RESEARCH ARTICLE

BIP4COVID19: Releasing impact measures for articles relevant to COVID-19

Thanasis Vergoulis[1], Ilias Kanellos[1], Serafeim Chatzopoulos[1,2], Danae Pla Karidi[1,3], and Theodore Dalamagas[1]

[1]IMSI, "Athena" RC, Athens, Greece
[2]Dept. of Informatics and Tele/tions, University of the Peloponnese, Tripolis, Greece
[3]NTU Athens, Athens, Greece

Keywords: scientometrics, impact measures, covid-19, open scholarly data

ABSTRACT

Since the beginning of the coronavirus pandemic, a large number of relevant articles have been published or become available in preprint servers. These articles, along with earlier related literature, compose a valuable knowledge base affecting contemporary research studies or even government actions to limit the spread of the disease, and directing treatment decisions taken by physicians. However, the number of such articles is increasing at an intense rate, making the exploration of the relevant literature and the identification of useful knowledge challenging. In this work, we describe BIP4COVID19, an open data set that offers a variety of impact measures for coronavirus-related scientific articles. These measures can be exploited for the creation or extension of added-value services aiming to facilitate the exploration of the respective literature, alleviating the aforementioned issue. In the same context, as a use case, we provide a publicly accessible keyword-based search interface for COVID-19-related articles, which leverages our data to rank search results according to the calculated impact indicators.

1. INTRODUCTION

COVID-19 is an infectious disease caused by the coronavirus SARS-CoV-2, which may result, in some cases, in progressing viral pneumonia and multiorgan failure. After its first outbreak in Hubei, a province in China, it subsequently spread to other Chinese provinces and many other countries. On March 11, 2020, the World Health Organization (WHO) declared the 2019–20 coronavirus outbreak a pandemic. Up to October 2021 more than 237,000,000 cases had been recorded in more than 200 countries, counting more than 4,800,000 fatalities.[1]

At the time of writing, an extensive number of coronavirus-related articles have been published since the virus' outbreak (indicatively, our collected data contain about 294,619 articles published from 2020 onwards). Taking additionally into account previous literature on coronaviruses and related diseases, it is evident that there is a vast corpus on the subject. However, it is critical for researchers or other interested parties (e.g., government officers, physicians) to be able to identify high-impact articles. A variety of impact measures for scientific articles have been proposed in the fields of bibliometrics and scientometrics. Some of them rely on the analysis of the underlying citation network (Kanellos, Vergoulis et al., 2019). Other approaches utilize measures commonly known as "altmetrics" (Piwowar, 2013), which analyze data from

[1] https://covid19.who.int/

an open access journal

Citation: Vergoulis, T., Kanellos, I., Chatzopoulos, S., Pla Karidi, D., & Dalamagas, T. (2021). BIP4COVID19: Releasing impact measures for articles relevant to COVID-19. *Quantitative Science Studies*, *2*(4), 1447–1465. https://doi.org/10.1162/qss_a_00169

DOI: https://doi.org/10.1162/qss_a_00169

Corresponding Author: Thanasis Vergoulis
vergoulis@athenarc.gr

Copyright: © 2021 Thanasis Vergoulis, Ilias Kanellos, Serafeim Chatzopoulos, Danae Pla Karidi, and Theodore Dalamagas. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

social media and/or usage analytics in online platforms (e.g., publishers' websites). Both approaches have their benefits and shortcomings, each capturing different aspects of an article's impact. Thus, by considering a wide range of different measures we can better uncover a comprehensive view of each article's impact.

In this context, the main objective of this work is to produce and publish *BIP4COVID19*, an openly available data set that contains a variety of different impact measures calculated for COVID-19-related literature. Four citation-based impact measures (Citation Count, PageRank (Page, Brin et al., 1999), RAM (Ghosh, Kuo et al., 2011), and AttRank (Kanellos, Vergoulis et al., 2021)) were chosen to be calculated, as well as an altmetric indicator (Tweet Count). The selected measures were chosen so as to cover different impact aspects of the articles. Furthermore, to select a representative set of publications, we rely on two open data sets of COVID-19-related articles: the CORD-19 (Wang, Lo et al., 2020a) and the LitCovid (Chen et al., 2020a, 2020b) data sets. BIP4COVID19 data are updated on a regular basis and are openly available on Zenodo[2] (Vergoulis, Kanellos et al., 2021b). Since its initial launch in March 2020, 63 versions of BIP4COVID19 have been released, with Zenodo reporting more than 160,000 views and 23,000 downloads.

The BIP4COVID19 data set may be useful in various applications. Indicatively, the calculated indicators can be incorporated in academic search engines for COVID-19 literature, such as LitCovid and COVIDScholar (a comprehensive list of such search engines is given in Section 5.3), helping users to prioritize their reading by ranking search results according to scientific impact. The scores may also be valuable for monitoring research output impact and as a basis for various scientometrics analyses (see Section 4.4 for details). In fact, in this work, apart from the data set itself, we have also developed and released a publicly accessible web interface[3], which functions as a keyword-based search engine on COVID-19-related literature. This search engine exploits the calculated measures to offer relevant functionalities (e.g., impact-based ranking for search results) that facilitate the exploration of the COVID-19-related literature.

The remainder of this paper is structured as follows. In Section 2 we present the processes of data collection, cleaning, and integration, as well as the calculation of the impact indicators. Section 3 presents the created data set and the search engine built on top of it. In Section 4 we present analytics related to the data set produced, we discuss its possible usages and consider its limitations. Section 5 presents related work and, finally, Section 6 concludes the work.

2. MATERIALS AND METHODS

BIP4COVID19 is a regularly updated data set.[4] Data production and update are based on the semiautomatic workflow presented in Figure 1. In the following subsections we describe the processes involved.

2.1. Article Data Collection and Cleaning

The list of COVID-19-related articles is created based on two data sources: the *CORD-19*[5] Open Research Dataset (Wang et al., 2020a), provided by the Allen Institute for AI, and the *LitCovid*[6] collection (Chen et al., 2020a) provided by the NLM/NCBI BioNLP Research Group. CORD-19 offers a full-text corpus of more than 780,000 articles on coronaviruses and COVID-19, collected based on articles that contain a set of COVID-19-related keywords

[2] https://zenodo.org/record/4555117
[3] https://bip.covid19.athenarc.gr/
[4] Details on our update policy can be found in Section 3.1.
[5] https://pages.semanticscholar.org/coronavirus-research
[6] https://www.ncbi.nlm.nih.gov/research/coronavirus/

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

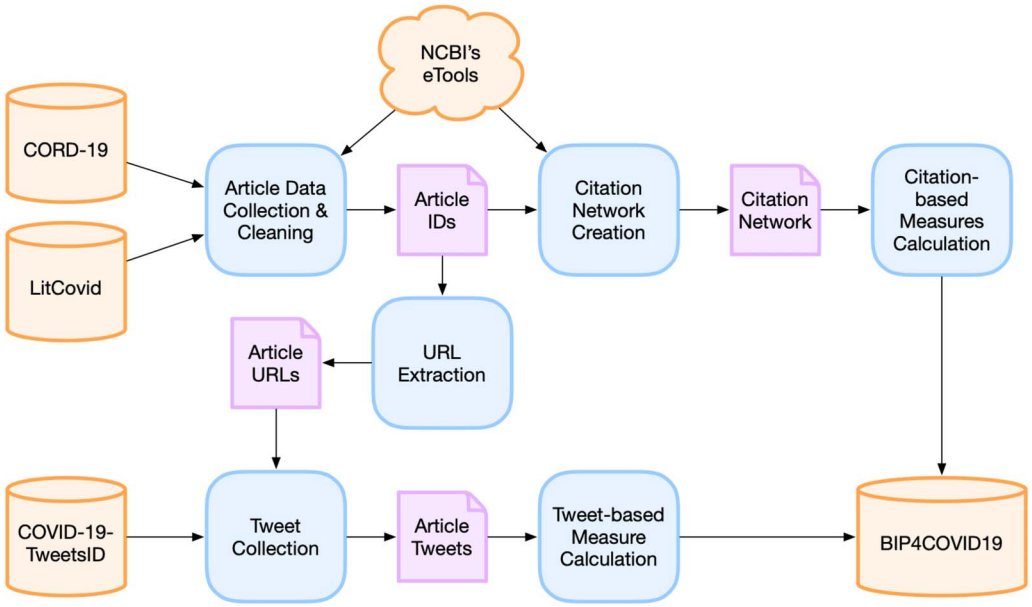

Figure 1. The data update workflow of BIP4COVID19.

from PMC, arXiv, biorXiv, and medRxiv and the further addition of a set of publications on the novel coronavirus, maintained by the WHO. LitCovid is a curated data set which currently contains more than 180,000 papers on the novel coronavirus, which are indexed by PubMed.

To produce the BIP4COVID19 data set, we combine data from both of the two aforementioned sources (CORD-19 and LitCovid). Both sources maintain publication data records by providing paper identifiers in different databases. For our purposes we only collect, from both sources, the identifiers given from the PubMed (pmids) and PubMed Central (pmcid) databases. Then, for each distinct article identified, we use NCBI's eSummary tool[7] to collect each paper's title, publication venue, and publication year, along with three identifiers: its pmid, its pmcid, and its DOI.[8] While some of these data are included in the original sources, this is not the case for all of them. For example, LitCovid's downloadable TSV file provides only the pmids, titles, and publication venues. Hence, by directly collecting the data from NCBI we enrich the source data, while avoiding potential differences between records in the two sources (e.g., differences in titles). The metadata collected are required to uniquely identify each article, providing useful information to be displayed by BIP4COVID19's web search interface (see Section 3.2). Additionally, the paper publication years are a prerequisite for the calculation of popularity-based impact measures (see Section 2.2).

At this point, inconsistencies may still be found in our collected data, due to entries collected from both PubMed and PubMed Central. Such entries refer to the same article, but there are small differences in the metadata provided between PubMed and PubMed Central (e.g., title, publication year). Our workflow includes steps that deduplicate and clean most of these cases; however, a small number remain that requires manual curation. These correspond to less than 0.01% of the data[9] and are manually examined and homogenized. After this

[7] https://www.ncbi.nlm.nih.gov/books/NBK25500/

[8] eSummary is part of NCBI's eTools suite. We also use the eLink tool from the same suite to gather article citation data and build the citation network required by our citation-based analysis (see Section 2.2).

[9] For reference, version 63 of our data set had about 80 such records.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

postprocessing step, the resulting data set contains one entry for each distinct article. Each entry contains the pmid, the DOI, the pmcid, and the publication year of the corresponding article. This collected information is the minimum required for the calculation of the selected impact measures.

The final BIP4COVID19 data set contains fewer records than the original individual data sources. This is due to two reasons: First, many of the recorded data in CORD-19 do not have a pmid or pmcid identifier. Because we need these identifiers to create a citation network on which to calculate the impact measures provided (see Section 2.2), these records are not included in BIP4COVID19. Second, a number of records overlap between CORD-19 and LitCovid. As we record each article only once, the total number of records in BIP4COVID19 is by default less than the sum of the records in the two source data sets. Indicatively, in the 63rd release of BIP4-COVID19, we record 173,717 articles found in both CORD-19 and LitCovid. This corresponds to about 98% of the papers we collect from LitCovid and about 44.5% of papers we collect from CORD-19. On the other hand, version 63 of BIP4COVID19 contains 215,811 distinct articles found only in CORD-19 and 3,503 articles found only in LitCovid. Hence, combining both of the two sources allows for the collection of a most comprehensive data set.

2.2. Calculation of Citation-Based Measures

A prerequisite for calculating the citation-based impact measures of the collected articles is the compilation of their citation network (i.e., the network that has articles as nodes and citations between them as directed edges). The citations of the articles required to construct this network are not provided by the original data sources, and hence are gathered using NCBI's eLink tool. For a given article, the tool returns the identifiers (pmids/pmcids) of all articles that cite, or are cited by it. Using this information for all the articles in our data set we are able to build a citation network for coronavirus-related articles on which our citation-based measures are calculated.

During the design of our data set, we chose to include the scores of a variety of citation-based impact measures. This choice was based on the following arguments: First of all, any impact measure has inherent drawbacks. For example, citation counts are unable to differentiate citations according to the importance of the citing articles, making them vulnerable to citation cartels and other related malpractices. Moreover, relying on a single measure to quantify scientific impact is an oversimplification, as scientific impact can be defined in many different ways (Bollen, Van de Sompel et al., 2009). Hence, it is essential to study impact from a broader set of perspectives. In addition, taking into consideration that unique and widely applied assessment measures are likely to be abused (according to a generalization of Goodhart's law), providing a variety of impact measures (each capturing a different perspective) can work as countermeasures to reduce the effects of such exploits.

Based on the previous rationale, four citation-based impact measures are calculated for each article using the constructed citation network:

- the *Citation Count*, which sums all citations received by the article.
- the PageRank (Page et al., 1999) scores, based on the well-known web page ranking algorithm, which has been useful in ranking papers in citation networks (e.g., Chen, Xie et al. (2007)).
- the RAM (Ghosh et al., 2011) scores, which are weighted citation counts, where recent citations are considered as more important.
- the AttRank (Kanellos et al., 2021) scores, which are based on a PageRank variant that focuses on overcoming PageRank's bias against recent papers.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

The Citation Count was selected because it is the most widely used bibliometric measure; the other three measures were selected based on the results of a recent experimental study (Kanellos et al., 2019) and subsequent work (Kanellos et al., 2021), which found them to perform best in capturing the overall and the current citation-based impact of an article (i.e., its "influence" and its "popularity"), respectively. According to the definitions provided in Kanellos et al. (2019), the current popularity of an article is reflected in the citations it receives in the near future[10], while its influence refers to the overall attention it has received diachronically. In other words, the current impact ("popularity") implies that a particular article is at the present focus of the scientific community, while the overall impact ("influence") implies that a paper is well established in its field, but not necessarily at the current center of attention of the scientific community. The relative advantages of our selected impact measures can be summarized as follows: PageRank evaluates the overall impact of articles by differentiating their citations based on the importance of the articles making them. However, it is biased against recent articles that have not accumulated many citations yet, but may be the current focus of the research community. RAM alleviates this issue by considering recent citations as being more important, while AttRank modifies PageRank so that it promotes recently published or recently cited papers, an approach which has been found to perform more effectively in ranking papers based on their current impact (Kanellos et al., 2021). Based on the previous discussion, and the results in Kanellos et al. (2019), we consider the Citation Counts, as well as the PageRank score, as indicators that capture influence, while we consider the RAM and AttRank scores as indicators of popularity.

2.3. Counting Recent Tweets

Following the rationale of providing as many measures as possible in an attempt to capture diverse aspects of each article's impact (see also Section 2.2), we also include in our data set a Twitter-based altmetric that quantifies the attention each article received in this social networking platform during a recent time period[11]. In other words, our intention was to include an altmetric that could capture (to an extent) the current hype of each article in Twitter, an aspect of social media attention that we believe may give useful insights about the current trends in the coronavirus-related literature when combined with the other calculated measures[12]. The calculation of this measure is essentially equivalent to counting the number of appearances of article-related URLs in the examined COVID-19-related tweets. In the following paragraphs we elaborate on the respective workflow.

The *COVID-19-TweetIDs* [13] data set (Chen, Lerman, & Ferrara, n.d.) is used for the collection of COVID-19-relevant tweets. This data set contains a collection of tweet IDs, each of them published by one of nine predetermined Twitter accounts (e.g., @WHO) and containing at least one out of 71 predefined coronavirus-related keywords (e.g., "Coronavirus," "covid19," etc). At the time of writing, a subset of this data set containing tweets posted from September 4 to September 10, 2021 (20,852,233 unique tweet IDs) has been integrated in

[10] The length of the considered future period is a problem parameter, which depends on various factors, such as the publication life cycle of a particular discipline (for manuscript writing, peer-review, publication).

[11] Currently, our policy is to collect the Twitter posts that mention each article and that have been published during the last available week for which data are included in the COVID-19-TweetIDs data set at the time of processing (see below).

[12] Of course, this is not a unique option: One could also calculate other types of Twitter-based measures such as the overall tweet count (which captures a different social media attention aspect). The extension of our data set with additional impact measures is in our future plans, because we strongly believe that a wide range of measures may provide a more comprehensive picture of article impact.

[13] https://github.com/echen102/COVID-19-TweetIDs

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

BIP4COVID19. The corresponding Tweet objects were collected using the Twitter API. The result was a collection of 19,567,301 tweet objects. The difference between the number of IDs and collected objects is due to circumstances, such as the deletion of tweets in the meantime, that make some tweets impossible to retrieve.

To find those tweets that are related to the articles in our database (i.e., to find the subset of the approximately 19 million tweets examined that which do mention articles in our collection), we rely on the URLs of the articles in doi.org, PubMed, and PubMed Central (PMC). These URLs are easily produced based on the corresponding identifiers. In addition, when possible, the corresponding page in the publisher's website is also retrieved based on the doi.org redirection. After the collection of the URLs of all articles, the number of appearances of the URLs related to each one (which corresponds to the respective recent tweet count) is produced. However, as the Twitter API returns either shortened or not fully expanded URLs, the fully expanded URLs are collected using the unshrtn[14] library.

3. RESULTS

By applying the processes outlined in Section 2, we compile the BIP4COVID19 data set, the characteristics and format of which are described in the following subsections. We further demonstrate the search interface developed on top of our data.

3.1. Data Set Details

The BIP4COVID19 data set, produced by the previously described workflow, is openly available on Zenodo (Vergoulis et al., 2021b) under the Creative Commons Attribution 4.0 International license. Its first version was released on March 21, 2020. At the time of writing, the 63rd release of this data set (v63) is available, counting 393,031 records. Of these, 353,976 have an identifier in PubMed, 295,705 have an identifier in PMC, and 379,842 have an associated DOI. All publications included were published from 1856 to 2021. The distribution of publication years of the articles recorded in the data set is illustrated in Figure 2. A total of 294,619 of these articles were published in 2020–2021 (i.e., after the start of the coronavirus pandemic outbreak), while 93,160 were published from 1825–2019. Moreover, the numbers of articles per venue for the top 30 venues (in terms of relevant articles published) are presented in Figure 3.

The BIP4COVID19 data set is comprised of five files in tab-separated (TSV) format. The files contain identical information, but in each of them the records are ordered based on a different impact measure. At the time of writing, the BIP4COVID19 data set has already been downloaded more than 23,000 times (according to the metrics provided from Zenodo for all its versions), and it appears as the most viewed data set (among 516 data sets) in the "Coronavirus Disease Research Community—COVID-19" collection in Zenodo.

Finally, we follow an approach of regular batch updates, similarly to our data sources (CORD-19 and LitCovid)[15]. This policy of frequent batch updates is essential because article additions in most scholarly data sources come in bulk, by default, and citation-based measures cannot be easily calculated incrementally (i.e., the whole citation network must be reprocessed each time). This is not an important limitation, though, because due to our efficient citation network analysis workflow[16], it is possible to apply a policy of very frequent updates,

[14] https://github.com/docnow/unshrtn

[15] Most updates so far have been carried out on a weekly basis.

[16] Our implementations require less than three hours on a typical computer to calculate all the scores for the whole data set.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

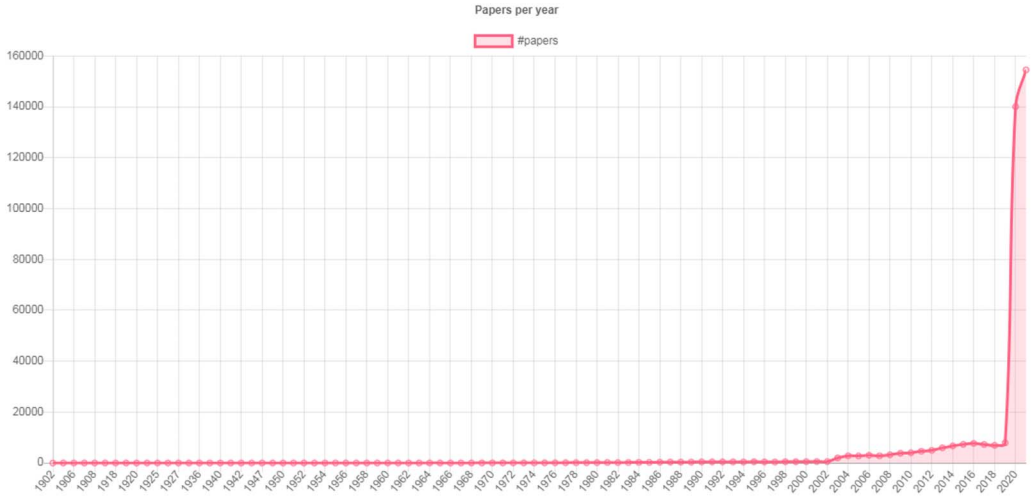

Figure 2. COVID-19 related articles per year. Papers published before 1902 are excluded for presentation reasons.

ensuring that the latest articles will be included in our data set shortly after their publication. In addition, the inclusion of a small number of additional articles for each update is not expected to have a significant effect on the calculated impact scores of the previously included articles, in any case.

3.2. Web-Based Search Engine

A web interface has been developed on top of the BIP4COVID19 data. Its aim is to facilitate the exploration of COVID-19-related literature. Apart from a basic keyword search functionality, which is common to other search services for COVID-19-related articles (see Section 5.3), the option to order articles according to different impact measures (or a combination

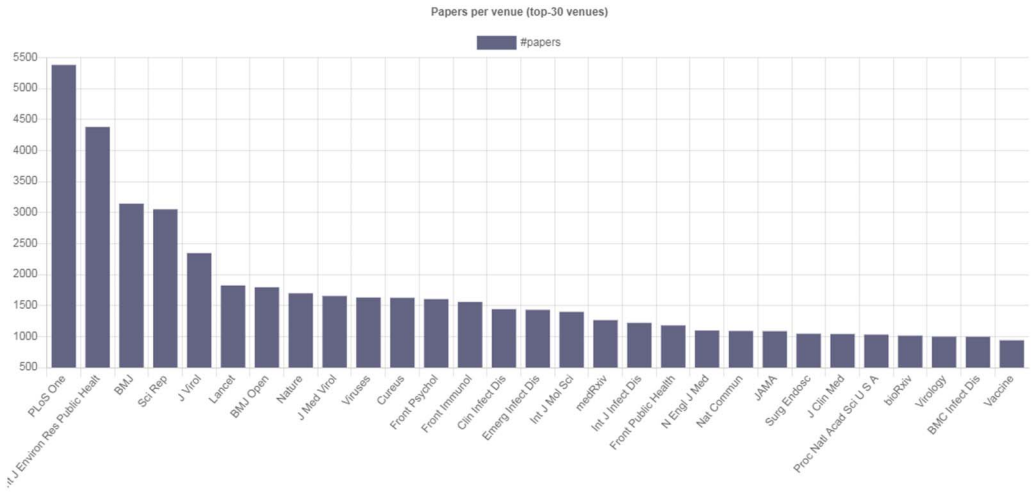

Figure 3. Top 30 venues in terms of published COVID-19-related articles.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

of them with keyword relevance[17]) is provided. This is expected to be useful, as users can better prioritize their reading based on their needs. For example, a user that wants to delve into the background knowledge about a particular COVID-19-related subtopic could select to order the articles based on their influence[18]. On the other hand, another user that needs to get an overview of the latest trends in the same topic could select to order the articles based on their popularity. It is worth mentioning that, to avoid user confusion, the web interface incorporates only one influence measure (PageRank) and one popularity measure (AttRank); the selection was based on their ranking effectiveness based on previous studies (Kanellos et al., 2019, 2021). Finally, the web user interface takes advantage of the Mendeley API[19] to also display the current number of Mendeley readers, as an extra altmetric measure.

The information shown to users, per publication, includes its title, venue, year, and the source data set where it was found (see Figure 4). Moreover, each result is accompanied by color-coded icons that denote the publication's importance based on each calculated impact measure. In this way, users can easily get a quick insight into the different impact aspects of each article. The tooltips of these icons provide the exact scores for each measure. Each publication title functions as a link to the corresponding article's entry in its publisher's website, or to PubMed. Additionally, if an article has been retracted, the corresponding entry in the results page is flagged accordingly (we collect this information from NCBI's esummary tool[20], which provides summary information for publications, including their possible retraction status, given their identifier in PubMed). Moreover, our interface allows users to download the results of a particular search query by clicking on the download button to the right of the search bar. Query results are downloaded in tab-separated format, ordered by the selected impact measure and providing a file which contains the pmid, pmcid, DOI, title, venue, and publication year of each search result. Finally, a page containing various summary statistics about the BIP4COVID19 data is provided. This page contains various charts that visualize, for example, the number of articles per year, or the number of articles that have substantial impact based on each of the provided impact measures, per year.

To assess the usefulness of BIP4COVID19's search engine and the calculated impact indicators, we conducted a preliminary user evaluation study based on six participants who are involved in research concerning the novel coronavirus, or have a general scientific background from life sciences at the postgraduate level. Each participant provided us with up to three keyword search queries (many of them were multikeyword ones), presented in Table A1. We then provided a set of articles for each of the provided queries. To gather the results of each query, we performed the provided keyword search on the search interface of BIP4COVID19 and collected the shuffled union of the top 5 results according to popularity and influence (with the combination with keyword relevance enabled)[21]. We then asked them

[17] To combine impact scores with keyword relevance we use of the CombMIN rank aggregator (Fox & Shaw, 1994), applied on normalized values of relevance and impact scores on search results. Relevance scores are provided by a MySQL fulltext index on titles and abstracts. Impact scores, however, are further processed, as using them as is would result almost in an impact-based ranking only, due to their power law distribution and value range (i.e., in (0, 1)). Hence, we use their square root to increase their value.

[18] For an introduction to the notions of "influence" and "popularity" refer to Section 2.2.

[19] Mendeley API, https://dev.mendeley.com/

[20] https://www.ncbi.nlm.nih.gov/books/NBK25499/

[21] To focus on current research trends, for the purposes of this experiment we only kept results that are found in LitCovid, because this curated data set only contains articles published after the pandemic's outbreak; in most cases no papers found only in CORD-19 were returned in the top 5 set in any case. Note that this exclusion of CORD-19 papers does not concern the published data set, but only this experiment.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

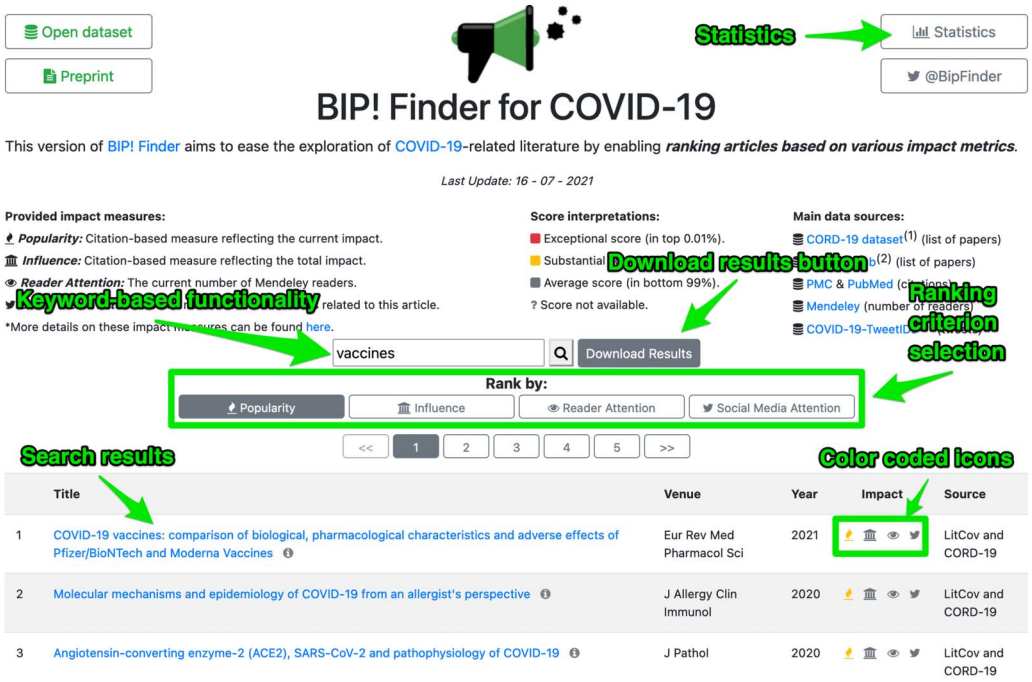

Figure 4. A screenshot of the BIP4COVID19 web interface.

to grade each article from 0–2, with 0 denoting that the article is irrelevant, 1 denoting it is likely relevant and 2 denoting it is extremely relevant. Finally, we calculated the Discounted Cumulative Gain for the top 5 results (DCG@5) produced by BIP4COVID19 for influence and popularity, when keyword relevance is taken into account. The DCG is a common measure of ranking quality for search engines, which scores search results based on their perceived usefulness (i.e., the grades assigned by the experts), and discounts the scores based on rank position. In addition to the DCG@5, we count the number of results in the top 5 that were not considered irrelevant (i.e., results rated "Likely Relevant," or "Extremely Relevant"). We present the results calculated from the ratings received in Table 1.

Overall, users report useful results in the top 5 articles returned for both of the examined ranking criteria. Note that the maximum DCG@5 for the relevance scores provided is 5.897 (i.e., all results are rated as "extremely relevant"), which is a value achieved in two queries when ranking by popularity and two when ranking by influence, respectively. The top 5 results are all considered relevant for 7/16 queries using popularity and 9/16 queries using influence. No query returned fewer than 3/5 relevant results.

Further note that, based on the calculated DCG, half of the queries benefit more from ranking based on popularity while the other half benefit more from ranking based on influence. This can be seen as further evidence regarding the usefulness of providing different result ranking options, as apart from a different result order based on scientific impact, different sets of relevant articles may be returned in the top results (see the last column in Table 1).

At this point, we should note that we have applied a simple mechanism to rank search results taking into account both impact and keyword relevance (see also the relevant discussion above). Additional in-depth tuning procedures to rank results may further increase the usefulness of our engine, but this optimization is out of the scope of the current work.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

Table 1. DCG@5 and number of relevant results per expert query,when ranking by influence and popularity, taking keyword relevance into account

Researcher	Query	Popularity DCG@5	Relevant results	Influence DCG@5	Relevant results	Overlap
R1	Q1	2.448	4	2.335	4	4
R2	Q1	5.266	5	5.510	5	3
	Q2	4.397	5	4.266	5	4
	Q3	2.510	3	2.149	3	4
R3	Q1	2.818	4	3.379	5	1
	Q2	4.897	4	4.135	4	3
	Q3	3.397	4	2.193	3	2
R4	Q1	2.861	3	2.773	3	5
	Q2	5.510	5	5.897	5	3
	Q3	5.897	5	5.897	5	4
R5	Q1	3.067	4	3.405	4	3
	Q2	5.123	4	5.897	5	3
	Q3	5.897	5	5.897	5	4
R6	Q1	4.966	5	4.579	5	4
	Q2	3.379	5	3.879	5	4
	Q3	2.274	3	3.405	4	3

4. DISCUSSION

In this section, we present analytics related to the BIP4COVID19 data set, elaborate on processes followed to ensure its data integrity, consider our analysis' limitations, and discuss our data set's possible usages.

4.1. Data Set Analytics

In Figures 5–7 we present various analytics on the BIP4COVID19 data set. Figure 5 presents the timeline of the data set size (in thousands of papers recorded). Since its first versions the data set has shown a steady linear increase in the number of papers recorded, which follows the corresponding increase in the size of the CORD-19 and LitCovid data sets.

In Figure 6 we examine the overall correlation of the influence- and popularity-based rankings. The rankings examined are based on PageRank for influence and RAM for popularity.[22] We measure correlation using Spearman's ρ, which ranges in $[-1, 1]$, where $\rho = 1$ denotes perfect correlation and $\rho = -1$ denotes perfect inverse correlation. From Figure 6, we observe that the popularity- and influence-based rankings are relatively highly correlated ($\rho \sim 0.7$ throughout the data set's timeline). While popularity and influence are highly correlated, they do not, however, correlate perfectly.

[22] We chose these measures because citation counts and AttRank scores have only been included in the most recent versions.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

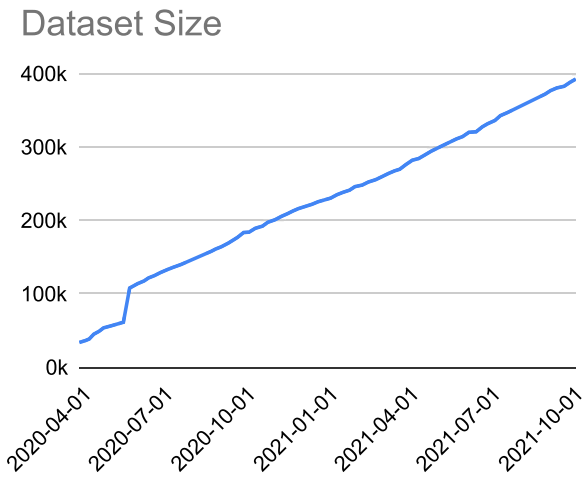

Figure 5. BIP4COVID19 data set size timeline.

In Figure 7 we present the timeline of the intersection of the top 100 ranked papers based on influence and popularity (i.e., the proportion of overlap between the top 100 set of each of the two rankings). We observe that for all versions of the BIP4COVID19 data set the overlap of top-ranking papers is less than 65%, while for early versions there was no overlap. These results complement Figure 6 in highlighting that popularity- and influence-based rankings, while correlated, do indeed constitute rankings with distinct semantics, which correspond to different types of impact.

We note that the popularity- and influence-based rank scores follow a typical power law distribution, which is commonly observed for scores based on citation network centrality variants (i.e., PageRank for influence and AttRank, a PageRank variant for popularity). To showcase, however, how popularity and influence differ in the "head" of their corresponding distributions (i.e., the papers they consider as having the highest impact), we illustrate in Figure 8 the number of top 200 ranking papers based on their publication year. While both distributions have the majority of their papers published in 2020 (an expected outcome, because the majority of overall papers on the subject were published in that year), the popularity-based ranking includes only papers published in 2020–2021. The influence-based

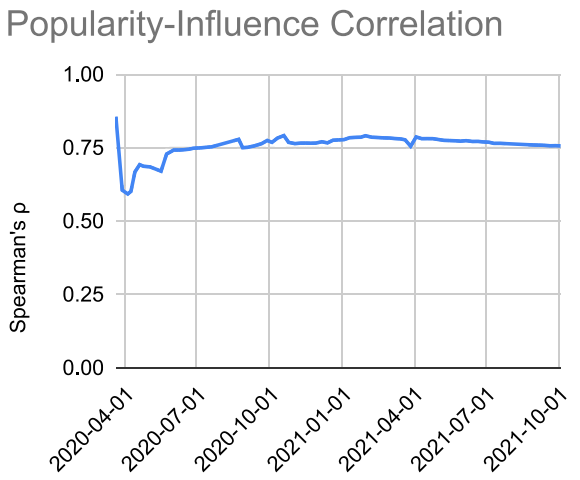

Figure 6. Popularity-Influence ranking correlation timeline.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

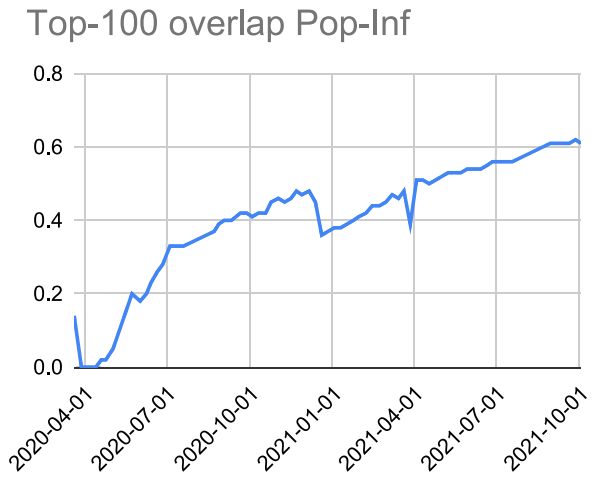

Figure 7. Timeline of top 100 ranking papers by Influence and Popularity.

ranking, on the other hand, includes older papers, published as far back as 1965. Note further that the popularity based ranking includes 21 papers published in 2021 in its top results, while the influence-based ranking includes only seven. This is practical evidence of our claims that influence-based rankings favor well-established but older papers, while popularity-based rankings favor papers currently at the center of attention of the scientific community.

4.2. Ensuring Data Integrity

To ensure the proper integration and cleaning of the CORD-19 and LitCovid data sets, we rely on NCBI's eTool suite. In particular, we collect pmids and pmcids from both data sets and use them as queries to gather each article's metadata. After cleaning the article title (e.g., removing special characters) we automatically identify duplicates by comparing each record's complete

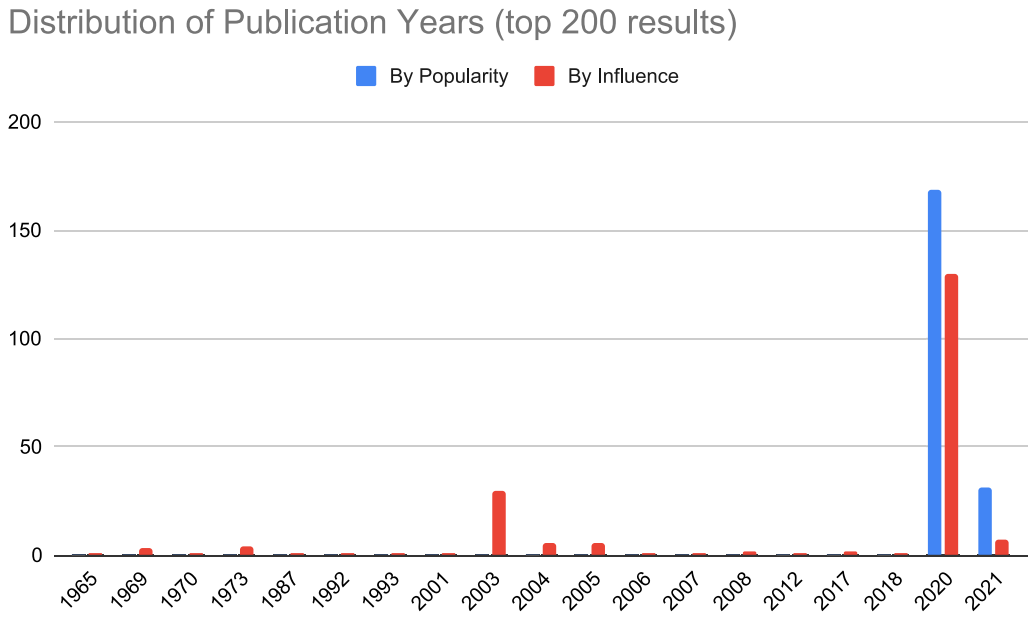

Figure 8. Distribution of top ranking papers by publication year.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

content and eliminate them. Finally, manual inspection is performed to produce the correct metadata for a limited number of duplicates that remain (e.g., duplicate records containing the title of the same publication in two different languages).

Further, to guarantee the correctness of the compiled citation graph we apply the following procedures. After gathering all citing–cited records using NCBI's eTools, those that include identifiers not found in the source data are removed. Because many citing–cited pairs may have been found both with pmids and pmcids, the resulting data may still contain duplicate records. These records are removed, after mapping all pmids/pmcids to custom identifiers, with pmid–pmcid pairs that refer to the same article being mapped to the same identifier. The final resulting citation graph is based on these mapped identifiers. As an extra cleaning step, any links in the graph that denote citations to articles published at a later time than the citing article are removed[23].

To ensure that we retrieve a set of tweets about each article that is as comprehensive as possible, we collect not only the URLs in doi.org, PubMed, and PubMed Central (PMC), but also the URL to the article on its publisher's website, where possible. These latter URLs are very important, as they are widely used in tweets. To collect them we utilize doi.org redirections. To avoid incorrect tweet counts due to duplicate tweets, we used a simple deduplication process after the Tweet object retrieval. Moreover, the use of the unshrtn library to expand the short URLs from tweet texts ensures that our measurements derive from all available URL instances of each publication record, no matter how they were shortened by users or Twitter.

4.3. Limitations

The following limitations should be taken into consideration with respect to the data: While we make an effort to include as many articles as possible, there are many cases where our source data do not provide any pmids or pmcids. As a consequence, no data for these articles are collected and they are not included in the BIP4COVID19 data set. Furthermore, with respect to the calculated impact scores, it should be noted that the citation analysis we conduct is applied on the citation graph formed by citations *from* and *to* collected publications only, that is, our analyses are not based on PubMed's complete citation graph, but on a COVID-19-related subgraph. Consequently, the relative scores of publications may differ from those calculated on the complete PubMed data. Finally, regarding the tweet-based analysis, because our data come from the COVID-19-TweetIDs data set which only tracks tweets from a predefined set of accounts and which is based on a particular set of COVD-19-related keywords, the measured number of tweets is only based on a subset of the complete COVID-19-related tweets.

4.4. Usage Notes

Our data are available in files following the TSV format, allowing easy import to various database management systems. They can be conveniently opened and edited by any text editor or spreadsheet software. We have been regularly updating the BIP4COVID19 data since March 2020, and we plan to continue providing regular updates, incorporating any additions and changes from our source data sets. Additionally, we plan to incorporate any further sources

[23] Such references to future articles are often observed in citation data. Hence, a common practice is to remove them (e.g., Ghosh et al., 2011).

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

on coronavirus-related literature that may be released and which will index the literature based on pmids and/or pmcids.

The contents of the BIP4COVID19 data set may be used to support multiple interesting applications and added-value services. For instance, the calculated scores for each impact measure could be used to rank articles based on their impact to help researchers prioritize their reading. In fact, we used our data to implement such a demo as previously described (see Section 3.2), while a similar use case has been implemented in the "Publications" section of the Greek COVID-19 data portal[24], where popularity and influence scores from BIP4-COVID19 are used to rank COVID-19-related publications that involve scientists affiliated to a Greek institution. Additionally, the impact scores may be useful when combined with topic modeling approaches for monitoring trends in the impact of particular subtopics of the coronavirus-related literature. Furthermore, the calculated scores could be utilized as weighted features in machine learning applications that apply data mining on publications related to coronavirus, or in learning-to-rank scenarios to optimize results of academic search engines based on multiple factors (i.e., combinations of traditional keyword relevance with publication impact). Finally, the set of impact measures provided may be useful as a basis for various scientometric analyses, e.g., for studying the relations between the various impact measures, studying the patterns connecting them, or gaining other insights hidden in these data (if combined with scholarly knowledge graphs and processing techniques from the fields of Text Mining and/or Semantic Web).

It should be highlighted that the previous examples are only indicative exploitation scenarios for which the BIP4COVID19 data set could serve; providing an exhaustive list of potential applications and added-value services that could leverage our data set is out of scope of the current work. As a final remark, it is worth mentioning that the large number of downloads of our data set in Zenodo (> 20,000) is indicative of the potential interest for developing such applications and services.

5. RELATED WORK

Since the outbreak of the pandemic, a number of studies, data sets, and search interfaces on COVID-19 and COVID-19-related literature have been published. Although BIP4COVID19 includes a keyword-search-based interface, the main aim is to publish a data set of impact scores for COVID-19-related literature, with different potential applications and uses in research scenarios. In the following we outline the related work.

5.1. Studies of the COVID-19 Literature

During recent months, various works studying COVID-19-related scientific literature have been published. Kousha and Thelwall (2020) evaluated the coverage of various scholarly data sources regarding COVID-19-related articles during the time period between March 21, 2020 and April 18, 2020. They also report how many articles received a significant number of citations or a substantial mass and social media attention during the same period. An interesting finding of their analysis is that, for COVID-19 papers, the convergence between citation counts and social media attention appears to be high, something that has not been observed for other fields of study.

Horbach (2020) investigated whether medical journals have managed to accelerate their publication processes for COVID-19-related articles. He studied the duration time of the

[24] https://covid19dataportal.gr/

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

publication process of 14 medical journals before and after the start of the COVID-19 pandemic and found that the time between submission and publication for COVID-19-related articles was 49% shorter, on average, than the corresponding times of other articles during the same period or in the past. In addition, in a subsequent work (Horbach, 2021), he studied review reports and editorial decision letters published during or before the COVID-19 pandemic, identifying indications that reviewers and journal editors implicitly and explicitly may use different quality criteria to assess COVID-19-related and nonrelated articles.

Thelwall (2020) used Mendeley reader counts to reveal that studies analyzing SARS and MERS data to provide useful intuition about COVID-19 have gathered more academic interest than primary studies of SARS and MERS, on average. In another study (Lucas-Dominguez, Alonso-Arroyo et al., 2021), the authors moved the focal point to the availability of COVID-19-related data sets. They found that although many journals released a large number of publications in open access, the deposit of supplementary material and data in open repositories was very limited, revealing that data sharing is not common even during health emergencies, such as the COVID-19 pandemic. Finally, Colavizza (2020) focused on COVID-19-related Wikipedia pages and investigated the pace with which the results of new research are incorporated into them and which proportion of the relevant literature is covered in Wikipedia pages.

Although we provide a small set of statistics about the recorded COVID-19-related articles (see Section 4.1), the main focus of this work is to provide an open, frequently updated data set of impact measures for the relevant literature. Nevertheless, the statistics we provide are complementary to those given by the aforementioned studies.

5.2. Available Data Sets

A number of openly available data sets have been released and actively maintained to facilitate the COVID-19-related literature exploration and research. The most popular example is the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020a). CORD-19 is a frequently updated data set of publications, which, apart from COVID-19 research articles, contains current and past research related to the coronavirus family of viruses. It combines articles from different sources (PMC, arXiv, biorXiv, medRxiv, and the World Health Organization) and offers deduplicated metadata and a full-text corpus for a large number of articles. In another line of work, Rohatgi, Karishma et al. (2020) further enhanced a version of CORD-19 with additional abstracts, tables, and figures also providing *COVIDSeer*, a keyword search engine on top of the enhanced data set (see Section 5.3). Another very popular example is LitCovid (Chen et al., 2020a, 2020b), an openly available curated literature resource of COVID-19-related articles, further categorized by research topics and geographic locations for improved access. Moreover, there are various attempts to create of knowledge graphs that capture information contained in the COVID-19-related literature (Wang, Li et al., 2020b; Wise, Ioannidis et al., 2020). Finally, there are also various large-scale data sets collecting COVID-19-related tweets (Banda, Tekumalla et al., 2020; Chen et al., n.d.), while Twitter itself released its own API endpoint to provide real-time access to COVID-19 related tweets.

Our own data set, BIP4COVID19, focuses on providing a range of different impact measures on COVID-19-related publications. This line of work aims at providing a complete picture about a publication's impact, in line with the best practices in research assessment. In fact, our work is a COVID-19-specific version of BIP! DB (Vergoulis, Kanellos et al., 2021a). To the best of our knowledge, BIP4COVID19 is the only data set of this scale to provide a fair range of different impact measures on COVID-19-related literature.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

5.3. COVID-19 Search Services

Following the beginning of the COVID-19 outbreak, due to the extremely large interest in COVID-19-related scientific articles, various scholarly search engines, tailor-made for the COVID-19-related literature, have been developed. First of all, the teams that developed and maintain the major literature, data sets provide their own search engines: Allen Institute for AI released *SciSight* (Hope, Portenoy et al., 2020), a tool for exploring the CORD-19 data, while *LitCovid* (Chen et al., 2020a, 2020b) provides a search engine featuring basic functionalities (e.g., keyword search, facets). *COVIDScholar* (Trewartha, Dagdelen et al., 2020) is another search engine that uses natural language processing (NLP) to power search on a set of research papers related to COVID-19. *iSearch COVID-19 Portfolio*[25] is a tool that provides search functionality and faceting for COVID-19 articles. *KDCovid*[26] is a tool that retrieves papers by measuring similarity of queries to sentences in the full text of papers in the CORD-19 corpus using a similarity metric derived from BioSentVec. *COVIDSeer* (Rohatgi et al., 2020) is another tool that was built on top of a CORD-19 extension and which provides keyword filtering and paper recommendation functionalities. *Vilokana* (Panja, Maan, & James, 2020) is a semantic search engine on top of the CORD-19 data set that exploits a set of modified TFIDF features and cosine similarity with ontology maps to provide semantic search functionalities. *CAiRE-COVID* (Su, Xu et al., 2020) is a tool that answers priority questions and summarizing salient question-related information. It combines information extraction with state-of-the-art QA and query-focused multidocument summarization techniques, selecting and highlighting evidence snippets from existing literature based on a given query. Finally, *AWS CORD-19 Search* (*ACS*) (Bhatia, Arumae et al., 2020) is a COVID-19 specific, ML-based search engine that supports natural language based searches providing document ranking, passage ranking, question answering, and topic classification functionalities.

To the best of our knowledge, apart from a couple of engines that display citation count scores for each article in the result list, or use them to rank the results, most of them do not leverage impact measures. Incorporating the impact measures, which are openly available through the BIP4COVID19 data set, may unlock various opportunities towards providing advanced search, filtering, monitoring, and other valuable features. Recognizing these opportunities, we have developed our own prototype search engine (see Section 3.2). It is worth mentioning that this tool is a COVID-19-specific variant of *BIP! Finder* (Vergoulis, Chatzopoulos et al., 2019), which allows users to choose the impact measure based on which their search results will be ordered. Currently, the BIP4COVID19 search engine allows ordering based on a combination of keyword relevance with popularity, influence, Mendeley reader counts, or recent tweet counts. Further, it allows users to download the search results of particular queries, ordered based on the selected impact measure.

6. CONCLUSION

We presented BIP4COVID19, an openly available data set, providing impact scores for coronavirus-related scientific publications. Our data set is potentially useful for many applications such as assisting researchers to prioritize their reading of COVID-19-related papers, providing study material for scientometricians, etc. We have additionally built on our data set, providing a web-based article search engine which exploits the calculated impact measures to offer relevant functionalities, such as impact-based ranking of the keyword-search results. As future work, additional impact measures can be integrated into our data set; for instance

[25] https://icite.od.nih.gov/covid19/search/
[26] https://kdcovid.nl/

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

additional altmetrics, such as the overall tweet count, or paper download statistics, could be included in BIP4COVID19. Furthermore, our web interface can be expanded to support faceted search, facilitating users to navigate quickly in the result set. Last but not least, BIP4COVID19 can be interlinked (e.g., by providing additional article identifiers) with existing knowledge graphs (e.g., the Open Research Knowledge Graph (Jaradeh, Oelen et al., 2019) and the Microsoft Academic Knowledge Graph (Färber, 2019)) to allow for seamless inclusion of our provided impact indicators into other data sets.

ACKNOWLEDGMENTS

We would like to thank Fotis Psomopoulos, Stamatia Laidou, Nikolaos Pechlivanis, Iro (Argyro) Sfoungari, Rania Theologi, and Ioanna Kiourti for their participation in the user study of the BIP4COVID19 search engine.

AUTHOR CONTRIBUTIONS

Thanasis Vergoulis: Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Ilias Kanellos: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Serafeim Chatzopoulos: Data curation, Investigation, Software, Validation, Writing—original draft, Writing—review & editing. Danae Pla Karidi: Data curation, Investigation, Software, Validation, Writing—original draft, Writing—review & editing. Theodore Dalamagas: Supervision, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

We acknowledge support of this work by the project "Moving from Big Data Management to Data Science" (MIS 5002437/3) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure," funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and cofinanced by Greece and the European Union (European Regional Development Fund).

DATA AVAILABILITY

BIP4COVID19 data are openly available on Zenodo (https://zenodo.org/record/4555117) under a CC BY 4.0 license.

REFERENCES

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., ... Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration. *ArXiv*, arXiv:2004.03688v1. https://doi.org/10.3390/epidemiologia2030024

Bhatia, P., Arumae, K., Pourdamghani, N., Deshpande, S., Snively, B., ... Kass-Hout, T. A. (2020). AWS CORD-19 search: A scientific literature search engine for COVID-19. *arXiv*, arXiv.2007.09186. https://arxiv.org/abs/2007.09186

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS ONE*, *4*(6), e6022. https://doi.org/10.1371/journal.pone.0006022, PubMed: 19562078

Chen, E., Lerman, K., & Ferrara, E. (n.d.). Covid-19: The first public coronavirus Twitter dataset. arxiv 2020. *ArXiv*, arXiv:2003.07372.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8–15. https://doi.org/10.1016/j.joi.2006.06.001

Chen, Q., Allot, A., & Lu, Z. (2020a). Keep up with the latest coronavirus research. *Nature*, *579*(7798), 193–193. https://doi.org/10.1038/d41586-020-00694-1, PubMed: 32157233

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

Chen, Q., Allot, A., & Lu, Z. (2020b). LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, *49*(D1), D1534–D1540. https://doi.org/10.1093/nar/gkaa952, PubMed: 33166392

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

Färber, M. (2019). The Microsoft Academic Knowledge Graph: A linked data source with 8 billion triples of scholarly data. *International Semantic Web Conference* (pp. 113–129). https://doi.org/10.1007/978-3-030-30796-7_8

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. *The Second Text REtrieval Conference (TREC-2)* (pp. 243–252).

Ghosh, R., Kuo, T.-T., Hsu, C.-N., Lin, S.-D., & Lerman, K. (2011). Time-aware ranking in dynamic citation networks. *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 373–380). https://doi.org/10.1109/ICDMW.2011.183

Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., ... West, J. (2020). SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020* (pp. 135–143). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.18

Horbach, S. P. J M (2020). Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, *1*(3), 1056–1067. https://doi.org/10.1162/qss_a_00076

Horbach, S. P. J M (2021). No time for that now! Qualitative changes in manuscript peer review during the Covid-19 pandemic. *Research Evaluation*, *30*(3), 231–239. https://doi.org/10.1093/reseval/rvaa037

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., ... Auer, S. (2019). Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 243–246). https://doi.org/10.1145/3360901.3364435

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., & Vassiliou, Y. (2019). Impact-based ranking of scientific publications: A survey and experimental evaluation. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1567–1584. https://doi.org/10.1109/TKDE.2019.2941206

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., & Vassiliou, Y. (2021). Ranking papers by their short-term scientific impact. *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 1997–2002). https://doi.org/10.1109/ICDE51399.2021.00190

Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. *Quantitative Science Studies*, *1*(3), 1068–1091. https://doi.org/10.1162/qss_a_00066

Lucas-Dominguez, R., Alonso-Arroyo, A., Vidal-Infer, A., & Aleixandre-Benavent, R. (2021). The sharing of research data facing the COVID-19 pandemic. *Scientometrics*, *126*(6), 4975–4990. https://doi.org/10.1007/s11192-021-03971-6, PubMed: 33935332

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Panja, S., Maan, A. K., & James, A. P. (2020). Vilokana—Lightweight COVID19 document analysis. *63rd IEEE International Midwest Symposium on Circuits and Systems* (pp. 500–504). https://doi.org/10.1109/MWSCAS48704.2020.9184598

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

Rohatgi, S., Karishma, Z., Chhay, J., Keesara, S. R. R, Wu, J., ... Giles, C. L. (2020). COVIDSeer: Extending the CORD-19 data set. *ACM Symposium on Document Engineering 2020* (pp. 21:1–21:4). https://doi.org/10.1145/3395027.3419597

Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., & Fung, P. (2020). CAIRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. *Proceedings of the 1st Workshop on NLP for COVID-19*. https://doi.org/10.18653/v1/2020.nlpcovid19-2.14

Thelwall, M. (2020). Coronavirus research before 2020 is more relevant than ever, especially when interpreted for COVID-19. *Quantitative Science Studies*, *1*(4), 1381–1395. https://doi.org/10.1162/qss_a_00083

Trewartha, A., Dagdelen, J., Huo, H., Cruse, K., Wang, Z., ... Ceder, G. (2020). COVIDScholar: An automated COVID-19 research aggregation and analysis platform. *arXiv*, arXiv.2012.03891. https://arxiv.org/abs/2012.03891

Vergoulis, T., Chatzopoulos, S., Kanellos, I., Deligiannis, P., Tryfonopoulos, C., & Dalamagas, T. (2019). BIP! Finder: Facilitating scientific literature search by exploiting impact-based ranking. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2937–2940). https://doi.org/10.1145/3357384.3357850

Vergoulis, T., Kanellos, I., Atzori, C., Mannocci, A., Chatzopoulos, S., ... Manghi, P. (2021a). BIP! DB: A dataset of impact measures for scientific publications. *arXiv*, arXiv:2101.12001. https://doi.org/10.1145/3442442.3451369

Vergoulis, T., Kanellos, I., Chatzopoulos, S., Pla Karidi, D., & Dalamagas, T. (2021b). *BIP4COVID19: Impact metrics and indicators for coronavirus related publications* (Version 36). Zenodo. https://doi.org/10.5281/zenodo.4555117

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., ... Kohlmeier, S. (2020a). CORD-19: The Covid-19 Open Research Dataset. *arXiv*, arXiv:2004.10706v2.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., ... Onyshkevych, B. A. (2020b). COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv*, arXiv:2007.00576. https://doi.org/10.18653/v1/2021.naacl-demos.8

Wise, C., Ioannidis, V. N., Calvo, M. R., Song, X., Price, G., ... Karypis, G. (2020). COVID-19 knowledge graph: Accelerating information retrieval and discovery for scientific literature. *arXiv*, arXiv:2007.12731.

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023

APPENDIX: RESEARCHER QUERIES

Table A1 presents the query literals used in the evaluation study of BIP4COVID19's search engine.

Table A1. Queries provided by the participants of the user evaluation study of the BIP4COVID19 search engine

Researcher	Query	Query literal(s)
R1	Q1	"SARS-CoV-2," "wastewater," "NGS," "Next generation sequencing," "co-occurring mutation," "bioinformatics," "method"
R2	Q1	"reinfection"
	Q2	"variants of concern," "variants of interest"
	Q3	"antibodies"
R3	Q1	"wastewater"
	Q2	"variants of concern"
	Q3	"bioinformatics"
R4	Q1	"Pfizer BioNTech," "vaccine effectiveness"
	Q2	"infectiousness"
	Q3	"SARS-COV-2 Mutations"
R5	Q1	"Vaccine side-effects"
	Q2	"myocarditis"
	Q3	"troponin"
R6	Q1	"mortality rate"
	Q2	"cases"
	Q3	"symptoms"

Downloaded from http://direct.mit.edu/qss/article-pdf/2/4/1447/2007834/qss_a_00169.pdf by guest on 07 September 2023