ELEPHANT
IN THE LAB

OPINION

Why we need a public infrastructure for data on open access

Short title	Why we need a public infrastructure for data on open access
Long title	Why we need a public infrastructure for data on open access
Authors	Mikael Laakso [1]
Author affiliation	[1] Hanken School of Economics, Helsinki, Finland
Author bios	Mikael Laakso is an Associate Professor at Hanken School of Economics, Helsinki Finland. He has been researching the changing landscape towards openness in scholarly publishing by studying combinations of bibliometrics, web metrics, business models, science policy, and author behavior. He is also active in national and international working groups (e.g. as a member of the European Commission's Horizon 2020 expert group on Future of Scholarly Publishing and scholarly Communication, and a member of the national strategy groups for open science and national library consortium publisher negotiations).
Author social links	Mikael Laakso: ORCID – ResearchGate – Twitter
Date published	16th of January 2019
DOI	10.5281/zenodo.2540472
Cite as (APA)	Laakso, M. (2019). Why we need a public infrastructure for data on open access. *Elephant in the Lab*. DOI: https://doi.org/10.5281/zenodo.2540472

During the last decade I have been conducting research on scholarly communication, primarily focusing on how open access in various forms has been introduced into an environment traditionally supported by subscription-based distribution models. Establishing the historical development and current status of journals and articles publishing open access still requires a lot of manual data collection. Insights on open access, and the development of scholarly publishing in general, is not only something of merely bibliometric research interest. In 2019, readily-available data on the state of open access is still limited, even though open access

publishing has become a crucial goal promoted by funders and policy-makers worldwide. The scientific enterprise at large would benefit from more informed science policy, and by having access to better data and metrics about the journal publishing landscape; metrics that would be standardized and could be followed-up.

From my point of view, this is a symptom of a much larger issue - that obtaining a comprehensive bird's-eye view of journal publishing often relies on using the data and services of commercial bibliographic database providers. This not only risks vested commercial interests informing science policy (Tennant 2018), but also maintains the underlying data infrastructure as proprietary and closed. With only limited information on the research landscape – and thereby limited information on the effects of policy interventions – as factors influencing and steering policy is, to put it mildly, not optimal.

Even though journals and the articles they publish can be considered digital artefacts, their indexing and semantic interlinking is still in need of development. Before explaining some of the primary limitations and problems that I have come across during the data collection for my studies, I want to highlight that Crossref has been doing important improvements in this area, e. g., by launching and maintaining the DOI system and providing open APIs and common standards for journals to deposit metadata (incl. Open Citations). Similarly the Directory of Open Access Journals (DOAJ), which was founded in 2003, is doing an important job; without which we would be completely lost. However, there is still a major gap in what we know (and what can be known) about the historic and current state of journal publishing due to the lack of an open and public infrastructure to track and summarize journal activities. The technical implementation can be realized in many different ways, but the end-result should be able to provide an up-to-date and historically accurate representation of how the journal landscape evolves over time. In this post I focus primarily on identifying the deficiencies that this blind spot has for knowledge about open access publishing, but the utility of an improved information environment for journal tracking for both practice (i.e. policy-making, infrastructure development) and research goes far beyond that.

Current gold standards and why improvement is needed

Indicative of the lack of information on journals is the fact that the most comprehensive mapping of the longitudinal development of open access journals has been put together manually by visiting over 10 000 journal websites and counting the number of articles published (Crawford 2018). Though an important contribution to the knowledge available, even this massive undertaking can only provide answers to the "simple" questions that pertain to existing and active open access journals, for example how many journals and articles are published open access annually per discipline, and what their pricing levels for article processing charges are. In

another recent landmark study on the longitudinal growth of open access in journals and elsewhere on the web, Piwowar et al (2018) analyzed data collected from users of the Unpaywall browser extension which is used to facilitate finding open access copies of articles. While providing a huge jump in scale and fidelity of what is known about the prevalence of open access provision mechanisms in and out of journal webpages, having to rely on bottom-up user-provided observations–rather than having first-hand sources available at both the article and journal level–shows that there are limitations in the availability of data concerning the current environment for disseminating scholarly journal articles.

Improved precision and coverage of bibliometric data would allow us to begin answering the following questions:

- How is open access publishing growing in comparison to the overall growth of science, in terms of number of journals and number of articles?
- How many journals publishing open access started as open access journals?
- How many journals publishing open access have flipped to open access from first being subscription-based?
- How many open access journals have become subscription-based?
- How have article processing charges developed over time? Is this development uniform across research disciplines/publishers/countries?
- How many journals (open access or not) have became inactive?
- How many articles have been published as hybrid open access in subscription-journals?
- How many articles were published as delayed open access in subscription-journals last year?
- How has the market for open access publishing changed since the introduction of a specific policy intervention (e.g. Plan S)?
- How has the development and adoption of public open access infrastructures progressed?

(Since this post is primarily focused on the journal landscape these questions do not even go into the data and monitoring problems specific to green open access/self-archiving which has a lot of specific limitations as well.)

There are services like Web of Science, Scopus, Ulrichsweb, Dimensions, The Lens, the ISSN Portal, DOAJ, and Crossref that are built on large volumes of publication data and can be queried

in very advanced ways, why is this still a problem? I will review some of the central shortcomings I have come across here.

Three key obstacles in current indexing

1. **Amnesia**

   One key methodological issue is that current bibliometric databases can only deliver a snapshot of results for any query one composes, **they are not designed to deliver time-series data that would account for classification and status changes of individual journal metadata**. If one is interested in changes one would need to download, or obtain an old archived version of the database filtered by the same query, and make comparisons to detect things that have been added, removed, or modified over time. This is simply not realistic for most use cases. The problems with keeping track of journals are in many ways similar to the problems related to monitoring any longitudinal changes happening on the Internet. Service providers rarely retain or archive website versions as new ones get rolled out, with hyperlinks and content at risk of also getting lost in the shuffle. The Internet Archive is a great service, but its snapshot coverage is limited as I have personally come to realize as I have sifted through thousands of archived journal websites during the last two years. It's a good service for many purposes but is not an optimal solution to keep track of changes to scholarly publication outlets.

2. **Selective coverage**

   Another issue is created by each bibliometric database coming with its own **biases and limitations in how comprehensively journals across disciplines, countries, and languages are selected for inclusion**. Web of Science is widely considered to be the most restrictive database for including journals, while Urichsweb is more inclusive but lacks e.g. article counts for journals. DOAJ have their own set of requirements for journals that apply to become included, if journals go through the process of applying in the first place. Many indexes only include active journals, where the records of journals that become discontinued might be removed as time goes on. Crossref with its DOI system has widespread adoption but not all journals are enrolled in the DOI system. To get the most comprehensive perspective on the landscape which data sources would one select? How then to establish the baseline population of journals and also obtain e.g., article counts for said population?

3. **Commercial dominance**

   A central issue is the commercial nature of many of the most comprehensive databases for these purposes, e.g. Web of Science, Scopus, and Ulrichsweb. Being designed primarily for institutional subscribers **access to them is limited**, and **datasets created on**

the basis of such data can rarely be freely redistributed in their most usable form. The databases might change owners, be discontinued, or radically change with short notice. Bibliometric research is often conducted on top of these commercial databases but having better open alternatives would enable more comprehensive and reproducible research as well removing potential biases and commercial incentives from influencing science policy.

Conclusion

This post is intended to raise awareness about the current drawbacks of the information landscape around journals, and particularly how it relates to our limited knowledge about the history, current status, and trajectory of open access journal publishing. Unfortunately, there are no easy solutions but the next step would be to initiate a wider discussion about potential ways to provide access to journal metadata as well as web services to aggregate and present the data in an usable way.

As we are speaking, a lot of movement in the journal landscape is going unnoticed, perhaps forever, since retrospective representation of what happens on the Internet is so incomplete. Journals are not static; they are living and breathing things which commonly switch publishers, merge together, change mode of access, or can even vanish from the web one day. Most of the questions listed earlier about the open access journal landscape cannot be answered retrospectively, there would need to be registration at the time any changes are made. The solutions that could facilitate a more open and comprehensive infrastructure is not something any single actor can solve alone. Rather requires a concerted effort including policy makers, infrastructure developers, libraries, and most of all funders who already play an integral part in conjunction with their overall mandating of open access publishing. In my mind, it is not unrealistic to design targeted calls for research and calls for tenders for a dual purpose; uncovering new knowledge about scientific communication and knowledge-building at the same time as contributing to a collective open data infrastructure. The EU already has had many key initiatives in place, e.g. the European Open Science Cloud, the Open Science Monitor, and OpenAire but their synergies with each other and with Horizon Europe could leveraged further to work towards the goal of an open data infrastructure for open access.

For the immediate future it would be important that resourcing of vital infrastructures for open access data would be secured through The Global Sustainability Coalition for Open Science Services (SCOSS), as that enables further expansion and development of e.g. DOAJ which has already accumulated and continues to accumulate a lot of key data points for open access journals. DOAJ already has collaboration with the ISSN organisation (DOAJ.org 2018) but more could likely be done in order to create data that would answer some of the open questions stated

earlier. Further development and temporal archival of Crossref journal-level metadata is also a promising proposition for reaching better immediate and open insight on the living and breathing scholarly communication landscape. Unless open data and tools for keeping track of scholarly journal activity is provided commercial players will remain the dominant providers of analytics based on proprietary pools of data that can only tell a partial picture about the scholarly journal publishing. And since these commercial players are also active in other parts of the scholarly landscape, e.g. by being heavily invested in journal publishing themselves, reliance on that partial picture is very problematic.

Placing a finger on the pulse of open access journal publishing should not require months of experimental research or manual data collection. It should be one click away in order to facilitate improved discoverability, measurability, and decision-making. Science policy has been pushing hard for open access, but open data and tools for measurement and follow-up are still missing. Many pieces of the puzzle are already in place, but more work is needed to provide services that would mitigate the three central obstacles outlined in this post.

References

Crawford, W. (2018). GOAJ3: Gold Open Access Journals 2012-2017 LINK.

DOAJ.org (2018). ISSN and DOAJ: a renewed partnership LINK.
(accessed: 16th January, 2019)

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

SCOSS (2018). The Global Sustainability Coalition for Open Science Services (SCOSS) LINK.
(accessed: 16th January, 2019)

Tennant, J. (2018). Elsevier are corrupting open science in Europe. The Guardian. 29 June 2018 LINK.