google Scholar*

Academic publications containing the URL string "vosviewer.com" were retrieved from *Google Scholar* using the *Publish or Perish* v7 software[13] (patent documents included). To do this, the direct query "vosviewer.com" was performed, excluding protocols (https and www) to improve recall as much as possible. The search was not limited to a specific period in order to retrieve all publications in *Google Scholar*, regardless the year of publication, language or document type.

The search returned 1,190 records as of 31 March 2020. These records were subsequently exported including the following bibliographic fields: publication author(s), publication title, publication year, publication source, publication URL, and the number of citations received. Due to the unsupervised indexing process carried out by *Google Scholar*, bibliographic errors were found (Orduna-Malea, Martín-Martín & Delgado López-Cózar 2017), including multiple records for one same publication and incomplete/erroneous authorship.[14] To solve this limitation, data was manually cleansed, and multiple copies were merged, obtaining a final amount of 988 records. The document type (journal article, book, book chapter, conference proceeding, working papers and reports) was directly determined from the information contained in the publication source field. All dubious cases were manually checked.

The publication URL was used to manually access each of the publications. The search functionalities of web browsers (for HTML publications) and *Adobe Acrobat* (for PDF publications) were used to locate each URL mention in each of the publications, according to the available full text format. The publication section where the string "vosviewer.com" appeared (introductory sections, method, conclusions, or references) was manually checked and annotated throughout this process. When no clear structure (IMRaD type) was found, the category 'unstructured' was assigned.

13 https://harzing.com/resources/publish-or-perish.

14 For example, one author field found was: <I Kellevezir, G Özdağoğlu, M Damar…>, and a manual inspection revealed a missing author. The field was completed as follows: <I Kellevezir, G Özdağoğlu, M Damar, A Özdağoğlu>. In other cases, author names and surnames were altered. For example, the author field <VA Vasco López, M Moreno Mejía, PA Reyes Gavilán…> was updated to <M Moreno Mejía, P Reyes Gavilán, V Vasco López, A Aroca Mejía, N Herrera>.

Table 1 Categories used to typify the purpose of those tweets mentioning the software's URL (vosviewer.com)

Category	Description
Discovery	It includes all tweets in which authors express they have found out the software and they are testing it (e.g., expressions such as 'playing with', 'just found out' or 'starting with'
Diffusion	It includes all tweets in which authors just inform about the existence of the software (e.g., 'VOSviewer', 'map tool')
Upgrade	It includes all tweets in which authors inform about the existence of a new version of the software released (e.g., expressions such as 'new version', 'update', 'new release', etc.)
Recommend	It includes all tweets in which authors express some emotion about the software, recommending its use (e.g., expressions such as 'like', 'recommend', 'love it', 'try it', etc.)
Errors	It includes all tweets in which authors indicate some technical failure of the software (e.g., expressions such as 'bug', 'error', 'incorrect', 'problems', etc.)
Operating	It includes all tweets in which authors indicate or explain technical features of the software (e.g., expressions covering functionalities names)
Discussion	It includes all tweets in which the software is mentioned as part of a discussion or conversation between authors, including questions and answers
Use example	It includes all tweets in which the authors share a map created with the software, either to show the capabilities of the software or to disseminate the information embedded in the map
Learning	It includes all tweets in which the authors share learning materials about the software (e.g., expressions such as 'manual', 'tutorial', etc.)
Presentation	It includes all tweets in which the authors inform about a demo or presentation of the software (e.g., expressions such as 'demo', 'presentation', 'workshop', etc.) in one specific event

Along this process, full text access was not possible for 51 publications. No URL mentions were found for 32 publications (due to *Google Scholar* parsing errors[15]), and document typology could not be properly determined for 18 publications (due to lack of information on the full texts available), thus resulting in a final set of 887 publications mentioning 'vosviewer.com'.

URL mentions in webpages via *Majestic*

Link data related to the *VOSviewer* website was gathered from *Majestic*[16] through the site explorer feature (historic index). To do this, the direct root domain query "vosviewer.com" was carried out, obtaining a total of 17,261 mentioning webpages belonging to 1,330 distinct websites were gathered as of 12 April 2020. To characterize those webpages including a URL mention of "vosviewer.com", additional web metrics related to each of these webpages were also directly obtained from *Majestic* (see Table 1 for a detailed description of these metrics), including the IP address, website language, and flow metrics (*Trust Flow*

[15] For example, the sentence "formatados e importados para o software VosViewer com o intuito de esboçar a rede de conexão dos termos pesquisados" does not include a URL mention but a text-mention.

[16] https://majestic.com.

and *Citation Flow*)[17] (Jones, 2012). These flow metrics are meant to capture some idea of the "prestige" or reputation of the linking URLs.[18]

Finally, each website was categorized. To this end, a bottom-up process based on a previous work oriented to scientific-related websites categorization (Orduna-Malea 2021) was carried out. *First*, each website was accessed to and manually classified according to their functional nature. A total of 46 academic-related categories were identified, whilst all non-academic related websites (casinos, adult content, etc.) were discarded. *Second*, an external researcher with expertise in websites classification was asked to carry out an inter-coder reliability test through a random sample of the 10% of academic-related websites. The percentage of agreement achieved was 80%, and the Krippendorff's alpha (nominal) achieved was 0.92, which is considered acceptable.

It should be noted that only URLs at the web domain level were considered for website categorization. As blogs from generic blog providers (under 'wordpress.com', 'blogspot' or similar web domains) employ generic web domains provided by the blog service, several different blogs may be linked to one same web domain. In these cases, no academic-related category has been assigned, as particular blogs under generic blog providers have not been checked.

URL mentions in tweets via *Twitter*

All tweets containing the string "vosviewer.com" until 31 March 2020 were gathered from *Twitter* using the *TweetDeck* dashboard application,[19] without any time restriction. A corpus of 267 tweets mentioning the URL "vosviewer.com" was finally retrieved.

For each tweet, the username, date of publication, tweet text, number of replies, likes, and retweets were obtained. Likewise, the tweet type (original tweet, retweet, and reply) and the inclusion of images or videos were also collected.

Finally, each tweet was categorized according to the main purpose. To this end, a bottom-up process was carried out. *First*, each tweet content was accessed to, and manually classified according to the general purpose perceived, considering the words, textual signs, visual signs, multimedia, and tweet type. *Second*, all categories identified were grouped, standardized, and defined, achieving a total of 10 general categories (see Table 1). *Third*, a second classification round was performed to reassign the standardized category to each tweet. *Fourth*, an external researcher with expertise in tweets classification was asked to carry out an inter-coder reliability test through a random sample of the 10% of tweets. The percentage of agreement achieved was 80%, and the Krippendorff's alpha (nominal) achieved was 0.76, which is considered acceptable. This test was used to reclassify few tweets and improve the definition of categories.

At the user-level, all users providing likes, retweets and replies to each of the original 267 tweets were also obtained, and categorized (female, male, institutional, unknown). The *social authority*[20] of each *Twitter* user was obtained from *Followerwonk*.[21] This metric

[18] The incorporation of this metric only plays a role to illustrate the relevance of characterizing linking websites by their "prestige", but this does not represent a validation of this metric (which at best must happen in future research) neither a recommendation to be incorporated as a fix element of the analytical framework proposed.

[19] https://tweetdeck.twitter.com.

[20] https://followerwonk.com/social-authority.

[21] https://followerwonk.com.

[17] Table 2 includes formal definitions of these two metrics.

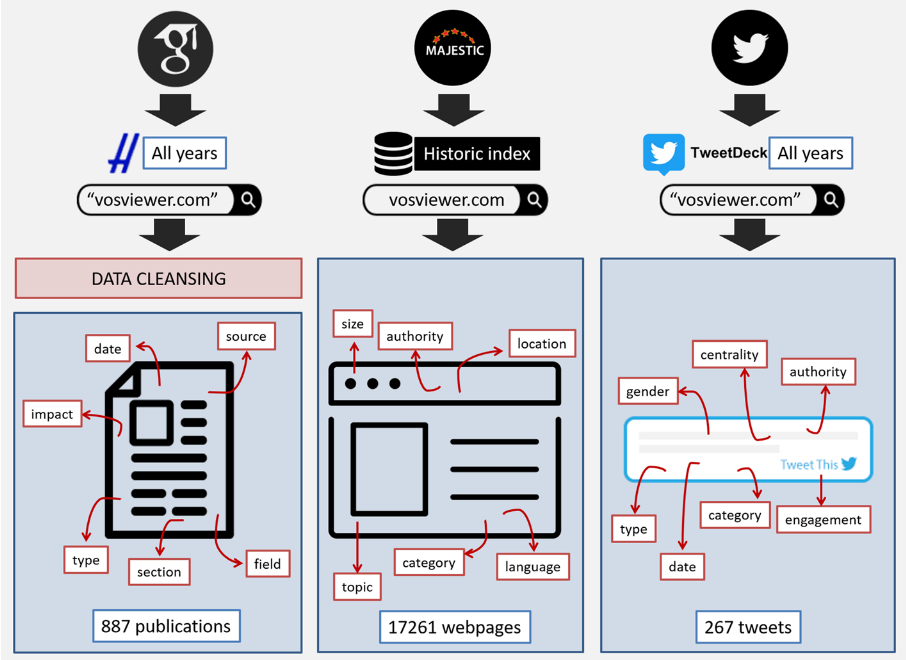

Fig. 2 Data gathering process for publications (*Google Scholar*), webpages (*Majestic*) and tweets (*Twitter*)

recursively measures the prestige of a *Twitter* account based on the prestige of the followers who retweet the tweets by the given account.[22] Specifically, it includes the following three components: a) the retweet rate of a few hundred of the measured user's last non-@ mention tweets; b) a time decay to favor recent activity versus old activity; c) other data for each user (such as follower count, friend count, and so on) that are optimized via a regression model trained to retweet rate. Social authority metrics score from 0 (no authority) to 100 (maximum authority).[23]

Different networks (connecting the user who publishes a tweet with the user who likes/retweets/replies the original tweet published) were generated with *Gephi* v0.9.1.[24] Finally, data was statistically analyzed with *XLStat*.[25]

Data processing and metrics

At the end of the process, a total of 887 academic publications, 17,261 webpages and 267 tweets are considered. The overall process followed is illustrated briefly in the Fig. 2.

[22] https://moz.com/blog/social-authority.

[23] Like before, the incorporation of this metric only plays a role to illustrate the relevance of characterizing Twitter users on their "prestige" or "social media capital" (see Díaz-Faes, Bowman, & Costas 2019), but this does not represent a validation of this metric (which at best must happen in future research) neither a recommendation to be incorporated as a fix element of the analytical framework proposed.

[24] https://gephi.org.

[25] https://www.xlstat.com.

As we can observe in the Fig. 2, each *scenario* operates with one specific *source*, each providing a different set of mentioning events: publications, webpages, and tweets. From each of these three document bodies, several context metrics can be obtained.

Based on software diffusion indicators (paper diffusion breadth and journal diffusion breadth) proposed by Pan et al., (2018), the following tracking and context metrics have been obtained (Table 2). These are grounded on *diffusion breadth metrics* (number of elements mentioning the software) and *impact breadth metrics* (attention achieved by elements mentioning the software). Elements can be authors, journals, webpages, countries, languages, etc.

For the sake of clarity, in this work webpage will refer to any document displayed to a user in a web browser, regardless its format and represented by a URL (e.g., vosviewer.com/vosviewer.php). Likewise, a website will refer to a collection of webpages linked together in a coherent fashion, also represented by a URL which nests hierarchically all related webpages' URLs (e.g., vosviewer.com).

Results

URL mentions in academic literature via *Google Scholar*

A total of 1,144 *publication URL mentions* from 887 different publications (*publication diffusion breadth*) were found. Most publications (79%) include only one URL mention, although few publications include up to ten URL mentions, denoting a strong importance of the software in that publications (Fig. 3). In any case, the *publication URL mention intensity* is low (average of 1.29 URL mentions per publication).

The *publication impact breadth* achieves an i-10 index of 188 (and an i-100 index of 22), showing a significant number of publications mentioning the software's URL achieving citation-based impact. 36.4% of this corpus of URL mentioning documents had not received any citations at the time of data gathering. The *publication diffusion breadth* has increased over the years, especially in 2019 (Fig. 4). About two-thirds of these publications are journal articles (70.1%), while other categories such as theses and Master theses (10.8%) and Conference papers (9.2%) also show a remarkable presence (Fig. 5).

The references section is the most frequently location where *VOSviewer*'s URL mentions were found (30.8% of all URL mentions found), followed by the methodology Sect. (29.7%) and results (18.1%). Otherwise, 108 documents (providing 144 URL mentions) did not exhibit a standard structure (Fig. 6). These results suggest a preference to mention *VOSviewer*'s URL as part of the bibliographic references, and as a methodological item in the mentioning publications.

Mentioning journal articles come from 499 different academic journals (*source diffusion breadth*), out of which 16% belong to the Library and Information Sciences field. The distribution of URL mentions per journal shows a highly skewed distribution (431 journals appear with just one publication each including at least one URL mention). Scientometrics, JASIST, and Journal of Informetrics are the principal sources (Table 3), being all of them core journals in the Library and Information Sciences field, area in which the *VOSviewer* software has been applied.

At the author-level, a total of 2,130 authors are found as author(s) or co-author(s) of publications including at least one mention to the software's URL (*author diffusion breadth*). The authors who have mentioned the software's URL the most times are shown

Table 2 List of tracking and context metrics used to measure the software use

Variables	Scope
Publications	
Author diffusion breadth	Total number of authors authoring or co-authoring at least one publication which includes a mention to the software's URL
Publication diffusion breadth	Total number of publications including at least one mention to the software's URL (e.g., number of mentioning publications)
Publication impact breadth (ix-index)	Number of publications including at least one mention to the software's URL, with at least i-x citations received each (e.g., i-10 index or i-100 index)
Publication URL mention intensity	Total number of URL mentions to the software's website divided by the number of mentioning publications
Source diffusion breadth	Total number of sources including at least one mention to the software's URL. It may include all or specific sources types (e.g., journals)
Specific publication URL mention counts	Number of URL mentions to the software's website from publications according to specific variables, such as publication type, section, etc
Total publication URL mention counts	Number of URL mentions to the software's website from publications
Webpages	
Academic website diffusion breadth	Number of academic-related websites which include at least one mention to the software's URL
Citation Flow	Score on a scale between 0 and 100 achieved by one website, based on the number of hyperlinks it receives. It measures how often a URL is linked (Jones, 2012). Therefore, it measures quantity of links received
Citation Flow impact breadth	Number of websites mentioning the software's URL which achieves a Citation Flow score equal or above 50
Specific website URL mention counts	Number of URL mentions to the software's website from external websites, according to specific variables, such as website category, language, location, etc
Total website URL mention counts	Number of URL mentions to the software's website from external websites
Trust Flow	Score on a scale between 0 and 100 achieved by one URL. It is based on the number of hyperlinks (and clicks on these links) from trusted seed sites that the URL receives. Therefore, it measures authority and ability to generate web traffic (Jones, 2012)
Trust Flow impact breadth	Number of websites mentioning the software's URL which achieves a Trust Flow score equal or above 50
Webpage diffusion breadth	Total number of webpages which include at least one mention to the software's URL (i.e., number of mentioning webpages)
Webpage language diffusion breadth	Number of different languages in which webpages including at least one mention to the software's URL are written
Webpage location diffusion breadth	Number of different countries in which webpages including at least one mention to the software's URL are located
Webpage URL mention intensity	Number of URL mentions to the software's website divided by the number of mentioning webpages

Table 2 (continued)

Variables	Scope
Website diffusion breadth	Number of websites which include at least one mention to the software's URL (e.g., number of mentioning websites)
Website impact breadth (ix-index)	Number of websites including at least one mention to the software's URL, with at least i-x hyperlinks received each
Tweets	
Like diffusion breadth	Number of likes received by a tweet which includes one mention to the software's URL
Liking author diffusion breadth	Number of authors putting at least one like on a tweet which includes a mention to the software's URL
Network average degree	Average number of edges (i.e., likes, retweets, and replies) per node (*Twitter* users) in the network
Network density	Number of connections the network has, divided by the total possible connections the network could have. Each node corresponds to a *Twitter* user and each edge to one type of interactivity (liking, retweeting, and replying)
Number of followers	Number of users who follow the activities of one Twitter user
Number of friends	Number of users followed by one *Twitter* user
Reply diffusion breadth	Number of replies received by a tweet which includes one mention to the software's URL
Retweet diffusion breadth	Number of retweets received by a tweet which includes one mention to the software's URL
Retweeting author diffusion breadth	Number of authors putting at least one retweet on a tweet which includes a mention to the software's URL
Specific author diffusion breadth	Number of authors publishing at least one tweet which includes a mention to the software's URL, according to specific variables, such as user gender
Specific tweet URL mention counts	Number of URL mentions to the software's website from tweets, according to specific variables, such as the tweet type, tweet category, tweet publication year
Total author diffusion breadth	Number of authors publishing at least one tweet which includes a mention to the software's URL
Total tweet URL mention counts	Number of URL mentions to the software's website from tweets
Tweet diffusion breadth	Number of tweets which include at least one mention to the software's URL (i.e., number of mentioning tweets)
Tweet like-based impact breadth (ix-index)	Number of tweets including at least one mention to the software's URL, with at least i-x likes received each
Tweet reply-based impact breadth (ix-index)	Number of tweets including at least one mention to the software's URL, with at least i-x replies received each
Tweet retweet-based impact breadth (ix-index)	Number of tweets including at least one mention to the software's URL, with at least x retweets received each
User centrality	Score that measures the prestige of a node (*Twitter* user) if it is connected to many other nodes who themselves have high scores and vice versa
User degree	Number of edges (likes, retweets or replies) directed into a node (*Twitter* user) in a directed graph

Table 2 (continued)

Variables	Scope
User social authority	Score from 0 (no authority) to 100 (maximum authority) which recursively measures the prestige of a Twitter user based on the prestige of the followers who follow said account

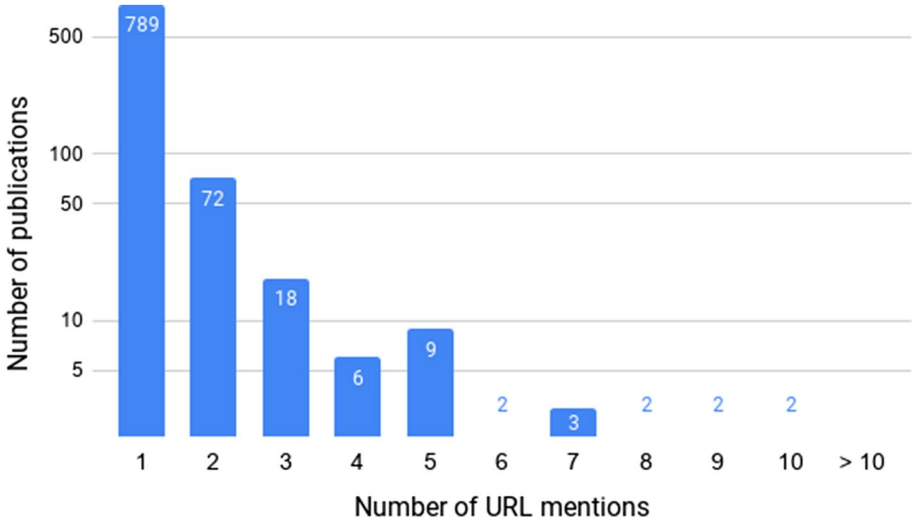

Fig. 3 Histogram showing the number of publications (y-axis) according to the number of mentions to the *VOSviewer*'s URL (vosviewer.com) (x-axis)

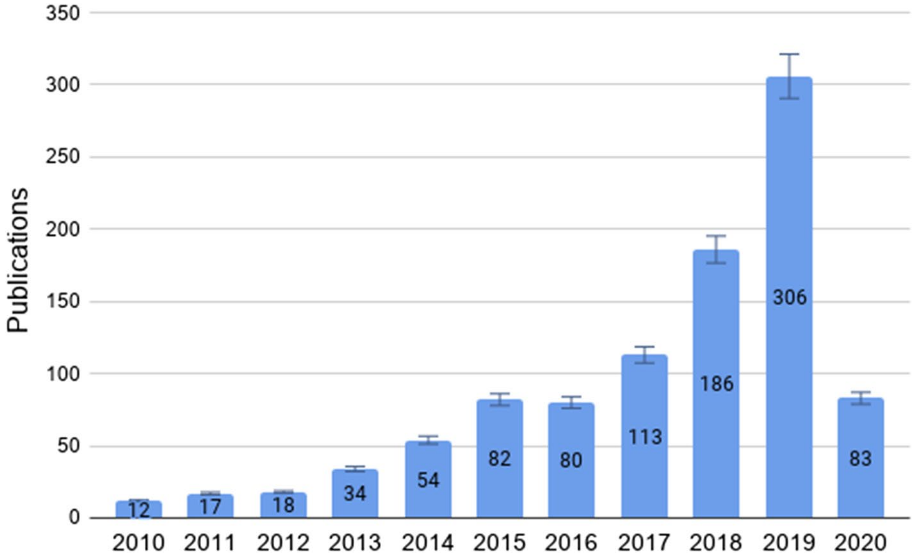

Fig. 4 Number of publications indexed in *Google Scholar* mentioning the *VOSviewer*'s URL (vosviewer.com) over the years

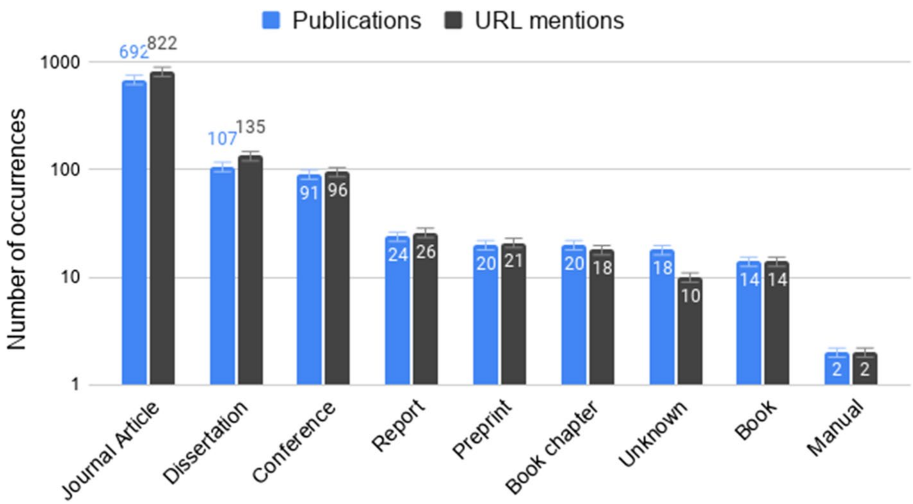

Fig. 5 Number of publications indexed in *Google Scholar* mentioning the *VOSviewer*'s URL (vosviewer.com) and total number of URL mentions included, according to the publication type

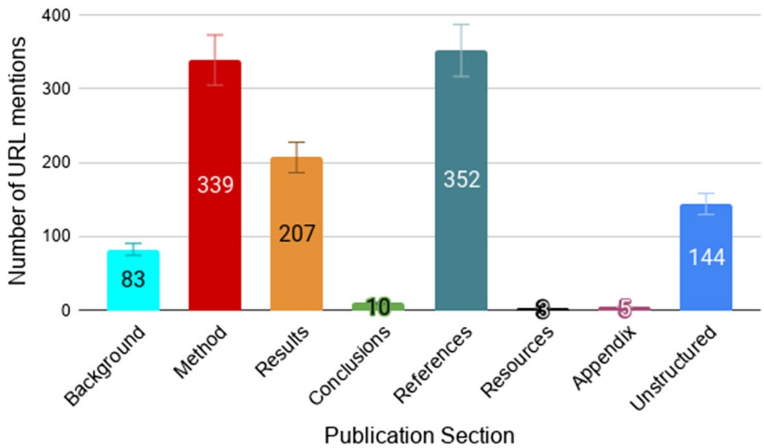

Fig. 6 Number of *publication URL mentions* to the *VOSviewer*'s URL (vosviewer.com) according to the publication section where located. *Note*: excluding documents retrieved but without mentioning 'vosviewer.com' (N = 32), and documents without full text access (N = 51)

Table 3 List of academic journals according to the number of mentions to the *VOSviewer*'s URL (vosviewer.com)

Journal	Number of Publications	Publication URL mention counts
Scientometrics	48	54
JASIST	15	53
Journal of Informetrics	12	15
PLoS ONE	8	12
International Journal of Environmental Research and Public Health	7	7
Journal of Cleaner Production	6	6
El Profesional de la Información	5	8
Dental Hypotheses	5	5
Sustainability	5	5
Borås Journal of Science	5	5
Journal of Business Research	5	5
图书情报工作 (Journal of Information Service)	5	6

JASIST Includes journal name changes over time

in the Table 4, where the co-developers of the software as well as other eminent researchers in the field of Scientometrics can be distinguished.

URL mentions in websites **via** ***Majestic***

VOSviewer's official website has accumulated 21,440 *website URL mentions* since 2014 (no data is available before this date in majestic), out of which 99.5% appear in internal pages (only 110 URL mentions come from websites' homepages).

These *website URL mentions* come from 17,261 webpages (*webpage diffusion breadth*) belonging to 1330 websites (*website diffusion breadth*). The *website impact breadth* of this corpus of mentioning websites is elevated (119 of these websites receive hyperlinks from at least 100,000 different external websites, while 244 of these websites receive hyperlinks from at least 10,000 different websites).

The number of new webpages mentioning *VOSviewer*'s URL increases over time. From 2015 onwards, the monthly average of new mentioning webpages is 246.7. During 2019, this average value increases to 467.9 (Fig. 7). The monthly average of new mentioning websites is 17.8 (Fig. 8).

The web authority of this corpus of websites mentioning the software's URL is diverse. 136 websites (10.2%) achieve a *Trust Flow* score equal to or greater than 50 (out of 100), and 181 websites (13.6%) achieve a *Citation Flow* score equal to or greater than 50 (also, out of 100), while most mentioning websites achieve lower scores, specially *Trust Flow* scores (926 mentioning websites achieve scores lower than 10), being websites with low web authority (Fig. 9).

The nature of mentioning websites is also diverse. About 28.5% of these 1330 websites (379) correspond to academic-related websites (*academic website diffusion breadth*),

Table 4 List of authors according to the number of mentions provided to the *VOSviewer*'s URL (vosviewer.com)

Author	Number of URL mentions
L Leydesdorff	34
NJ van Eck	24
L Bornmann	17
L Waltman	15
I Ràfols	13
E Şenel	11
R Haunschild	10
W Marx	9
DC Benton	9
J Kolahi	8
J Li	8
WM Sweileh	7
SH Zyoud	7
JM Merigó	7
AFJ van Raan	7

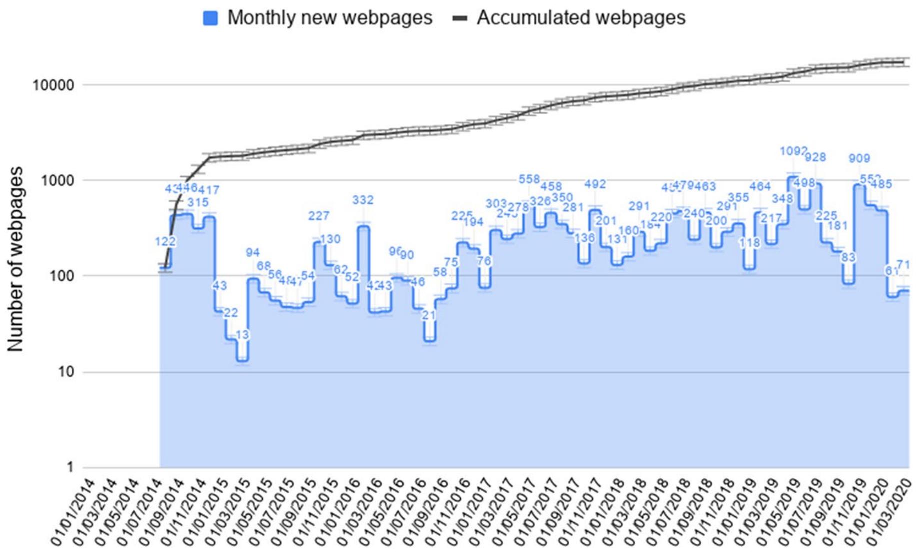

Fig. 7 Number of webpages (both total accumulated–line–and monthly–bars) mentioning *VOSviewer*'s URL (vosviewer.com) over time

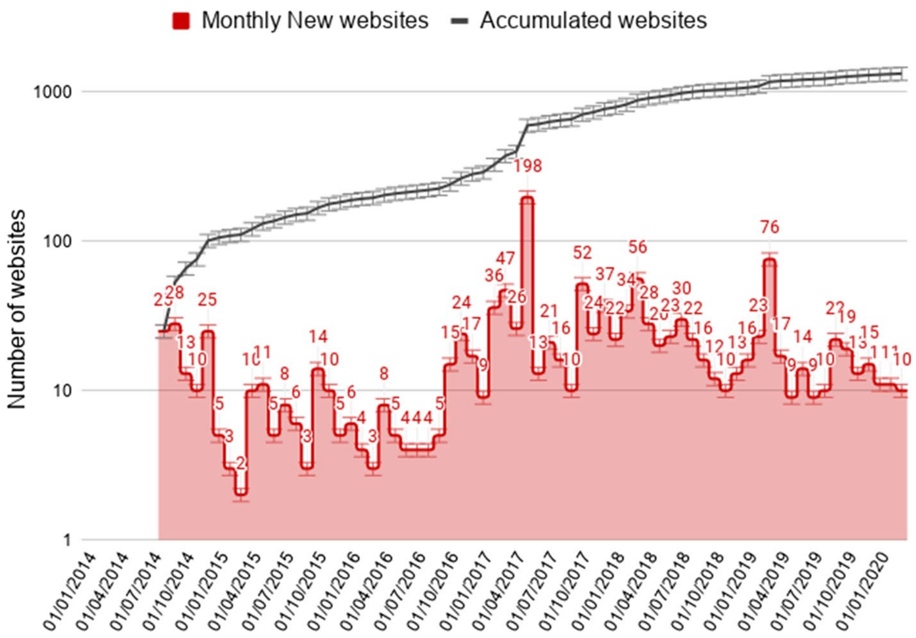

Fig. 8 Number of websites (both total accumulated–line–and monthly–bars) mentioning *VOSviewer*'s URL (vosviewer.com) over time

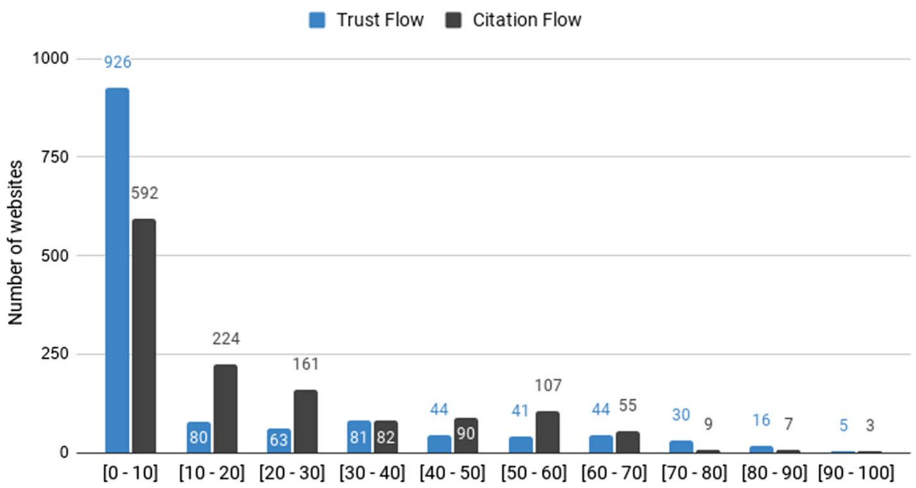

Fig. 9 Distribution of *Trust Flow* and *Citation Flow* scores for those websites mentioning *VOSviewer*'s URL (vosviewer.com)

whereas 23.8% (317) include dubious websites related to casinos, gambling, online bets, and even pornographic websites.

It is note to worth that a significant number of websites were not available two months after data gathering (473 websites, containing 2,906 mentioning webpages). Consequently, they could not be categorized. Likewise, 95 websites (containing 215 mentioning webpages) were *parked*[26] (Orduna-Malea, 2021), and 76 domains (208 mentioning webpages) automatically redirected to dubious websites.

Notwithstanding, the total percentage of mentioning webpages from non-academic websites was small (5.1%; 877 mentioning webpages), while URL mentions from academic-related webpages were majority (74.1%; 12,799 mentioning webpages). Therefore, its incidence on *VOSviewer*'s website overall online impact is limited (Fig. 10).

At the webpage level, the origin of URL mentions from academic-related webpages to *VOSviewer*'s URL is mainly from personal blogs (5,352 webpages), academic information products (1,883), and research groups (1,450). At the website level, universities and academic journals stand out (101 and 32 URL mentions, respectively). Most of websites (91.9%) provide just one URL mention to the software. In addition, non-academic websites have been included by way of illustration (Table 5). Of these, 473 websites have expired, 95 were parked and 76 redirected to other web locations. Other significant non-academic categories included websites with tricks for SEO professionals (44), non-academic companies (40) and mentions from online messages groups (57 links from 34 websites).

The language has been identified for 71% of all 17,261 webpages mentioning the software's URL, covering 26 different languages (*webpage language diffusion breadth*). URL mentions come mainly from webpages written in English (5,894) and French (5,216).

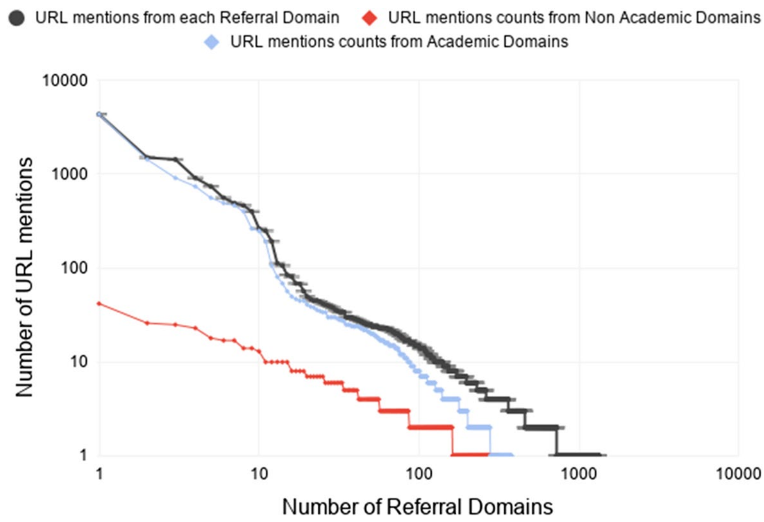

Fig. 10 Distribution of the number of mentioning webpages by mentioning websites

[26] A parked web domain is a domain name which has been registered but it is not associated with any service (email, website, etc.). Sometimes, a dummy webpage is artificially introduced to avoid a blank page on the browser.

Table 5 Ranking of website categories according to the number of academic webpages and websites mentioning *VOSviewer*'s URL (vosviewer.com)

Academic-related Category	Number of websites	Number of webpages	Non-Academic-related Category	Number of websites	Number of webpages
Universities	101	583	No available	473	2906
Academic journals	32	229	Parked websites	95	215
Academic information products	31	1,883	Redirects	76	208
Personal websites	28	656	Web SEO tricks	44	141
Academic Publishers	21	177	Companies	40	92
Personal blogs	17	5,352	Messages Group	34	57
Companies	15	849	Web services & tools	26	99
Thematic blogs	13	68	Blogs providers	17	301
Research institutes	10	63	Media	15	65
Blogs networks	9	141	Incomplete websites	15	22
Applications	9	47	Bet & Gambling	14	20
Associations	9	35	Porn	13	27
Apps directories	8	21	No access	12	15
Academic portals	7	9	Portal	10	113
Research centers	6	111	Thematic blogs	9	22
Research groups	4	1,450	Unknown	6	18
Research councils	4	13	Suspended	6	12
Research projects	4	12	eCommerce	5	19
Libraries	4	53	Q&A sites	4	26
Academic networks	4	5	Personal blogs	4	9

Analyzing the geo-location of IP addresses of each website's web domain, we find 3,849 webpages from 91 websites hosted in Netherlands (which is coherent as it is the place where *VOSviewer*'s developers work). However, only 46 webpages written in Dutch were identified (Table 6). A similar issue is found with Germany, from which 539 webpages from 94 different websites geo-located in this country are found, but only 21 webpages are written in German language. The use of English in most academic web environments may explain these results.

Table 6 Ranking of languages according to the number of webpages mentioning *VOSviewer*'s URL, written in the corresponding language

Language	Number of webpages	Language	Number of webpages
English	5,894	Croatian	32
French	5,216	Turkish	24
Spanish	398	German	21
Chinese	341	Slovenian	17
Swedish	62	Norwegian	16
Russian	61	Italian	8
Indonesian	53	Hungarian	8
Dutch/Flemish	46	Korean	8
Portuguese	40	Japanese	8

The IP address geo-location has identified webpages placed in 54 different countries (*webpage location diffusion breadth*), mainly from websites located at United States (53.9%; 717 websites) and United Kingdom (11%; 146 websites) (Fig. 11). At the level of mentioning webpages, Argentina (255 webpages from 3 websites) and Indonesia (113 webpages from 3 websites) stand out due to the existence of specific websites with many mentioning webpages (mainly from 'r020.com.ar', an Argentinean website dedicated to Library and information sciences resources with 249 mentioning webpages, and 'dasaptaerwin.net', an Indonesian personal academic website providing 113 mentioning websites, respectively).

67.1% of webpages mentioning the software's URL link directly to *VOSviewer*'s homepage. In addition, we can find a significant number of URL mentions linking to specific sections of the software's website, especially maps created by the software as use examples, and the page where the software is available to download (Table 7).

URL mentions in tweets via *Twitter*

A total of 267 tweets containing a URL mention to the *VOSviewer*'s official website have been identified in the period (*tweet diffusion breadth*). As all tweets include just one URL mention to the software's URL, the *tweet diffusion breadth* and the *total tweet URL mention counts* are the same.

This corpus of mentioning tweets has originated a further engagement of 89 replies (*reply diffusion breadth*), 646 retweets (*retweets diffusion breadth*), and 1109 likes (*like diffusion breadth*) (Table 8). The *tweet impact breadth* is limited; the *like-based impact breadth* achieves an i10-index of 26 (i100-index of 1) and the *retweet-based impact breadth*

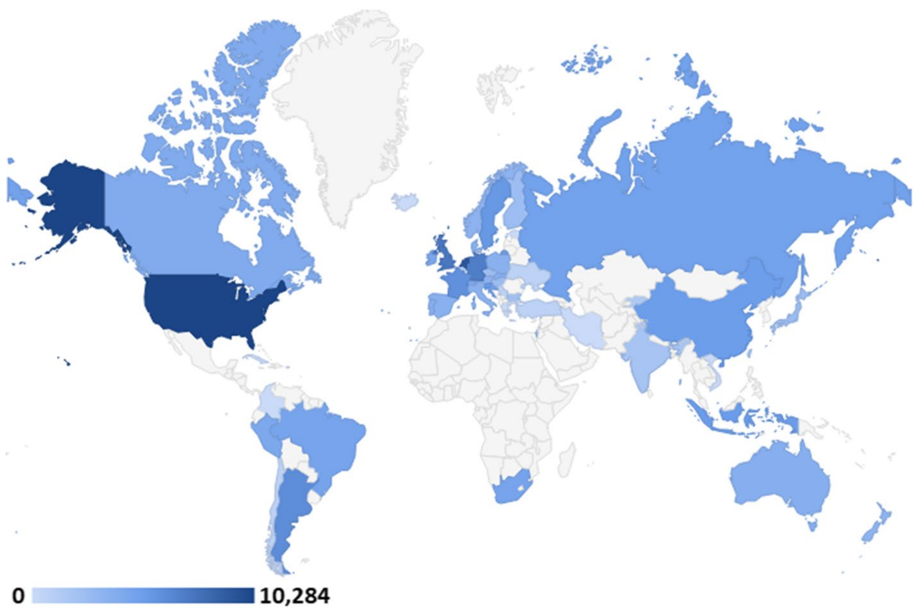

Fig. 11 Distribution of webpages mentioning *VOSviewer*'s URL (vosviewer.com) according to the website's web domain IP address location

Table 7 *VOSviewer*'s website specific URLs most mentioned by external webpages

Internal URL	Number of webpages
http://www.vosviewer.com/	11,588
http://vosviewer.com/	962
http://www.vosviewer.com/media/images/content/54e2cf50fbcad09b2ab812ccc6a17c72_large.png	377
http://www.vosviewer.com/media/images/content/3a62b50f9c73942203b27298338aa9c1_large.png	283
http://www.vosviewer.com/download	184
http://www.vosviewer.com/media/images/content/4c3e84a99b8dacaddff199ceaf66db90.png	179
http://www.vosviewer.com/publications	152
http://www.vosviewer.com/Home	144
http://www.vosviewer.com/vosviewer.php?map=http://www.sussex.ac.uk/Users/ir28/patmap/KaySupplementary3.txt	123
http://www.vosviewer.com/maps/	109
https://www.vosviewer.com/	87

Table 8 Number of tweets mentioning the software's URL (vosviewer.com) over the years, and the engagement achieved

Year	Tweets	Likes/Tweets	Retweets/Tweets	Replies/Tweets
2009	1	2.0	0.0	0.0
2010	3	5.7	3.7	0.0
2011	14	3.4	0.9	0.5
2012	20	19.4	8.9	1.0
2013	22	3.2	1.3	0.3
2014	13	5.5	2.8	0.5
2015	33	7.2	4.1	0.4
2016	41	1.7	1.6	0.2
2017	36	3.3	3.9	0.3
2018	37	1.4	0.8	0.2
2019	39	0.9	0.2	0.2
2020	8	0.0	0.4	0.0

The year 2020 is incomplete; it just covers from January to March

achieves an i10-index of 15 (i100-index is null). The low number of replies make *reply-based impact breadth* null (i10-index is 0; i1-index is 65).

75% of the 267 tweets gathered are original tweets, while 20% (53) are replies. Conversely, retweets containing URL mentions are scarce (5%; 14 tweets). Most tweets (80%) do not include images, and only two include media.

Original tweets related both with diffusion (21.3% of all tweets) and software recommendation (16.9%) are the most frequent. *VOSviewer*'s URL is also frequently mentioned in conversations about bibliometric maps generation (17.2%). However, tweets related with new versions releases and use examples are those achieving the largest engagement, both in terms of number of likes and retweets received (Fig. 12).

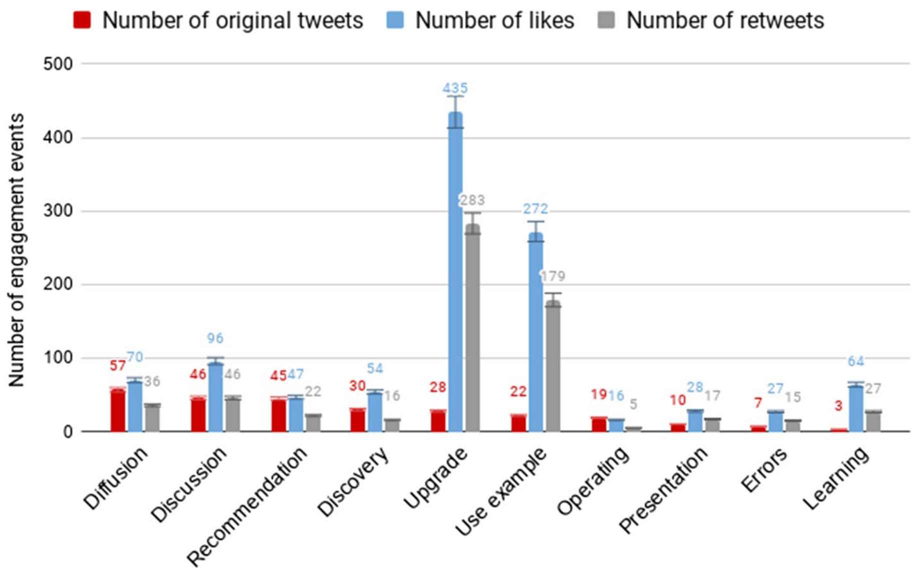

Fig. 12 Distribution of tweets mentioning *VOSviewer*'s URL (vosviewer.com) according to the category of each tweet Note: due to the low volume of the number of replies, this metric has been excluded

As regards to tweets' authorship, 200 users have been identified (*total author diffusion breadth*), 56% of which are male, 30% female, and 12% institutional accounts (Table 9). The community of users giving likes (*liking author diffusion breadth* of 748) and retweets (*retweeting author diffusion breadth* of 428) follow similar gender distribution, except for institutional profiles, which exhibit a lower participation in the generation of likes.

The overall community of attention of users who have published at least a tweet mentioning *VOSviewer*'s URL (200 users) can be characterized as having a high level of *Twitter* reputation, with a social authority median value equal to 47 (given that only few users in the world achieve a score of 100, this median value is considered substantial), and with a significant number of followers and total tweets published (Table 10). However, tweeting shows a skewed distribution of tweets per author, as only 18 users have published more than one tweet mentioning *VOSviewer*'s URL.

Table 9 Gender analysis of the users who tweeted, retweeted, and liked tweets mentioning the *VOSviewer* software's website

Distribution	Users originally tweeting	Users liking	Users retweeting
Male	111	435	233
Male (%)	55.5%	58.2%	54.4%
Female	59	243	126
Female (%)	29.5%	32.5%	29.4%
Institutional	24	52	62
Institutional (%)	12.0%	7.0%	14.5%
Undefined	6	18	7
Undefined (%)	3.0%	2.4%	1.6%
TOTAL	200	748	428

Table 10 Descriptive statistics related to users mentioning the *VOSviewer* software's URL (vosviewer.com) ($n = 200$)

Statistic	Followers	Friends	Tweets	Social-Authority
Minimum	1	0	34	1
Maximum	19,403	17,638	377,200	74
Range	19,402	17,638	377,166	73
1st Quartile	379.8	325.8	1,204.0	36.0
Median	1,087.5	847.0	4,207.0	47.0
3rd Quartile	2,313.5	1,762.0	12,900.0	57.5
Mean	2,067.4	1,417.1	15,607.6	44.2
Standard deviation (n-1)	3,087.6	2,000.0	38,018.6	17.8

The network of the community of attention of *VOSviewer* is a very sparse user (tweet creator) to user (like generator) network (Fig. 14 up), which shows one large node, few significant nodes, and lots of small nodes involved in sporadic interactions (one user generating a tweet and few liking it). This network (with low density and average degree and large diameter considering the number of nodes involved) shows the dependence on the *Twitter* activity of one user (van Eck, one of the co-developers of *VOSviewer*). In addition, a strong unbalanced user behavior is detected. Users who receive a great number of likes do not provide likes to those users including a URL mention in their tweets, reflecting a lack of interactivity in this community of users. The network based on retweets (Fig. 13 bottom) shows similar patterns, being reduced, sparse, and dependent on van Eck's activity.

Combining scenarios

Each of the three scenarios analyzed (academic literature, websites and *Twitter*) provides a complementary story about the use and interest in *VOSviewer*. Each scenario is determined by the different available context metrics provided in the corresponding *analytical framework*, each of which covers specific document bodies (academic publications, webpages and tweets, respectively).

The academic literature scenario allows checking the use of the software in publications. This way, *VOSviewer*'s *diffusion breadth* can be determined in terms of the number of publications, sources or authors mentioning the software's URL, and the intensity of this mentioning *event*. Moreover, driving the analysis to a greater detail, the appearance of mentions in different sections—beyond the references—can inform about the nature of the mentions. Thus, URL mentions can appear in the method section (as part of describing procedures and tasks performed), results (mainly to show data created with the software to illustrate direct findings related to objectives), and introductory sections (mainly to supplement literature reviews).

The web scenario has allowed determining *VOSviewer*'s interest through webpages linking to the software's URL. This way, *VOSviewer diffusion breadth* can be determined in terms of the number of webpages and websites mentioning the software, and the number of different languages and countries from which the software is mentioned. Moreover, the appearance of URL mentions in specific academic-related webpages can potentially inform about the interest of software in academic spaces beyond publications.

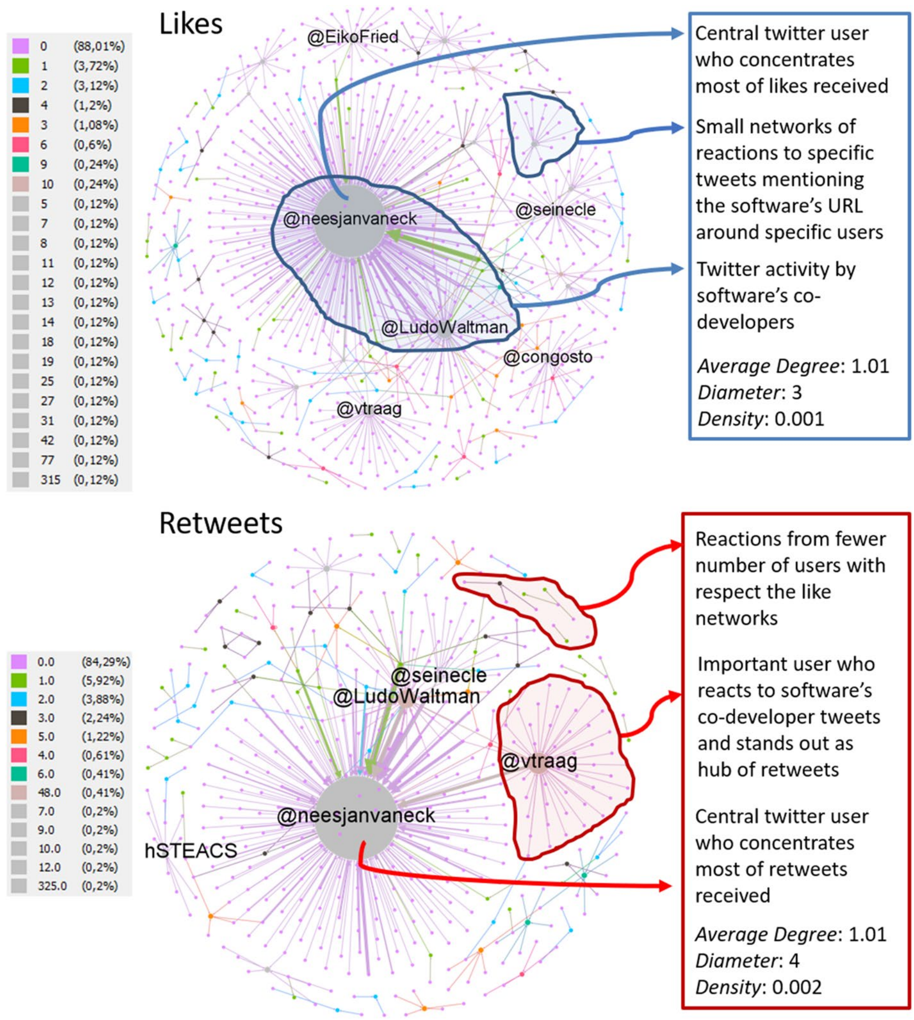

Fig. 13 Network of users providing likes (up) and retweets (bottom) to tweets mentioning the software's URL (vosviewer.com). Node color and node size according to the weighed InDegree value of each node. *Note*: due to the low volume of the number of replies, this metric has been excluded

The social media scenario has allowed checking the academic software's use and interest through tweets linking to the software's URL. This way, *diffusion breadth* can be determined in terms of the number of tweets and users mentioning the software's URL. Moreover, the appearance of URL mentions in different tweets allows detailed analyses at the tweet-level, such as the type of tweet (distinguishing replies—as part of discussions—and retweets—for mere diffusion purposes—from original tweets), the motivation of the tweet (e.g., software releases, use examples, etc.) and the engagement of the tweet (number of likes and retweets achieved by tweets mentioning the software's URL). In addition, data captures attention characteristics of the audience interested in the software, for example users' likes. The conversational nature of Twitter can also be useful in detecting user

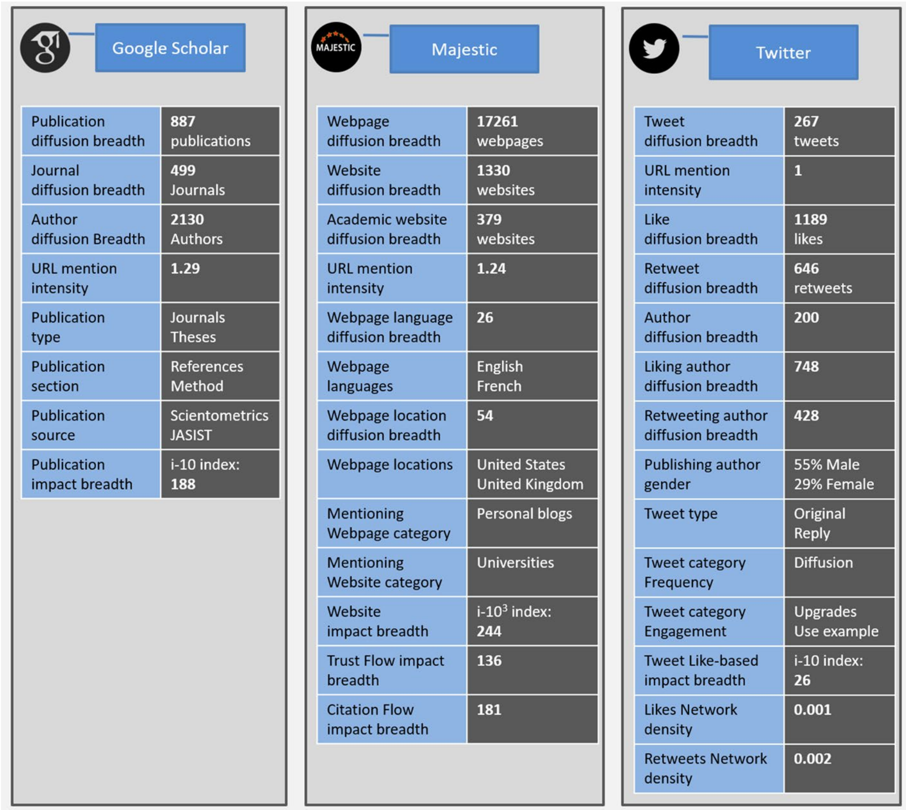

Fig. 14 Example of webometric technical evaluation sheet for *VOSviewer* software Note: *object*: URL (vosviewer.com); *event*: URL mentions

networks who may potentially employ the software, favoring the interaction and circulation of scientific knowledge across different communities of attention (Díaz-Faes et al., 2019).

All metrics gathered from each of the scenarios can be combined in technical analytical sheets (Fig. 14) to characterize the software's use in a multidimensional way. It can be highlighted that all data is captured externally in a non-invasive way, and without access privileges as webmasters, allowing its use to evaluate any other software in a similar fashion, as long as the software counts with an unambiguous URL.

Data included in the evaluation sheet is aimed at giving a detailed overview of software's diffusion breadth. Obviously, we do need to compare these results with other digital *objects* to determine whether the results obtained show an outstanding diffusion or not.

Discussion

The suitability of the approach presented in this case study depends mainly on the *sources* employed and the *object* selected. In this work, three sources have been selected to exemplify the proposed approach (*Google Scholar*, *Majestic*, and *Twitter*). The coverage, accuracy and availability of these *sources* determine the comprehensiveness and suitability of tracking and context metrics.

All three *sources* have as main advantages the coverage and the wide variety of available *tracking metrics*. *Google Scholar* covers academic publications regardless the type, language, and discipline, enabling the discovering of mentions in a larger number of publications and sources not available in other bibliographic *sources*. *Majestic* covers millions of webpages regardless the webpage type, language, or location, enabling the discovering of mentions in the global open online sphere, where academics and practitioners also participate. *Twitter* covers millions of tweets, enabling the discovering of mentions in the Twittersphere (the total universe of *Twitter* users and their publishing/following/reacting habits), which allows measuring interactivity between users (e.g., discussions threads or engagement).

Conversely, all these sources show limitations on availability, accuracy, and volatility. *Google Scholar* does not offer data export facilities (the *Publish or Perish* software helps but it is also limited to *Google Scholar* requirements). It also introduces errors in the bibliographic records that require manual examination, which in turn impairs the analysis of large sets of publications. Moreover, unlike other bibliographic databases, *Google Scholar* is not an accumulative database, and documents can be unindexed if they stop meeting *Google Scholar*'s technical requirements (Delgado López-Cózar, Orduña-Malea & Martín-Martín 2019). *Majestic* operates with highly-volatile data, as webpages are continuously changing. For example, 4,365 URL mentions to *VOSviewer* came from one specific personal blog (culturalibre.ca). Most of these hyperlinks were deleted just few months after data gathering, and the overall domain was lately disabled. Therefore, web data should be treated as dynamic and fluid impact instead of current impact. Moreover, *Majestic* data is offered under a paid license –which limits its use for large scale research endeavors– and its *flow metrics* are composite indicators whose full methodology is unknown, jeopardizing its transparency. Taking apart the volatility of the data, *Twitter* also deals with the limitations derived from compliance with its terms of use together with the *General Data Protection Regulation* (GDPR) as regards demographic data of users. Availability is also a problem as public *Twitter* API is not enough to carry out large data analyses and full API functionalities are offered under different paid licenses. The recent *Twitter* academic research API[27] might solve partially these limitations.

The *object* used to represent the software is another important aspect of the proposed approach. In this work, the *object* corresponds to the URL of the software's official website. The choice of a URL as a traceable digital object has important advantages. For example, the mentions to the URL can be unequivocally identified, avoiding the polysemy and synonymy of natural language. In addition, URLs have specific search filters to ease data retrieval in a multitude of sources. These characteristics allow the generation of faster, simpler and more refined information retrieval systems. Moreover, URLs also allow online navigation (and, therefore, web traffic), which potentially facilitates software usage, and constitutes an effective type of mentioning software for scientific purposes. Finally, we can find a wide variety of URL metrics, which are not available for other textual metrics or, in some cases, obtaining them would require much more complex computing needs.

Despite the benefits of using URL as an *object*, the following limitations (and threats) should be acknowledged (Table 11):

[27] https://developer.twitter.com/en/portal/petition/academic/is-it-right-for-you.

Table 11 URL-based text mentions limitations

Limitation	Description	Potential solutions
A URL is mandatory	The link-based approach used in this work cannot be used to measure software without a URL as unique identifier or with very common names (e.g., 'R', 'SAS', 'EXCEL', 'WORD', etc.)	Additional *objects* (e.g., other software's text mentions or non-textual mentions) should be incorporated. However, using different *objects* might introduce complexity (noise and need to differentiate among distinct *events*) to the model
URL multiplicity	A piece of software can have more than one URL associated with it. In addition, this URL may correspond to different platforms (software repository, institutional repository, journal article, official website, etc.). Specific dynamics and demographics of each platform can produce different events and engagement from different types of users	All URLs related to the software analyzed should be identified and treated as aliases. Link-related metrics should be shown both at the aggregate level (aggregating the different links received for the software, regardless the URL accruing them) and individually (counting the links per differentiated URL). The individual analysis of the different URLs of a given software, each URL representing the software should be categorized accordingly
Data prevalence	There are text-mentions to the software that do not include its URL. As a reference, this study reports 1,190 hit counts in *Google Scholar* to the query "vosviewer.com", but 7,080 using the string "vosviewer" (as of 31 March 2020). Therefore, URL mentions counts may not always be enough to identify all the mentions and usage of a specific software	The use of URL mentions would improve over time as software citation developments claim that metadata should include a way to digitally locate the software (Niemeyer et al., 2016)
Data volume	When working with large amounts of cases or data volumes (e.g., analyzing many different software simultaneously), alternative strategies would be needed for data collection, especially regarding data related to mentions from publication sources	Initiatives such as *Softcite* (Du et al., 2021), which apply supervised machine learning techniques for automatic extraction of software mentions from PDF format research publications, will facilitate the data capturing processes
Data sources	The analytical framework presented in this paper relies on data from different data sources that are mainly available under paywalls, and offering non-transparent sources and metrics (e.g. Trustflow, etc.)	Open sources are advisable when available, whereas description of metrics (see Table 2) is a must to provide a technical evaluation sheet as much accurate as possible
Data volatility	Online data (both from web and social media sources) are highly volatile, since they may disappear, reappear, or change, thus affecting the counts and metrics reported	Data should be interpreted as rough estimates. Longitudinal data is necessary to avoid seasonal or technical biases

Table 11 (continued)

Limitation	Description	Potential solutions
Other forms of software impact missed	URL mentions cannot be expected to capture all forms of software impact. For example, clicks to URLs, downloads and real web traffic driven to the official website are impact-related metrics missed. Most importantly, neither the linking, visit nor the download of the software, can be taken as forms of digital acts capturing the actual use (or application) of the software, and they must be seen merely as metrics capturing the idea of "access" as discussed in Haustein et al. (2016)	Evidence of software usage would preferably be supplemented with other online metrics, such as the number of downloads, visits to the website, "appraisals" of the software (e.g., tweets praising the software), its application, etc. Official academic software websites could also openly offer online metrics from their analytical suites (e.g., their own websites), as external tools are not always accurate enough

Software can be mentioned by authors for a wide range of reasons (an instrument, an artifact, a scientific protocol, a method, or just an example—Howison & Bullard 2016; Li et al., 2019) regardless the type of text-mention used (URL, software's name, etc.). The publication sections where the mention appears or the category of tweets where the software is mentioned have been precisely studied in this work as exploratory signals. Further qualitative techniques would add more context to the Twitter conversations, which may enhance the software's technical evaluation sheet obtained. The taxonomy of tweets is only based on the collected tweets, but a more generic taxonomy would be advisable, including other potential categories or subcategories (e.g., software awards, datacamps, dedicated conferences and meetings, official use of scientific software in syllabi, mentions and reviews from professional specialized media, diffusion of books dedicated to the software, forks, etc.) not considered in this analysis but foreseeable necessary for other types of software.

VOSviewer has been used as a case study to exemplify the method. The manageable number of URL mentions found together with its online presence has made its choice an adequate one to test a wide number of metrics per *analytical framework*.

Conclusion

A webometric analytical approach to track scientific software use and interest has been proposed in this work. This approach is based on the definition of *scenarios*, *analytical elements*, *analytical frameworks*, *sources*, *objects*, *events*, *tracking metrics* and *context metrics*. The operationalization of the approach has been exemplified by analyzing one specific academic software (*VOSviewer*) and object (*VOSviewer*'s official URL).

Results show that the different *analytical frameworks* provide useful information about the usage of scientific software, expanding the notion of *usage of scientific software in research publications* to *dissemination and interest of scientific software in the research community*, illustrating how this usage information is relevant to fully comprehend the broader influence of scientific software in the whole research ecosystem.

The *Google scholar scenario* has shown *VOSviewer* as a research resource, whilst the *Majestic* and *Twitter scenarios* have shown the interest of *VOSviewer* as an information and a conversational resource, respectively.

Finally, the approach proposed in this study can be expanded by adding new *scenarios* and new *analytical frameworks*. Each *analytical framework* can also be expanded by considering additional *sources*, *objects*, *events*, *tracking metrics* and *context metrics*. In addition, data from different scenarios can be combined to create new indicators and to show added-value information (e.g., an integrated publication timeline for each specific event tracked).

As a matter of fact, the approach proposed in this study can actually be used to track any research *object* that can be enclosed in a specific URL (e.g., software, scientific conferences, presentations, online courses, videos, scientific exhibitions, research projects, academic websites, etc.), effectively expanding the analytical scope of the scientometric toolset by incorporating a novel digital dimension through methods that draw from the fields of Webometrics and Altmetrics.

Appendix 1 Data sources used in the study

In this section we discuss the main data sources selected for the three scenarios and the main methodological approach.

Google Scholar

Google Scholar[28] is a freely accessible academic search engine launched in 2004 and aimed at facilitating the discovery of academic literature worldwide (Ortega, 2014). To accomplish with this overarching goal, *Google Scholar* employs user agents (web applications) that automatically discover and scan websites by following hyperlinks from one webpage to another (Delgado López-Cózar et al., 2017). This way, *Google Scholar* parses the entire academic web (websites of universities, scientific publishers, repositories, aggregators, library catalogues, and any other web spaces where they might find academic-like materials).

Google Scholar indexes in a –mostly– unsupervised manner every scholarly document it finds as long as it meets a set of technical requirements, covering thus a whole range of academic document types (books, book chapters, journal articles, conference articles, teaching materials, theses, posters, presentations, reports, patents, etc.), from a wide range of disciplines (including Arts, Engineering, Humanities and Social Sciences) and regardless their language (Delgado López-Cózar et al., 2019).

These operating characteristics make *Google Scholar* the largest bibliographic database in the world today (Delgado López-Cózar et al., 2019; Gusenbauer, 2019; Orduna-Malea et al., 2015; Ortega, 2014), and therefore, a suitable source to measure software text-mentions in the academic literature.

Majestic

Majestic[29] is a link intelligent tool launched in 2008 by *Majestic-12* and oriented to massive link analysis and search engine optimization (SEO). While *Google Scholar* indexes academic documents, *Majestic* indexes URLs, providing a wide range of tailored metrics to determine the impact and authority of these URLs –and the webpages they represent– on the Web, especially the number of webpages including text mentions to the URL analyzed.

Currently, *Majestic* is one of the most comprehensive sources of web data on the Web, declaring 2,482 billion unique URLs indexed in its historic database (coverage from 2015 to 2020) and 947 billion unique URLs indexed in its fresh database (last five months), as of 21 January 2021. This database includes all kind of websites (blogs, portals, wikis, fora, etc.) from all kind of users (personal websites, company websites, organizations, institutions, etc.).

Given its functionalities and coverage, *Majestic* has been used as a data source in webometrics contributions (Jansen, Jung & Salminen, 2020; Lepori et al., 2014; Orduna-Malea & Regazzi, 2014) and it is also used as a data *source* in the *Ranking Web of universities*.[30]

[28] https://scholar.google.com.

[29] https://majestic.com.

[30] http://www.webometrics.info/en.

For these reasons, *Majestic* constitutes an authoritative source to measure software text-mentions on the Web-at-large.

Twitter

Twitter[31] is a microblogging service created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass. The main feature of this platform is to instantly create and publish short messages called tweets (originally up to 140 characters, a size that was expanded to 280 in 2017) in which diverse files can be embedded (static or moving images, hyperlinks, hashtags and mentions to other users). Readers can also interact with these messages in different ways, expressing that they find interesting it (like), spread it to their contacts (retweet) or give a direct answer (reply) generating thus discussion threads (a conversational feature). In a complementary way, users generate social networks by following the publications and activities of other users. They can also follow topics of interest (through hashtags), make user lists or even communicate with other users through private direct messages. *Twitter* also offers a wide number of metrics both at the user-level (e.g., number of tweets published, number of followers achieved, user demographic data [gender, location, interests…]) and at the tweet-level (e.g., number of likes, retweets and replies received, tweet demographic data [who created the tweet, when, where…]).

Twitter generates about 700 million tweets per day approximately (as of January 2020). It is estimated that since its inception, around 1.3 billion accounts have been created (Smith, 2020), maintaining some 330 million active monthly users and 145 million daily users. The *Twitter* website is visited by 6.54 billion users (both registered and unregistered) monthly according to December 2021 data provided by the analytics tool *SimilarWeb*.[32]

Given the huge amount of data generated on *Twitter*, this platform has been widely used not only as a source of data in the scientific literature but also as an object of study on its own (Bruns et al., 2014; Ovadia, 2009; Stewart, 2017; Williams et al., 2013). For these reasons, *Twitter* constitutes an authoritative source to measure software text-mentions on social media.

Supplementary Information The online version supplementary material available at https://doi.org/10.1007/s11192-021-04082-y.

Acknowledgements Special thanks to Ariadna Fernández-Planells for the Twitter intercoder reliability test, and to Cristina Font-Julián for the website intercoder reliability test.

Authors' contributions Rodrigo Costas is member of the Distinguished Reviewers Board of Scientometrics, and is employed at the Centre for Science and Technology Studies (CWTS) that is host of the VOSviewer software. However, he is not involved in its development. All authors contributed to the study conception and design. Data collection and analysis were performed by Enrique Orduña-Malea. The manuscript was written by all authors. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding RC was partially funded by the South African DST-NRF Center of Excellence in Scientometrics and Science, Technology, and Innovation Policy (SciSTIP).

[31] https://twitter.com.

[32] https://www.similarweb.com/website/twitter.com.

References

Bruns, A., Weller, K., Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management, 66*(3), 250–261.

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science, 49*(14), 1319–1328.

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2019). Google Scholar as a data source for research assessment. In W. Glänzel, H. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 95–127). Springer.

Delgado López-Cózar, E., Orduna-Malea, E., Martín-Martín, A., & Ayllón, J. M. (2017). Google Scholar: The big data bibliographic tool. In F. J. Cantú-Ortiz (Ed.), *Research analytics: Boosting university productivity and competitiveness through scientometrics* (pp. 59–80). Taylor and Francis.

Díaz-Faes, 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

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24454

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.

Hafer, L., & Kirkpatrick, A. E. (2009). Assessing open source software as a scholarly contribution. *Communications of the ACM, 52*(12), 126–129.

Halavais, A. (2008). The hyperlink as organizing principle. In J. Turow & L. Lokman (Eds.), *The hyperlinked Society: Questioning connections in the digital age* (pp. 39–55). The University of Michigan Press.

Hannay, J. E., MacLeod, C., Singer, J., Langtangen, H. P., Pfahl, D., & Wilson, G. (2009). How do scientists develop and use scientific software? *Proceedings of the 2009 ICSE workshop on software engineering for computational science and engineering*, SECSE 2009, 1–8. https://ieeexplore.ieee.org/abstract/document/5069155.

Haustein, S., Bowman, T. D., & Costas, R. (2016). Interpreting "altmetrics": Viewing acts on social media through the lens of citation and social theories. In C. Sugimoto (Ed.), *Theories of informetrics and scholarly communication* (pp. 372–406). De Gruyter Saur.

Hey, T., Tansley, S., & Tolle, K.M. (Ed.) (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft research. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology, 67*(9), 2137–2155.

Howison, J., & Herbsleb, J. D. (2011). Scientific software production: incentives and collaboration. *Proceedings of the ACM 2011 conference on computer supported cooperative work –CSCW '11*, 513–522. https://doi.org/10.1145/1958824.1958904

Howison, J., Deelman, E., McLennan, M. J. M., Da Silva, R. F., & Herbsleb, J. D. (2015). Understanding the scientific software ecosystem and its impact: Current and future measures. *Research Evaluation, 24*(4), 454–470.

Jansen, B. J., Jung, S.G., & Salminen, J. (2020). *Data Quality in Website Traffic Metrics: A Comparison of 86 Websites Using Two Popular Analytics Services*. http://www.bernardjjansen.com/uploads/2/4/1/8/24188166/traffic_analytics_comparison.pdf.

Jones, D. (2012). *Flow Metrics™ will change the way you look at links. Majestic Blog*. https://blog.majestic.com/development/flow-metrics.

Katz D. S., Choi S-. C. T., Niemeyer, K. E. et al. (2016). *Report on the third workshop on sustainable software for science: practice and experiences* (WSSSPE3). https://arxiv.org/abs/1602.02296.

Li, K., Chen, P. Y., & Yan, E. (2019). Challenges of measuring software impact through citations: An examination of the lme4 R package. *Journal of Informetrics, 13*(1), 449–461.

Li, K., & Yan, E. (2018). Co-mention network of R packages: Scientific impact and clustering structure. *Journal of Informetrics, 12*(1), 87–100.

Li, K., Yan, E., & Feng, Y. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. *Journal of Informetrics, 11*(4), 989–1002.

Lepori, B., Aguillo, I. F., & Seeber, M. (2014). Size of web domains and interlinking behavior of higher education institutions in Europe. *Scientometrics, 100*(2), 497–518.

Niemeyer, K. E., Smith, A. M., & Katz, D. S. (2016). The challenge and promise of software citation for credit, identification, discovery, and reuse. *Journal of Data and Information Quality, 7*(4), 1–5.

Orduna-Malea, E. (2021). Dot-Science Top Level Domain: Academic websites or dumpsites? *Scientometrics*, *126*(4), 3565–3591. https://doi.org/10.1007/s11192-020-03832-8

Orduna-Malea, E. (2020). Investigando con Twitter: una mirada según el Reglamento General de Protección de Datos. In Francisca Ramón-Fernández (Ed.). *Marco jurídico de la ciencia de datos* (pp. 331–378). Valencia: Tirant lo Blanch.

Orduna-Malea, E., & Alonso-Arroyo, A. (2017). *Cybermetric techniques to evaluate organizations using web-based data*. Chandos Publishing.

Orduna-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics, 104*(3), 931–949.

Orduna Malea, E., Martín-Martín, A., & Delgado-López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista Española De Documentación Científica, 40*(4), 1–33.

Orduna-Malea, E., & Regazzi, J. J. (2014). US academic libraries: Understanding their web presence and their relationship with economic indicators. *Scientometrics, 98*(1), 315–336.

Ortega, J. L. (2014). *Academic search engines: A quantitative outlook*. Elsevier.

Ovadia, S. (2009). Exploring the potential of Twitter as a research tool. *Behavioral & Social Sciences Librarian, 28*(4), 202–205.

Pan, X., Cui, M., Yu, X., & Hua, W. (2017). How is CiteSpace used and cited in the literature? An analysis of the articles published in English and Chinese core journals. *ISSI 2017–16th International conference on Scientometrics and Informetrics*. http://issi-society.org/proceedings/issi_2017/2017ISSI%20Conference%20Proceedings.pdf.

Pan, X., Yan, E., & Hua, W. (2016). Disciplinary differences of software use and impact in scientific literature. *Scientometrics, 109*(3), 1–18.

Pan, X., Yan, E., Cui, M., & Hua, W. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. *Journal of Informetrics, 12*(2), 481–493.

Pan, X., Yan, E., Cui, M., & Hua, W. (2019). How important is software to library and information science research? A content analysis of full-text publications. *Journal of Informetrics, 13*(1), 397–406.

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics, 9*(4), 860–871.

Park, H. W., & Thelwall, M. (2003). Hyperlink analyses of the World Wide Web: A review. *Journal of computer-mediated communication*. https://doi.org/10.1111/j.1083-6101.2003.tb00223.x

Park, H., & Wolfram, D. (2019). Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse. *Journal of Informetrics, 13*(2), 574–582.

Pia, M. G., Basaglia, T., Bell, Z. W., & Dressendorfer, P. V. (2009). Geant4 in scientific literature. *IEEE Nuclear Science Symposium Conference Record*, 189–194. https://ieeexplore.ieee.org/document/5401810.

Piwowar, H. A. (2013). Value all research products. *Nature, 493*, 159.

Pradal, C., Varoquaux, G., & Langtangen, H. P. (2013). Publishing scientific software matters. *Journal of Computational Science, 4*(5), 311–312.

Smith, K. (2020). 58 Incredible and Interesting Twitter Stats and Statistics. *Brandwatch.* https://www.brandwatch.com/blog/twitter-stats-and-statistics.

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://peerj.com/articles/cs-86/.

Soito, L., & Hwang, L. J. (2016). Citations for Software: Providing identification, access and recognition for research software. *IJDC, 11*(2), 48–63.

Stewart, B. (2017). Twitter as method: Using Twitter as a tool to conduct research. *L. Sloan, & A. Quan-Haase, Social Media Research Methods*, 251–266.

Thelwall, M. (2004). *Link Analysis: An information science approach*. Elsevier.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*(1), 60–68.
Thelwall, M., & Kousha, K. (2016). Academic software downloads from google code. *Information Research*, *21*(1). http://informationr.net/ir/21-1/paper709.html#.XzelJ-gzbIU.
Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.
Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation, 69*(3), 384–410.
Wouters, P., Zahedi, Z., & Costas, R. (2019). Social media metrics for new research evaluation. In W. Glänze, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 687–713). Springer.
Yang, B., Rousseau, R., Wang, X., & Huang, S. (2018). How important is scientific software in bioinformatics research? A comparative study between international and Chinese research communities. *Journal of the Association for Information Science and Technology, 69*(9), 1122–1133.