Received 15 April 2022
Revised 29 July 2022
13 August 2022
Accepted 21 August 2022

Who are the 100 largest scientific publishers by journal count? A webscraping approach

Andreas Nishikawa-Pacher
TU Wien Bibliothek, Vienna, Austria;
Vienna School of International Studies, Vienna, Austria and
Department of Legal and Constitutional History, University of Vienna, Vienna, Austria

Abstract

Purpose – How to obtain a list of the 100 largest scientific publishers sorted by journal count? Existing databases are unhelpful as each of them inhere biased omissions and data quality flaws. This paper tries to fill this gap with an alternative approach.

Design/methodology/approach – The content coverages of Scopus, Publons, DOAJ and SherpaRomeo were first used to extract a preliminary list of publishers that supposedly possess at least 15 journals. Second, the publishers' websites were scraped to fetch their portfolios and, thus, their "true" journal counts.

Findings – The outcome is a list of the 100 largest publishers comprising 28.060 scholarly journals, with the largest publishing 3.763 journals, and the smallest carrying 76 titles. The usual "oligopoly" of major publishing companies leads the list, but it also contains 17 university presses from the Global South, and, surprisingly, 30 predatory publishers that together publish 4.517 journals.

Research limitations/implications – Additional data sources could be used to mitigate remaining biases; it is difficult to disambiguate publisher names and their imprints; and the dataset carries a non-uniform distribution, thus risking the omission of data points in the lower range.

Practical implications – The dataset can serve as a useful basis for comprehensive meta-scientific surveys on the publisher-level.

Originality/value – The catalogue can be deemed more inclusive and diverse than other ones because many of the publishers would have been overlooked if one had drawn from merely one or two sources. The list is freely accessible and invites regular updates. The approach used here (webscraping) has seldomly been used in meta-scientific surveys.

Keywords Bibliographic systems, Data collection, University presses, Journals, Online databases, Journal publishers, Predatory publishers

Paper type Research paper

Corrigendum: It has come to the attention of the publisher that the article: Nishikawa-Pacher, A. (2022), "Who are the 100 largest scientific publishers by journal count? A webscraping approach", *Journal of Documentation*, Vol. 78 No. 7, pp. 450-463. https://doi.org/10.1108/JD-04-2022-0083 mistakenly labelled IOS Press as a predatory publisher in Table 2. Amendments have been made to Table 2 and throughout the text to correct this issue. The authors sincerely apologise to IOS Press and the readers for any inconvenience caused.

A preprint version of this paper appeared as "Who are the 100 Largest Scientific Publishers by Journal Count? A Webscraping Approach" and has been posted on the SocArXiv repository.

Funding: The author acknowledges TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Journal of Documentation
Vol. 78 No. 7, 2022
pp. 450-463
Emerald Publishing Limited
0022-0418
DOI 10.1108/JD-04-2022-0083

Introduction

There is no complete and freely accessible catalogue of all scientific publishers and their journals. Since there may be tens of thousands of active publishers, a project that uses a sample of journals to assess meta-scientific trends could be content with analyzing only the *largest* publishers. This superlative can be defined by the yearly volume of paper outputs, by the annual profit margin, by the size of the publishing company, by the reputation among the academic community, or by the number of journals published. The present paper is interested in the latter; for, while publishers with high journal counts are believed to amount only to a tiny share of the scientific publication ecosystem, they are nevertheless assumed to process the vast majority of the scholarly output (Pollock, 2022, based on data from *OpenAlex*, cf. Priem *et al.*, 2022). But how would one proceed to identify the, say, hundred largest academic publishers by their journal counts?

Bibliographic platforms may offer a first solution; but while there are indeed large databases of scientific outlets, they usually do not aim at comprehensiveness. As a result, their samples of publishers and their journal counts diverge significantly. *Web of Science*, for instance, is more exclusive than *Scopus* which, however, is likewise not promiscuous (Mongeon and Paul-Hus, 2016); like other databases, it instead proclaims a set of criteria that are to be fulfilled before a publisher can have its journals indexed, leading to potentially large-scale omissions. Smaller catalogues follow specific rationales and thus do not intend to achieve an all-inclusive overview of the landscape of scientific publishing. The lists at the *Directory of Open Access Journals* (DOAJ) or the *Open Access Scholarly Publishers Association* (OASPA), for instance, only record journals and publishers that fulfil conditions pertaining to open access policies. The possibly most comprehensive dataset, the one crawled by *Google Scholar*, may have harvested an impressive directory of publisher-level and journal-level data, but it is not available openly, thereby remaining invisible to the public (Harzing, 2014). Other options do not offer viable alternatives either; *Sci-Hub* (Himmelstein *et al.*, 2018) does not transparently disclose its coverage source, and only comprises articles with a digital object identifier (DOI) – but not all publishers necessarily use DOIs. *CrossRef* faces the same issue regarding DOIs, and adds to the difficulty by not listing "publishers", but rather "members" which may or may not overlap with the legal entity of a publisher. For instance, among the largest CrossRef members are *Cairn*, *JSTOR*, *African Journals Online* and others, all of which are not publishers themselves, but rather offer "digital library" platforms harbouring works from various sources pertaining to multiple publishers. Browsing through the list of members already indicates that the share of non-publisher organizations is so large that filtering them would require an immense amount of detailed, manual labour [1]. The same issue of "over-inclusion" applies to *Scilit.net*, a database maintained by *MDPI* – it likewise includes *Cairn* or *African Journals Online* erroneously as "publishers". Other web platforms, such as *JournalTOCs*, exhibit the same issue, as they list *SciELO*, *RMIT Publishing (Informit)*, *Project MUSE*, *Sabinet Online, Redalyc, Érudit*, *Nepal Journals Online*, or *Bangladesh Journals Online* among their (largest) publishers despite their character as data aggregators rather than actual publishers. The most promising development with regards to high quality meta-scientific data, *OpenAlex* (Priem *et al.*, 2022), is still in its early days under construction as of mid-2022; it remains to be seen how well the publisher-level data will be curated. Finally, *Ulrichsweb* remains a commercial database that is inaccessible to a broader audience, and even with a subscription, users cannot download a holistic catalogue of publishers and their journals; instead, the online platform only offers results based on specific user-inputs. Using *Ulrichsweb*, one could obtain a glance regarding the largest publishers based on journal counts when one queries for active scholarly journals – the query would be `Status:("Active") Serial Type:("Journal") Content Type:("Academic/Scholarly")`. But this glance remains limited to the few dozens of top options, and already this limited list contains multiple variations of publisher names (Figure 1). In brief, in searching for a list of the largest academic publishers by journal count, one will only encounter a heterogeneous, often incomplete blend of noisy and fragmentary numbers.

Figure 1.
Screenshot of *Ulrichsweb*'s filter option regarding publishers, sorted by count, after the search query Status:("Active") Serial Type:("Journal") Content Type:("Academic / Scholarly") on 14 May 2021

An authoritative list of the largest academic publishers, however, could be helpful in many ways. It would aid in achieving robust analyses regarding various aspects of scholarly publishing, such as on the implementation of research ethics policies (Gardner *et al.*, 2022); on the prices of Article Processing Charges, or APCs (Asai, 2020; Schönfelder, 2019); on peer review practices (Besançon *et al.*, 2020; Hamilton *et al.*, 2020; Spezi *et al.*, 2018); on journals' social media presence (Ortega, 2017; Zheng *et al.*, 2019); on their profit-orientation (Beverungen *et al.*, 2012); on their open access and pre-print policies (Laakso, 2014; Laakso *et al.*, 2011); on "editormetrics" (Mendonça *et al.*, 2018; Pacher *et al.*, 2021); on community engagement through paper awards (Lincoln *et al.*, 2012) or through podcasts (Quintana and Heathers, 2021); on data-sharing policies (Holt *et al.*, 2021); on their efforts in fostering diversity (Metz *et al.*, 2016) or in supporting early career researchers (O'Brien *et al.*, 2019); on their rate of ORCID adoption (cf. Porter, 2022); and so forth.

But without a near-complete catalogue of publishers and journals, any researcher risks omissions. An analyst who usually covers STEM (science, technology, engineering and math) disciplines may overlook, for example, the publisher *Philosophy Documentation Center* which possesses 249 journals; a social scientist may not know of the *World Scientific* despite its portfolio size of 204 journals; and a Western scientist may easily miss the Chinese company *KeAi* (with 130 journals) or the Indonesian press of *Universitas Gadjah Mada* (with 123 journals).

To fill this gap, a webscraping approach could aid in generating a list of major academic publishers as well as their journals. Due to coverage biases inherent to every platform, this approach should webscrape not just a single, but rather *multiple* research-related sources. The underlying rationale thus resembles a "Swiss cheese model", where a given layer (or platform) has various holes (or flaws and omissions), but if multiple layers are stacked together side by side, losses can be prevented since the holes (or flaws and omissions) differ in their position. Accordingly, the project presented here first fetches data from four large research-related platforms to obtain a list of publishers that are supposed to be mid-sized or

large according to each platform respectively. As a second step, it accesses each of these publishers' websites to scrape their journal count, so as to filter out only the largest publishers among the collected sample.

The aim is thus to generate a catalogue of major academic publishers and their scholarly journals, a list that is supposed to be more comprehensive, accessible and inclusive than any of the existing ones – while still being focused only on publishers with voluminous portfolios (to reduce the data-collection burden). Moreover, the list should not merely offer a snapshot of a specific moment but be adaptable over time; this possibility of always having the data up-to-date is guaranteed by a public sharing of the codes so as to enable extensions and reiterations of the webscraping process.

The following describes the methodical approach in greater detail. The chapter afterwards presents the results of the top 100 academic publishers, sorted by the number of serial titles they publish, with interesting findings regarding the relatively high shares of Global South university presses on the one hand, and of allegedly predatory publishers on the other hand. The discussion section then outlines various limitations encountered during the research process, including issues of data quality due to the non-uniform data distribution, or the difficulty of disambiguating imprints. The paper concludes with a possible guidance on how the limitations nevertheless point towards future research paths so as to reach the wider goal of a complete overview of academic publishers and their scholarly journals that could serve as a starting point for broad meta-scientific investigations.

Methods

To generate a comprehensive list of academic publishers and their scholarly journals, two separate methodical steps were necessary. The first one comprised data collection on the *publisher*-level. Based on the preliminary results of that first step, the second one proceeded with gathering *journal*-level data, or at least the respective journal count. The following will describe the respective approach in sequence.

The data and the codes are available in a Zenodo repository at https://doi.org/10.5281/zenodo.7081147 under a Creative Commons-license (CC0).

Publisher-level data

Data sample and data collection. One single data source seems insufficient when one seeks to attain a complete overview over the landscape of scholarly publications; for each source inheres its own biases and indexing criteria. Instead, one should draw from *multiple* platforms. While heterogenous in character and scope, they may, taken together, provide a more complete menu of publishers than if one merely used a single database.

The present project thus uses four data samples, each of which comprises not only a large list of academic publishers, but also (at least implicitly) the number of journals assigned to them.

The first one is *Scopus*, a large-scale database of scientific publications that provides an openly available source title list. Using their source list from October 2020 comprising 40.804 journals in total, the names of the publishers were extracted and their frequency (i.e. journal count) counted.

The second data sample, *Publons*, is a platform designed to document and verify peer reviews. It allows anyone to register a referee report conducted for any journal from any publisher (Van Noorden, 2014). It thus follows a "bottom-up" approach which potentially covers even publishers that tend to be invisibilized in other indexing services. Using webscraping with *R*'s *rvest* library (Wickham and RStudio, 2020), this project accessed *Publons*' directory of publishers ("All publishers", n.d.).

The third source is *DOAJ*, a directory of open access journals aiming at a global coverage of scholarly publishers and journals that adhere to standards of open access publishing.

To fetch the relevant information, this project used the JSON-formatted journal metadata from *DOAJ*'s public data dump.

The final source of publishers used was *Sherpa Romeo*, a website which aggregates open access archiving policies from a growing number of more than 4.000 publishers. Their publisher list was scraped with *R*.

All these data were collected on 11. December 2020.

Data analysis. Having collected four datasets comprising publisher names and their number of journals according to each respective platform, this project joined these datasets together, harmonized some publisher names, and extracted the highest journal count per publisher. For example, if the publisher *Copernicus Publications* had 41 journals in *Scopus*, 47 in *Publons*, 40 in *DOAJ*, and 71 in *Sherpa Romeo*, that publisher was assigned the maximum journal count of 71. This count was only a preliminary one; the real number of journals would be verified later (as will be outlined below).

After garnering these data, the list was sorted by the preliminary number of journals in descending order. In total, there were 24.722 distinct publisher names. As resource constraints made it impossible to look at each of the publisher distinctly and thoroughly, a threshold was chosen that would leave one with a still-manageable sample while ensuring that the result would still be a plausible list of the largest publishers. With that threshold, only publishers that supposedly carried at least 15 titles according to any of the four data sources were kept – for example, since *Copernicus Publications* had been assigned the preliminary count of 71 journals (above the threshold of 15), it remained in the sample for further validation of its journal count. The threshold was chosen because it seemed low enough to ensure that all publishers that would make it into the final list would pass that threshold, even if the four data sources did not have a complete portfolio of these publishers; in this sense, the lower the threshold, the more complete will be the final data. However, the threshold should not be too low – it should rather be high enough to yield a sample that would be manageable for a manual verification of each publisher's journal count. In other words, as one lowers the threshold, the sample size increases, and thereby the likelihood of detecting yet another large publisher that will make it into the final list becomes greater. However, larger sample sizes require more resources, and there may be "a point where an effect [of increasing the sample size] becomes so minuscule that it is meaningless in a practical sense" (Alba-Fernández *et al.*, 2020, p. 14). The threshold of 15 journals may have allowed for sufficient data to create a reliable top 100 list (cf. the superficial assessment in the Results section below).

Preliminary publisher-level results. A preliminary result extracted 568 distinct publisher names that supposedly published at least 15 journals, according to any of the four data sources DOAJ, Publons, Scopus or Sherpa Romeo.

This preliminary list was then cleaned manually, as there were obvious data quality issues such as inflated numbers and unharmonized publisher names. The manual refinement also got rid of duplications, discontinued presses and non-publishers (e.g. *Egyptian Knowledge Bank* or *SciELO*), resulting in a preliminary list of 414 academic publishers.

Journal-level data

Based on the preliminary list that resulted from the publisher-level data collection, the next step was to visit each listed publisher's website to find the respective portfolio of journals. In order to webscrape each publisher's respective journal list, the so-called CSS [2] selectors that harbour the names and the links of the journals were required. The manual collection of these CSS selectors for each of the 414 publishers was undertaken in January 2021 (and updated in mid-2022). The respective publisher websites were then scraped between March and July 2022, fetching data about journal names and journal counts [3], finally filtering the 100 largest publishers according to these webscraped journal counts.

Figure 2 offers a diagram of the methodical approach taken.

Results

The outcome of the data-collection resulted in a catalogue of the 100 largest academic publishers (comprising 28.060 serial titles) based on journal counts. Summary statistics are visible in Table 1.

Ordered by journal counts, the top ones resemble the prominent "oligopoly" of academic publishing (Larivière *et al.*, 2015) – Springer, Taylor & Francis, Elsevier, Wiley, and SAGE lead the list. Many of the middle-ranging ones, however, may offer surprisingly unknown or only faintly familiar names to researchers whose usual range is confined to just a single, specific discipline or to a single, specific region.

Of the 100 largest publishers, 17 are university-based presses headquartered in research institutions at the Global South (perhaps surprisingly; cf. Collyer, 2018). Eight of them are from Latin America (cf. Delgado-Troncoso and Fischman, 2014), while seven are based in Indonesia (cf. Irawan *et al.*, 2021; Wiryawan, 2014) – including the largest among them, the *Universitas Pendidikan Indonesia* that publishes 177 journals. One press from Iran and Malaysia each round up this subset of Global South university presses.

Another possibly surprising result is that the list contains a large share of so-called predatory publishers – namely, 30 out of 100 [4]. Most of the allegedly predatory publishers in the present list even publish more than one hundred titles; the largest one, *OMICS*, even has 705 journals in its portfolio, propelling it into the sixth place of the overall ranking. In total, they publish 4.517 outlets, or more than 16% of all journals covered by the 100 publishers – roughly every sixth journal of a major publisher is a predatory one. Admittedly, the attribute of predatoriness is a contested one, but in its core, the term denotes organizations that publish seemingly scientific articles against monetary charges without offering an authentic peer-review, while at the same time conducting dishonest practices such as deceiving the public of

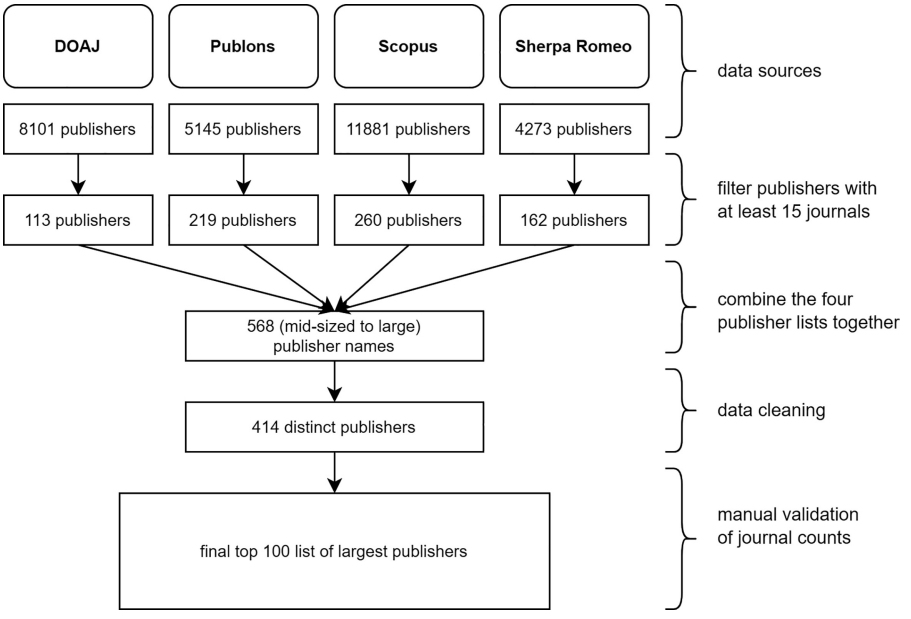

Figure 2.
The methodical approach that led to the final list of the 100 largest academic publishers

wrong impact factors, or listing researchers as editorial board members without their knowledge (Cobey *et al.*, 2018, p. 8). Such (allegedly) predatory publishers are usually left out by curated databases for ethical reasons, but for comprehensive meta-scientific surveys, it may be useful to not exclude them.

The top 100, sorted by journal count, is visible in Table 2.

Some of the publishers listed are not indexed in all four data sample platforms, meaning that they would have been overlooked if this project merely drew from one or two sources. This is especially the case for the so-called predatory publishers; for instance, *OMICS* (with 705 titles) was missing at both *DOAJ* and *Sherpa Romeo*; or, if one only used *DOAJ* and *Scopus* as relevant sources, then one would have omitted *Gavin Publishers* (with 168 journals) and *Scientific and Academic Publishing* (comprising 149 titles); and if one drew from just *Publons* and *Scopus*, then *Open Access Pub* (boasting 198 journals in its portfolio) would not have been found.

However, non-predatory publishers like university presses would have suffered a similar fate; for example, the press of *Universitas Negeri Semarang* which has 120 journals would not have been found if one merely collected publishers that had any reviews verified at *Publons*.

The "Swiss cheese model" approach of using various layers, or multiple research-related platforms for data-collection, thus helped to prevent potential data losses.

This is not to claim that the result is exhaustive and accurate, as the Discussion section will consider below. There still may be omissions, especially in the lower ranks of the list – the distribution is so non-uniform that the upper "cloud" of the ranking is likely accurate, while the "tail" is rather noisy. To give a rough impression of how accurate the ranking is, at least with regards to the four data sources used here, one can slice the original sample (the unharmonized one comprising the 414 publishers that had at least 15 journals according to either of our four data sources) into ten deciles, with the tenth decile showing the largest publishers and the first decile the smallest ones. Each decile contains 41 or 42 publisher names. In the tenth decile, the vast majority of the publishers (87.8%) made it into the final top 100 list; in the ninth decile, that share fell to roughly a half (48.8%). The eigth decile was down to less than a fourth (22.0%). In general, there is a clear downward trend (with a few exceptions) until the first decile, which had just 2.4% of its publishers in the final list (see Table 3). With each decile, the median decline in percentage points was −7.1%, so that one could except a further quantile to have an even lower probability that any of the listed publishers there would make it into the final list. While such statistical numbers do not guarantee that the final top 100 list is accurate, they do provide confidence that the probability of errors is not overly high, at least given the four data sources here; and even if one demanded higher precision, the paper's purpose was primarily to demonstrate the utility of a method (webscraping) rather than to execute it until perfection.

Discussion

Webscraping, first, multiple databases of scientific indexing services, and second, the publishers' websites themselves offers an effective way to obtain a comprehensive overview

Table 1.
Descriptive data about the number of journals (grouped by publisher) in the one hundred largest publishers in the webscraped dataset

Total nr. of journals	Mean nr. of journals per publisher	Median	Mode	Std.Dev.	Minimum	Maximum
28.060	281	124	92	553	76	3.763

Rank	Publisher	Journals	Predatory	Global South Univ. Press
1	Springer	3,763		
2	Taylor & Francis	2,912		
3	Elsevier	2,674		
4	Wiley	1,691		
5	SAGE	1,208		
6	OMICS	705	Yes	
7	De Gruyter	513		
8	Oxford University Press	500		
9	InderScience	472		
10	Brill	461		
11	Cambridge University Press	422		
12	Thieme	407		
13	Medknow	386		
14	Emerald	377		
15	MDPI	376		
16	Lippincott, Williams & Wilkins	375		
17	BioMedCentral	306		
18	IEEE	294		
19	Science Publishing Group	273	Yes	
20	Philosophy Documentation Center	249		
21	SCIRP	247	Yes	
22	IRMA	244		
23	Hindawi	243		
24	IGI Global	238		
25	World Scientific	204		
26	Austin Publishing Group	202	Yes	
27	Bentham	201	Yes	
28	Universidade de Sao Paulo	200		
29	Open Access Pub	198	Yes	
30	Longdom	190	Yes	
31	Universitas Pendidikan Indonesia	177		Yes
32	Gavin Publishers	168	Yes	
33	Universidad de Buenos Aires	168		Yes
34	iMedPub	163	Yes	
35	Nauka	162		
36	Schweizerbart	158		
37	Fabrizio Serra	157		
38	Scientific and Academic Publishing	149		
39	JSciMedCentral	147	Yes	
40	Frontiers	138		
41	Hans Publishers	137	Yes	
42	Advanced Research Publications	135	Yes	
43	Open Access Text (OAT)	134	Yes	
44	KeAi	130		
45	eScholarship Publishing	128		
46	Universidad Nacional Autonoma de Mexico	127		Yes
47	Intellect Books	126		
48	Hilaris	125	Yes	
49	Academic Journals	125	Yes	
50	Science and Education Publishing	125	Yes	
51	Universitas Gadjah Mada	123		Yes
52	Conscientia Beam	122		
53	Universitas Negeri Semarang	120		Yes
54	Pleiades	119		

(*continued*)

Table 2.
The final list of the 100 largest academic publishers ordered by their journal counts

Rank	Publisher	Journals	Predatory	Global South Univ. Press
55	University of Tehran	115		Yes
56	Sciencedomain International	112	Yes	
57	Karger	105		
58	Polish Academy of Sciences	102		
59	IOP Publishing	102		
60	Peertechz Publications	101	Yes	
61	Chinese Academy of Sciences	101		
62	Mary Ann Liebert	101		
63	Universidad Nacional de La Plata	100		Yes
64	John Hopkins University Press	100		
65	Universitas Airlangga	99		Yes
66	Universitat de Barcelona	98		
67	University of Malaya	94		Yes
68	Universitas Negeri Yogyakarta	93		Yes
69	Universidade Federal do Espirito Santo	93		Yes
70	Medcrave	93	Yes	
71	Universidad Nacional de Cordoba	92		Yes
72	APA	92		
73	SciTechnol	92	Yes	
74	University of Chicago Press	92		
75	Universitas Negeri Surabaya	91		Yes
76	Ubiquity Press	91		
77	University of Hawaii Press	90		
78	John Benjamins	90		
79	Jagiellonian University Press	89		
80	Dovepress	89		
81	IOS Press	89		
82	Universidade Federal do Rio Grande do Sul	88		Yes
83	Universitas Diponegoro	87		Yes
84	University of Alberta Press	87		
85	Universidade de Brasilia	86		Yes
86	Internet Scientific Publications	86	Yes	
87	Adam Mickiewicz University	86		
88	Penn State University Press	84		
89	Franco Angeli Edizioni	83		
90	International Scholars Journals	83	Yes	
91	Annex Publishers	82	Yes	
92	Open Access Journals	81	Yes	
93	Pontificia Universidad Javeriana, Bogota	81		Yes
94	Herbert Publications	81	Yes	
95	Il Mulino	80		
96	Medwin Publishers LLC	79	Yes	
97	Premier Publishers	78	Yes	
98	Pulsus Group	76	Yes	
99	Scholarena	76	Yes	
100	Editura Academiei Romane	76		

Table 2.

of the landscape of academic publishing, at least when it comes to *large* publishers in terms of the number of journals in their portfolio. The present project utilized data from *Scopus*, *Publons*, *DOAJ* and *Sherpa Romeo* to automatically enumerate a list of major academic publishers and their scholarly journals as complete as possible. It first gathered a list of publishers that allegedly published at least 15 journals, before validating each publisher's

Decile	Journals (min.)	Journals (max.)	Publishers (sample)	Publishers (final top 100)	Share of publishers in the final top 100 list (%)
10	117	3,920	41	36	87.8
9	63	115	41	20	48.8
8	46	63	41	9	22.0
7	36	46	41	4	9.8
6	29	36	41	4	9.8
5	24	28	41	8	19.5
4	20	24	42	7	16.7
3	18	20	42	1	2.4
2	16	18	42	4	9.5
1	15	16	42	1	2.4

Note(s): The data are based on the preliminary list of 414 publishers; accordingly, the journal counts refer not necessarily to the 'true' count, but to the maximum value according to any of the four data sources (DOAJ, Publons, Romeo Sherpa, or Scopus)

Table 3. How many publishers in the original sample made it into the final top 100 list?

journal count that resulted in a catalogue of the 100 largest academic publishers comprising 28.060 scholarly periodicals.

Many of these publishers, especially in the mid- and smaller range, would have been omitted if one had drawn only from a subset of the databases. This is especially pertinent to those that are either located in the Global South (Collyer, 2018; Jimenez *et al.*, in press, pp. 4–5; Okune *et al.*, 2018; Teixeira da Silva *et al.*, 2019) or that publish articles in languages other than English ("LOTE") (Ren and Rousseau, 2002; Vera-Baceta *et al.*, 2019). They are not always indexed in the major scientific databases, and some of them do not issue DOIs for various reasons, making it easy to overlook them in conventional searches. Examples include the Iranian press of the *University of Tehran* (with 115 journals), the Chinese one of *KeAi* (130 journals), the major Indonesian players like the presses of *Universitas Gadjah Mada* (123 journals), *Universitas Negeri Semarang* (120 journals) and *Universitas Diponegoro* (87 journals), Eastern European publishers like the *Editura Academiei Romane* (76 journals), or Latin American entities belonging to the *Universidade de Brasília* (86 journals) or to the *Universidad Nacional Autónoma de México* (127 journals). The fact that the present project did not omit them indicates that the catalogue gathered here might be less susceptible to systemically biased omissions than if one had used merely one or two sources.

The list generated by this project thus offers a gateway towards large-scale analyses regarding macro-scale engagements, actions and policies of publishers and journals. May they relate to open access aspects, to the conduct of peer review, to article processing charges, to the availability of metadata or to editorial boards – whatever the use case, a webscraping approach that gathers meta-scientific information seems to offer a viable path for alternative and inclusive samples. And it is on the basis of these samples that one can thoroughly investigate existing research cultures in all their diversity.

In addition, as all the present paper's codes and data are shared publicly, they can find extension so as to cover further data sources, and they me be executed repeatedly to update the catalogue over time.

However, there are various weaknesses and limitations to be discussed. First and foremost, while the upper "cloud" of the dataset may accurately depict the league of the largest academic publishers, the mid- and lower ranges (or "tail") may be more susceptible to noisy errors and omissions. In other words, the dataset is most likely an imbalanced one due to the non-uniform distribution of the underlying data (Kotsiantis *et al.*, 2006). That is, there is a high probability of the largest publishers to occur in any of the four samples, but the smaller the publisher, the less

likely it is that one identifies them through webscraping the four sources (a problem of undersampling). After all, the use of multiple platforms does not dispense with the necessity to be aware of inherent biases; it is possible that there are still enough publishers that have not made it into any of the four data samples used for this project. Such biases could be mitigated by drawing from more and more sources. *CORE* (Makhija *et al.*, 2018), *JSTOR* (Schonfeld, 2012), *BASE* (Pieper and Summann, 2006), *OpenAIRE Explore* (Alexiou *et al.*, 2016), the *Directory of Free Arab Journals* (DFAJ) (2021), *SciELO* (Packer, 2009), the Iranian *Scientific Information Database* (SID.ir), or *African Journals OnLine* (AJOL) may serve as likely candidates, though one would first need to ensure that one can indeed obtain structured data from them.

Other data difficulties remain. The issue of disambiguating publisher names and their imprints is one that may lead to arbitrary definitions (e.g. differentiating *Springer* from *Springer Nature* and *BioMedCentral*, but not from *Demos Medical Publishing*, even though they all share the same parent companies). A related problem arises when the samples used aggregators or information retrieval platforms (such as *SciELO* or the *Egyptian Knowledge Base*) erroneously as publishers. This is one reason why *CrossRef*'s member list or *Scilit* could not be used as data sources for the present project. A further limitation lies in the fact that some of the journals listed in the publisher's online catalogues may be discontinued or inactive (Cortegiani *et al.*, 2020). The next step should thus necessarily entail a closer and possibly manual assessment of each publisher's precise journal count.

Once these limitations are addressed, the webscraping approach outlined here may fill a gap in the meta-scientific literature, especially with regards to exhaustive surveys of university presses, scholarly publishers and scientific journals. Without a reliably and freely available comprehensive list, scientometric examinations would risk an incomplete coverage of the diverse landscape of academic publishing, leading to a structural invisibilisation of underrepresented journals or an underestimation of the extent to which predatory publishers have occupied the scientific ecosystem.

With additional data refinements and even more encompassing, alternative sources, the list may finally attain a satisfying degree of saturation and accuracy. Once one can be certain that there is a complete and inclusive catalogue of academic publishers and scholarly journals from all around the world without any blind spots, this cannot but benefit the whole science of science.

Notes

1. *CrossRef* itself does not have data about whether and which of their members are (non-)publishers; private communication from 26 April 2021 (internally saved at *CrossRef* as request #364948).
2. Cascading Style Sheets, a computer language used for layouting and structuring websites (usually in conjunction with HTML, or Hypertext Markup Language).
3. Due to technical errors (e.g. outdated security certificates of the respective host server) or due to improperly structured websites, some journal counts had to be collected manually.
4. Despite controversies (Koerber *et al.*, 2020), this paper defines predatoriness largely by the inclusion of the respective publisher in the updated version of Beall's list as of December 2021 ("Beall's List of Potential Predatory Journals and Publishers", 2021). There are two exceptions – *Frontiers* is not marked as predatory in the present paper because its inclusion into Beall's List has always remained highly contested (Kendall, 2021, p. 382); but *Annex Publishers* is marked as predatory even though it was not in Beall's List for the following reasons: it refers to a bogus version of the Impact Factor ("CiteFactor") as a reference, promises rapid peer reviews (21 days), a publication within 24 h after acceptance, a high visibility due to its inclusion on *Google Scholar* (which is trivial); furthermore, it is not indexed in the DOAJ and demands quite high Article Processing Charges (between USD 1.200 and USD 3.600, as of July 2022).

References

"All publishers" (n.d.), "Publons", available at: https://publons.com/publisher/?page=1 (accessed 4 January 2021).

Alba-Fernández, M.V., Ariza-López, F.J., Rodríguez-Avi, J. and García-Balboa, J.L. (2020), "Statistical methods for thematic-accuracy quality control based on an accurate reference sample", *Remote Sensing*, Vol. 12 No. 5, p. 816.

Alexiou, G., Vahdati, S., Lange, C., Papastefanatos, G. and Lohmann, S. (2016), "OpenAIRE LOD services: scholarly communication data as linked data", in González-Beltrán, A., Osborne, F. and Peroni, S. (Eds), *Semantics, Analytics, Visualization. Enhancing Scholarly Data*, Springer International Publishing, Cham, pp. 45-50.

Asai, S. (2020), "Market power of publishers in setting article processing charges for open access journals", *Scientometrics*, Vol. 123 No. 2, pp. 1037-1049.

"Beall's List of Potential Predatory Journals and Publishers" (2021), 8 December, available at: https://web.archive.org/web/20220727081817/https://beallslist.net/ (accessed 29 July 2022).

Besançon, L., Rönnberg, N., Löwgren, J., Tennant, J.P. and Cooper, M. (2020), "Open up: a survey on open and non-anonymized peer reviewing", *Research Integrity and Peer Review*, Vol. 5 No. 1, p. 8.

Beverungen, A., Böhm, S. and Land, C. (2012), "The poverty of journal publishing", Organization, Vol. 19 No. 6, pp. 929-938.

Cobey, K.D., Lalu, M.M., Skidmore, B., Ahmadzai, N., Grudniewicz, A. and Moher, D. (2018), "What is a predatory journal? A scoping review", F1000.

Collyer, F.M. (2018), "Global patterns in the publishing of academic knowledge: global North, global South", *Current Sociology*, Vol. 66 No. 1, pp. 56-73.

Cortegiani, A., Ippolito, M., Ingoglia, G., Manca, A., Cugusi, L., Severin, A., Strinzel, M., Panzarella, V., Campisi, G., Manoj, L., Gregoretti, C., Einav, S., Moher, D. and Giarratano, A. (2020), "Citations and metrics of journals discontinued from Scopus for publication concerns: the GhoS(t)copus Project", F1000.

Delgado-Troncoso, J.E. and Fischman, G.E. (2014), "The future of Latin American academic journals", in Cope, B. and Phillips, A. (Eds), *The Future of the Academic Journal*, 2nd ed., Chandos Publishing, pp. 379-400.

Gardner, V., Robinson, M. and O'Connell, E. (2022), "Implementing the declaration on research assessment: a publisher case study", *Insights*, Vol. 35 No. 0, p. 7.

Hamilton, D.G., Fraser, H., Hoekstra, R. and Fidler, F. (2020), "Journal policies and editors' opinions on peer review", *eLife*, Vol. 9, e62529.

Harzing, A.-W. (2014), "A longitudinal study of Google Scholar coverage between 2012 and 2013", *Scientometrics*, Vol. 98 No. 1, pp. 565-575.

Himmelstein, D.S., Romero, A.R., Levernier, J.G., Munro, T.A., McLaughlin, S.R., Greshake Tzovaras, B. and Greene, C.S. (2018), "Sci-Hub provides access to nearly all scholarly literature", *eLife*, Vol. 7, e32822.

Holt, J., Walker, A. and Jones, P. (2021), "Introducing a data availability policy for journals at IOP Publishing: measuring the impact on authors and editorial teams", *Learned Publishing*, Vol. 34 No. 41, pp. 478-486.

Irawan, D.E., Abraham, J., Zein, R.A., Ridlo, I.A. and Aribowo, E.K. (2021), "Open access in Indonesia", *Development and Change*, Vol. 52 No. 3, pp. 651-660.

Jimenez, A., Vannini, S. and Cox, A. (In press), "A holistic decolonial lens for library and information studies", *Journal of Documentation*.

Kendall, G. (2021), "Beall's legacy in the battle against predatory publishers", *Learned Publishing*, Vol. 34 No. 3, pp. 379-388.

Koerber, A., Starkey, J.C., Ardon-Dryer, K., Cummins, R.G., Eko, L. and Kee, K.F. (2020), "A qualitative content analysis of watchlists vs safelists: how do they address the issue of predatory publishing?", *The Journal of Academic Librarianship*, Vol. 46 No. 6, 102236.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2006), "Handling imbalanced datasets: a review", *GESTS International Transactions on Computer Science and Engineering*, Vol. 30 No. 1, pp. 25-36.

Laakso, M. (2014), "Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed", *Scientometrics*, Vol. 99 No. 2, pp. 475-494.

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C. and Hedlund, T. (2011), "The development of open access journal publishing from 1993 to 2009", *PLOS ONE*, Vol. 6 No. 6, e20961.

Larivière, V., Haustein, S. and Mongeon, P. (2015), "The oligopoly of academic publishers in the digital era", *PLOS ONE*, Vol. 10 No. 6, e0127502.

Lincoln, A.E., Pincus, S., Koster, J.B. and Leboy, P.S. (2012), "The matilda effect in science: awards and prizes in the US, 1990s and 2000s", *Social Studies of Science*, Vol. 42 No. 2, pp. 307-320.

Makhija, V., Kumar, V., Tiwari, A. and Verma, A. (2018), "Knowledge management system using CORE repository", *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*, pp. 59-64.

Mendonça, S., Pereira, J. and Ferreira, M.E. (2018), "Gatekeeping African studies: what does 'editormetrics' indicate about journal governance?", *Scientometrics*, Vol. 117 No. 3, pp. 1513-1534.

Metz, I., Harzing, A.-W. and Zyphur, M.J. (2016), "Of journal editors and editorial boards: who are the trailblazers in increasing editorial board gender equality?", *British Journal of Management*, Vol. 27 No. 4, pp. 712-726.

Mongeon, P. and Paul-Hus, A. (2016), "The journal coverage of Web of Science and Scopus: a comparative analysis", *Scientometrics*, Vol. 106 No. 1, pp. 213-228.

Okune, A., Hillyer, R., Albornoz, D., Posada, A. and Chan, L. (2018), "Whose infrastructure? Towards inclusive and collaborative knowledge infrastructures in open science", *ELPUB 2018*.

Ortega, J.L. (2017), "The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations)", *Aslib Journal of Information Management*, Vol. 69 No. 6, pp. 674-687.

O'Brien, A., Graf, C. and McKellar, K. (2019), "How publishers and editors can help early career researchers: recommendations from a roundtable discussion", *Learned Publishing*, Vol. 32 No. 4, pp. 383-393.

Pacher, A., Heck, T. and Schoch, K. (2021), "Open editors: a dataset of scholarly journals' editorial board positions", *SocArXiV*. doi: 10.31235/osf.io/jvzq7.

Packer, A.L. (2009), "The SciELO open access: a gold way from the South", *Canadian Journal of Higher Education*, Vol. 39 No. 3, pp. 111-126.

Pieper, D. and Summann, F. (2006), "Bielefeld Academic Search Engine (BASE): an end-user oriented institutional repository search service", *Library Hi Tech*, Vol. 24 No. 4, pp. 614-619.

Pollock, D. (2022), "News and views: publishers and market consolidation – Part 1 of 2", *Delta Think*, 21 June, available at: https://deltathink.com/news-views-publishers-and-market-consolidation-part-1-of-2/ (accessed 6 July 2022).

Porter, S.J. (2022), "Measuring research information citizenship across ORCID practice", *Frontiers in Research Metrics and Analytics*, Vol. 7, 779097.

Priem, J., Piwowar, H. and Orr, R. (2022), "OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts", arXiv 2205.01833. doi: 10.48550/arXiv.2205.01833.

Quintana, D.S. and Heathers, J.A.J. (2021), "How podcasts can benefit scientific communities", *Trends in Cognitive Sciences*, Vol. 25 No. 1, pp. 3-5.

Ren, S. and Rousseau, R. (2002), "International visibility of Chinese scientific journals", *Scientometrics*, Vol. 53 No. 3, pp. 389-405.

Schönfelder, N. (2019), "Article processing charges: mirroring the citation impact or legacy of the subscription-based model?", *Quantitative Science Studies*, Vol. 1 No. 1, pp. 6-27.

Schonfeld, R.C. (2012), *JSTOR: A History*, Princeton University Press, New Jersey.

Spezi, V., Wakeling, S., Pinfield, S., Fry, J., Creaser, C. and Willett, P. (2018), "Let the community decide'? The vision and reality of soundness-only peer review in open-access mega-journals", *Journal of Documentation*, Vol. 74 No. 1, pp. 137-161.

Teixeira da Silva, J.A., Adjei, K.O.K., Owusu-Ansah, C.M., Sooryamoorthy, R. and Balehegn, M. (2019), "Africa's challenges in the OA movement: risks and possibilities", *Online Information Review*, Vol. 43 No. 4, pp. 496-512.

Van Noorden, R. (2014), "The scientists who get credit for peer review", *Nature*, doi: 10.1038/nature.2014.16102.

Vera-Baceta, M.-A., Thelwall, M. and Kousha, K. (2019), "Web of science and Scopus language coverage", *Scientometrics*, Vol. 121 No. 3, pp. 1803-1813.

Wickham, H. and RStudio (2020), "Rvest: easily harvest (scrape) web pages", available at: https://CRAN.R-project.org/package=rvest (accessed 4 January 2021).

Wiryawan, K.G. (2014), "The current status of science journals in Indonesia", *Science Editing*, Vol. 1 No. 2, pp. 71-75.

Zheng, H., Aung, H.H., Erdt, M., Peng, T.-Q., Raamkumar, A.S. and Theng, Y.-L. (2019), "Social media presence of scholarly journals", *Journal of the Association for Information Science and Technology*, Vol. 70 No. 3, pp. 256-270.

Corresponding author
Andreas Nishikawa-Pacher can be contacted at: andreas.pacher@tuwien.ac.at