Research in progress

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

Citation: Sanford, H. (2022). The State of Unpaywall: Analyzing the Consistency of Open Access Data. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22127). https://doi.org/10.5281/zenodo.6975430

Collection: https://zenodo.org/communities/sti2022grx/

The State of Unpaywall: Analyzing the Consistency of Open Access Data

Henning Sanford*

*sanford @dzhw.eu
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr. 6A, 10117 Berlin, Germany

Introduction

With the continued progression of the open access (OA) transformation, the topic of OA itself is gaining traction and prominence. Numerous efforts attempt to measure and monitor the success of the transformation. For their studies, OA monitors and academic studies analysing the prevalence of OA often rely on data offered by the OA discovery service Unpaywall[1] (Jahn et al., 2021). As a source of up-to-date publication-level OA information, Unpaywall data are now incorporated by commercial bibliometric database providers (Basson et al., 2022). The data is made available by Unpaywall with the promise of highly reliable and accurate data that allows to assess the state of OA (Piwowar et al., 2018).

This marks a remarkable transition from a tool that was intended to provide access to OA publications to the role of a data provider. Despite the frequent use of Unpaywall data, the quality of the data has only received scant attention in the academic discourse (Robinson-Garcia et al., 2020). Over the past years, the data provided by Unpaywall to its users has undergone several structural changes.[2] The goal of this study is to highlight uncertainty associated with Unpaywall's OA status information.

The analysis compares the open access status classification of a random sample of publications drawn from the Unpaywall database over time. This approach allows to uncover a potential source of uncertainty by showing how classifications for the same articles change. Studies relying on Unpaywall data should be aware that the reproducibility and comparability of their results are time dependent.

Data

The study is based on a comparison of 13 snapshots of the Unpaywall database. In regular intervals, Unpaywall makes complete dumps of their database publically accessible.[3] These snapshots were curated and made accessible for this study by the research group at the SUB Göttingen (Jahn et al., 2021). Unpaywall makes snapshots available about twice a year. The

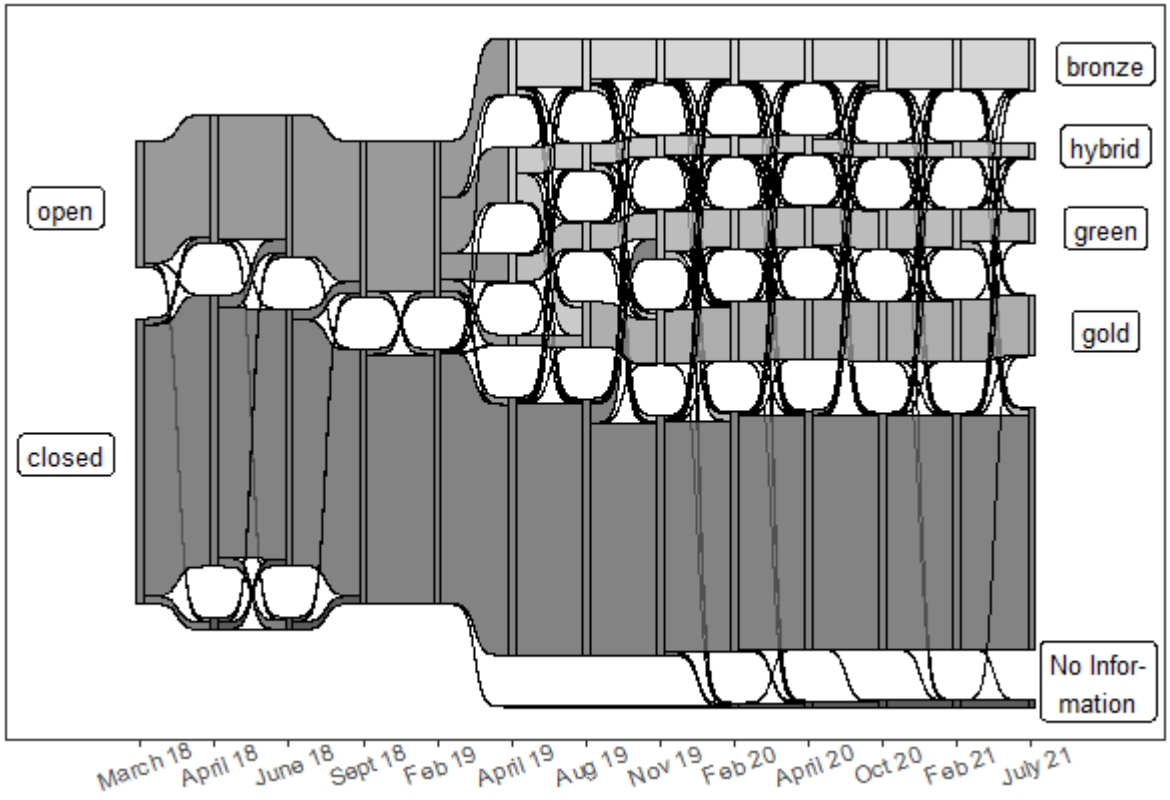
[1] https://unpaywall.org.
[2] A list of major changes implemented since November 2019 can be found here: https://support.unpaywall.org/support/solutions/articles/44001867302. Changes made prior to this date are not documented in the same fashion.
[3] https://unpaywall.org/products/snapshot

first snapshot that is used for the study was published in March 2018 and the most recent one in July 2021. For reasons related to computational resources as well as data quality the items recorded in the snapshots are limited to journal articles published after 2008. The resulting snapshots include between 26.75 Million and 28.2 Million publications. A random subset of 10,000 publications is drawn from the first snapshot. For each publication of the sample, the OA status information is recorded across all snapshots. Only the OA status that Unpaywall identifies as the best fitting description of the OA status is used. For publications that can be found in more than one OA location, Unpaywall uses a deterministic algorithm to choose the OA status.[4]

Results

Figure 1 shows the recorded OA classifications with a node for each snapshot. Transitions between nodes are made visible as streams between nodes. The Sankey plot gives a qualitative impression of the classification system, classification changes over time, as well as infrastructural changes.[5] The figure shows the transition from a binary classification system (open vs. closed) to a more granular approach with 5 classes by April 2019. The sizable stream from Hybrid OA to Gold OA between April 2019 and August 2019 is the result of changes to the classification algorithm that were implemented by August 2019. The changes were made in order to prioritize Gold OA over Hybrid OA.

Figure 1. OA Status Classification for a Random Sample of Articles across all Snapshots



[4] The full description of the algorithm can be found here: https://support.unpaywall.org/support/solutions/articles/44001943223-how-is-the-best-oa-location-determined-

[5] It is not possible to derive the exact point in time at which a change is implemented from the data. Rather it is possible to show that by the time a new snapshot is published, changes to the infrastructure took place.

Lastly, the figure shows that while classifications stabilized towards the most recent snapshot, there was substantial variation in classifications for the same publications over time. These changes include transitions from Closed OA to Open Access as well as vice versa. There seems to be a small share of publications that are present in the March 2018 snapshot but fall out of later snapshots. Figure 1 seems to indicate that these publications start missing in the February 2020 snapshot. Starting with this snapshot, Unpaywall allows to filter out paratext (i.e. front matter, letters to the editor, etc.). It is likely that the missing publications are in fact extraneous material.

Table 1 approaches the question of consistency from the perspective of classification volatility. The table shows the number of classification changes leading up to the classification in the most recent snapshot. The underlying assumption of the analysis is that the current classification is the best possible guess of the publication's true OA status. Reclassifications leading up to this guess can either be corrections of prior misclassifications or reflect true changes in the OA status of the publication. As the classification system changed from a two class system to a 5 class system, it is not possible for a publication to be categorized as Open in the most recent snapshot. Further, all OA color classifications have to change at least one time from Open to a more granular OA color.

Closed Access appears to be a very stable category with almost 90% of publications that are categorized as closed access never changing their class. Gold OA also appears to be very stable with about 95% of publications changing classes only once or twice. Publications that are reclassified two times and end on Gold OA were likely initially classified as Hybrid OA. Following the changes to the classification system these publications then changed to Gold OA. With about 70% of publications changing only one time, Green and Hybrid OA pose classes that are slightly less stable than Gold OA and Closed Access. Both categories show a wider spread over publications that were reclassified two or three times before ending on Green or Hybrid OA respectively. Lastly, publications that are categorized as Bronze OA in the most recent snapshot are likely to have changed classes several times. About 79% of items that are placed in the Bronze OA category have previously changed classes more than once and up to five times.

Table 1. Volatility of Open Access Classification. Volatility measured as number of classification changes before final classification across all snapshots

Number of Classification Changes	**No Information**	**Closed**	**Gold**	**Green**	**Hybrid**	**Bronze**
0	0%	89.534%	0%	0%	0%	0%
1	38.095%	0.843%	50%	69.928%	73.750%	21.429%
2	4.762%	8.426%	45.690%	10.145%	10%	32.540%
3	19.048%	0.488%	2.586%	15.217%	16.250%	26.190%
4	38.095%	0.488%	1.724%	2.899%	0%	15.079%
5	0%	0.133%	0%	1.087%	0%	4.365%
6	0%	0.089%	0%	0.362%	0%	0.397%
7	0%	0%	0%	0.362%	0%	0%

Discussion

The underlying reasons for classification changes cannot be determined directly from the data. The changes could be due to a previous classification error or reflect a true change in the open access status of the publication. Nevertheless, the reliability of the classification system in terms constituency and volatility is also depend on the design choices of the classification infrastructure. The change from a two-class system to a more granular system allows for deeper insight into the OA status of publications. But this change also opens up the data for reliability issues.

A crucial difference that divides reliable and volatile OA groups appears to be the actors that decides about the OA status of the publication. Author-based OA decisions (Closed Access, Gold, Green, or Hybrid OA) appear far more stable than publisher-based decision (Bronze OA). After authors decide to publish their research as open access publication, the publications are reliably identifiable. With publisher-choice based contributions, however, the availability of the publication does not appear similarly stable.

Conclusion

These result highlight the difficulties of identifying the open access status of a publication. Especially for less rigid OA subgroups like Hybrid and Bronze OA the classification task is a process of iterating over improved algorithms. Generally, it can be assumed that these iterations lead towards a more accurate reflection of the true OA status. This process, however, has implications for the academic users of Unpaywall data. Studies that use these data to analyze OA status and especially OA subgroups should be aware that the reliability of the data and reproducibility of the results are dependent on time and infrastructural design choice. This observation poses essential background information for OA studies that rely on Unpaywall data at a single point in time.

For the OA transformation, the results also highlight the importance of author-choice based contributions. Publisher-choice based contributions appear to be harder to identify but also volatile in their status over time. For Open Access studies, these findings provide empirical reasons for caution when including data on Bronze OA into their analysis. For the OA transformation in general, the findings highlight authors as the key contributors to a successful transformation.

References

Basson I., Simard, M.A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2022). The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. *PLOS ONE, 17*(3), e0265545. https://doi.org/10.1371/journal.pone.0265545

Jahn, N., Hobert, A., & Haupka, N. (2021). Entwicklung und Typologie des Datendiensts Unpaywall. *Bibliothek Forschung und Praxis*, *45*(2), 293-303. https://doi.org/10.1515/bfp-2020-0115

Piwowar, H., Priem, J., Larivière, V., Alperin, J., 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

Robinson-Garcia, N., van Leeuwen, T. N., & Torres-Salinas, D. (2020). Measuring Open Access Uptake: Data Sources, Expectations, and Misconceptions. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.23