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

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. Overall, xxxx hybrid journals from xxx agreements that published at least one open access article between 2019 and 2023. formed the basis of this study.

3.1 Data sources

hoaddata. hoaddata, developed and maintained by the author, 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 (`https://doi.org/10.31222/osf.io/smxe5`), hoaddata sources first-author affiliations from OpenAlex. While the country in which a first-author was located was used to aggregate country-level statistics, Research Organization Registry (ROR) identifiers (ROR-ID) were used in conjunction with ESAC registry information on the duration of an agreement was used to estimate whether an article was published under active transformative agreements. This matching benefited from the availability of the ROR-IDs in both sources. To improve the matching, JCT data were enriched to include associated institutions, such as unviersity hospitals.

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 package has been regularly updated since 2022 and the version including the computation log is available on GitHub.

hoaddata is used as a data basis of the Hybrid Open Access Dashboard, a data analytics services for library consortia and publishers to track the uptake of hybrid open access through transformative agreements. It was also used in bibliometric research (Jahn 2025).

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).

The WoS provides important data points for studying spending on 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, the database of the Kompetenznetzwerk Bibliometrie (KB) in Germany is used to access bibliometric data. 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, 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 Core collection (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. 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

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 journal data source used was a unified dataset covering various snapshots of the Transformative Agreements Public Data from the cOAlition S Journal Checker Tool (JCT), a well-established source to help authors identify suitable open access publishing venues. These snapshots, archived weekly by the author using a cron job via a publicly available GitHub repository, cover agreements active from July 2021 to July 2024. After enriching the JCT journal data with the linking ISSN (ISSN-L) according to the ISSN Registry Agency, a comprehensive exclusion of fully open access journals, including flipped journals covering multiple journal lists (DOAJ, OpenAlex, Bielefeld) and article-level investigations, was applied as described elsewhere. In total, xxxx hybrid journals could be linked to xxx agreements listed in the ESAC Transformative Agreements Registry, which formed to basis for obtaining article-level data including author affiliations, open access status ans document types.

For each of the databases examined, article metadata such as publication identifier, document types (where available), author roles and institutional affiliations, publication date and open access information were retrieved for each journal, using all ISSN variants linked to an ISSN-L according to the ISSN Registry Agency. The publication period was determined using the earliest year of publication. In the case of Crossref, the article data source for hoaddata, this was the publication 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.

Affiliation data, including author roles, are essential for linking articles to transformative agreements. Therefore, author affiliations were retrieved for the first and corresponding author at the article level for each of the databases examined. In order to account for different address variants in bibliometric databases, the database-specific IDs for affiliations were used. For hoaddata these were the ROR IDs from OpenAlex, for Web of Science the affiliation enhanced names and for Scopus the Scopus Affiliation Identifier, an eight-digit ID number. In addition, the country of the author's address, represented by ISO country codes, was retrieved.

At the time of our study, there was no bibliometric database that included information on whether an article was published under a transformative agreement. Here, eligible articles were identified by matching hybrid journals and participating institutions from the Transformative Agreement Data Dump with the databases examined per active agreement. This matching process also took into account the duration of the agreements according to the ESAC registry. While in the case of hoaddata, which uses OpenAlex as an affiliated data source, matching was straightforward due to the availability of the ROR ID in both sources, neither Web of Science nor Scopus supported ROR at the time of the study. Instead, xxxx articles identified under a transformative agreement in hoaddata were processed to map the ROR ID associated with the first author to the proprietary affiliation identifier in Web of Science and Scopus representing the affiliation of the first author. Due to multiple affiliations, different organisational hierarchies encoded in the different organisational ID systems, and to deal with conflicting matches, an algorithm then sorted the ROR-ID and proprietary affiliation identifier pairs in descending order and selected the most frequent pair as a match.

On the basis of the matching tables thus compiled, 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.

Identifying eligible articles under transformative agreements.

adn which of them were os.

3.3 Data records

Table 1

Coverage of hybrid journals in transformative agreements 2019-23.

	HOAD	**Web of Science**	**Scopus**
Hybrid journal metrics			
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 metrics			
Total published articles	9,740,015	8,616,053	8,117,644
Core 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
Core 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
Core OA articles	969,817	1,019,784	922,578
Core 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

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

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

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