The Canadian Journal of Information and Library Science - La Revue canadienne des sciences de l'information et de bibliothéconomie (CJILS-RCSIB)
Vol. 48, No. 1 (2025)
DOI: 10.5206/cjils-rcsib.v48i1.22290
Research Article

Bibliodiversity of Small Academic Publishers: The Role of Open Access for Impact and Visibility

Roberto Cruz Romero [1], Dimity Stephen [1], and Stephan Stahlschmidt [1,2]
[1]German Centre for Higher Education Research and Science Studies (DZHW), Research System and Science Dynamics
[2]EC3 Research Group, Universidad de Granada

Large bibliographic databases highlight tangible and symbolic differences regarding the standards of quality attached to them, underlining diverging incentive structures for small and large academic publishers. To assess the academic differences associated with these publishers, we explore bibliometric data for small publishers' journals from the Web of Science and Scopus. We then discuss the visibility and impact of highly cited literature in small open access journals in relation to their cited references from indexed and non-indexed sources. We find that non-indexed references are consistently relevant for highly cited literature, yet the share of items that obtain high citation counts is rather small and uneven across disciplines. In general, we identify regional and linguistic specificities, whilst there are some observable thematic differences compared to more mainstream publications. In particular, we underline that a healthy bibliodiversity can, depending on language or regional contexts, shape epistemic and scientific practices and narratives.

Keywords: scholarly publishing, bibliodiversity, small publishers, open access, references, bibliometrics

Introduction

Tensions around productivity, visibility, and impact characterize academic publishing. Visibility and impact are shaped by indexing and citations, linked to a) epistemic practices in academic fields (Leonelli, 2022) and b) material determinants like costs to publish or costs to access (cf. Krawczyk & Kulczycki, 2021a). As a result, access restrictions are increasingly debated amid internal and external pressures to deepen the open access (OA) transformation.[1] The OA transformation thus reflects diverging interests within the publication landscape.

One set of actors – mainly policy-making and funding agencies like the German Research Foundation (DFG) or the Canadian Tri-Agency[2] – advocates for the development of publicly-funded research that prioritizes OA publishing (Deutsche Forschungsgemeinschaft, 2022). The goal is to promote transparency and the broad dissemination of publicly funded research. Yet, it also externalizes commercial incentives for publishers regarding their OA options (e.g., through the push for broader, larger transformative agreements). Thus, OA publishing represents a new pathway towards profit for academic publishers (McGuigan & Russell, 2008), rather than a completely new mode of operation.

Other actors – scholars and academic communities – seek to contribute to academic discussions whilst building their scholarly profiles. Journal selection then involves a balance of several competing priorities, such as (perceived or quantified) quality, reputation, reach, institutional evaluation criteria, OA options, and publishing costs (Rowley et al., 2020). Despite general support for OA publishing (Nobes & Harris, 2023), scholars' preferences are likely discipline-specific and, commonly, outweighed by visibility, reputation, and prestige (Nobes & Harris, 2023; Rowley et al., 2020). This mismatch is accentuated as research communities connect through spaces and platforms beyond the academic jour-

Correspondence concerning this article should be addressed to Roberto Cruz Romero: cruzromero@dzhw.eu

[1]See for example the Open Science in Horizon Europe framework, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/other/events/20210421/open-science_en.pptx.

[2]The Tri-Agency is a joint structure of government funding agencies for science and research established by the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Social Sciences and Humanities Research Council of Canada (SSHRC). The Tri-Agency's Open Access policy explicitly states that its goal is "to ensure that all agency-funded, peer-reviewed research articles are immediately and freely available online to the research community, readers in the public, private and not-for-profit sectors, and the general public". See: https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/open-access/draft-revised-tri-agency-open-access-policy-publications.

nal (e.g., workshops and conferences, or institutional agreements). Journals, then, may only serve as an official outlet rather than the core of academic activities.

Additionally, many institutions have decided to move beyond quantitative evaluation practices (e.g., h-indices or impact scores), supporting alternatives for evaluating diversity and innovation qualitatively.[3] The DFG has expressed its support for this shift,[4] as well as the Canadian Tri-Agency Framework, which has underlined its commitment to transparency, accountability, and integrity (Responsible Conduct of Research, 2021). These norms introduce further tensions in the publication system, particularly between the scientific and administrative spheres of academic work. This tense duality underscores the role of established structures (e.g., learned societies), as they may well also function as gatekeepers or enablers. This underlying dynamic determines which publishing models are perceived as standards in particular fields.

Following, the tension between visibility and impact is itself greatly influenced by commercial interests and incentives, evidenced in the material differences between large publishing houses and small scholarly societies. These different incentives underscore the gaps between the academic, economic, and policy dimensions of publishing, amplified by publication-focused evaluation practices. Negotiating visibility (through indexation in prominent bibliometric databases) and the consolidation of scholarly networks becomes the source of conflicting patterns in OA adoption. This article focuses on the activities and roles of the publishers and is concerned with differentiating between the dimensions of publishers (more on this in the Methods section). Our claim is that different incentive structures may exist between publishers, and they may help propagate different types of scholarly output. In the midst of a structural transformation, such as the OA movement, these incentives may also play a determining role in a publisher's outlook over the scholarly landscape.

Moreover, the question of OA is often seen as of secondary importance by smaller publishers in particular (Leonelli et al., 2015; Severin et al., 2020; Tenopir et al., 2016; cf. Knöchelmann, 2023), for whom economic sustainability is of foremost importance, whilst larger publishers have adapted more easily to the policy and market characteristics of the OA transformation (L.-A. Butler et al., 2023; Nishikawa-Pacher, 2022). Size becomes a transversal factor confounding visibility and impact, as it can determine the position that journals (from small publishers - SPs) occupy in the publishing landscape. These distinctions are compounded by OA options (and their technicalities) that justify the alleged necessary quantitative and qualitative standards for indexation in the large bibliometric databases (e.g., Web of Science [WoS] or Scopus), which are frequently applied in quantitative evaluations.

Representation and diversity (both regional and linguistic) escalate the tensions for research communities and publishers alike (Giménez Toledo et al., 2019; Ma et al., 2023). Diversity, as the reflection of representation – and hence linked to the visibility aspect of the OA transformation – is discussed in publishing under the term of bibliodiversity. Visibility is mainly defined via indexation in databases, which acts as a mechanism to incorporate (or exclude) as many voices as possible into the scientific discourse. As the Jussieu Call for Open Science and Bibliodiversity states, OA must be complemented by "support for the diversity of those acting in scientific publishing (. . .) putting an end to the dominance of a small number among us imposing their terms to scientific communities" (Jussieu Call for Open Science and Bibliodiversity, 2017). Hence, we view the OA dimension of scholarly publication as a mechanism that alleviates the structural conditions of inaccessible research.

Open access transformation and the academic journal landscape

The OA movement has ignited a profound shift in the academic journal landscape, reshaping the traditional boundaries of scholarly communication. Unrestricted access to scholarly literature has challenged the prevailing subscription-based publishing model and prompted a fundamental revaluation of the way research is produced, disseminated, and consumed. Regardless of size and profit orientation, most publishers have seen their business models transformed. Subsequently, OA has also affected how publishers market their journals within specific scientific milieus (Knöchelmann, 2023). A key issue relates to journals' field classifications and their engagement with OA standards. For example, this applies to disciplines where OA is a commonplace, such as medical and health sciences, but not as much in the social sciences and humanities, as our results will show. Hence, disciplinarity may determine the extent to which incentives and motivations to participate in OA channels gain or lose traction.

Despite its contribution to bibliodiversity, the OA transformation has also opened the door for doubt to be cast upon the integrity of OA publishers: detractors argue that the new

[3] For example, the San Francisco Declaration on Research Assessment (DORA). Additionally, the cOAlition S, a coalition of funders including the likes of the European Commission and the Wellcome Trust, commits to DORA in its Plan S: https://www.coalition-s.org/about/.

[4] Maßnahmenpaket zum Wandel der wissenschaftlichen Bewertungskultur [Package of Measures to Change the Scientific Assessment Culture], https://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_61/index.html. Also: The Commission signs the Agreement on Reforming Research Assessment and endorses the San Francisco Declaration on Research Assessment, https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/commission-signs-agreement-reforming-research-assessment-and-endorses-san-francisco-declaration-2022-11-08_en.

author-pays model incentivized unscrupulous publishers to publish material of any quality for financial gain via Article Processing Charges (APCs). This accusation was first levelled by Jeffrey Beall (2015) – a staunch opponent of the OA movement who labelled the Latin American OA publishing landscape a "publication favela" on the basis of North American researchers' ignorance of its existence – but quickly gained traction amongst the academic community, spurring much of the research to date on predatory publishing practices (Krawczyk & Kulczycki, 2021a). Subsequent poor practices by (typically large) OA publishers have continued to threaten the reputation of OA journals. Notable examples include the presumably profit-driven production of an excessive number of articles by MDPI (e.g., 17,000 in a single journal, Sustainability, in 2022 alone) (Exclusive, 2024), and the well-known OMICS scandal, which culminated in a lawsuit brought by the US Federal Trade Commission that held OMICS liable for US$50 million for deceptive practices (Manley, 2019).

As a consequence, the landscape of OA has become muddled and problematic dynamics regarding the visibility of smaller scholarly publishers vis-à-vis OA are reproduced. Fully OA journals are sometimes regarded as inferior to other subscription-based journals with hybrid OA options (Pinfield et al., 2021); the same applies to young or independent journals. This can be observed in the difference between journal types in APCs, which may be viewed as a proxy for the attractiveness of a journal to authors. For instance, the median APC set for gold OA journals (US$2,000) is substantially lower than the median of hybrid journals (US$3,230) of several large publishers, including Elsevier, MDPI, PLOS, and Springer Nature (Brainard, 2024; L.-A. Butler et al., 2024). This situation creates a paradoxical dynamic that makes it difficult for newer journals to make their way into the narratives that dominate certain disciplines and contribute to the overall academic debate (Severin et al., 2020), or to even contribute to a broader (biblio-)diverse landscape (Berger, 2021; Ma et al., 2023). Consequently, newer and OA journals are frequently overlooked by researchers selecting venues and are thus unable to reach the necessary publication and citation standards for indexation in larger databases, reducing their overall visibility. In contrast, more established journals may rely on their academic communities, establishing a feedback loop that excludes and hinders the inclusion of smaller, newer, OA journals.

Visibility of open access publishers

Yet, publishers still occupy a dominant position within the academic landscape. On one side, large editorial corporations (or entities, such as large university presses) tend to assert their dominance of the publication spectrum (L.-A. Butler et al., 2023). On the other side, most universities or learned societies represent the smaller end of the publishing landscape, commonly representing the (un)official outlet of their respective field or organization. With increasing frequency, however, larger publishers acquire some field- or entity-specific journals (from a smaller publisher), allowing the journals to retain editorial control (Knöchelmann, 2023) whilst taking advantage of the larger publisher's established business practices (see Matthias et al., 2019). This acquisition trend may lead to a concentration of the business side of publishing (again, see Nishikawa-Pacher, 2022). This same concentration also shapes the (overwhelmingly technical) possibilities for OA licensing models, as well as the possibilities for accessing diverse research paradigms.

In this sense, OA represents a business opportunity and becomes a systemic condition in the publishing landscape, highlighting the role that specific editorial structures provide. These structures, particularly in relation to OA, underline the pressure on the publishing side, where resources and infrastructure are crucial factors in smaller publishers' decisions to implement OA options and streamline editorial decisions regarding topics and content. For instance, the adoption of an OA model becomes a determining factor for the visibility and potential impact of journals from SPs, as they can reach their audiences irrespective of any pay-to-read costs.[5] Similarly, OA also represents a landmark for scholars; it becomes an opportunity to reach a broader readership, potentially amassing more citations.[6] Further, from a macro perspective, OA can also represent a medium through which narratives and epistemic practices are either consolidated or innovatively explored within and between academic communities.

Small and scholar-led journals

Journals edited and managed by SPs (see Methods section for an explanation of this difference) represent dissimilar faces in relation to the question of independence and profit orientation. For instance, small journals are, by definition, small in scale and operation compared to the more dominant publishers (cf. Nishikawa-Pacher, 2022), and they are managed by publishers that may or may not be associated with academic institutions or learned societies. Scholar-led journals, on the other hand, are strictly linked to research communities or learned societies; in the latter case, usually representing national communities of academic disciplines. However, scholar-led journals are neither always OA nor exclusively small: see, for example, the Johns Hopkins Univer-

[5] Further, diamond open access would signify zero costs for authors, which would represent a truly inclusive standard of publication. See Piwowar et al. (2018) for details on the colour classification and their scholarly and policy impacts.

[6] For instance, Langham-Putrow et al. (2021) evidences the positive but inconclusive effect of the so-called "open access citation advantage". In other terms, Huang et al. (2024) show how, in general, OA offers more citations with a greater variety (diversity) of citing origins – though the latter is subject to field and regional specificities.

sity Press (USA), the Chinese Academy of Sciences (China), or the Universidade de Sao Paulo (Brazil), which each manage at least 100 journals (Nishikawa-Pacher, 2022). Similarly, small journals are neither exclusively scholar-led nor strictly OA. SPs navigate their involvement in the OA paradigm differently, influenced by various factors. For instance, platform management and digital communication issues influence decisions on flipping journals from closed to open access. Conversely, in certain social science and humanities sub-fields, for example, maintaining a non-OA approach is not uncommon (see Knöchelmann, 2023). Nonetheless, in general, OA remains highly relevant for the visibility and potential impact of SPs' journals among wider academic audiences, beyond financial barriers.

Small and scholar-led journals have great potential for enhancing the bibliodiversity of academic narratives. They are able to embrace unconventional or niche topics that might not find a home in larger, mainstream journals (Giménez Toledo et al., 2019). Small and scholar-led journals can thus include a broader range of ideas and voices, promoting a more inclusive and (biblio-)diverse academic communication landscape. Small journals also encourage the intangible diversity of language (Ma et al., 2023). Multilingualism is also a factor highlighted institutionally within the OA movement,[7] as it is a driver of regionally- and or culturally-specific knowledge not captured in mainstream academic conversations underpinned by the standardized large bibliographic databases. Hence, visibility, impact, and diversity highlight the relevance of smaller journals.

Research aims and approach

Given this background, in this article we seek to highlight discipline-specific differences regarding OA uptake, publishers' size, and epistemic practices, focusing on the (biblio)diversity of the publication landscape at the publisher level (Estelle, 2021; Kaier & Lackner, 2019; Matthias et al., 2019). We conduct an exploratory analysis into the composition of the publishing landscape and the trends regarding the space and role small (OA) journals have in specific scientific discourses. From this, we find that small academic publishers are underrepresented in the main bibliometric databases (Clarivate's WoS and Elsevier's Scopus), and that regionally and linguistically diverse literature remains marginal. The transfer mechanism from specific to general manages only to bring limited scholarship to the fore, as most of it replicates general publication patterns (as in meta-analyses, systematic reviews, and such). These differences are further amplified along disciplinary lines, as the already under-representation of the social sciences and humanities (SSH) in the main databases is intensified at the small OA level. We further discuss the role of different small academic publishers and their journals, motivated by the degree of distance to their larger, oligopolist counterparts regarding publication processes and bibliographic visibility (Larivière et al., 2015; cf. L.-A. Butler et al., 2023).

We address these elements with data from two of the main bibliometric databases available: WoS and Scopus. Hence, the questions at the centre of the analyses are whether, and to what extent, SPs' journals a) contribute to the diversity of the (OA) publishing landscape and b) play a role within specific scientific narratives? In this sense, visibility (indexation in the bibliographic databases) and relevance (impact) will be taken as proxies for the diversity of different academic discourses that coexist at various levels. We take the mainstream and the English-speaking scholarship as the dominant academic discourse, as well as the most cited literature in each field, as a reflection of these features. We focus on these databases due to their centrality in the academic publishing landscape, at least in terms of structured data availability. We recognize the relevance and role that other repositories and databases play in diversifying the publishing landscape.

Methods

Identifying small publishers and their open access status

As introduced, the analyses in this article focus on the bibliometric characteristics of OA journals from SPs, including their publication counts, citations, and references. Taking these values into the larger discussion about visibility and diversity in scholarly output, it is possible to see the parallels and contradictions that these publications share in terms of their input to the overall academic discourses. Figure 1 offers a visual description of the methodological steps taken in this direction.[8]

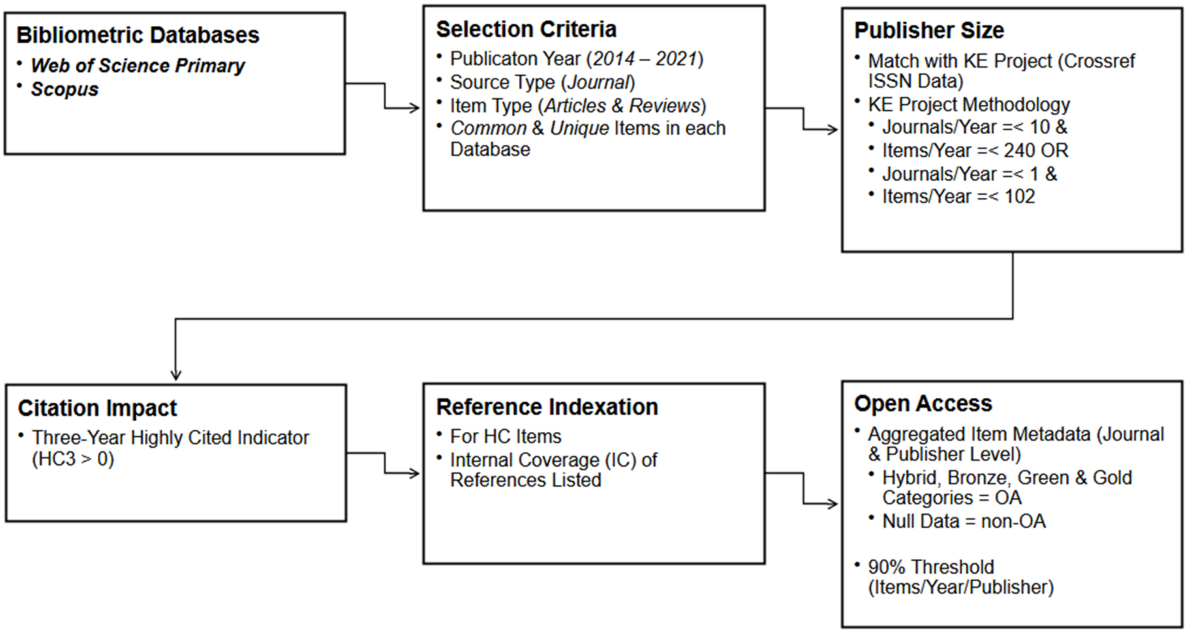
The first step in the analysis was to identify SPs and their corresponding journals. We applied the methodology used by Stephen and Stahlschmidt (2022) in a study commissioned by the Knowledge Exchange (KE) to identify SPs in bibliometric databases (Stephen & Stahlschmidt, 2022). Classifications of what constitutes a small publisher often rely on the number of journals published, but no standard definition is applied. For instance, small- and mid-size publishers have been classified as those producing up to 150 journals (Laakso & Multas, 2022), while publishers of more than thirty journals were elsewhere classified as "major" (Nishikawa-Pacher, 2022). Stephen & Stahlschmidt (2022) classified publishers into three categories (very small, small, and rather small) based on the number of articles and journals they published. However, in this study, we dissolve these categories and examine these three groups together as SPs. Based on the same precedent established by Stephen and Stahlschmidt (2022), SPs are defined here as publishers that produced 10 or fewer

[7] See the Helsinki Initiative on Multilingualism in Scholarly Communication: https://www.helsinki-initiative.org/.

[8] Data and codes (as well as reviews) can be found at the GitLab repository: https://gitlab.com/roberto_cr/indioa

Figure 1

Methodological Flow Chart for Bibliometric Data



journals with 240 or fewer articles per year, or 1 journal with up to 102 articles per year.[9]

Additionally, for the Stephen and Stahlschmidt (2022) study, Crossref was used as the basis for identifying publishers, as it constitutes an extensive and publicly available bibliometric dataset of scholarly content. Crossref's membership is continuously growing, primarily in the lower fee tiers, indicating that it likely contains a large number of SPs (Crossref, 2019, p. 7). From it, journal titles and annual article counts were extracted for all publishers active in 2019-2021. Publishers were disambiguated by their membership identifier, and journals were disambiguated by title, after manual cleaning (for a full description of these lists, please see Stephen & Stahlschmidt (2022)). After applying the aforementioned journal and article count criteria, the SPs dataset consisted of 23,058 journals published by 14,328 publishers.

We matched this dataset using ISSN identifiers to the Competence Network for Bibliometrics' (KB) WoS Primary and Scopus databases.[10] We filtered the matches solely to articles and reviews published in peer-reviewed journals between 2014 and 2021, from which we then differentiated large publishers as non-SPs as a comparison group. We examine articles and reviews as they carry the largest share of original scientific contributions and, hence, the largest share of citations (see Donthu et al., 2021; Ellegaard & Wallin, 2015). We also used the title text strings to match items between the WoS Primary and Scopus databases to identify the common items in both databases and the items exclusive to each database. Fields of Science (FoS) were then defined for each item based on disciplinary metadata from WoS and Scopus that was matched to the Organization for Economic Cooperation and Development's (OECD) six FoS classification, and aggregated to the journal and publisher levels.[11] Thus, publishers are multi-coded in FoS, since a single publisher may edit many journals in diverse fields.

Moreover, we also obtained the total citation counts for the period after three years of publication from WoS Primary and Scopus. From these, an indicator for highly cited items was also obtained, and the internal reference coverage (IC)

[9]This threshold generates a double filter: a) if publishers manage 10 or fewer active journals in a year, and b) if from these journals (<10) there are 240 or fewer published articles in a year. These parameters filter out outliers, such as mega-journals (e.g., PLOS One) and any other exceptionally large single-journal publisher.

[10]We make use of the April 2022 snapshots (https://bibliometrie.info/). WoS data includes publications from the Science Citation Index-Expanded, Social Sciences Index, and the Arts & Humanities Citation Index (Clarivate names this WoS Primary, so we follow this convention). As the snapshot provides incomplete data for 2022, we limit the period to 2021. WoS and Scopus constitute structurally different perspectives from the totality of the global academic system and the examination of publishers in both sources underlines the outcomes of differences in indexation standards between databases (See Stahlschmidt & Stephen, 2022); thus, we focus on the intersecting share of items found both in WoS Primary and Scopus (see Table 1).

[11]This classification consists of six fields: agricultural sciences, engineering and technology, humanities, medical and health sciences, natural sciences, and social sciences. See (OECD, 2007) for further details.

in each database was calculated. The highly cited (HC) measure indicates whether a publication was in the top 10% most cited publications of its field and publication year (Donthu et al., 2021). We used a three-year window for HC (HC3 > 0), as published scientific works tend to reach a stable level of citations indicative of long-term impact three years after publication (Hyland & Jiang, 2017; Moed, 2005). The IC measure identifies the percentage of cited references that are also indexed in each database and thus is a measure of the databases' coverage of a field's literature.

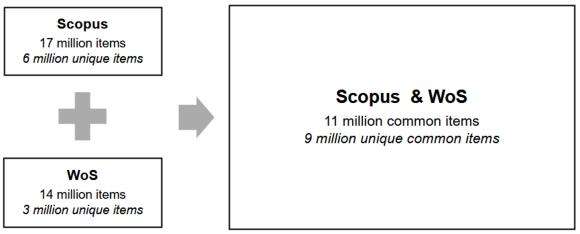
Similarly, using WoS Primary and Scopus metadata on the OA status of each item, we classified journals and publishers as either OA or non-OA.[12] A relevant methodological step here is that we take the latest available type listed for each item and define the OA vs non-OA distinction based on these. We group every item with OA status data (non-null) into the OA category (primarily for Scopus, as there are many OA labels for a single item). Admittedly, this step implies some granularity loss, but the focus of our analysis is set on the differences introduced by a) the size of the publishing entity and, secondarily, b) the imposition of restrictions to access (regardless of the mode through which the access is opened). For broadness, we then grouped items by journal and then by publisher, setting a threshold at 90% OA items, so that any yearly output (per journal per publisher) with at least 90% of its published items in OA format was considered to be OA (journal and, hence, publisher), as there are few pure OA journals and publishers in each database. Finally, we narrowed our sample to highly cited publications from small OA publishers (SOAP) and the references they cite. The latter also serve as input for textual analyses (item and journal titles) for both indexed and non-indexed items.

Text analyses

We examine three dimensions of textual relevance in this study: a) term co-occurrence, b) language identification, and c) entity recognition. The first dimension is based on term frequency counts, which are calculated based on item and journal titles (for both indexed and non-indexed references) of the HC-OA subset.[13] We make use of both the common and unique items (see Figure 2). We then generate a network of term co-occurrence frequencies by estimating a term frequency-inverse document frequency (tf-idf) matrix for the titles of both citing and cited (reference) items, thereby creating a reference link between them. This is done for items HC-OA from SPs' journals and their indexed and non-indexed references. This is methodologically and analytically relevant, as it is the primary link that connects the literature and, with it, the respective metadata. Finally, we employ a natural language processing (NLP) approach utilizing annotators from the Apache OpenNLP Tools Interface R package (Hornik, 2022), wrapped in the entity package (Rinker, 2023). The package calls functions from other NLP-based toolkits and thus identifies the type of entity to be annotated. In this case, we focus on location entities, which refer to geographic locations (countries or cities) matched from the references item and journal titles from small OA items.

Figure 2

Bibliometric Data Composition



All analyses are based on the intersection sample (common items), along which the exclusive items are also visualized and listed. Data wrangling and plot design were undertaken using R, with the dplyr (Wickham et al., 2023), stringr (Wickham, 2022) and ggplot2 (Wickham, 2016) packages. The term frequency analyses were carried out using stringr (Wickham et al., 2023) and cld2(Sites, 2013); the frequency estimations and entity recognition were completed using tm, stopwords and entity (Feinerer et al., 2023; Muhr et al., 2022; Rinker, 2023); and visual presentations were generated with igraph (Csárdi et al., 2024; Wickham, 2016). For the tf-idf, we also

[12] For robustness purposes, we checked the WoS and Scopus' OA status against the OA data from OpenAlex, which includes regular snapshots of the Unpaywall data. We used the August 31st, 2024 snapshot of OpenAlex and matched the individual items according to their (unique) DOIs. We first found that around 2 million items did not have any DOI (1.5 in Scopus and 0.5 in WoS Primary). Further, when looking at the filtered and matched OA categories, we found only of a 0.8% smaller share of OA items, on average (in their respective categories) in WoS Primary (0.6%) and Scopus (1.0%) in relation to the OpenAlex metadata. Moreover, as we do not consider the OA categories for our analyses, the ratio of correctly assigned OA statuses appears to be within a reliable margin – the share of OA vis-a-vis non-OA items is 49%, on average (with a slight overestimation in the OpenAlex data when respectively checked against Scopus and WoS Primary data: 52% vs. 45%, Scopus, and 51% vs. 47%, WoS Primary).

[13] Term frequencies represent the density distribution of the most commonly used words in a text string. In this case, we use item and journal titles as the source and decompose the strings into words, or tokens, and divide the sum of these between the total number tokens, per FoS and database (i.e., for t1 = nt1/N, where n is the sum for t1 and N is the sum of all terms in a FoS and database). Tokens can then be grouped and listed in reverse order, showing the proportion of appearance as a proxy for the generality or specificity of term use.

estimated these using bigrams (see Appendix).

Results

Overview of small publishers and their OA status

As seen in the diagram above, we obtained around 17 million articles and reviews in Scopus and around 14 million in WoS Primary (for a combined total of 31 million items). Of these, only 11 million are common to both databases (i.e., Scopus WoS), leaving 6 million unique items in Scopus and 3 million unique items in WoS Primary. Our focus primarily lies on this 11-million common subset, although Table 1 details both common and exclusive data subsets. Our final sample comprises 11,018,829 (9,736,441) articles and reviews from 11,400 (23,107) journals published by 357 (12,963) publishers– the values in parentheses (also in cursive) reflect the unique items from each database.[14] Table 1 further reveals strong concentration patterns: a limited group of dominant publishers shapes the landscape, reflecting both indexation bias and access type. Notably, the visibility of pure OA publishers in the common subset is low, accounting for 28% of the items (the exclusive items from WoS and Scopus account for 8% and 19%, respectively),[15] whilst articles from SOAP only comprised 0.5% of all common items (0.3% in WoS and 1.7% in Scopus).

Figure 3 (top panel) shows the distribution of publishers by database, size, access type, and field. Here, disciplinary differences are observable, which align with expectations already found in the literature (e.g., Mongeon & Paul-Hus, 2016). Unlike in Stephen & Stahlschmidt (2022), here the concentration of publishers is given in the health and medical and natural sciences instead of the social sciences and humanities (SSH). This distribution appears consistent across size and field, though SOAP shows more erratic patterns. These discrepancies suggest that larger publishers are more prevalent in specific fields like engineering, technology, or the humanities. Introducing the OA dimension amplifies variation across disciplinary subfields, particularly in SSH and natural sciences. Some of this inconsistency may stem from how publishers are classified. The relatively low number of publishers matched in both databases (357) could limit visual coherence and comparability across subsets.

Temporal fluctuations in publisher size are also evident and may result from changes in publication volume or inaccuracies in database records. SP, in particular, often face challenges sustaining regular publication flows—whether due to inconsistent submissions, issue scheduling, or other editorial burdens (again, see Knöchelmann (2023) for an overview of the editorial perspectives in independent journals). The OA dimension also imposes some restrictions on publishers that have not adopted cooperative agreements with libraries, societies, or a third-party funder, as reader payment (or, on the authors' side, article processing charges – APCs) may remain an operational necessity.

Figure 3 (bottom panel) shows a discipline-based representation of the fluctuations of publishers. The plot is divided into the levels of interest for our perspective, namely, the entire common subset (as a point of reference), and the SPs and SOAP subsets. It is clear from the patterns that disciplinary predominance breaks down the more restricted the subset becomes. For instance, in the more specific levels (e.g., small OA), publishing regularity may not follow common periodizations (monthly, bi-monthly, quarterly, biannual, or yearly). In addition to these and the other issues listed above, the small OA category is limited to a few topic-specific publishers (per year and per discipline) - a relevant epistemic practice that seems to determine the visibility and impact of smaller, independent publishers.

Figure 4 presents the relative changes in publishing output (unique items) by subset and category. A general slowdown is evident across all groups since 2014, but the most striking trend is the persistent decline in output from small publishers, both OA and non-OA. This confirms patterns seen in previous figures, particularly regarding the divergent indexation paths for SOAP across Scopus and WoS. Whilst the common subset points generally upwards in the later periods of the sample, the small (OA and non-OA) subset not only points mostly downwards but also shows the largest divergence. Scopus data registers a positive trend for small OA publisher items, whilst WoS records a negative one.

These contrasting patterns reflect deeper structural dynamics. As Knöchelmann (2023) notes, indexation serves as a visibility proxy, shaping editorial workflows and determining which forms of "process optimization" are pursued (p. 402). Our argument revolves around the pressure small, independent publishers have in relation to these practices and how visibility and impact coalesce to generate incentives (positive and/or negative) for editors and publishers alike to either adopt visibility-enhancing mechanisms or maintain a reliance on scholarly networks for academic impact and reputation. The trends observed thus far point towards the latter, as visibility within these databases is marginally limited for SP, especially for those editing purely OA content. Some reasons for these trends are arguably related to indexation practices and data completeness, which seem to negatively affect small (OA and non-OA) publishers.

[14] As seen in Table 1, these are the subsamples ScopusWoS' as well as Scopus'WoS. The nomenclature follows the structure: common items (sum of unique items) – e.g., 11,018,829 (9,736,441).

[15] In the whole sample, the majority of OA items belong, primarily, to large publishers with hybrid offerings. Hence the reduced numbers in Table 1. Our analyses further focus only on the strictly OA category based on the 90% threshold.

Figure 3

Publisher Distribution and Output

Table 1

*Item, Journal and Publisher Counts in WoS and Scopus by size and access type.**

Database	Common Items (Unique Items)†	Journals	Publishers‡	%
All				
Total	11,018,829 (9,736,441)	11,400 (23,107)	357 (12,963)	100.0
WoS	(3,023,539)	(2,854)	(3,120)	31.0
Scopus	(6,719,902)	(20,253)	(9,843)	69.0
Open Access Publishers				
Total	3,118,692 (2,692,403)	3,041 (9,765)	166 (6,160)	100.0
WoS	(798,712)	(1,100)	(1,435)	29.6
Scopus	(1,893,691)	(8,665)	(4,725)	70.4
Non-Open Access Publishers				
Total	7,755,515 (7,345,180)	9,838 (19,001)	303 (10750)	100.0
WoS	(2,371,252)	(2,760)	(2,710)	32.2
Scopus	(4,973,928)	(16,241)	(8,040)	67.8
Large Open Access Publishers				
Total	3,045,456 (2,524,415)	2,724 (8,452)	125 (4,311)	100.0
WoS	(784,395)	(1,018)	(1,075)	31.0
Scopus	(1,740,020)	(7,434)	(3,236)	69.0
Small Open Access Publishers				
Total	58,008 (198,677)	309 (1,678)	40 (1,969)	100.0
WoS	(29,561)	(148)	(378)	15.1
Scopus	(169,116)	(1,530)	(1,591)	84.9
Large Non-Open Access Publishers				
Total	7,648,362 (7,128,350)	9,356 (17,587)	243 (8,725)	100.0
WoS	(2,323,857)	(2,599)	(2,180)	32.6
Scopus	(4,804,493)	(14,988)	(6,545)	67.4
Small non-Open Access Publishers				
Total	82,559 (267,172)	458 (2,058)	54 (2,255)	100.0
WoS	(72,080)	(292)	(575)	27.1
Scopus	(195,092)	(1,766)	(1,680)	72.9

*The size classification is applied at the publisher level, hence, journals and items classified as small belong to the publishers classified as such. OA Publishers (small and large) refer to publishers within our 90% threshold. Percentages for each database are calculated regarding the total number of unique items per category (values in parentheses).
†Due to the lax OA classification (90% threshold), the sum of sub-categories is over-, underestimated with respect to the total of items, journals, and publishers.
‡Publisher overlap between both databases is low overall due to inconsistencies in main commercial and/or subsidiary names.

Assessing bibliodiversity

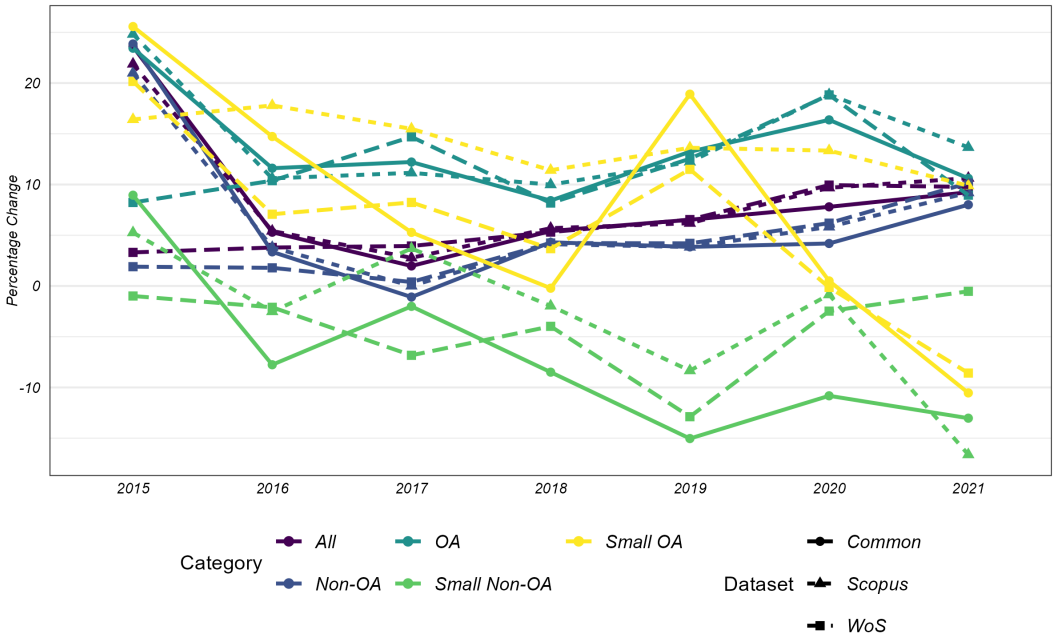
References and their coverage

Referencing practices are one of the key dimensions with which to identify knowledge circulation patterns within and between research communities. Internal coverage (IC) represents the share of "references cited within the studied publication sample" (Jappe, 2020, p. 12). High IC implies engagement with core disciplinary literature, while low IC suggests ties to peripheral or underrepresented scholarly domains. This is particularly relevant in disciplines like the SSH, where citation practices often include books, non-textual media, and local sources not indexed in bibliometric databases.[16] This

[16] As noted above, other repositories and databases (e.g., see SciELO in Latin America), highlight their relevance in promoting bib-

Figure 4

Yearly Changes in Publication Outputs, by Size and OA Type



is true in discipline-specific citing practices, such as in the SSH with the use of books, dissertations, historical texts, and non-written media.

Table 2 shows IC by publisher group. While small and large OA publishers cite a similar number of references per article, IC diverges: 66% of the references in small OA items are indexed, compared to 80% for large OA publishers. These dynamics are suggestive of diverse epistemic reasoning, where large OA publishers favour disciplinary discourses already found in mainstream journals (which they own or administer). Further, these coverage values also speak of specific citation patterns that point towards particular dimensions of knowledge production and transfer, such as the latent relevance of non-indexed sources for a specific category of publications (for reference, see both Hyland, 1999; Hyland & Jiang, 2017).

From a bibliodiversity lens, small journals have a particular function in promoting specific 'repertoires' that involve a sense of locality, both regional and disciplinary (Leonelli, 2022). Additionally, due to this locality (or localities), the referencing patterns of items published in small (OA) journals diverge from those found in journals from large publishers. The cited works may or may not differ regarding their scientific narratives. The crucial difference is not only their presence in bibliographic databases, but their field thematic relevance. This factor carries implications for their visibility and potential impact.

Highly cited literature

Citations embody the visibility and overall impact of scholarship when they are directly compared to similar items published in a particular year (here, we use a three-year window after publishing). The HC indicator (top 10% of cited publications per field/year) captures visibility and impact (Donthu et al., 2021). HC articles in small OA journals mediate scholarly communication by linking local knowledge to broader networks. Their impact stems from both the topic's relevance and its contribution to disciplinary discourse, especially for under-represented communities. Thus, we assume that HC items carry a specific value in relation to the topic they are presenting and to the field they are addressing. This approach is twofold: a) it is a reflection of visibility and impact, indicative of relevance, and b) it enables under-represented literature to enter into discussions that are potentially formative in disciplinary narratives (as well as contribute to consolidation).

Table 3 presents a descriptive summary of the HC from the small OA subset. The data shown are the total number of HC items, the counts of total and unique references in these articles, and the number and percentage of these references that are (non-)indexed in WoS and Scopus. The number of HC articles in each database is, in comparison to the total number of items in the overall sample, rather reduced; we tallied 58,008 (common) items in small OA publications, and

liodiversity through the same dimensions we consider: visibility and impact.

Table 2

Internal Reference Coverage in WoS and Scopus, by Publisher Size and Access Type.

Database	Total References*	Indexed References	Mean†	Median	IC‡
All					
Total	513,701,377	397,858,848	46.5	39	75.5
WoS	(134,890,768)	(101,855,712)	(44.6)	(39)	(73.5)
Scopus	(242,652,550)	(157,096,695)	(35.8)	(29)	(62)
Open Access Publishers					
Total	152,344,242	124,133,769	48.8	41	79.3
WoS	(37,931,026)	(31,091,347)	(47.4)	(40)	(79.8)
Scopus	(71,474,850)	(48,502,113)	(37.5)	(31)	(65.2)
Non-Open Access Publishers					
Total	355,027,526	268,962,426	45.7	38	74.1
WoS	(103,289,351)	(75,527,018)	(43.5)	(38)	(71.3)
Scopus	(177,724,788)	(113,368,980)	(35.3)	(29)	(61)
Large Open Access Publishers					
Total	149,457,956	122,230,750	49.0	41	79.7
WoS	(37,334,122)	(30,756,674)	(47.5)	(40)	(80.2)
Scopus	(66,300,865)	(45,933,206)	(37.8)	(31)	(66.7)
Small Open Access Publishers					
Total	2,287,291	1,496,407	39.4	32	64.5
WoS	(1,195,899)	(741,285)	(40.4)	(34)	(61.1)
Scopus	(5,782,152)	(2,975,823)	(33.8)	(28)	(50.3)
Large Non-Open Access Publishers					
Total	350,791,834	266,655,253	45.8	39	74.4
WoS	(101,662,667)	(74,897,216)	(43.7)	(38)	(72)
Scopus	(172,163,611)	(111,087,419)	(35.5)	(29)	(61.7)
Small Non-Open Access Publishers					
Total	3,221,020	1,748,187	39	32	53.5
WoS	(2,641,356)	(1,188,788)	(36.5)	(30)	(44.1)
Scopus	(6,644,401)	(2,838,763)	(32.6)	(26)	(41.5)

**Total references are taken from the intersection sample; references for each database are from unique exclusive items.*
†Average number of references per item (from Table 1).
‡IC: internal reference coverage rate.

the HC share of these, as in the table below, amounts to 1,760 (3% of the items in the small OA category). This number diverges from the logic of the HC, as it is not the expected ten percent value. However, this is the case due to the non-systematic classification in disciplines and years, as well as the sub-sample in general. The HC items from SOAP make up around 0.1% of the total HC items in the whole sample (around 1.1 million items, roughly 12% of the common subset, deviating again from the expected HC value for the same reasons).

Disciplinary distribution aligns with earlier figures: medical, health, and natural sciences dominate. However, WoS's HC-OA subset shows notable representation from the humanities (24%), while SSH accounts for 44% in Scopus—suggesting that small OA formats serve as crucial platforms for diverse epistemic contributions. Roughly 60% of all references in HC items are unique, meaning 40% are cited repeatedly.[17] This underscores the anchoring role of certain sources in knowledge consolidation (see Park et al., 2023),

[17]The FoS distribution of the OA-HC items is as follows: Agricul-

Table 3

Reference Description in HC Small OA Publications.

	WoS	%	Scopus	%	Total	%[1]
HC Items[2]	1,991	-	6,983	-	1,760	-
Total References	81,598	-	274,202	-	195,286	-
Unique References	76,697	100	252,687	100	116,049	100
Indexed References	53,643	70	179,529	71	85,772	74
Non-Indexed References	23,132	30	73,411	29	35,082	26

[1] *Percentages for each category (i.e., WoS, Scopus, Total)*
[2] *The Total column refers to the intersection sample, and each database count refers to the unique exclusive items in these. All items are counted at the three-year highly cited window.*

as well as in their disciplinary specificity, their entrenchment, permeability, and demarcation (2023, p. 993). To assess these characteristics, in the following analyses, we focus on the non-indexed v. indexed references distinction in order to shed light on the communicative dynamics that foster the inclusion of diverse voices in the most impactful scholarship of their respective areas.

The panels in Figure 5 show a) the references' indexation shares for each subset at the SOAP level (WoS, Scopus, and their common intersection); and b) the most common document types referenced for each field (for the common and exclusive items) and according to their (non-)indexation.[18] These results demonstrate that peer-reviewed materials represent the most relevant data sources for the HC items in every FoS. Non-peer-reviewed sources, however, are present in the SSH, as well as in natural and health sciences, where non-article material in journals, as well as online content, appears relevant. In addition, the plot presents several sources (or document types) that contribute to different disciplinary directions, as the field classification from the micro (item) level is aggregated to the macro one (document type classification). Similarly, a number of other document types are listed (also beyond the scope of the plot), which begs to question the type of referencing practices that each disciplinary convention requires. The presence of this type of seemingly marginal material supports our transmission mechanism claim, bridging the context-specific literature to the fore through citation. However, the extent to which this transmission actually contributes to diversifying the narratives and discourses is unclear, as the share of these sources is rather small.

Further to the point of language diversity, Table 4 shows the number of identified languages in the (non-) indexed references from SP' HC-OA items. The title's languages (from both item and journal) were automatically detected using the cld2 R package (Sites, 2013). For summary purposes, we aggregate the variety of languages listed as an output, and highlight the overwhelming distinction between English language references and those in other languages (non-English). We still acknowledge the inherent diversity as indicative of "the preservation of national languages in scholarly communication, and the societal usefulness and impact of new knowledge" (Giménez Toledo et al., 2019). Yet, the dominant Anglophone trend characterizes English as currently the lingua franca of academia.

Additionally, the long tail of diverse languages also illustrates the potential representation of specific regional academic communities. A linguistic variety of cited items is also indicative of a process of transmission, by which local knowledge is carried into the mainstream English-speaking scholarship. Nonetheless, we warn that the precision of the language identification must be taken cautiously, as the underlying algorithm used in cld2 "is not designed to do well on very short text, lists of proper names, part numbers, etc." (Sites, 2013), and is more precise with elements of at least 200 characters (approximately 30-40 words), while the cited titles vary from short phrases to long sentences.

Textual Analysis

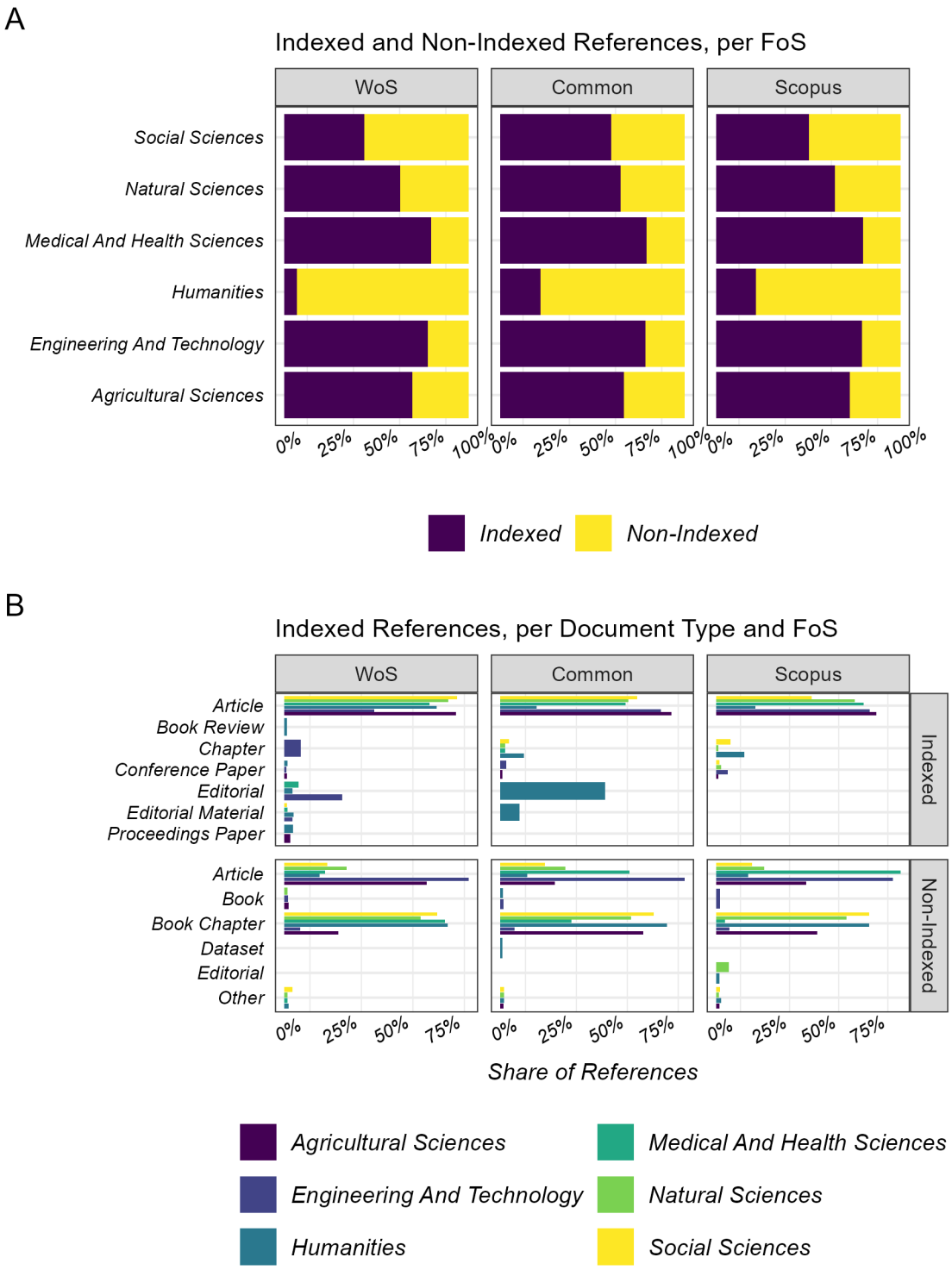
This section provides exploratory analyses beyond the journal or publisher containers and offers a disciplinary comparison of textual structures. From a bibliodiversity perspective,

tural Sciences (1.5% Scopus / 3.8% WoS), Engineering and Technology (1.6% Scopus / 12.7% WoS), Humanities (1.1% Scopus / 22.5% WoS), Medical and Health Sciences (46.2% Scopus / 15.1% WoS), Natural Sciences (43.5% Scopus / 33.1% WoS), and Social Sciences (6.2% Scopus / 12.7% WoS).

[18] For this step, we matched out SOAP items subset with metadata from OpenAlex' August, 2023 snapshot from the KB, to filter-match the non-indexed references in either WoS or Scopus. We queried their titles' text strings (as their identifiers are database-specific) in OpenAlex in order to get the document type information. We then classified the items according to our analytical structure, i.e., by Field of Science, by data subset (WoS, Scopus, or their intersection), and by indexation status. We matched 9.3% of the non-indexed references (41,764) in total, with 7% of those unique to WoS, 9.7% unique to Scopus, and 10% of the intersection (common) subset.

Figure 5

Indexed and Non-Indexed References by Field of Science in Common and Unique Items in WoS and Scopus



we underline the relevance of exploring latent thematic connections in journal and item titles. These connections are indicative of closeness and similarity in both field-specific works and linguistic and territorial dimensions. We base this approach on textual structures, focusing not only on which terms appear but also on the diversity with which these are utilized. As mentioned, we base this approach on two dimensions: a) a term frequency network, and b) an entity recognition. The network also links the term frequencies to a much clearer view of the centrality of specific topics and their relation to surrounding themes. Entity recognition provides valuable input for identifying regional distributions in both citing and cited items (and journals), which are often overlooked at first glance.

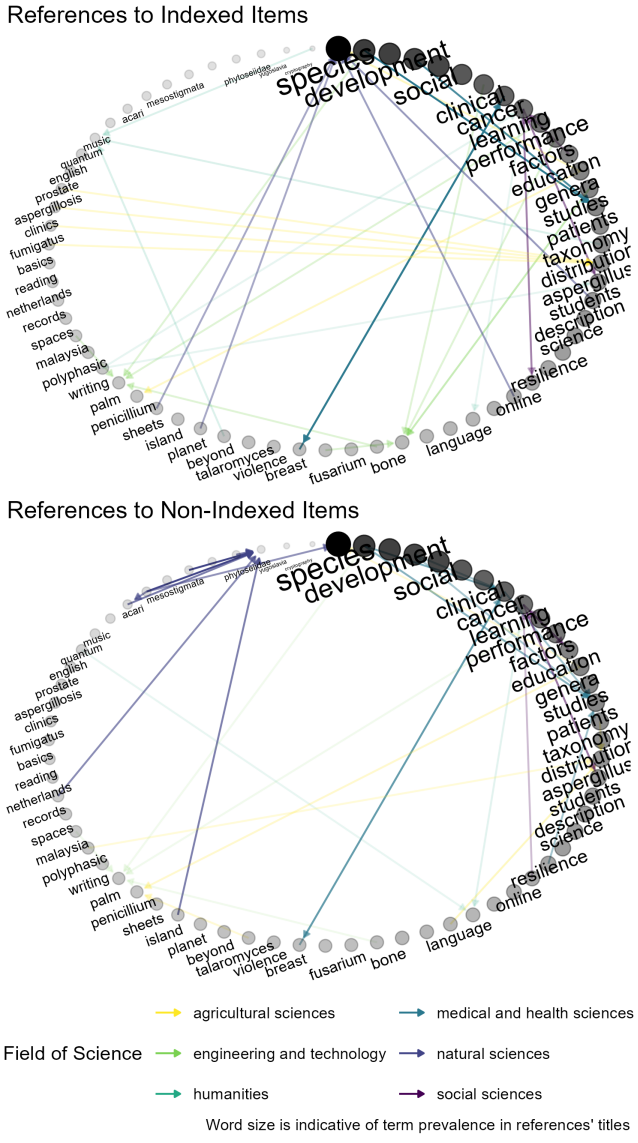
From a network perspective, we estimated the tf-idf for the titles of both citing and cited items common to both Scopus and WoS Primary, also subdividing them into the indexed and non-indexed categories, creating a reference link between them. Thus, we can observe directed relations between these terms (nodes) and the number of occurrences amongst them (edges - links or arrows joining them, i.e., the darker, the more common). Node and label size also indicate the frequency and relevance of the respective nodes. Figure 6 illustrates only the most common connections between these terms, highlighting their textual and narrative relevance. Data-wise, these nodes represent the top-10 words in each FoS by indexation category (indexed vs. non-indexed), resulting in 120 most salient terms and the links between them (i.e., bigrams). This visualization also highlights the apparent diversity of topics, which is partially lost when introducing the HC dimension. The topics range from socio-economic studies to biomedical research and specific biochemical areas.

Moreover, since we sampled the most common bigram pairings by discipline, we can maintain a certain representation of topics found in less-cited material (e.g., from the SSH). These disciplines are generally underrepresented in the databases, and this underrepresentation is magnified at the small journal and small HC-OA item levels. Additionally, in order to have more meaningful connections, we removed key terms that reflect specific modalities of knowledge (re-)production, indicative of scientific narratives and practices; viz., the terms "model", "analysis", and "review" (which would have occupied the top places in the visualization). These are arguably correlated with general epistemic practices, whereby replication and comparability are requisites for the acceptance and circulation of knowledge.

For our argument, the spread and sheer quantity of nodes is indicative of a broad palette of topics and approaches. Yet, the predominance of an apparent structure of knowledge (re-)production presents a Janus-faced dynamic in light of our transmission mechanism argument (visibility and impact). That is, these types of publications replicate established knowledge and discourses whilst effectively serving as

Figure 6

Term Network in Small and HC-OA Item and Reference Titles, in WoS and Scopus (Top 10 Terms per FoS)



vehicles for regional or language-specific publications to become part of a relevant scientific discourse. In general, these two aspects may indicate the hurdles to entering mainstream academic discourses, as consolidating research is much more prone to higher visibility and impact.

Now, we look at entities named in both the items' and the journals' names from the references (both indexed and non-indexed). Table 5 presents a list of the top locations named in each item title. We further match the name data with a list of countries sorted by their political-economic classification, i.e., the so-called Global South v. Global North categories.[19] In general, Global North entities are more commonly identified in the titles, indicative of a centrifugal tendency. However, there is a relevant count of title entities mentioning regions located in the Global South, reflecting specific contextual relevance.[20] We also highlight the presence of regional variety in the case of the non-indexed items' titles, actually pointing towards a transfer mechanism. Since other regions are also identified in the titles, we narrow down the classification (for instance, allocating the category "Global" to the entity "Earth" and "Other" to the entity "Arctic"). We take the entire subset of unique references (670k) from HC items in journals from SOAP, and group the first five entity pairings – for some fields of science, the match only results in four.

For our purposes, focusing on indexed and non-indexed references, the Global South→Global North dynamic points towards a pattern of bridging in academic outputs. That is, Global South topics are mostly addressed in Global North journals (regarding the global dynamics of knowledge, see Ma (2025), and the relevance of materially bridging spaces, or "mislocated centres" by Krawczyk & Kulczycki (2021b)). Now, whether this regional orientation is actually reflected in highly cited literature focused on these locations requires further research. Another plausible argument may be that authors choose to publish their work in these smaller journals due to the journal's general thematic focus or its representation within a specific scholarly community. The latter is inferable from the disconnection of entities identified in journal and item titles, which may go against a claim of regional representation in specific small OA journals. If this were the case, then further inspection would be necessary to adequately survey the small OA publisher landscape and identify the editorial parameters they follow (Knöchelmann [2023] offers a similar perspective on the diverging incentives of specific academic fields).

Discussion

The role of small OA academic publishers is multidimensional. They mostly represent particular research communities and, likewise, face challenges tied to tensions around visibility and impact. These challenges are both material (economic sustainability) and symbolic (recognition and relevance). Impact may be local or global, as well as rich in interdisciplinary perspectives; whilst visibility represents a mediated process, determined by editorial standards, indexation, scores, and rankings, amongst others. This interplay of incentives and demands creates tensions that involve the academic and commercial aspects of publishing.

We have explored how SPs contribute to a diverse publication landscape and whether their outputs shape dominant academic narratives. That is, fostering and integrating a "host of varied voices [which] can be heard" (Hawthorne, 2014, p. 12) within the most influential literature. Again, this issue encapsulates the clashing tensions between academic (reputational) and commercial (publishing business) incentive structures. We thus characterize SOAP as transmitters, bridging community- or topic-specific knowledge into a more mainstream array of academic discourses.

The argument is two-pronged: on one hand, highlighting and arguing the potential of OA as a tool for democratization (Knöchelmann, 2021) and expanded access (Aspesi & Brand, 2020) to literature on both the supply (authorship) and demand (readership) sides. On the other hand, taking small publishing organizations as key stakeholders in the transmission of specific content from fringe regional, linguistic, or disciplinary contexts to the mainstream academic discourses (Estelle, 2021; cf. Stahlschmidt & Stephen, 2022). As stated throughout the text, we refer to this approach as bibliodiversity (cf. Jussieu Call for Open Science and Bibliodiversity, 2017), and have emphasized its potential for promoting visibility and, hence, increasing the scholarly impact. Then, "bibliodiversity [. . .] seeks to empower the [Global] South in taking ownership of open access for knowledge creation and support through mutual assistance" (Berger, 2021, p. 385). Hence, visibility and impact have an incremental potential for both OA and diversity.

Our analyses evidence that the share of small OA publications is a dramatically slender portion of the overall sample; even fewer belong to the top performing items (HC). The HC items are further skewed in favour of the 'hard' sciences, likely due to distinct collaboration and publishing patterns and, presumably, more available funds as part of OA institutional agreements.[21] National and regional funders have more decisively opted to fund OA in their respectively fi-

[19] See the Network for International Policies and Cooperation in Education and Training's (NORRAG) document: https://www.norrag.org/wp-content/uploads/2023/02/List-of-Global-South-and-Global-North-Countries.pdf.

[20] In most cases, the title's reference to a geographic location represents a case study or application. This trend, however, is counteracted by journal titles, which are often broadly related to a country or region.

[21] Recent research has also shown that OA costs (in terms of APCs) become particularly prohibitive and increasing over time (L.-A. Butler et al., 2023). However, in contrast to the approach of the latter study, we focus only on SPs, which function under different incen-

nanced projects.[22] However, SPs get locked out of publishing agreements, leading them to often choose paywalled models, limiting their implementation of OA, evidenced here in the very low counts of published items.

By focusing on HC items, we examine how this metric reflects the scholarly relevance of certain subfields and narratives, as well as the epistemic practices within them. The HC literature helps us understand SPs' bridging function and the role of citation practices across fields of study (FoS). We highlight the role of referencing patterns as particular practices that, depending on topics, regions and, as shown, languages, help shape epistemic narratives. Thus, the HC-OA items represent a relevant microcosm for making certain disciplinary differences evident, while highlighting them as core characteristics of epistemic practices.

The indexing patterns in the references cited by the HC-OA items appear to be discipline-dependent for a larger proportion of non-indexed sources (e.g., SSH). In our analysis of visibility and impact, the use of non-indexed references is indicative of precise thematic foci that disregard age and access characteristics of the items cited. Top performing and thus highly visible items make constant use of preprints and grey literature, affecting indexation rates and complicating field-wide epistemic analysis.

Yet, as we have shown, there are numerous data inconsistencies at the more precise level of small OA publications. Compared to large publishers' publications, there are evident issues of data harmonization, particularly regarding metadata that can help identify OA and bibliodiversity. Currently, the discussion on database standardization is crucial, as other alternatives flourish (e.g., OpenAlex or OpenAIRE) and comparisons regarding completeness are coupled with quantity, breadth and, relevantly, representation.

Principles of representation, quality, visibility, and impact are cast into doubt given the overshadowing of SOAP by its larger counterparts. Larger publishers dominate the publishing game, both in general terms and also when only focused on OA publications (Shu & Larivière, 2024; cf. L.-A. Butler et al., 2023). We observe a rather limited impact of small OA literature within the greater scope of academic publishing. This impact is driven by the focus on citations and the incentives they generate regarding epistemic practices (including, but not limited to, internationalism, interdisciplinarity, and thematic diversity). HC items tend to reproduce and consolidate established knowledge, thus hindering the bridging potential given by existing linguistic representation and diverse references. However, the mere presence of these journals provides a silver lining of diversity and representation of different scholarly backgrounds.

Limitations

Our analysis presents several key limitations – some structural, others methodological. Structurally, both WoS Primary and Scopus reflect systemic biases favouring English-language and Western scholarship. These databases disproportionately represent core geopolitical and epistemic spaces, reinforcing global academic inequalities(Krawczyk & Kulczycki, 2021b)]. While this centrality skews our findings on bibliodiversity, it also serves as evidence of the databases' limitations: in their pursuit of quality and precision, they often neglect representation and diversity. Though initiatives like the Emerging Sources Citation Index and the inclusion of SciELO and the Korean Journal Database by Clarivate show progress, these indices, outside the WoS Core Collection and unavailable due to licensing constraints, were not part of our analysis.

From a coverage perspective, both databases lack a comprehensive representation of publishers, especially small OA ones. Since our study centers on indexation and citation patterns, this gap underscores how large bibliometric platforms can exclude or consolidate particular scholarly perspectives. Future research—especially quality audits—could further assess these omissions.

A second limitation relates to our simplified treatment of OA categories. We focused solely on the open–closed access distinction, omitting "OA colour" licenses (e.g., green, gold, hybrid). While this decision streamlined the analysis, it reduced our ability to assess how specific licensing affects citation and indexation. Given the small size of the HC-OA subset, however, a more granular licensing breakdown would likely yield limited insights. Future research could extend this focus.

Linguistic diversity also presents a challenge. Our language analysis reveals a heavy English-language bias, difficult to counter given the nature of the databases and the metadata available. Table 4 highlights this imbalance across all fields. Finally, by emphasizing highly cited (HC) literature as a proxy for impact, we focus on works that meet prevailing visibility norms. These are often systematic reviews or meta-analyses—document types that attract more citations and, in turn, enhance the visibility of smaller journals. This introduces a potential selection bias, as impact is narrowly defined by visibility within dominant scientific discourses.

Conclusion and outlook

This article investigates the role of SPs in the diversity of scientific publication landscapes, represented in bibliographic databases. By arguing that the visibility of scholarly output is determined by the presence of published works in databases such as WoS or Scopus, we problematize the role of

tives and commercial logic, for the most part (cf. Knöchelmann, 2023).

[22] In addition, the ERC Frontier Research programme has informed that the physical sciences and engineering have obtained €5.58 billion in H2020 funding, whilst the social sciences and humanities have about half of that, €2.80 billion.

specific journals in amplifying the reach of certain academic topics, and their inherent tensions between academic and commercial incentives and practices. We base our argument on the idea of bibliodiversity and expand on the characteristics that small, OA journals have vis-à-vis journals from larger (mostly commercial) publishers. Observing this bi-dimensional nature allows us to present and discuss the data from a perspective that characterizes the (epistemic) relevance and narrative of this specific type of scholarship.

We have shown that SPs, through their more direct contact with academic communities, can effectively amplify the diversity of the overall publishing landscape. Indeed, varied linguistic representation, as well as diverse geographical directions, are indicative of a biblio-diverse landscape. Yet, the impact, in terms of share of HC literature, represents a considerable hurdle for smaller OA publishers. Despite access to the scholarship, the works published in the journals of smaller publishers do not reach the scale of impact that the more visible journals of larger publishers can achieve.

We, however, have data limitations regarding the proportional presence of SPs in the databases utilized. As shown, their share of indexation is particularly low. This pattern skews the generalization of our findings, but it also underlines our analytical framework and thesis: as a matter of visibility and impact, small (OA) publishers face disadvantages relative to large (OA) publishers. Further research could thus pick up on the themes discussed here and present more specific bibliometric approaches to disentangle them. Finally, we argue that bibliodiversity should be the focus of more studies of this kind, focusing on science communication, publishing models, and innovation possibilities. Bibliodiversity can then become a common narrative that fosters and deepens open science.

References

Aspesi, C., & Brand, A. (2020). In pursuit of open science, open access is not enough. *Science*, *368*(6491), 574–577. https://doi.org/10.1126/science.aba3763

Beall, J. (2015). *Is SciELO a publication favela?* Emerald City Journal. https://www.emeraldcityjournal.com/2015/07/is-scielo-a-publication-favela/

Berger, M. (2021). Bibliodiversity at the Centre: Decolonizing Open Access. *Development and Change*, *52*(2), 383–404. https://doi.org/10.1111/dech.12634

Brainard, J. (2024). *Open for business: Authors are increasingly paying to publish their papers open access. But is it fair or sustainable?* https://www.science.org/content/article/pay-publish-model-open-access-pricing-scientists

Butler, L.-A., Hare, M., Schoenfelder, N., Schares, E., Alperin, J. P., & Haustein, S. (2024). An open dataset of article processing charges from six large scholarly publishers (2019-2023). *arXiv*. https://doi.org/10.48550/arXiv.2406.08356

Butler, L.-A., Matthias, L., Simard, M.-A., Mongeon, P., & Haustein, S. (2023). The oligopoly's shift to open access. How the big five academic publishers profit from article processing charges. *Quantitative Science Studies*, 1–33. https://doi.org/10.1162/qss_a_00272

Crossref. (2019). *Fact file: 2018-2019 annual report*. https://www.crossref.org/pdfs/annual-report-factfile-2018-19.pdf

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., Müller, K., Salmon, M., Antonov, M., & details, C. Z. I. igraph author. (2024). *Igraph: Network analysis and visualization*. https://cran.r-project.org/web/packages/igraph/index.html

Deutsche Forschungsgemeinschaft. (2022). *Wissenschaftliches Publizieren als Grundlage und Gestaltungsfeld der Wissenschaftsbewertung*. https://doi.org/10.5281/zenodo.6538163

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*(3), 1809–1831. https://doi.org/10.1007/s11192-015-1645-z

Estelle, L. (2021). Enabling smaller independent publishers to participate in open access transformative arrangements. *Septentrio Conference Series*, *4*. https://doi.org/10.7557/5.6220

Exclusive: MDPI journal undergoing reevaluation at scopus, indexing on hold. (2024). Retraction Watch. https://retractionwatch.com/2024/01/02/exclusive-mdpi-journal-undergoing-reevaluation-at-scopus-indexing-on-hold/

Feinerer, I., Hornik, K., Software, A., & Ghostscript), I. (pdf_info. ps taken from G. (2023). *Tm: Text mining package*. https://cran.r-project.org/web/packages/tm/index.html

Giménez Toledo, E., Kulczycki, E., Pölönen, J., & Sivertsen, G. (2019). *Bibliodiversity What it is and why it is essential to creating situated knowledge*. https://blogs.lse.ac.uk/impactofsocialsciences/2019/12/05/bibliodiversity-what-it-is-and-why-it-is-essential-to-creating-situated-knowledge/

Hawthorne, S. (2014). *Bibliodiversity: a manifesto for independent publishing* (1st publ). Spinifex Press.

Hornik, K. (2022). *openNLP: Apache OpenNLP tools interface* (Version 0.2-7) [Computer software]. Apache Software Foundation. https://cran.r-project.org/web/packages/openNLP/index.html

Huang, C.-K., Neylon, C., Montgomery, L., Hosking,

R., Diprose, J. P., Handcock, R. N., & Wilson, K. (2024). Open access research outputs receive more diverse citations. *Scientometrics*, *129*(2), 825–845. https://doi.org/10.1007/s11192-023-04894-0

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367. https://doi.org/10.1093/applin/20.3.341

Hyland, K., & Jiang, F. (Kevin). (2017). Points of Reference: Changing Patterns of Academic Citation. *Applied Linguistics*, *40*(1), 64–85. https://doi.org/10.1093/applin/amx012

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 20052019. *PLoS ONE*, *15*(4), e0231735. https://doi.org/10.1371/journal.pone.0231735

Jussieu call for open science and bibliodiversity. (2017). https://jussieucall.org/jussieu-call/

Kaier, C., & Lackner, K. (2019). Open Access aus der Sicht von Verlagen: Ergebnisse einer Umfrage unter Wissenschaftsverlagen in Deutschland, Österreich und der Schweiz. *Bibliothek Forschung und Praxis*, *43*(1), 194–205. https://doi.org/10.1515/bfp-2019-2008

Knöchelmann, M. (2021). The Democratisation Myth: Open Access and the Solidification of Epistemic Injustices. *Science & Technology Studies*, *34*(2), 65–89. https://doi.org/10.23987/sts.94964

Knöchelmann, M. (2023). Herausgeberschaft und Verantwortung: Über die Un-/Abhängigkeit wissenschaftlicher Fachzeitschriften. *Bibliothek Forschung Und Praxis*, *47*(2), 393–406. https://doi.org/10.1515/bfp-2022-0090

Krawczyk, F., & Kulczycki, E. (2021a). How is open access accused of being predatory? The impact of beall's lists of predatory journals on academic publishing. *The Journal of Academic Librarianship*, *47*(2), 102271. https://doi.org/10.1016/j.acalib.2020.102271

Krawczyk, F., & Kulczycki, E. and. (2021b). On the geopolitics of academic publishing: The mislocated centers of scholarly communication. *Tapuya: Latin American Science, Technology and Society*, *4*(1), 1984641. https://doi.org/10.1080/25729861.2021.1984641

Laakso, M., & Multas, A.-M. (2022). *European scholarly journals from small- and mid-size publishers in times of open access: Mapping journals and public funding mechanisms*. Zenodo. https://doi.org/10.5281/ZENODO.5909512

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PLOS ONE*, *16*(6), e0253129. https://doi.org/10.1371/journal.pone.0253129

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLOS ONE*, *10*(6), e0127502. https://doi.org/10.1371/journal.pone.0127502

Leonelli, S. (2022). Open Science and Epistemic Diversity: Friends or Foes? *Philosophy of Science*, *89*(5), 991–1001. https://doi.org/10.1017/psa.2022.45

Leonelli, S., Spichtinger, D., & Prainsack, B. (2015). Sticks *and* carrots: encouraging open science at its source: Encouraging open science at its source. *Geo: Geography and Environment*, *2*(1), 12–16. https://doi.org/10.1002/geo2.2

Ma, L. (2025). The limits of openness: Global knowledge production and its boundaries. *Journal of Documentation*, *81*(7), 121–134. https://doi.org/10.1108/JD-09-2024-0237

Ma, L., Buggle, J., & O'Neill, M. (2023). Open access at a crossroads: Library publishing and bibliodiversity. *Insights*, *36*, 1–8. https://www.proquest.com/docview/2816149846/abstract/A13183CCC43E4801PQ/1

Manley, S. (2019). Predatory journals on trial: Allegations, responses, and lessons for scholarly publishing from FTC v OMICS. *Journal of Scholarly Publishing*, *50*. https://doi.org/10.3138/jsp.50.3.02

Matthias, L., Jahn, N., & Laakso, M. (2019). The Two-Way Street of Open Access Journal Publishing: Flip It and Reverse It. *Publications*, *7*(2), 23. https://doi.org/10.3390/publications7020023

McGuigan, G., & Russell, R. (2008). The business of academic publishing: A strategic analysis of the academic journal publishing industry and its impact on the future of scholarly publishing. *Electronic Journal of Academic and Special Librarianship*, *9*. https://digitalcommons.unl.edu/ejasljournal/105/

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Muhr, D., Benoit, K., & Watanabe, K. (2022). *Quanteda/stopwords*. Quanteda Initiative. https://github.com/quanteda/stopwords

Nishikawa-Pacher, A. (2022). Who are the 100 largest scientific publishers by journal count? A webscraping approach. *Journal of Documentation*, *78*(7), 450–463. https://doi.org/10.1108/JD-04-2022-0083

Nobes, A., & Harris, S. (2023). Open access in low- and middle-income countries: Attitudes and experiences of researchers. *Emerald Open Research*, *1*. https://doi.org/10.1108/EOR-03-2023-0006

OECD. (2007). *Revised field of science and technology (FoS) classification in the frascati manual* (DSTI/EAS/STP/NESTI(2006)19/FINAL). Organ-

isation for Economic Cooperation; Development. https://www.oecd.org/science/inno/38235147.pdf

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x

Pinfield, S., Wakeling, S., Bawden, D., & Robinson, L. (2021). *Open Access in Theory and Practice | The Theory-Practice Relationship*. Routledge. https://www.taylorfrancis.com/pdfviewer/

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375. https://doi.org/10.7717/peerj.4375

Rinker, T. (2023). *Trinker/entity*. https://github.com/trinker/entity

Rowley, J., Sbaffi, L., Sugden, M., & Gilbert, A. (2020). Factors influencing researchers' journal selection decisions. *Journal of Information Science*, *48*, 321–335. https://doi.org/10.1177/0165551520958591

Secretariat on Responsible Conduct of Research. (2021). *Tri-agency framework: Responsible conduct of research*. Canadian Institutes of Health Research (CIHR); Natural Sciences; Engineering Research Council of Canada (NSERC); Social Sciences; Humanities Research Council of Canada (SSHRC). https://rcr.ethics.gc.ca/eng/documents/framework-cadre-2021-en.pdf

Severin, A., Egger, M., Eve, M. P., & Hürlimann, D. (2020). Discipline-specific open access publishing practices and barriers to change: an evidence-based review [Version 2, peer review: 2 approved, 1 approved with reservations]. *F1000Research*, *7*(1925). https://doi.org/10.12688/f1000research.17328.2

Shu, F., & Larivière, V. (2024). The oligopoly of open access publishing. *Scientometrics*, *129*(1), 519–536. https://doi.org/10.1007/s11192-023-04876-2

Sites, D. (2013). *cld2: Compact language detector 2*. CLD2Owners. https://github.com/CLD2Owners/cld2

Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of science, scopus, and dimensions as varying resonance chambers. *Scientometrics*, *127*, 2413–2431. https://doi.org/10.1007/s11192-022-04309-6

Stephen, D., & Stahlschmidt, S. (2022). *Landscape study of small journal publishers for the knowledge exchange task & finish group for "small publishers and the transition to open access"*. Zenodo. https://doi.org/10.5281/ZENODO.7258048

Tenopir, C., Dalton, E., Fish, A., Christian, L., Jones, M., & Smith, M. (2016). What Motivates Authors of Scholarly Articles? The Importance of Journal Attributes and Potential Audience on Publication Choice. *Publications*, *4*(3), 22. https://doi.org/10.3390/publications4030022

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer Verlag. https://ggplot2.tidyverse.org

Wickham, H. (2022). *Stringr: Simple, consistent wrappers for common string operations* (Version 1.5.1) [Computer software]. https://github.com/tidyverse/stringr

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation* [Computer software]. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr

Table 4

Language Prevalence in Indexed and Non-Indexed References[1]

		WoS		Scopus		Common	
Fields of Science	Coverage	n	%	n	%	n	%
English							
Agricultural Sciences	Indexed	2,405	99.54%	12,108	99.82%	11,880	100.00%
Engineering and Technology	Indexed	2,660	99.59%	13,338	99.20%	5,344	100.00%
Humanities	Indexed	779	75.85%	9,709	78.93%	705	95.01%
Medical and Health Sciences	Indexed	17,311	99.08%	72,969	97.81%	57,541	99.67%
Natural Sciences	Indexed	29,276	99.49%	49,904	99.02%	57,722	99.98%
Social Sciences	Indexed	3,798	87.21%	39,275	93.97%	17,760	97.48%
Agricultural Sciences	Non-Indexed	637	86.90%	2,718	82.76%	3,188	80.93%
Engineering and Technology	Non-Indexed	457	92.70%	1,935	96.03%	716	92.63%
Humanities	Non-Indexed	1,940	41.68%	10,846	58.52%	515	59.95%
Medical and Health Sciences	Non-Indexed	2,270	95.82%	10,441	94.64%	6,728	96.27%
Natural Sciences	Non-Indexed	8,647	79.24%	13,678	83.73%	12,590	79.65%
Social Sciences	Non-Indexed	2,084	78.85%	16,298	79.55%	4,318	85.71%
Non-English							
Agricultural Sciences	Indexed	11	0.46%	22	0.18%	0	0.00%
Engineering and Technology	Indexed	11	0.41%	107	0.80%	0	0.00%
Humanities	Indexed	248	24.15%	2,592	21.07%	37	4.99%
Medical and Health Sciences	Indexed	160	0.92%	1,637	2.19%	189	0.33%
Natural Sciences	Indexed	149	0.51%	496	0.98%	13	0.02%
Social Sciences	Indexed	557	12.79%	2,522	6.03%	459	2.52%
Agricultural Sciences	Non-Indexed	96	13.10%	566	17.24%	751	19.07%
Engineering and Technology	Non-Indexed	36	7.30%	80	3.97%	57	7.37%
Humanities	Non-Indexed	2,715	58.32%	7,688	41.48%	344	40.05%
Medical and Health Sciences	Non-Indexed	99	4.18%	591	5.36%	261	3.73%
Natural Sciences	Non-Indexed	2,266	20.76%	2,657	16.27%	3,217	20.35%
Social Sciences	Non-Indexed	559	21.15%	4,190	20.45%	720	14.29%

[1]References in *Small HC-OA* Items in Exclusive and Common Data Subsets. WoS and Scopus refer to each exclusive data subset.

Table 5

Top Five Locations Identified in References from HC-OA Items (Indexed and Non-Indexed References in WoS and Scopus, per Field of Science)[1]

Coverage in Databases	Locations	n
Agricultural Sciences		
Indexed	Country in Global North	802
Indexed	Country in Global South	701
Non-Indexed	Country in Global North	216
Indexed	Region in Global North	175
Indexed	Other	146
Engineering and Technology		
Indexed	Country in Global South	307
Indexed	Country in Global North	193
Indexed	Region in Global South	47
Non-Indexed	Country in Global South	47
Indexed	Global	30
Humanities		
Non-Indexed	Country in Global South	269
Indexed	Country in Global South	206
Indexed	Region in Global North	191
Non-Indexed	Region in Global North	172
Indexed	Region in Global South	161
Medical and Health Sciences		
Indexed	Country in Global North	4,204
Indexed	Country in Global South	1,668
Indexed	City in Global North	867
Indexed	Region in Global North	699
Non-Indexed	City in Global North	276
Natural Sciences		
Indexed	Country in Global South	3,283
Indexed	Country in Global North	2,316
Indexed	Region in Global North	1,597
Non-Indexed	Country in Global North	1,276
Non-Indexed	Country in Global South	1,227
Social Sciences		
Indexed	Country in Global North	1,210
Indexed	Region in Global North	984
Indexed	Country in Global South	840
Non-Indexed	Country in Global South	498
Non-Indexed	Region in Global North	293

[1]Total number of unique references (*N*) = 673,652 (351,540 in Scopus and 100,798 in WoS — 221,314 are common to both databases).