


# Replication: Hybrid Open Access in Transformative Agreements

Najko Jahn<sup>1\*</sup> ( 0000-0001-5105-1463)

<sup>1</sup> 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 rolled 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 verrechnungsformen 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

The aim of this study is to demonstrate the suitability of open scholarly data sources for assessing the impact of transformative agreements on hybrid open access. To achieve this, results from hoaddata, an openly available collection of open research information regarding hybrid open access, was compared with the established bibliometric databases Web of Science and Scopus. After describing the initial data sources used, the necessary pre-processing steps to obtain eligible articles from transformative agreements using open access evidence, author roles (first and corresponding) and affiliation data are presented.

### 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 comprising information about the uptake of hybrid open access since 2017 from several openly available data sources. It combines article-level metadata from Crossref and OpenAlex with

transformative agreement information from the cOAlition S Journal Checker Tool (JCT), which links journal and institutional data to agreements in the ESAC registry. More specifically, hoaddata uses Crossref, a DOI registration agency, for obtaining journal publication volume and open access status through Creative Commons licence information relative to the published version (“version of record”). Because of limited affiliation metadata in Crossref (Eck & Waltman, 2022), hoaddata sources first-author affiliations from OpenAlex.

hoaddata follows good practices for computational reproducibility using R. 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 interfaces with 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. 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. 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). 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), collectively referred to as Web of Science Primary (WoS Primary). The WoS Primary 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, the same provider as Openalex. However, it 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 provided as an in-house PostgreSQL database under a uniform schema. 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 coverage is substantially more extensive than that of the Web of Science collections considered in this study (Singh et al., 2021; 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.** This study examined the adoption of open access for hybrid journals included in transformative agreements, which published at least one journal article in the five-year period 2019-2023. The initial data source was the Transformative Agreements Public Data from the cOAlition S Journal Checker Tool (JCT), which helps authors identify suitable open access publishing venues. Weekly snapshots of this data were archived using an automated GitHub Actions workflow ([https://github.com/njahn82/jct\\_data](https://github.com/njahn82/jct_data)) and into `hoaddata`. The resulting JCT data cover agreements active from July 2021 to July 2024. The JCT journal data was enriched with the linking ISSN (ISSN-L) according to the

ISSN Registry Agency. A comprehensive exclusion of fully open access journals was performed, following the methodology described in (jahn\_2025?). This involved checking multiple journal lists (DOAJ, OpenAlex, Bielefeld) and article-level investigations. The 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.

For each database, article metadata was retrieved using all ISSN variants linked to an ISSN-L. The metadata included DOIs, document types (where available), author roles, institutional affiliations, publication dates and open access information. Publication years were determined using the earliest known date of publication in a journal. In the case of Crossref, the article data source for hoaddata, this was the issued date, although variations were observed (check ref). For Web of Science and Scopus, the earliest publication date was used where available. In the latter case, the KB determined the earliest known publication date by tracking versions of the raw data.

**Open Access.** Only published versions with Creative Commons licenses available on publisher platforms were considered as hybrid open access. hoaddata derived this information from Crossref’s license metadata, while Web of Science and Scopus used data from Unpaywall. In addition to Crossref, Unpaywall also parses publisher websites directly, because some publishers do not provide machine-readable Creative Commons license information (Piwowar et al., 2018), although transformative agreement workflows requiring this information during DOI registration (Geschuhn & Stone, 2017).

**Author affiliations.** Author affiliations at the article level were retrieved for both first and, if available, corresponding authors across all databases. 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 Identifiers for Scopus. 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 identify articles covered by transformative agreements. 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 organizational hierarchy differences.

This process linked 6,375 ROR IDs to 4,894 Scopus Affiliation IDs, and 6,034 ROR IDs to 4,894 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 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.

On the basis of these matching tables, articles eligible under transformative agreements could also be retrieved from the Web of Science and Scopus, although they did not contain the ROR IDs used by the JCT.

### **Estimating open access in hybrid journals covered by transformative agreements.**

#### **3.3 Data records**

As a result of the above-described comprehensive data processing, data-sets were prepared for each database on open access in hybrid journals included in transformative agreements at the country and journal level by year. Table 1 presents an general overview of the coverage of hybrid journals in transformative agreements 2019-23 per dataset. It shows that the majority of hybrid journals published at least one original article or review (core) in the five-years period. These journals constituted the basis for the subsequent aggregation of publication metrics. While hoadd only covers articles with DOI, the aggregation of Scopus and Web of Science was carried out using the database

identifier. A subsequent comparison of DOU coverage illustrates that non-core articles in Web of Science lack DOIs. These are in particular meeting abstracts, a peculiarity of this database. Open access was aggregated using DOI, as all also Unpaywall only collects open access status information for articles with DOI. A closer view on core articles with affiliation data reveal a lack of data with regard to corresponding authors in the case of OpenAlex. Only around two third of investigated articles had at affiliation data linked to corresponding authors. In case of first authors, the share was higher. In contrast, web of science and scopus recorded a higher coverage. Accordingly, only first author evidence were presented for hard data in the following



Table 1

*Coverage of hybrid journals in transformative agreements 2019-23.*

	HOAD	Web of Science	Scopus
<b>Hybrid journal metrics</b>			
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
<b>Publication metrics</b>			
Total published articles	9,740,015	8,616,053	8,117,644
Core articles	8,158,425	6,708,083	7,317,703
<b>Digital Object Identifier (DOI) coverage</b>			
Articles with DOI	9,740,015	7,713,796	8,105,112
Core articles with DOI	8,158,425	6,695,661	7,314,327
<b>Open Access (OA) metrics</b>			
OA articles	998,699	1,112,758	974,099
Core OA articles	969,817	1,019,784	922,578
<b>Core articles with affiliation data</b>			
First author articles	7,242,542	6,294,855	7,232,017
Corresponding author articles	5,534,207	6,291,441	6,898,487

## 4 results

### discussion

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](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](https://doi.org/10.1162/qss_a_00018)
- Eck, N. J. van, & Waltman, L. (2022). *Crossref as a source of open bibliographic metadata*. <https://doi.org/10.31222/osf.io/smxe5>
- 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>
- 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](https://doi.org/10.1162/qss_a_00348)
- Piowar, 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>
- 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>
- 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>
- 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](https://doi.org/10.1162/qss_a_00112)