Replication: Hybrid Open Access in Transformative Agreements

Najko Jahn[1*] (0000-0001-5105-1463)

[1] Göttingen State and University Library, University of Göttingen, Germany.

[*] Correspondence: najko.jahn@sub.uni-goettingen.de

Abstract

Keywords: hybrid open access, transformative agreements, scholarly publishing, big deals, bibliometrics

1 Introduction

This study aims to demonstrate the suitability of open scholarly data sources for assessing the impact of transformative agreements on hybrid open access. To achieve this, a replication study was conducted by comparing results from hoaddata, an openly available and continuously updated dataset on hybrid open access uptake based on Crossref, OpenAlex, and the cOAlition S Journal Checker Tool, with the established bibliometric databases Web of Science and Scopus.

This study focuses on the coverage of hybrid journal portfolios included in transformative agreements between 2019 and 2023. Special attention is given to potential differences in open access uptake by country when comparing first-author affiliation data to corresponding authorships. This is crucial because the lack of publicly available invoicing data corresponding to authorships plays an essential role in determining whether an open-access article is supported through transformative agreements. Data on corresponding authorships have been available on the Web of Science and Scopus for much longer than in open databases such as OpenAlex, where this information is still being roled out at the time of writing. Because of this weakness, open approaches such as hoaddata and related research use first-authorship data instead.

By conducting a large-scale comparative analysis, this study aims to

1. Determine the strengths and weaknesses of using open data sources in monitoring the impact of transformative agreements on hybrid open access publishing.
2. Assess the coverage and accuracy of open data sources compared with established bibliometric databases.
3. Evaluate the reliability of first author affiliation data as a proxy for corresponding authorship in the context of open access uptake analysis.

2 Background – Evidence base to measure the effects of transformative agreements

2.1 Anforderungen an das Monitoring

- esac guidelines
- gemeinsamkeiten und unterschiede zu apc (listenpreise, tatsächliche zahlungen, zentrales invoicing, rabatte, waivers)
- insitutionen covern cas, jedoch kann es zu unterschiedlichen verrechnugnsformen führen (antielig mit förderer, splitting innerhaklbd er einrichtung)

2.2 Bibliometrische Evidenzen

- allgmeeiner uptake
- wachstum apcs
- wachstum verträge (konsortien, forschung)
- konsequenzen

3 Data and methods

This study investigates the suitability of open scholarly data sources for assessing the impact of transformative agreements on hybrid open access. As shown in Figure 1, the methodology involved comparing hoaddata, an openly available collection of open research information on hybrid open access, with the bibliometric databases Web of Science and Scopus. This section introduces the initial data sources, followed by a presentation of the necessary data processing steps to obtain eligible articles enabled by transformative agreements using author roles (first and corresponding) and harmonised affiliation data.

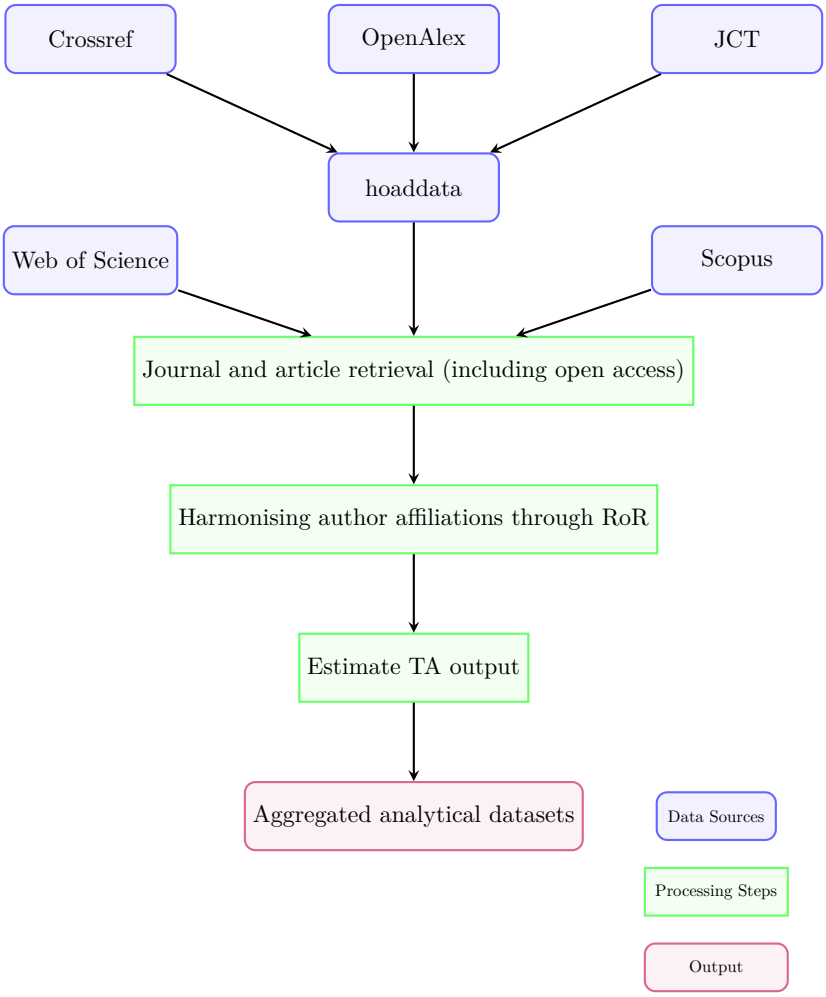

Figure 1. Data processing workflow for comparing hybrid open access uptake across bibliometric data sources. The workflow shows how data from different sources (hoaddata, derived from Crossref, OpenAlex, Transformative Agreement Data dump used by the cOAlition S Journal Checker Tool (JCT), the Web of Science, and Scopus) is processed and matched using ROR IDs to enable comparative analysis. TA = Transformative Agreement.

3.1 Data sources

hoaddata. hoaddata, developed and maintained by the author to support open access monitoring and research (Jahn, 2025), is an R data package that regularly collects information on hybrid open access uptake from multiple openly available data sources (Jahn, 2024). It combines article-level metadata from Crossref (Hendricks et al., 2020) and affilation metadata from OpenAlex (Priem et al., 2022) with transformative agreement information from Transformative Agreement Data dump used by the cOAlition S Journal Checker Tool (JCT)[1], which links journal and institutional data about participating research organisations to agreements in the ESAC registry.

hoaddata follows good practices for computational reproducibility using R (Marwick et al., 2018). The package, which includes data, code, a test suite and documentation, is openly available on GitHub. To ensure computational reproducibility while aggregating the data, a GitHub Actions continuous integration and delivery (CI/CD) workflow handles data retrieval from the SUB Göttingen's open scholarly data warehouse based on Google BigQuery, which provides high-performant programmatic access to monthly snapshots of Crossref and OpenAlex.[2] The workflow has run regularly to fetch updates from these data sources since 2022. The package version used in this study is 0.3, containing data from the Crossref 2024-08 dump provided to Crossref Metadata Plus subscribers and the OpenAlex 2024-08-29 monthly dump. It covers agreements collected between July 2021 to July 2024 from the JCT. This version including the computation log is available on GitHub (https://github.com/subugoe/hoaddata/releases/tag/v.0.3).

Web of Science. Clarivate Analytics' Web of Science (WoS) is a well-established proprietary bibliometric database consisting of several collections (Birkle et al., 2020). Web of Science is a selective, base research-focused database with dedicated panels to choose journals for inclusion (Stahlschmidt & Stephen, 2022; Visser et al., 2021). The collections considered in this study were the Science Citation Index Expanded (SCIE), the Social Sciences Citation Index (SSCI) and the Arts & Humanities Citation Index (AHCI).

These collections provides important data points for analysing open access: author affiliations and roles, differentiation of journal articles into document types representing different types of journal contributions, such as original articles or reviews, and open access status information derived from OurResearch's Unpaywall (Piwowar et al., 2018), the same provider as OpenAlex. However, the Web of Science lacks information about journals and articles under transformative agreements.

For programmatic access to article-level data, this study used the database of the Kompetenznetzwerk Bibliometrie (KB) in Germany. The KB processes raw XML data provided by Clarivate Analytics, which is ingested into an in-house PostgreSQL database under a uniform schema.

[1] `https://journalcheckertool.org/transformative-agreements/`

[2] `https://subugoe.github.io/scholcomm_analytics/data.html`

To support reproducibility, KB maintains annual snapshots of the database. Accordingly, this study used the annual snapshot from April 2024 (wos_b_202404), which is considered to cover almost the entire previous publication year (Schmidt et al., 2024).

Scopus. Elsevier's Scopus, launched in 2004, is another widely used proprietary bibliometric database for measuring research (Baas et al., 2020). Similar to the Web of Science, Scopus is selective with regard to the journals it indexes. However, its journal coverage is much broader than that of the Web of Science collections considered in this study, as it also indexes a wider range of applied research journals (Singh et al., 2021; Stahlschmidt & Stephen, 2022; Visser et al., 2021). With detailed metadata about article types, open access status information derived from Unpaywall, author roles, and disambiguated affiliations, Scopus also contains important data to assess open access uptake, although no direct information regarding transformative agreements was available at the time of the study.

This study used the Scopus annual snapshot of April 2024 as provided by the KB (scp_b_202404). The same KB curation effort was applied to the Scopus raw data as for the Web of Science (Schmidt et al., 2024).

3.2 Data processing steps

Determining hybrid journal publication volume. Following Jahn (2025), the starting point was a unified dataset of several safeguarded JCT snapshots[3], as provided by hoaddata. The JCT journal data in hoaddata were enriched with ISSN variants linked to an ISSN-L. To identify hybrid journals, a comprehensive exclusion of fully open access journals was performed using multiple journal lists (DOAJ, OpenAlex, Bielefeld). These data were used to determine the publication volume of hybrid journals for each database independently.

hoaddata relies on Crossref for obtaining journal publication volume and open access status through Creative Commons licence information relative to the published version ("version of record"). The article metadata included DOIs, publication dates, open access information as well as author roles and institutional affiliations. Publication years were determined using the earliest known date of publication in a journal. In hoaddata, this corresponded to Crossref's issued date. For Web of Science and Scopus, the earliest publication date was used where available, with Scopus dates specifically determined by the KB through version tracking of the raw data.

Many transformative agreements typically cover only certain types of journal articles, in particular original research articles including reviews (Borrego et al., 2021). Because of limited information on these document types in open scholarly data (Haupka et al., 2024), hoaddata used an extended version of Unpaywall's paratext recognition approach to exclude non-scholarly content (Piwowar et al., 2018). To exclude conference supplements, which are also often not covered by transformative agreements, only articles published in regular issues, indicated by numerical pagination,

[3] `https://github.com/njahn82/jct_data`

were considered. For Web of Science and Scopus, their established, mainly accurate document type classifications (Donner, 2017; Maisano et al., 2025) were used to identify original research articles and reviews, referred to as original articles throughout this study.

Identifying open access articles in hybrid journals. Articles in hybrid journals were considered open access when they were made freely available under a Creative Commons license on publishers' platforms. While hoaddata sourced this information from Crossref license metadata, Web of Science and Scopus relied on Unpaywall as evidence source. Unpaywall also uses Crossref license metadata, but supplements them by parsing publisher websites directly, addressing cases where publishers do not provide machine-readable Creative Commons license information (Piwowar et al., 2018). This additional parsing remains necessary despite transformative agreement workflows recommend the depositing license information during DOI registration (Geschuhn & Stone, 2017). Both Web of Science and Scopus defined hybrid open access consistently as content available under Creative Commons licenses on publisher platforms according to their documentations[4][5], distinguishing it from bronze open access that lack such explicit license information, or use publisher-specific licenses.

Harmonising author affiliations across databases. Author affiliations were retrieved for both first and, if available, corresponding authors to prepare the linking between articles and institutions covered by transformative agreements. To improve the data retrieval, JCT institution data was enriched with ROR-IDs from associated institutions, such as university hospitals or institutes of large research organisations such as the Max Planck Society, according to OpenAlex' institution entity. To handle different address variants, database-specific affiliation identifiers were used: ROR-IDs from OpenAlex for hoaddata, affiliation enhanced names for Web of Science, and Scopus Affiliation Identifier. Additionally, ISO country codes were retrieved for each author's address to compile country-level statistics. These country codes for Web of Science and Scopus were provided by the KB.

Because neither Web of Science nor Scopus support ROR IDs, the institution identifier used by the JCT, a two-step matching process was implemented to harmonise affiliation data. First, 2,782,540 articles from 6,457 institutions with ROR-IDs in the JCT data since 2017 (according to hoaddata) were processed to map first authors' ROR-IDs to corresponding proprietary affiliation identifier in Web of Science and Scopus using DOI matching. Then, an algorithm selected the most frequent ROR ID and proprietary identifier pairs to handle multiple affiliations and organisational hierarchy differences.

This process linked 6,375 ROR IDs to 4,894 Scopus Affiliation IDs, and 6,034 ROR IDs to 2,422 enhanced affiliation strings in the Web of Science. Quality evaluation through random sampling of 50 pairs revealed an error rate of 22% for Web of Science (11 mismatches) and 6% for Scopus (3 mismatches). Upon inspection, these mismatches primarily occurred with less-represented institutions

[4] https://webofscience.help.clarivate.com/en-us/Content/open-access.html

[5] https://blog.scopus.com/posts/scopus-filters-for-open-access-type-and-green-oa-full-text-access-option

having only a few publications, introduced through multiple affiliations of single authors. The difference between databases suggests that Scopus's affiliation control aligns more closely with ROR than that of the Web of Science.

Estimating open access in hybrid journals covered by transformative agreements. Based on these generated matching tables, articles eligible under transformative agreements could also be retrieved from Web of Science and Scopus, although they did not contain the ROR IDs used by the JCT. The estimation of eligible articles followed Jahn (2025) and included a matching of both journals and participating institutions. The matching also took into account the duration of agreements according to the ESAC registry, with only those matches where an agreement was actually in place being considered for subsequent analysis. A related study (Jonge et al., 2025), applied to publications funded by the Dutch Research Council (NWO) and validated against internal invoicing data, confirmed that such matching can accurately identify most articles under transformative agreements.

3.3 Data records

As a result of the comprehensive data processing described above, datasets on open access in hybrid journals included in transformative agreements were aggregated for each database at country and journal level by year. Table 1 provides a general overview of the coverage between 2019 and 2023. It shows that the majority of hybrid journals published at least one original research article or review marked as original during the five-year period. These journals formed the basis for the subsequent calculation of article-level indicators.

While hoaddata only covers articles with a DOI, Scopus and Web of Science publication indicators were calculated using their database identifier. A subsequent comparison of DOI coverage shows that non-original articles in Web of Science often lacked a DOI. This was particularly the case for meeting abstracts, which are notably prevalent in Health Sciences journals (Melero-Fuentes et al., 2025) and are not indexed by Scopus (Donner, 2017). Open access indicators were aggregated by DOI, as Unpaywall only collects information on open access status for articles with a DOI. A closer look at original articles with affiliation data, and in line with related research (Zhang et al., 2024), reveals a lack of affiliation data in the case of OpenAlex, the affiliation data source used by hoaddata, compared to Web of Science and Scopus. In particular, only about two-thirds of the articles examined provided corresponding author affiliations. For first authors, the proportion was 89%. At the time of writing, OpenAlex disclosed limited coverage of corresponding authorhip data [6]. Therefore, only first author data for hoaddata were considered in the following analysis.

[6] https://docs.openalex.org/api-entities/works/work-object/authorship-object#is_corresponding

Table 1

Coverage of hybrid journals in transformative agreements 2019-23.

	hoaddata*	Web of Science	Scopus
Hybrid journal coverage			
Active journals	12,890	8,655	11,888
Active journals (core)	12,888	8,655	11,878
Active journals (core) with OA	11,348	8,392	11,313
Publication volume			
Total published articles	9,740,015	8,616,053	8,117,644
Original articles	8,158,425	6,708,083	7,317,703
Digital Object Identifier (DOI) coverage			
Articles with DOI	9,740,015	7,713,796	8,105,112
Original articles with DOI	8,158,425	6,695,661	7,314,327
Open Access (OA) metrics			
OA articles	998,699	1,112,758	974,099
Original OA articles	969,817	1,019,784	922,578
Original articles with affiliation data			
First author articles	7,242,542	6,294,855	7,232,017
Corresponding author articles	5,534,207	6,291,441	6,898,487

* Journal article metadata from Crossref, except affiliations from OpenAlex

3.4 Data analysis

4 Results

This section first presents the indexing coverage of hybrid open access by comparing the data from open data sources with the proprietary bibliometric databases Scopus and Web of Science. Then, using the same methods, indicators at the publisher and country level are calculated independently for each database and compared with each other to assess the suitability of the bibliometric databases for investigating transformative agreements.

4.1 Coverage comparision

Overview. Figure 2 presents the coverage of hybrid journals included in transformative agreements, visualising the intersections of journals and articles across the examined databases as an UpSet graph (Krassowski, 2020; Lex et al., 2014). The analysis included hybrid journals that published at least one open access article between 2019 and 2023, based on open access status information from

each database. Only original research articles and reviews were considered in the analysis.

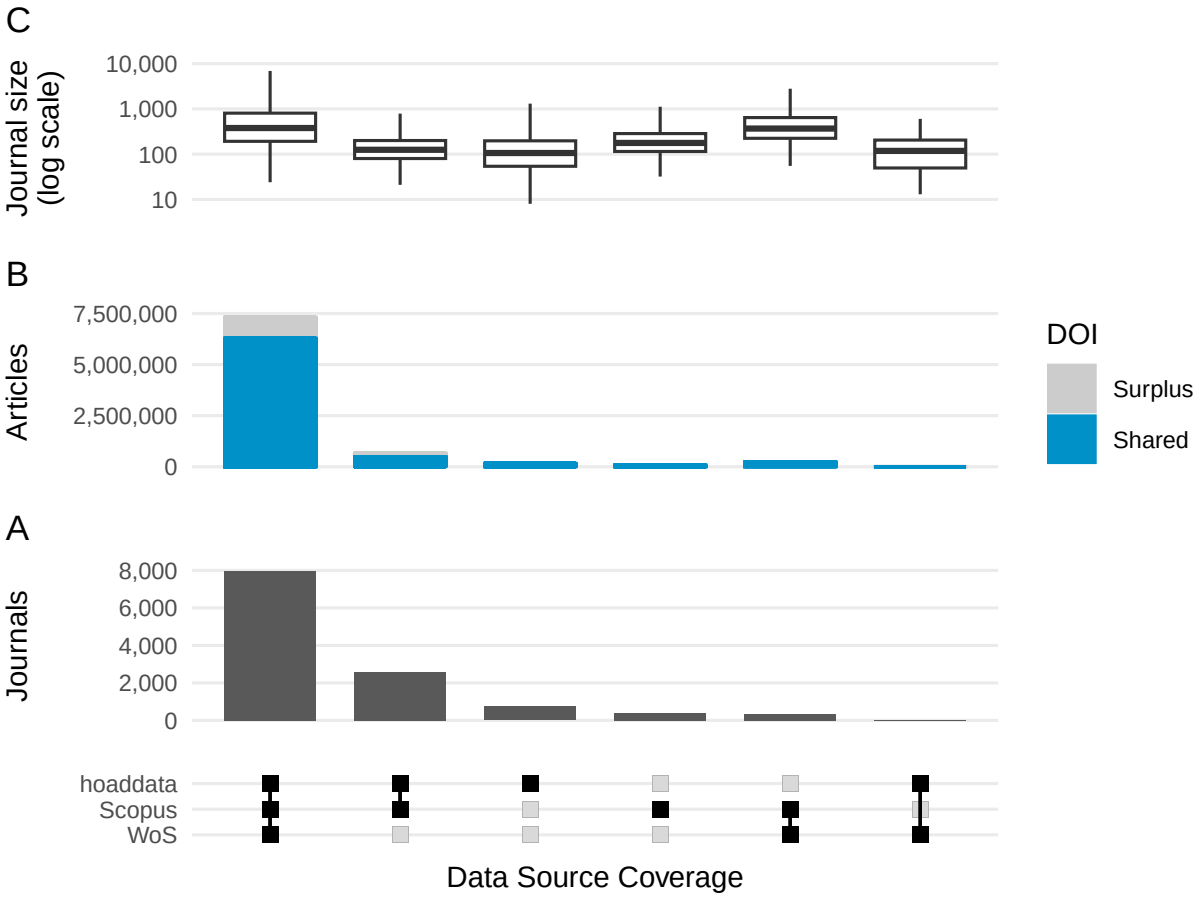

Figure 2. Upset graph

Journal coverage analysis revealed that 66% (n = 7,970) of hybrid journals included in transformative agreements were indexed in all three databases (Figure 2A). The second-largest set consisted of journals indexed in both hoaddata and Scopus, comprising 21% (n = 2,595) of hybrid journals. Notably, 6% (n = 739) of journals were exclusively contained in hoaddata, while another 6% (n = 748) were only found in the proprietory database Scopus. Of these, 354 were also available in the Web of Science. Upon inspection, this group of hybrid journals exclusively covered in the proprietary data sources mainly represented hybrid journals for which no open access evidence could be retrieved from Crossref, the open access evidence source for hoaddata.

In terms of article coverage, Figure 2B shows the total publication volume per combination in terms of DOI availability. The largest set of hybrid journals, which includes all three data sources, also contains the largest number of articles. In total, these journals recorded 6,289,687 overlapping articles, represented by the blue bar. They represented 94% of original articles with DOI indexed in the Web of Science sample, and 86% in Scopus. Another 657,697 articles were exclusive to both Scopus and hoaddata. Exclusively in hoaddata were 177,110 articles, and exclusively in the proprietary databases were 325,194 articles.

Figure 2B also shows the surplus of articles with DOI that were only available via hoaddata (grey area). In case of hybrid journals covered by all three data sources, 1,023,882 DOIs were only present in hoaddata. After validation at the DOI level using the KB databases and manual inspection, the main reasons for missing DOI coverage in the proprietary database were insufficient classification of journal content as original research articles and reviews during the compilation of hoaddata. Particularly, letters and editorials could not be fully detected. Paratext recognition failed in 37% of DOIs to identify non-scholarly content such as front matters or reviewer lists, which are generally not indexed by Scopus and Web of Science. To a lesser extent, differences in publication and indexing dates were a reason for non-overlapping DOIs.

Using overlapping DOIs, the publication volume between 2019 and 2023 was also calculated for each journal. Figure 2C illustrates the distribution for each combination. It shows that there is a large spread across the journals covered by all three data sources. Furthermore, journals in this set published more on average than journals covered by only one or two data sources. In particular, the journals covered exclusively by hoaddata were substantially smaller than those covered by all three sources. Upon inspection, these were often newly launched hybrid journals, which explains the relatively low five-year publication volume. An example is *Digital society* that published 86 articles. This hybrid journal was launched in 2022, being covered by various Springer Nature transformative agreements since then.

Coverage by publisher portfolio. Figure 3 presents the coverage of hybrid journals in transformative agreements across data sources from 2019 to 2023 with a focus on publisher portfolios. The analysis highlights the dominance of the three largest publishers–Elsevier, Springer Nature, and Wiley–, which collectively accounted for 47% of hybrid journals and 62% of articles published during this five-year period. In terms of article volume, Elsevier led with 2,441,358 articles (33% of the total) published across 1,951 hybrid journals (16% of the total). Springer Nature followed with 1,247,578 articles (17%) in 2,311, although recording the largest number of hybrid journals (19%). Wiley accounted for 858,939 articles (12%) in 1,382 hybrid journals (11%). The remaining 54 publishers collectively accounted for 2,749,847 articles (38%) in 6,476 hybrid journals (53%).

The three largest publishers–Elsevier, Springer Nature and Wiley–were best represented in the exclusive intersection of all three data sources (hoaddata, Scopus, and Web of Science). Together, they comprised 4,384 hybrid journals (55% of the intersectional set) and dominated article coverage (n = 4,174,315; 66%), as determined through shared DOIs. When examining publication volume per journal, (Figure 3C), Elsevier published, on average, the largest journals, followed by Springer Nature and Wiley.

Comparing publisher portfolios across different indexing sets demonstrates that publisher were not represented uniformly. Notably, Springer Nature exhibited 519 hybrid journals exclusively indexed

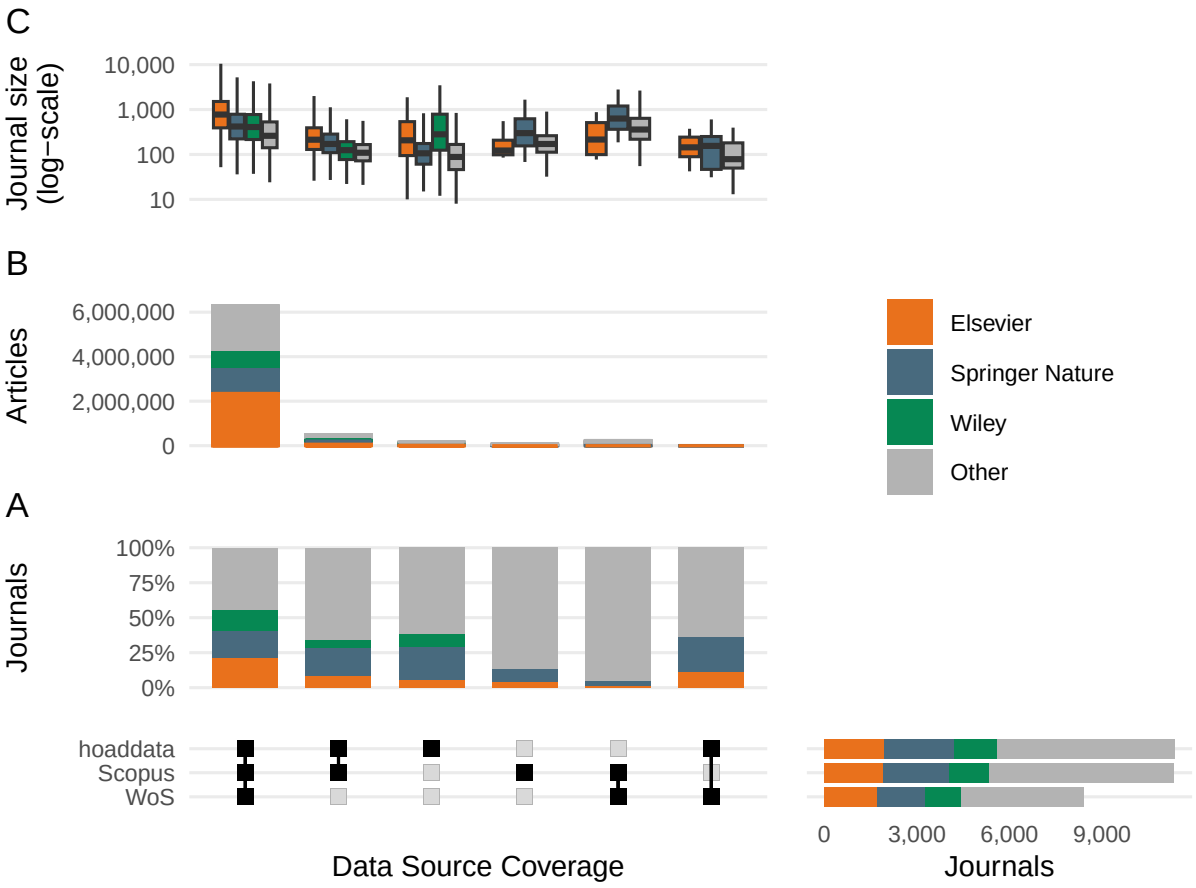

Figure 3. Upset graph publisher

in both hoaddata and Scopus. This set included journals from the Chinese Academy of Science, German-language medical journals, and Eastern European publications including the *Journal of Mathematical Sciences*, which also publishes English-language translations of Russian-language works. Additionally, this subset included titles with a more broader disciplinary focus such as *SN Computer Science* and newly launched hybrid journals like *Nature Computational Science*, which started in 2021 and was indexed in Scopus but not yet in Web of Science. The set also captured ceased journals, providing further insights into the dynamics of journal publishing.

Examining publisher portfolios not covered by hoaddata but present in Scopus or Web of Science identified several publishers with missing CC license information in Crossref. In particular, Emerald represented 322 journals with 86,409 articles, AIP Publishing accounted for 24 journals with 64,898 articles, and World Scientific recorded 87 journals and 42,531 articles. In total, 9 publishers were not represented in Crossref.

An inspection of individual journals also uncovered discrepancies in Unpaywall's open access identification for certain publishers that typically deposit CC licenses with Crossref. Notably, some subscription-only journals contained one or two articles erroneously tagged as hybrid open access by Unpaywall, which were subsequently reflected in Scopus and Web of Science. Examples of such

misclassifications include Elsevier's *Journal of Bioscience and Bioengineering* and Springer Nature's *Journal of Mechanical Science and Technology*.

4.2 Open Access Indicator Comparision

This section examines the uptake of open access in hybrid journals, focusing on the influence of transformative agreements across hoaddata, Scopus, and Web of Science. The aim was to assess whether consistent results can be derived from these data sources despite differences in coverage and methodologies. Following Jahn (2025), indicators were calculated for each data source and comprise the number and proportion of open access articles, including those enabled by transformative agreements, from 2019 to 2023. For Web of Science and Scopus, the impact of transformative agreements was estimated using both first and corresponding authorships, while hoaddata indicator calculation was limited to first author affiliations.

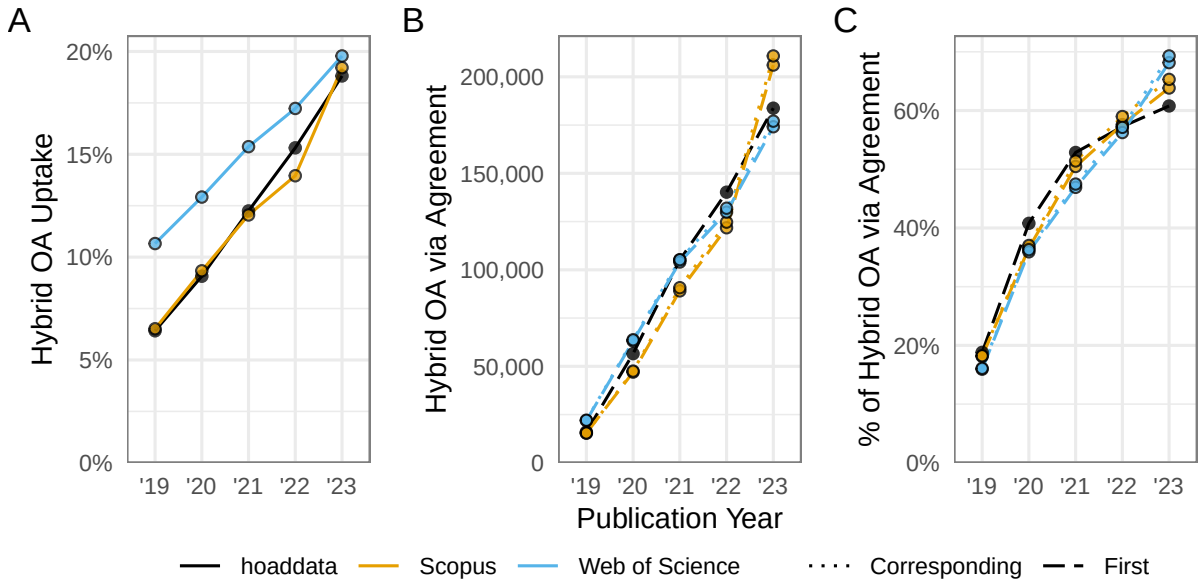

Figure 4. Open Access indicato

Overview. Figure 4A shows a moderate growth of open access in hybrid journals, which is consistent across hoaddata (black line), Scopus (orange line), and Web of Science (blue line). According to hoaddata, hybrid open access uptake increased from % (n =) in 2019 to % (n =) in 2023. Similarly, Scopus recorded an growth from 6.5% (n = 84,648) in 2019 to 19% (n = 322,850) in 2023. However, Web of Science recorded higher open access uptake in early years, before converging in 2023, from 11% (n = 137,202) in 2019 to 20% (n = 255,481) in 2023, suggesting an different approach towards labeling hybrid open access by the Web of Science.

Similarly, hybrid open access by transformative agreements increased between 2019 and 2023 (Figures 4B and C). Trends were consistent for first (dashed line) and corresponding author (dotted line) affiliations. According to Scopus, 479,297 open access articles could be attributed to transformative agreements based on first author metadata (increasing from 15,341 to 206,084) and

489,262 using corresponding author metadata (from 15,444 to 210,816). Web of Science recorded 493,028 open access articles via transformative agreements using first author metadata (increasing from 21,871 to 174,126) and 500,076 using corresponding author metadata (from 22,092 to 177,030). hoaddata, lacking corresponding author data, linked 0 articles to transformative agreements but showed slower growth than Scopus and Web of Science, from 0 in 2019 to 0 in 2023.

From 2021 (hoadata, Scopus) resp. 2022 (Web of Science), transformative agreements enabled the majority of hybrid open access. For first authors, the share ranged between % (hoaddata), 64% (Scopus), and 68 % (Web of Science) in 2023. For corresponding authors, the shares were slightly larger, with Scopus recording 65% and Web of Science 69% in 2023. However, substantial hybrid open access was still facilitated outside transformative agreements, likely through APCs paid from discretionary research funds (Jahn et al., 2022; Suber, 2012).

Open access by publishers. When considering open access trends by publisher (see Figure 5), the observed differences in early uptake rates between hoaddata and Scopus compared to Web of Science can be largely attributed to articles published in Elsevier hybrid journals, the largest publisher in our sample. Both hoadata and Scopus reported an steady increase in open access uptake between 2019 and 2023 (hoaddata from 4% to 13%; Scopus from 5% to 15%). In contrast, Elsevier's share remained relatively constant, increasing only slightly from 14% to 15% according to the Web of Science. Upon inspection, this discrepancy primarily stemmed from articles in the publisher's open archive. These articles, made freely available after an embargo period under Elsevier's user license, were tagged as hybrid open access in Web of Science, even though its documentation[7] specified that only articles under a CC license variant were considered. Previous research (Haustein et al., 2024; Jahn et al., 2022) has shown that Elsevier provided a substantial portion of its articles under this license, explaining the relatively large and stable share of open access over the years.

Differences in open access evidence are also evident for Wiley. Specifically, Web of Science and Scopus recorded a drop in 2021 and 2022 compared to hoaddata. For these two years, hoaddata reported 35,308 more open access articles than Scopus and 32,491 more than Web of Science. This discrepancy is presumably because of challenges in fetching full-texts by Unpaywall, the open access evidence source for Scopus and Web of Science. According to Unpaywall's software version history, Wiley's publisher platform redirects prevented Unpaywall from parsing license information from full-texts.[8] hoaddata, which relies solely on Crossref metadata for open access identification, was unaffected by these issues.

Despite these differences in open access evidence, the three data sources show consistent

[7] https://webofscience.help.clarivate.com/en-us/Content/open-access.html

[8] See Unpaywall version history related to Wiley fixes:
https://github.com/search?q=repo%3Aourresearch%2Foadoi+wiley&type=commits

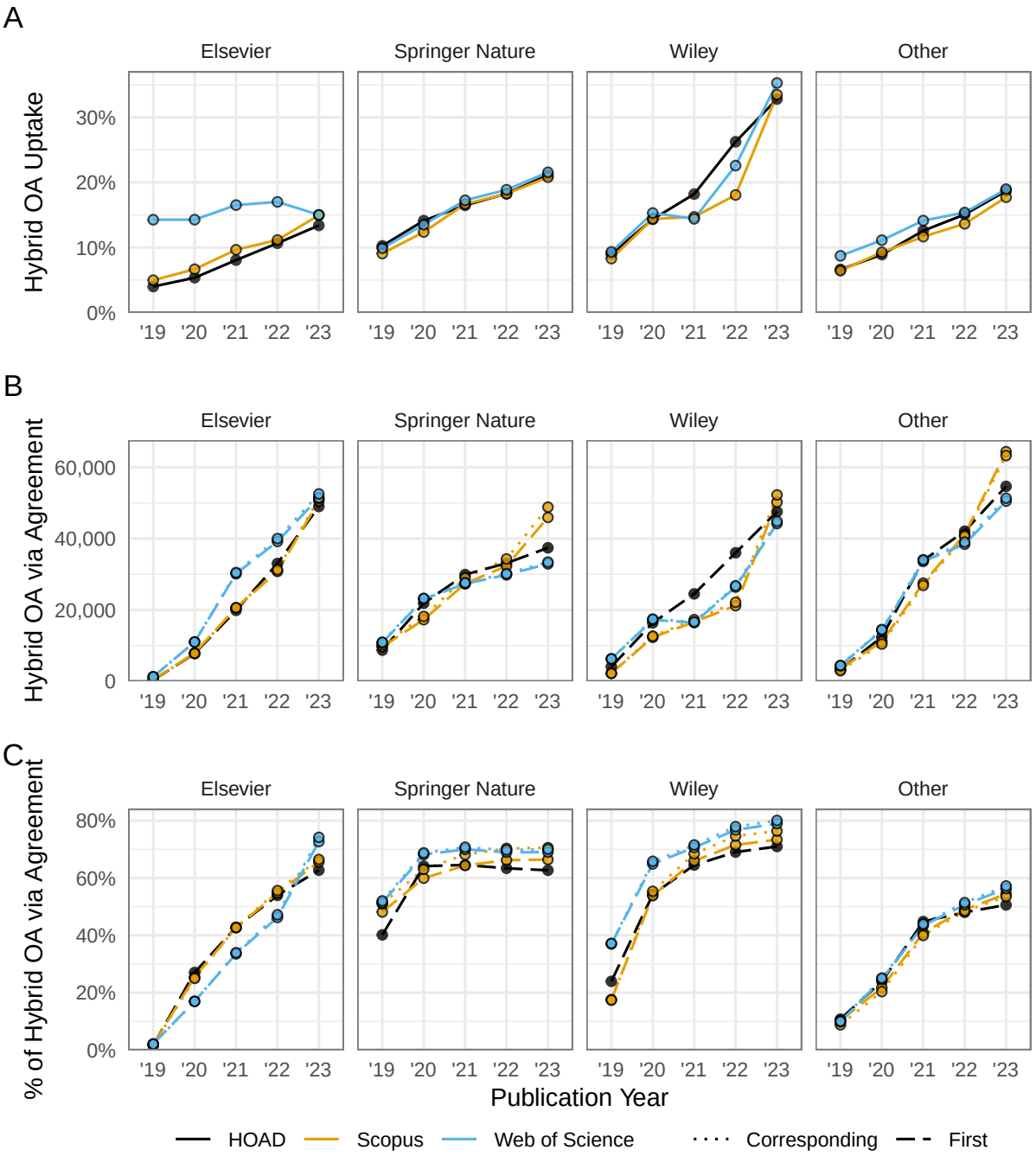

Figure 5. Open Access indicato

temporal trends in hybrid open access enabled by transformative agreements (see Figure 5B and C). Wiley emerged as the fastest-growing publisher in terms of open access uptake, with more than 30% of articles in hybrid journals reported as open access in 2023 across the examined data sources, followed by Springer Nature and Elsevier, recording a comparable late uptake, which is consistent with the publisher's historical reluctance to engage in negotations with library consortia. However, by 2023, the share of open access enabled by transformative agreements appeared to stabilise for all three publishers (see Figure 5C). Interestingly, the differences between first and corresponding author affiliations were more pronounced at the publisher level. In Scopus, for example, the share of open access via

transformative agreements measured by corresponding authorship was greater for Springer Nature in 2023 than when using first authorship.

Open access by country. When comparing countries, consistent patterns were observed across data sources for the five-year period 2019 to 2023. Figure 6 presents hybrid open access indicators by country, comparing hoaddata (x-axis) with Web of Science and Scopus (y-axis). Indicators calculated from these proprietary databases are shown for both first and corresponding authors, with full counting used to account for multiple country affiliations.

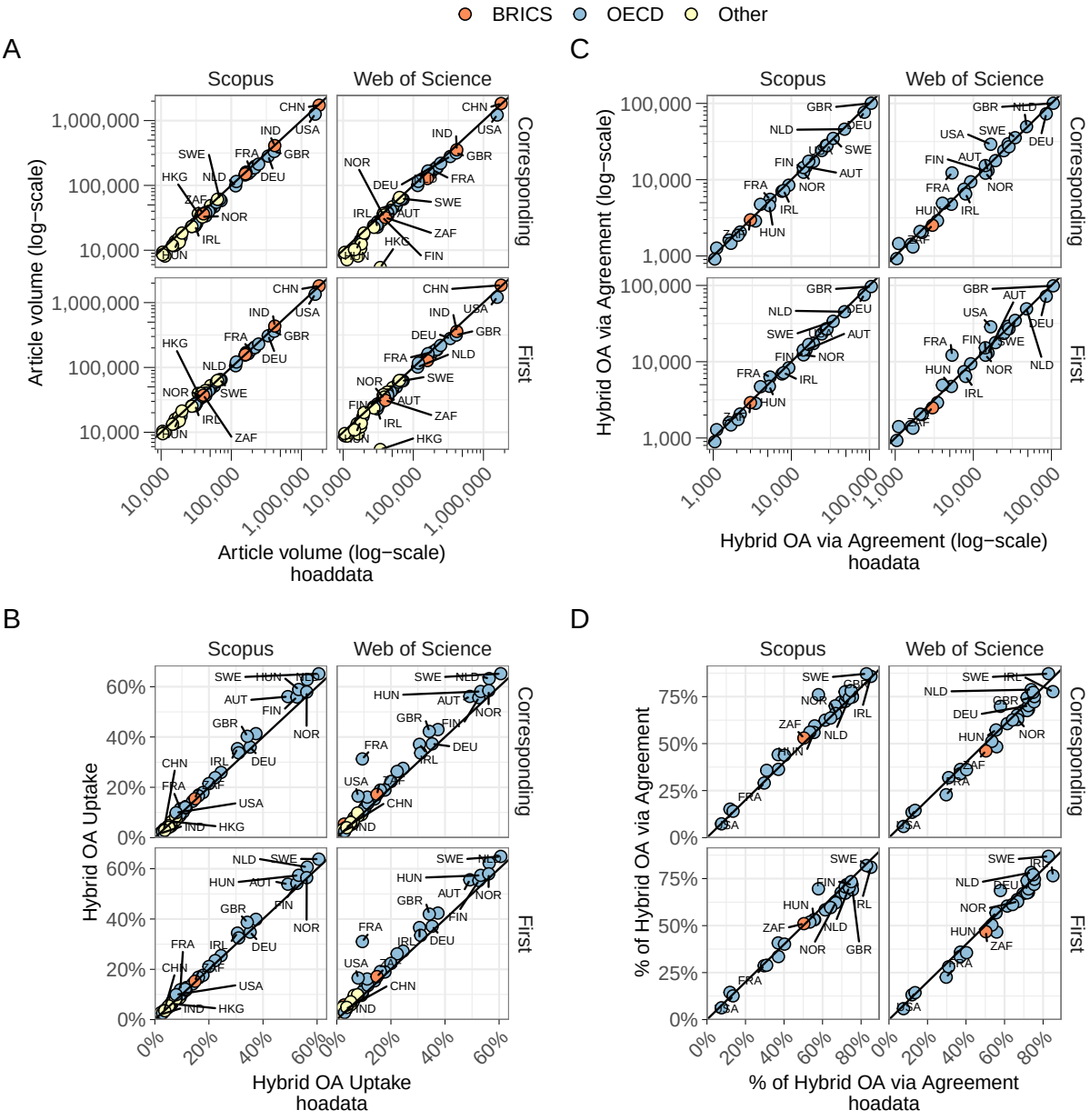

Figure 6. Open Access by country

In terms of article output by country (see Figure 6A), strong positive correlation was observed across the data sources and author roles (Spearman rank correlation $\rho > .9$, $p < 0.001$). Between 2019 and 2023, China was the most productive country, followed by the United States and, by a certain

margin, India, the United Kingdom, and Germany. Analysis of authorship roles revealed minimal variation, indicating that first and corresponding authors were typically from the same country.

When examining the percentage of open access articles in hybrid journals (see Figure 6B), a different pattern emerged. Authors affiliated with institutions from medium-sized European countries, such as Sweden, the Netherlands, Finland, and Hungary, provided a large proportion of their articles in open access. Germany and the United Kingdom also had approximately 40% of their output available as open access. In contrast, non-OECD countries showed notably lower adoption of hybrid open access, with South Africa being the only BRICS member well-represented in the data. The United States also demonstrated a relatively low proportion of open access articles. These findings were consistent across all databases. However, France was better represented in Web of Science, likely due to its agreement with Elsevier starting in 2019, which allowed delayed open access under the publisher's user license (Rabesandratana, 2019). This licence was not classified as hybrid open access in either Scopus or hoaddata. In all cases, Spearman rank correlations were $\rho > .9$, $p < 0.001$, showing a high level of correlations between the databases and authorship roles considered.

Transformative agreements appeared to be a key driver of national open access growth (see Figure 6C and D). OECD members accounted for the majority of open access articles enabled by transformative agreements. As a notable exception, South Africa also featured prominently, as the South African National Library and Information Consortium (SANLiC) successfully negotiated transformative agreements with major publishers from 2022 onward.[9] Results were consistent across data sources. However, Wiley's open access surplus in hoaddata led to better rankings for countries where Wiley played a substantial role, such as Germany, where the DEAL consortium negotiated its first transformative agreements with Wiley that started in July 2019.

A closer examination of the proportion of hybrid open access enabled by transformative agreements (see Figure 6D) revealed that the French Elsevier agreement on delayed open access was not included in JCT data, as it did not qualify as a transformative agreement. Medium-sized European countries again showed a high proportion of hybrid open access through transformative agreements, highlighting the impact of this licensing model across all three data sources. In contrast, the United States had a low proportion of hybrid open access enabled by agreements, suggesting that a substantial number of open access articles were likely financed through other means, such as article processing charges (APCs).

In all cases, strong positive correlations were observed using Spearman's rank correlation: $\rho > .9$, $p < 0.001$ between data sources and authorship roles, when considering countries with an minimum of 1,000 open access articles enabled by transformative agreements between 2019 and 2023. When this limitation was removed, the correlation remained strong but slightly lower ($\rho > .85$,

[9] `https://sanlic.ac.za/read-and-publish-agreements/`

$p < 0.001$). This difference may signal countries where only a few institutions had transformative agreements in place, as opposed to those participating in national consortia with broader participation.

5 Discussion

Investigating over 13,000 hybrid journals included in the cOAlition Journal Checker Tool shows a substantial increase in open access enabled by transformative agreements between 2019 and 2023, although most articles in these journals remained paywalled during this five-year period. While transformative agreements accounted for the majority of open access, a considerable number of open access articles are likely to have resulted from the payment of individual publication fees. The results confirm that transformative agreements and hybrid open access continue to be concentrated among a few large commercial publishers. They also highlight that high-income European countries in particular are using transformative agreements to make most of their article output open access, although South Africa showed similar levels of open access adoption. These trends were consistent across the investigated open data source hoaddata, derived from Crossref and OpenAlex, and the established proprietary bibliometric databases Scopus and Web of Science.

Comparing the data sources examined, the study results show strong correlations by country affiliation, despite differences in journal and article coverage, and in the availability of metadata on corresponding authorship and open access status. However, several notable discrepancies warrant discussion. They stem not only from known limitations of the open scholarly metadata sources Crossref and OpenAlex, but also from the proprietary databases Scopus and Web of Science.

The coverage analysis reveals that hybrid journals are well indexed in all three data sources, particularly in terms of article coverage. Differences can be found for journals targeting practitioners or local non-English language communties, with many such titles indexed exclusively in Crossref and Scopus. Using the open index Crossref demonstrated particular strength in identifying newly established hybrid journals, a notable finding given that transformative agreements primarily target existing subscription-based journals.

The situation of hybrid journal publishing is therefore different from that of fully open access journals. Comparing the coverage of OpenAlex, Scopus and Web of Science, Simard et al. (2024) indicate that only half of the fully open access journals listed in the DOAJ are also indexed in Scopus and Web of Science. Notably, journals that charge no publication fees (“diamond journals”) are absent from the selective Web of Science, which reinforces existing disparities in the indexing of underrepresented research communities and regions in selective bibliometric databases (ref).

A notable advantage of using Scopus and Web of Science to study the impact of transformative agreements is their comprehensive differentiation of journal articles by document type, as many agreements restrict funding to eligible article types. This resulted in slightly higher rates of open access uptake compared to hoaddata.

A similar limitation of studying open access with open scholarly data sources is the often reported lack of corresponding authorship information. However, indicators using first authors, which have often been used as a proxy for determining open access funding, and corresponding authors show a high level of correlation, which is not surprising given disciplinary norms in scholarly publishing. First authors are often considered to have carried out the main research underlying a paper, while the corresponding author is often recognised as having supervised a paper. It is therefore not surprising that both are strongly correlated, suggesting that first and corresponding authors share the same country affiliation. Moreover, in most cases the first author is indeed the same as the corresponding author. As such, first authorship can be used confidentially in large-scale studies and those focusing on national consortia to assess the impact of transformative agreements.

However, affiliation data is sparser in OpenAlex than in Scopus and Web of Science, resulting in slightly more articles that could be attributed to transformative agreements. However, in order to establish a link, a thorough mapping of the database affiliation identifier to the ROR ID, the JCT's identifier for institutional participation in transformative agreements, is required. In contrast, ROR IDs are the canonical organisation identifier in OpenAlex, allowing for less ambiguous matching between agreement and publication.

Comparison of scholarly databases is also valuable in identifying differences in open access identification. Not all publishers deposit a Creative Commons licence with Crossref, the citation authority. A collaboration with Scopus and Web of Science helped to identify such publishers, as their underlying unpaywall data also parses licence information from websites. However, there were significant differences between Scopus and Web of Science in terms of open access indexing, particularly with regard to the delayed open access provided by Elsevier. In the case of Wiley, however, there were also problems with the analysis of the Unpaywall. This suggests that hybrid open access studies should not only refer to a single source of evidence. Instead, open access tags can be validated by the combiantion of seevral data sources, as in our case where metadata from publishers was compared with data from open access discovery services such as Unpaywall. It is also particularly useful to include a curated list of journals under the hybrid model to avoid misclassifications based on a journal's business model (Visser et al., 2021).

6 References

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. https://doi.org/10.1162/qss_a_00018

Borrego, Á., Anglada, L., & Abadal, E. (2021). Transformative agreements: Do they pave the way to open access? *Learned Publishing*, *34*(2), 216–232. https://doi.org/10.1002/leap.1347

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, *113*(1), 219–236. https://doi.org/10.1007/s11192-017-2483-y

Geschuhn, K., & Stone, G. (2017). It's the workflows, stupid! What is required to make "offsetting" work for the open access transition. *Insights the UKSG Journal*, *30*(3), 103–114. https://doi.org/10.1629/uksg.391

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). *Analysis of the publication and document types in OpenAlex, web of science, scopus, pubmed and semantic scholar*. arXiv. https://doi.org/10.48550/ARXIV.2406.15154

Haustein, S., Schares, E., Alperin, J. P., Hare, M., Butler, L.-A., & Schönfelder, N. (2024). *Estimating global article processing charges paid to six publishers for open access between 2019 and 2023*. https://arxiv.org/abs/2407.16551

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

Jahn, N. (2024). *Hoaddata: Data about hybrid open access journal publishing*. https://github.com/subugoe/hoaddata

Jahn, N. (2025). How open are hybrid journals included in transformative agreements? *Quantitative Science Studies*, 1–39. https://doi.org/10.1162/qss_a_00348

Jahn, N., Matthias, L., & Laakso, M. (2022). Toward transparency of hybrid open access through publisher-provided metadata: An article-level study of Elsevier. *Journal of the Association for Information Science and Technology*, *73*(1), 104–118. https://doi.org/10.1002/asi.24549

Jonge, H. de, Kramer, B., & Sondervan, J. (2025). *Tracking transformative agreements through open metadata: Method and validation using dutch research council NWO funded papers*. https://doi.org/10.31222/osf.io/tz6be_v1

Krassowski, M. (2020). *ComplexUpset*. https://doi.org/10.5281/zenodo.3700590

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992. https://doi.org/10.1109/tvcg.2014.2346248

Maisano, D. A., Mastrogiacomo, L., Ferrara, L., & Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*. https://doi.org/10.1007/s11192-025-05244-y

Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, *72*(1), 80–88.

https://doi.org/10.1080/00031305.2017.1375986

Melero-Fuentes, D., Aguilar-Moya, R., Valderrama-Zurián, J.-C., & Gorraiz, J. (2025). Evolution and effect of meeting abstracts in JCR journals. *Journal of Informetrics*, *19*(1), 101631. https://doi.org/10.1016/j.joi.2024.101631

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

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. https://arxiv.org/abs/2205.01833

Rabesandratana, T. (2019). Elsevier deal with france disappoints open-access advocates. *Science*. https://doi.org/10.1126/science.aba5656

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The data infrastructure of the German Kompetenznetzwerk Bibliometrie: An enabling intermediary between raw data and analysis*. Zenodo. https://doi.org/10.5281/zenodo.13935407

Simard, M.-A., Basson, I., Hare, M., Lariviere, V., & Mongeon, P. (2024). *The open access coverage of OpenAlex, scopus and web of science*. https://doi.org/10.48550/ARXIV.2404.01985

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

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*(5), 2413–2431. https://doi.org/10.1007/s11192-022-04309-6

Suber, P. (2012). *Open access*. The MIT Press. https://doi.org/10.7551/mitpress/9286.001.0001

Visser, M., Eck, N. J. van, & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*. https://doi.org/10.1007/s11192-023-04923-y