Analysing and re-classifying open access information in OpenAlex

2023-11-07

Over the last few months, we have switched our data basis for open access analytics from [Unpaywall](https://unpaywall.org/) to [OpenAlex](https://openalex.org/). Both open scholarly data services are developed by [OurResearch](https://ourresearch.org/) and have a similar metadata format for describing open access full-texts. However, OpenAlex offers monthly data dumps, which we’ve found particularly helpful as the free snapshot releases from Unpaywall, which we used before, appear to have ceased as of March 2022.

While transitioning from Unpaywall to OpenAlex, we noticed a considerable number of around 4 million OpenAlex records with contradictory open access information. This blog post aims to explore this issue. To better understand it, we’ve re-implemented Unpaywall’s open access classification using OpenAlex data, and compared it against OpenAlex’s existing classification.

## What is the issue?

OpenAlex provides various methods for identifying open access literature. Within the [work object](https://docs.openalex.org/api-entities/works/work-object), the [open\_access](https://docs.openalex.org/api-entities/works/work-object#open_access) and [best\_oa\_location](https://docs.openalex.org/api-entities/works/work-object#best_oa_location) nodes contain information about the open access status at the article level. The [sources object](https://docs.openalex.org/api-entities/sources), on the other hand, gives information about the open access model of a journal.

The issue we’ve identified is that [filtering for open access works with is\_oa:true returns around 4 million records with the oa\_status marked as closed](http://api.openalex.org/works?filter=open_access.is_oa:true,open_access.oa_status:closed), a discrepancy we find inconsistent with OpenAlex’s own documentation. Accordingly, OpenAlex follows Unpaywall’s methodology tagging openly available works ([is\_oa](https://docs.openalex.org/api-entities/works/work-object#is_oa)) and qualifying their open access status ([oa\_status](https://docs.openalex.org/api-entities/works/work-object#oa_status)) using the following values:

* gold: Published in an open access journal.
* green: Toll-access on the publisher landing page, but there is a free copy in an OA repository.
* hybrid: Free under an open license in a toll-access journal.
* bronze: Free to read on the publisher landing page, but without any identifiable license.
* closed: All other articles.

## Understanding the issue

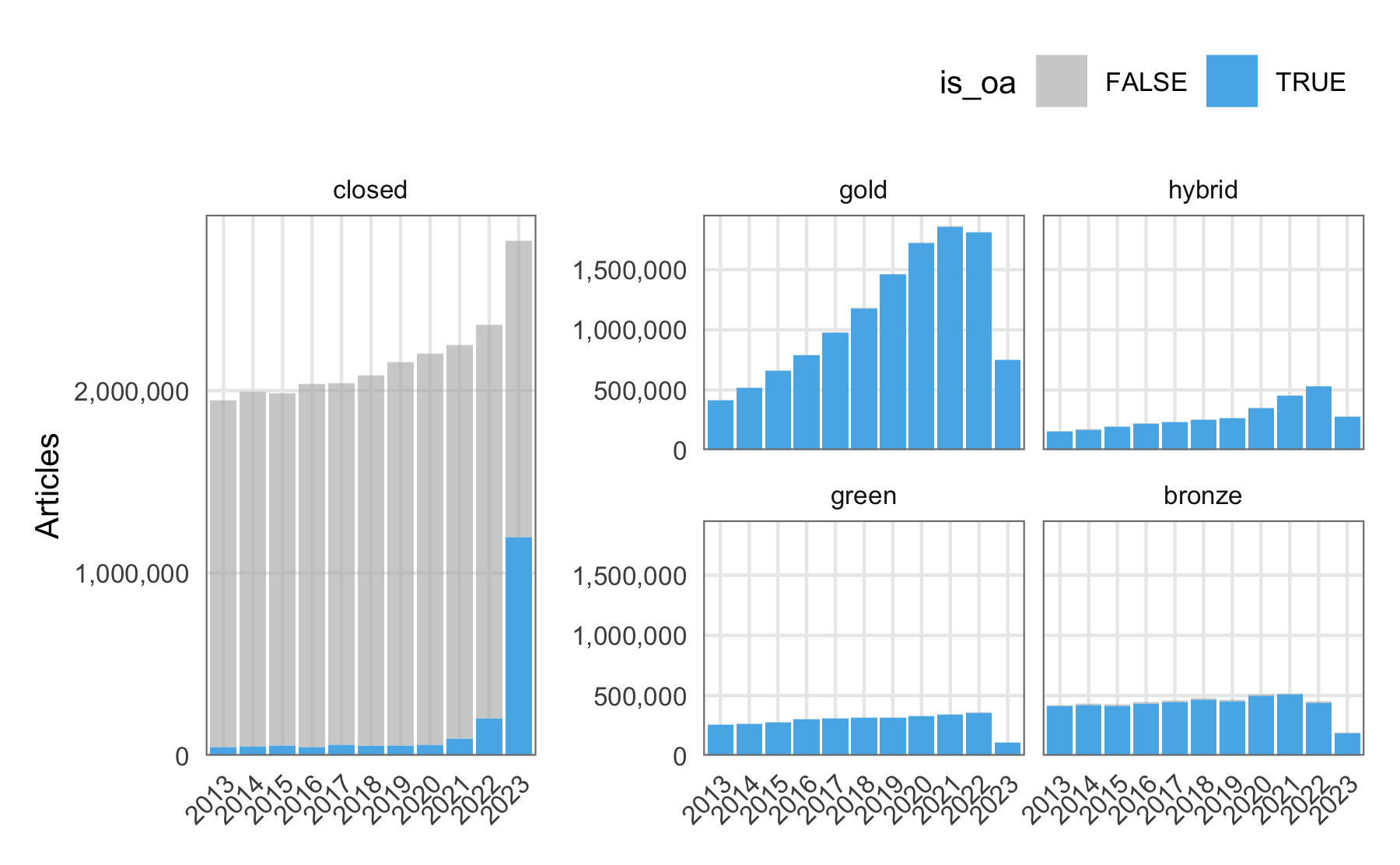
To better understand this issue, we analysed the most recent OpenAlex snapshot from October 2023. After importing this data into our [BigQuery data warehouse](https://subugoe.github.io/scholcomm_analytics/data.html), we created a subset with the focus on journal articles published since 2013, excluding retractions and non-scholarly content published in journals.

CREATE OR REPLACE TABLE  
 subugoe-collaborative.resources.oalex\_cr\_journal\_articles\_13\_23 AS (  
 SELECT  
 doi,  
 publication\_year,  
 open\_access,  
 best\_oa\_location,  
 sources.is\_oa AS journal\_is\_oa,  
 sources.is\_in\_doaj AS journal\_is\_in\_doaj,  
 sources.host\_organization\_name AS publisher\_name  
FROM  
 `subugoe-collaborative.openalex.works`  
LEFT JOIN  
 `subugoe-collaborative.openalex.sources` AS sources  
ON  
 primary\_location.source.id = sources.id  
WHERE  
 type\_crossref = "journal-article"  
 AND is\_paratext = FALSE  
 AND is\_retracted = FALSE  
 AND publication\_year BETWEEN 2013  
 AND 2023 )

We then analysed the open access prevalence over the years, aggregating the record counts across both is\_oa and oa\_status.

SELECT  
 COUNT(DISTINCT doi) AS articles,  
 publication\_year,  
 open\_access.is\_oa,  
 open\_access.oa\_status  
FROM  
 `subugoe-collaborative.resources.oalex\_cr\_journal\_articles\_13\_23`  
GROUP BY  
 open\_access.is\_oa,  
 open\_access.oa\_status,  
 publication\_year  
ORDER BY  
 publication\_year DESC

The resulting figure shows the distribution of open access evidence in OpenAlex over years. All possible open access status values as known from Unpaywall were also represented in OpenAlex. The figure also presents the number of records tagged as open (blue colored bar chart stacks) and closed (colored bar chart stacks) access in the is\_oa node. Notably, the bulk of contradictory open access information could be found in records representing journal articles published in 2023, with 1,197,013 articles tagged as open access, but assigned the open access status “closed”.



## Re-classification and analysis of changes

In order to address this inconsistency, we re-implemented [Unpaywall’s open access classification methodology](https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-). The following SQL code snippet shows how we approached the re-classification.

CREATE OR REPLACE TABLE  
 `subugoe-collaborative.resources.oalex\_reclassify\_oa` AS (  
 SELECT  
 DISTINCT doi,  
 publication\_year,  
 open\_access.is\_oa,  
 open\_access.oa\_status,  
 CASE  
 WHEN best\_oa\_location IS NULL THEN "closed"  
 WHEN best\_oa\_location.source.type = "repository" THEN "green"  
 WHEN (journal\_is\_in\_doaj = TRUE OR journal\_is\_oa = TRUE) THEN "gold"  
 WHEN (journal\_is\_in\_doaj = FALSE  
 AND journal\_is\_oa = FALSE )  
 AND best\_oa\_location.license IS NOT NULL THEN "hybrid"  
 WHEN (journal\_is\_in\_doaj = FALSE AND journal\_is\_oa = FALSE ) AND best\_oa\_location.license IS NULL THEN "bronze"  
 ELSE  
 NULL  
 END  
 AS oa\_new  
 FROM  
 `subugoe-collaborative.resources.oalex\_cr\_journal\_articles\_13\_23` )

We used the best\_oa\_location node to determine the availability of an open access full-text. If this node was absent, we categorised the work as ‘closed.’ For open access works not exclusively provided by a repository, we used open access journal information from the source object to distinguish between gold, hybrid and bronze.

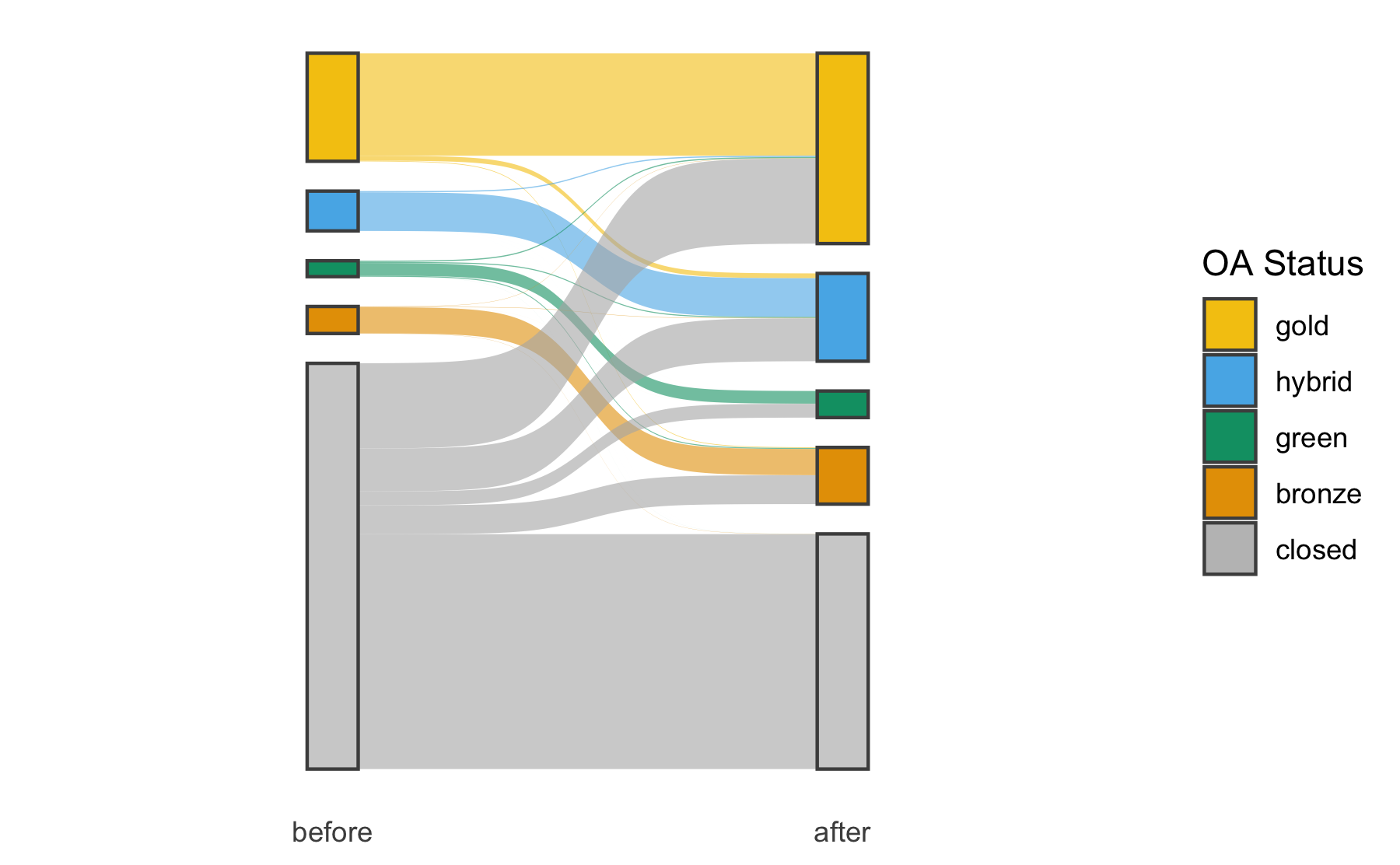
Following the re-classification, we calculated updated open access statistics.

SELECT  
 COUNT(DISTINCT doi) AS n,  
 oa\_status,  
 oa\_new,  
 publication\_year  
FROM  
 `subugoe-collaborative.resources.oalex\_reclassify\_oa`  
GROUP BY  
 oa\_status,  
 oa\_new,  
 publication\_year

The following figure compares OpenAlex open access classification (black-colored bars) with our approach (pink-colored bars). Notably, the re-classification resulted in many journal articles published in 2023 that were previously tagged as closed having one of the open access values such as gold, hybrid, green, or bronze.



Overall, we re-classified a total of 4,087,711 records representing journal articles published since 2013, with 1,257,175 of them being published in 2023. The following figure demonstrates changes in open access status after our re-classification for 2023. The “gold” category gained 607,896 additional records in 2023, “hybrid” gained 340,351, “green” gained 96,156, and bronze gained 211,065. The figure also highlights that we not only re-assigned records previously assigned to the ‘closed’ category, but we could also observe changes between other categories.



## Discussion and conclusion

Analysing and re-classifying open access data in OpenAlex has revealed inconsistencies in the actual implementation.

In response, we are sharing this detailed problem description in order to contribute to the ongoing improvement of OpenAlex, a scholarly data source that we enjoy working with on a daily basis. As a practical suggestion in the meantime, we recommend not to rely solely on the open access information provided. Instead, we suggest re-classifying open access status information.

## Funding

This work is funded by the Bundesministerium für Bildung und Forschung (BMBF) projects KBMINE (16WIK2101F) and KBOPENBIB (16WIK2301E). We acknowledge the support of the [German Competence Center for Bibliometrics](https://bibliometrie.info/).