Long Term Global Trends in Open Access. A Data Paper

Katja Heidbach, Johannes Knaus, Ingo Laut & Margit Palzenberger
Max Planck Digital Library (MPDL)
Big Data Analytics Group

Munich, 2022–01–12

RIO
Research Information Observatory

Contents

Abstract

Studies on long term trends in open access are of interest for the assessment of the evolution of scientific publishing and related markets. We therefore compiled and analysed a data set that integrated Web of Science as a global bibliographic data source on internationally relevant publications with data from Unpaywall, the primary provider of information related to open access at publication level. Data were captured in 2021 and show the open access categories as defined by Unpaywall for the publication years 2000 to 2020. In these two decades open access has gained substantial momentum. Starting with a few per cent, it now covers roughly half of the publications when embargo periods are over. The comparison of four variants of subsets of these data, however, show the wide variability in absolute and relative numbers. Results depend heavily on the characteristics of the data sources and the subsets selected within these. Major factors are listed and discussed. Aggregated data are provided in the MPG data repository.

Keywords: open access, global long term trends, quantitative analytics, bibliographic data sources, Web of Science, Unpaywall

DOI: 10.17617/2.3361428

1 Introduction

Long term global trends in open access are discussed by many actors and from various perspectives.

While the coarse general patterns are largely consistent across studies, there are considerable discrepancies with respect to absolute numbers and more granular patterns. The discrepancies often result from differences in the databases used and the specific definitions of metadata applied within these (Akbaritabar and Stahlschmidt, 2019; Basson et al., 2021; Visser et al., 2021). In this paper we present a comparison of four variants that might span the typical absolute value ranges usually found for summary statistics on the open access status of journal articles.

The data are compiled from global raw data sets of **Unpaywall** (Our Research) and **Web of Science** (Clarivate) including publications from the last two decades. Unpaywall was chosen as it is the only primary data source that provides a comprehensive coverage of open access status of articles worldwide and over the full time period examined. Web of Science is based on a more focused selection of journals and provides a more granular resolution of document types.

2 Data Sources and Methods

2.1 Unpaywall

Unpaywall[1] by Our Research is a nonprofit endeavor to make scholarly research more open and accessible. It links every publication that has been assigned a Crossref[2] DOI to the open access URLs where the paper can be read for free. It harvests publications from over 50,000 locations from all over the world and creates a set of open-access related metadata at publication level. Data are provided via a DOI based API or as snapshots of the full data set. The latter are available for free twice a year and as regular data feeds which are subject to a commercial license. Metadata include a field named ‘oa_status’, which provides a summary categorization of the licensing status encoded by the frequently used “open access colors”.

Open access categories provided by Unpaywall (Piwowar et al., 2018) at publication level:

- Gold: Published in an OA journal that is indexed by the Directory of Open Access Journals (DOAJ).
- Green only: Toll-access on the publisher page but is free in an OA repository.
- Hybrid: Free under an OA license in a toll-access journal.
- Bronze: Free to read on the publisher page, but without a clearly identifiable license.
- Closed: All other publications, including those shared only on an Academic Social Network (like ResearchGate and Academia.edu) or in Sci-Hub.

[1] https://unpaywall.org/
[2] https://www.crossref.org/

2.2 Web of Science

Web of Science[3] by Clarivate is a set of bibliographic database indices with global scope and broad subject coverage. All publications from internationally relevant publication sources that fulfill a set of criteria are included and indexed to a rich set of metadata.

XML raw data are provided in several products. The basic journal indices (SCI, SSCI, AHCI) and the indices dedicated to publications in conference proceedings (ISTP, ISSHP) are licensed by the Competence Center for Bibliometrics[4] via BMBF grant 01PQ13001. The Emerging Sources Citation Index (ESCI) is licensed by MPG and provides an additional set of publications from more than 7000 journals.

Since 2017 Web of Science includes data from Unpaywall (Bosman and Kramer, 2018). These data were not used for the current analyses, instead, the original Unpaywall data were matched with the Web of Science data, see Section 2.3.

2.3 Methods

Both data sources are regularly incorporated into the "Research Information Observatory" data lakehouse developed and run by the Max Planck Digital Library Big Data Analytics Group (MPDL.RIO).

Unpaywall data were ingested via a bulk download from a full data snapshot created in **July 2021**. It covers more than 126 mio scholarly publications. Records are provided in JSON lines format and thus were processed without further transformations.

Web of Science provides weekly data files in XML and CSV format. The data used for this paper included all indices licensed by CCB and MPG and accumulate data by **October 2021**. The data were converted to JSON lines format for further processing.

The JSON lines data were ingested into the PostgreSQL instance of MPDL.RIO and parsed into relational schemata. Cleaned data were integrated into a generalized metadata layer appropriate for quantitative analytics. Records from the two sources were matched via their DOI fields. Apart from setting DOIs to lower case no attempt was made to improve matching in case of missing or malformed DOIs in either source.

Basic analytics were run via PL/pgSQL pipelines in the database, any further processing was accomplished with the aid of the MPDL.RIO visualization and reporting framework based on Python and $\LaTeX$.

2.4 Data Reuse

Reuse of raw data is unproblematic for Unpaywall as the snapshot[5] is publicly available under a permissive license[6]. License restrictions, however, prohibit sharing of any Web of Science raw data at publication level.

Aggregated data used for plots in this paper are published in the MPG data repository (Heidbach et al., 2022).

[3] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[4] https://www.bibliometrie.info

[5] https://unpaywall.org/products/snapshot

[6] https://unpaywall.org/legal/terms-of-service

3 Results

3.1 Characteristics of Data Sources

The selection processes for publications and definitions of metadata differ considerably between the two data sources.

Unpaywall selects items primarily via **Crossref** and thus identifies only scholarly publications with a **DOI**. The data set covers a broad set of publication types, including not only journal-like sources but also books, book sections, reports and other types.

The **Web of Science** indices that are available for us basically select publications from an exhaustive list of **journal-like sources** and index any document type found therein including publications that do not have a DOI.

Document types are not clearly defined by the data providers and the same terms do not necessarily refer to the same concepts. An 'article' in Crossref-based data is basically any document type found in a journal or journal-like series, whereas Web of Science uses almost 50 different document types for journal publications. In Web of Science, an 'article' is defined in a much narrower sense, accompanied by 'review', 'editorial material', 'letter', 'book review' and other less frequent types.

Publication years are also subject to ambiguities. Web of Science seems to stick primarily to the date when a publication becomes part of a defined issue of the journal (only for very recent publication years they added so-called early access versions). Unpaywall seems to prefer the online publication date of the single article. Frequently this results in differences of one year, but can span up to 5 years. For the analyses based on Unpaywall data or Web of Science data we used the respective publication dates found in the data sets.

As can be seen in Figure 1, **absolute numbers** of publications differ substantially between the data sets chosen. Articles and reviews from Web of Science, a typical set often used in bibliometrics, yield a total of 2.5 million publications in 2020. If we add all other document types, we find 3 million publications in Web of Science. Unpaywall covers substantially more sources, a subset for the document type journal articles yields almost 5 million publications, in total more than 7 million scholarly publications are registered.

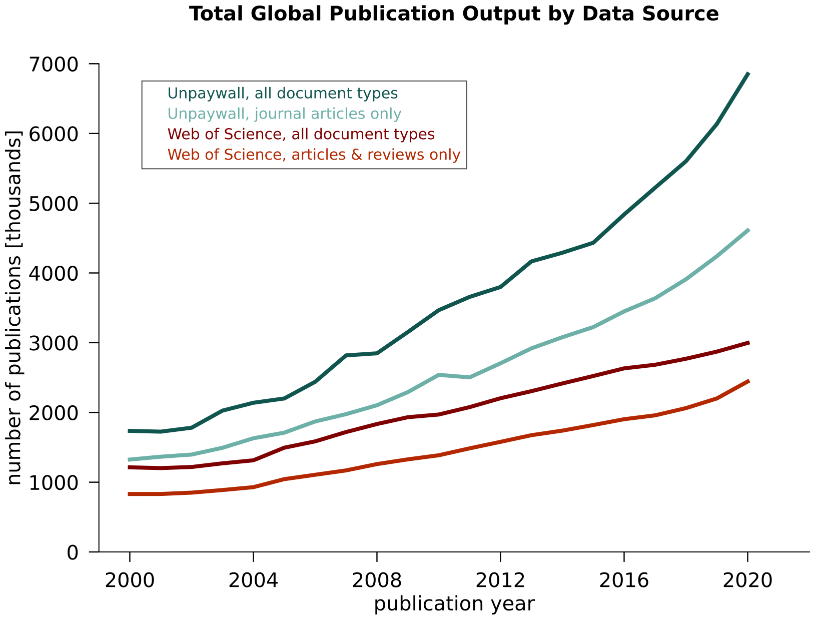

Figure 1 Total numbers for global publication output from the two data sources and subsets defined by document type.

3.2 Global Trends for Open Access Status

Figure 2 shows the open access shares for publications found in Web of Science with document types limited to 'article' and 'review' while Figure 3 shows the same data set for all document types but abstracts. In a similar manner, Figures 4 and 5 show the Unpaywall data for the 'journal article' document type and for all document types.

The open access shares show similar long term trends for all variants considered. The share of green open access has risen constantly from less than 5 per cent in the early years to now around 10 per cent after expiration of embargo periods. By 2010, gold open access had a share of around 6 per cent of the publications but then increased considerably to now more than 25 per cent. Hybrid as defined by Unpaywall is only rarely found for publications from the first decade and then rises continuously. Bronze, the category for openly available publications with no clear license for reuse, is found in moderate amounts for all publication years.

Looking at these trends, we have to take into account that the current situation for older publications does not fully reflect the situation at the time of publication.

So-called **embargo periods** are implemented by subscription journals, opening primarily closed publications to potential green, bronze or hybrid open access after a defined time span. These periods range between 6 and 48 months. This mechanism results in a typical decline in the share of green open access in the most recent publication years.

Even the situation for gold open access can change when subscription journals are **transformed** to gold open access journals and might or might not include older volumes. The Directory of Open Access Journals (DOAJ) has ceased to collect the starting date for gold open access due to the problems in clearly defining it.

Older publications might have been opened by publishers also for many other more reasons.

A considerable fraction of publications registered in Web of Science is not found in Unpaywall via a simple **DOI matching** process (termed 'unknown' in Figures 3 and 4). It amounts to 40 per cent for all document types and 25 per cent for 'article' and 'review' in early years and decreased to 10 and 6 per cent respectively for more recent publication years. This amount is generally decreasing in recent time as publishers add DOIs retrospectively and Web of Science is successively catching up with these dynamics.

3.2.1 Publication Subsets from Web of Science

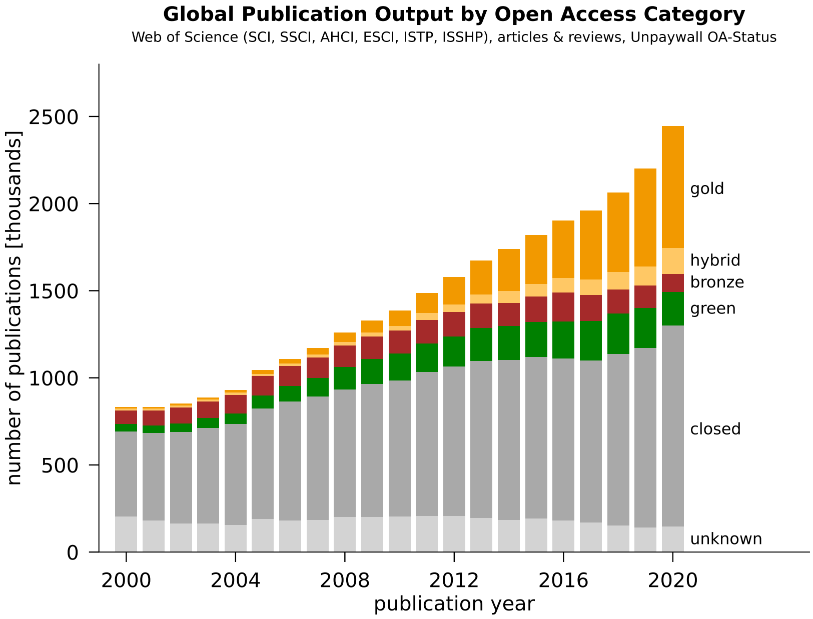

Figure 2 Global publication output based on Web of Science as available for MPDL (see Section 2.2 for details). Document types 'article' and 'review' only. Open access status as from Unpaywall field oa_status (see Section 2.1 for definitions). The light grey areas labeled 'unknown' represent Web of Science records that where not found in Unpaywall via DOI match to Web of Science. Time of observation: October 2021 for Web of Science and July 2021 for Unpaywall.

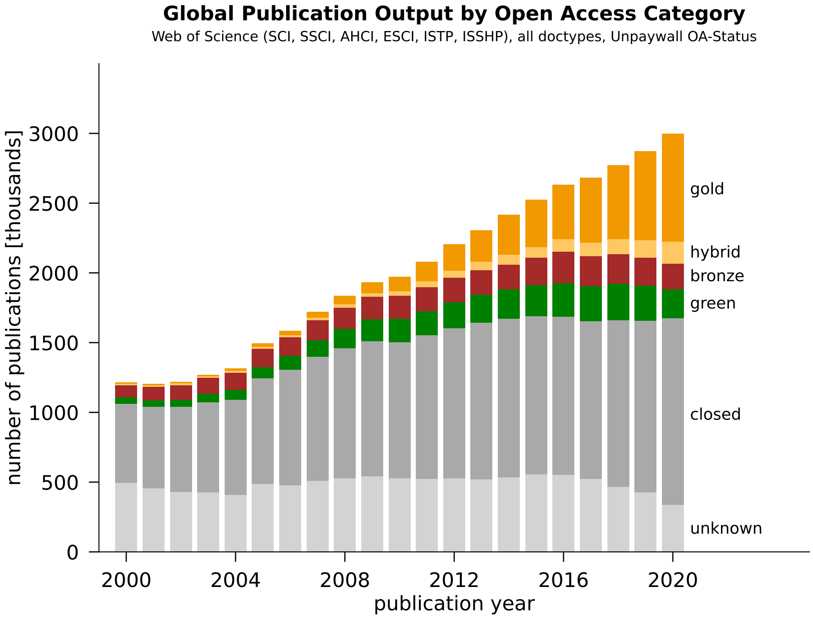

Figure 3 Global publication output based on Web of Science as available for MPDL (see Section 2.2 for details). All document types but abstracts. Open access status as from Unpaywall field oa_status (see Section 2.1 for definitions). The light grey areas labeled 'unknown' represent Web of Science records that where not found in Unpaywall via DOI match to Web of Science. Time of observation: October 2021 for Web of Science and July 2021 for Unpaywall.

3.2.2 Publication Subsets from Unpaywall

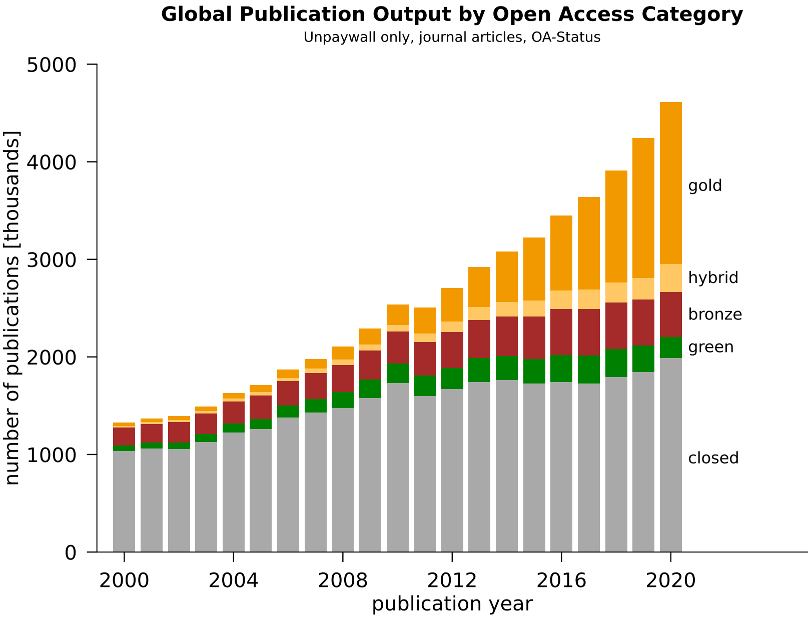

Figure 4 Global publication output based on Unpaywall only, document type 'journal article' only. Open access status as from Unpaywall field oa_status (see Section 2.1 for definitions). Time of observation: July 2021

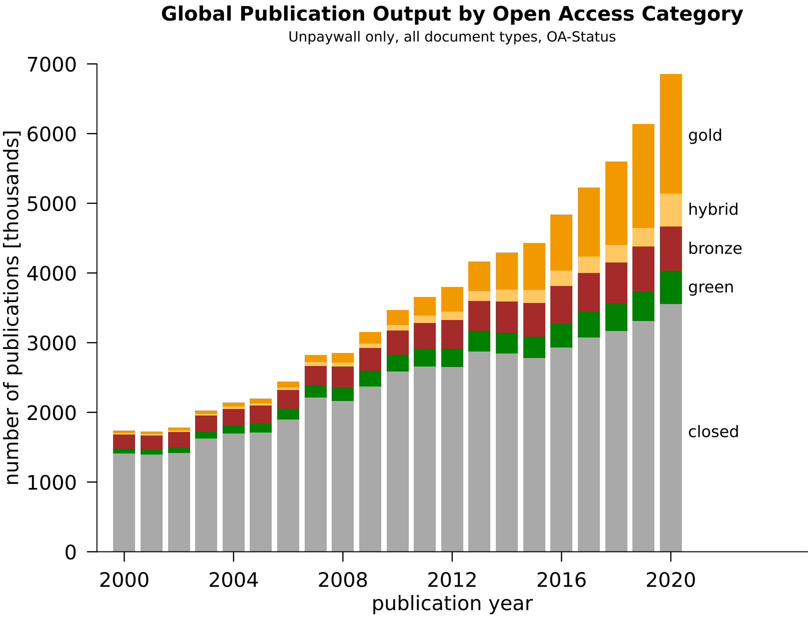

Figure 5 Global publication output based on Unpaywall only, all document types. Open access status as from Unpaywall field oa_status (see Section 2.1 for definitions). Time of observation: July 2021

4 Discussion

Open access has grown with considerable momentum during the last decade. While we can read around 30 per cent of the journal articles published in 2010 without any paywall restrictions, this fraction has grown to around 50 per cent for articles published in 2019. Any precise estimate of absolute or relative numbers, however, is dependent on the data subset chosen.

As might be expected, **total numbers** of publications vary widely between the bibliographic metadata sources available as they follow very different collection policies. In 2020, values range from a total of 2.5 million to 7 million publications per year in the four variants considered. Visser et al. (2021) give an overview on the size and overlap of five global data sources. Their data show that Dimensions[7] (Digital Science) has a very high overlap with Crossref. As the same holds true for Unpaywall, we can estimate that numbers based on Dimensions would be largely comparable to those we have seen for Unpaywall only. Manual checks of summary statistics at their website seem to confirm this conclusion. Scopus (Elsevier) and Microsoft Academic Net follow largely different collection policies. The selection of publications exerts considerable influence on open access shares and trends and introduces biases in the comparison of subsets like countries or institutions (Basson et al., 2021).

Whenever Unpaywall is used for open access information and is applied to a journal based data source like Web of Science we need to account for the coverage of **valid DOIs** in the latter. DOIs were introduced in the 1990s and gained widespread acceptance only by around 2010. Several publishers added DOIs for older publications retrospectively. Databases would need to cover these dynamics and check DOIs for validity. Beyond that, we still see publishers in some scientific domains that do not apply DOIs to their recent journal publications.

The concepts for the definition of **document types** vary widely between data providers and entail systematic effects on the outcome of open access statistics. Including all document types in Web of Sciences typically lowers the share of open access by five per cent as compared to 'article' and 'review' only. Editorials, letters and book reviews, the most common among the other types, might less often be deposited to repositories or might not be included in transformative agreements. These patterns cannot be accounted for when using data sources that rely primarily on Crossref, as there is no distinction of document types for journal publications.

Categories of open access status and their definitions are still under discussion in the bibliometric community (Taubert et al., 2019). The same terms used in different studies do not necessarily represent the same criteria. This is especially true for the category 'hybrid' which gains more attention with the emergence of big transformative agreements since around 2019 (Jahn et al., 2021). In this context, 'hybrid' refers to publications in subscription journals paid by the clients via article processing charges or publish-and-read fees. These publications are immediately open access without any embargoes and thus are only a subset of those categorized 'hybrid' by Unpaywall. Due to embargoes and other publisher decisions we also need to take into account category changes between the time of publication and the time of observation. Piwowar et al. (2019) present an in-depth analysis of these dynamics.

The choice of the data subset, the matching success, the definition of open access categories, and the time of observation have considerable influence on the details of global open access trends. We strongly encourage primary data providers to improve documentation of decisive characteristics of their data sets and data analysts to report these details. A more complete coverage of Crossref metadata by publishers would also help to get clearly defined and comparable open access analyses.

[7] https://www.dimensions.ai

References

Akbaritabar, A., & Stahlschmidt, S. (2019). Applying Crossref and Unpaywall information to identify gold, hidden gold, hybrid and delayed Open Access publications in the KB publication corpus. *SocArXiv*. https://doi.org/10.31235/osf.io/sdzft

Basson, I., Simard, M.-A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2021). Data sources and their effects on the measurement of open access. Comparing Dimensions with the Web of Science. In *18th international conference on scientometrics & informetrics*.

Bosman, J., & Kramer, B. (2018). Open access levels: a quantitative exploration using Web of Science and oaDOI data. *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.3520v1

Heidbach, K., Knaus, J., Laut, I., & Palzenberger, M. (2022). Long term global trends in open access. Supplementary material. *Max Planck Society*. https://doi.org/10.17617/3.8s

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. 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

Piwowar, H., Priem, J., & Orr, R. (2019). The future of OA: a large-scale analysis projecting Open Access publication and readership. *bioRxiv*, 795310. https://doi.org/10.1101/795310

Taubert, N., Hobert, A., Fraser, N., Jahn, N., & Iravani, E. (2019). Open Access–towards a non-normative and systematic understanding. *arXiv*, 1910.11568. https://arxiv.org/abs/1910.11568

Visser, M., van Eck, N. J., & 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