Research impact of paywalled versus open access papers

Éric Archambault, Grégoire Côté, Brooke Struck and Matthieu Voorons
Science-Metrix and 1science, 1335 Mont-Royal E., Montreal, Quebec, Canada, H2J 1Y6

This note presents data from the 1science OAIndx on the average of relative citations (ARC) for 3.3 million papers published from 2007 to 2009 and indexed in the Web of Science (WoS). These data show a decidedly large citation advantage for open access (OA) papers, despite them suffering from a lag in availability compared to paywalled papers.

Background

There is an abundant literature on the citation advantage of OA papers, starting with a succinct communication by Lawrence (2001). Several studies have been listed by SPARC, the majority of which support the idea that when papers are openly available, they are more cited than papers for which availability is restricted to those who pay for access (http://sparceurope.org/oaca/).

As noted by Diana Hicks on the ScienceMetrics.org blog (http://sciencemetrics.org/oaca-open-access-citation-advantage/), skeptics argue that the advantage of OA is partly due to citations having a chance to arrive sooner. Another purported artefact would be a selection bias according to which authors pick their best (hence most citeable) papers to make OA. This paper examines the first aspect and concludes that the purported head start of OA papers is actually contrary to observed data.

The limitation of the existing literature on the subject, whether supporting or refuting the OA citation advantage, is often the small number of articles analyzed, the limited size and diversity of the citing sources, and the short citation window considered. The present note examines 3,350,910 papers published between 2007 and 2009 and indexed in the WoS, with a citation window starting in 2007 and continuing up to the latest date possible (in practice, mid-2016). More than 12,000 journals indexed in the WoS were used to compute the citations for these papers, which have received a total of 34,865,430 citations from these citing journals.

Methods and definitions

Here, *open access* refers to Peter Suber's definition of *gratis OA* (http://sparcopen.org/our-work/gratis-and-libre-open-access/), while *libre OA* papers are considered a subset of gratis OA and are therefore also included in the present data. *Gold OA* refers to papers made available for free by the publishers themselves, be it on their website (e.g., in fully gold OA journals on Springer Open and BioMedCentral, or as *hybrid OA*, that is, OA papers from otherwise paywalled journals on, for example, Springer's website) or on the site of an aggregator (e.g., Scielo, and also PubMedCentral, on which the majority of papers are archived by the publishers themselves).

Green OA refers to papers made available for free by parties other than publishers, usually the authors themselves, who archive papers in institutional repositories, subject repositories such as arXiv, or commercial repositories such as ResearchGate. The scores for OA as a whole include uncategorized open access articles (for which we cannot determine whether the paper is gold or green).

Determination of whether a paper is gold or green relied on the manual coding of top-level domains (TLDs). Around 8,100 TLDs holding 26,500,000 URLs were manually checked and coded. The remaining 230,000 TLDs were tagged as unknown for one of three reasons: (1) they did not have a clear open access status for a variety of reasons (e.g., a language barrier meant the team could not determine the nature

of the documents on a site); (2) they could not be qualified with a unique tag (e.g., some repositories have both green and gold papers as they host journals published by the institution while also having green papers by their researchers; or (3) they had too few papers to be worthy of manual checking at this first stage of coding (TLDs with less than 500 URLs).

1science is working on algorithms to properly code hybrid papers, but these efforts will only come to fruition in the future, and hybrid papers are not considered as a distinct category here (as they are found on the publisher's site, they are coded as gold). Note that the coding was done at the URL level and as such a paper could be coded as both gold and green if a copy appeared on a publisher's website and, for example, also on an institutional repository. This note does not examine the overlap between green and gold.

The 1science OAIndx is a growing aggregate that currently comprises 18 million gratis OA papers and their metadata. All papers in OAIndx, regardless of whether they are green or gold OA, are published in peer-reviewed journals. OAIndx covers pre-prints and post-prints, but only to the extent that metadata triangulation can ascertain that these were actually published, and hence pre-prints that did not ultimately make their way to a published version are not included in OAIndx. A pre-print is defined as a paper that has yet to undergo peer review, while a post-print is a paper accepted for publication following peer review, or, in some fields, editorial review or other types of scholarly quality review.

Only records possessing enough metadata to produce a bibliographic reference (the minimal set is defined as the paper's first author, year of publication, article title and journal title) and having a directly downloadable PDF document are included. This therefore creates a number of false negatives: some papers are gratis OA but do not have all required metadata available, some are only available in other formats (EPUB, Word, etc.), some could be readable only online with no means to download (HTML), and some would say that registering to use a website does not constitute closed access.

The WoS comprises bibliographic records for 44 million articles published in scholarly journals between 1980 and 2016. The statistics presented here are based only on original contributions to knowledge (WoS document types: Articles and Reviews). All citations to these articles are counted (i.e., there is no "sliding citation window"). The impact is computed for papers that were published sufficiently long ago to have reached their peak citation number per year, although in the arts and humanities it is more challenging to determine what the peak citation year is, provided such a thing exists in these domains.

The average of relative citations (ARC) is an indicator of the impact of papers on the research community. The number of citations received by each of the 3.3 million papers is counted for the year in which it was published and for all subsequent years. For papers published in 2007, for example, citations received in the 2007–2016 period are counted.

To account for different citation patterns across fields and subfields of research and scholarly publications (e.g., there are more citations in biomedical research than in mathematics) and for differences in the age of publications (e.g., older papers have accumulated citations over a longer period), each publication's citation count is divided by the average citation count of all publications that were published the same year in the same subfield, to obtain a relative citation count (RC).

The ARC of gold OA, green OA, and not OA papers in a given subfield is the average of the RCs (i.e., if there are 4,000 publications in the subfield, the ARC is the average of 4,000 RCs, one per publication). An ARC value above 1 means that a class of papers is cited more frequently than the average citation level for all papers in that field; when it is below 1, it means the class is cited less frequently.

Findings

- Publishing in paywalled journals without green archiving is never an effective impact maximization strategy.
- In total, and for all these fields, publishing in paywalled journals with no additional green archiving always yields below average citedness (the average being 1.0).
- Publishing in paywalled journals is the least impactful strategy overall, and the least impactful in 16 out of 22 fields.
- On average, open access papers produce a 50% higher research impact than strictly paywalled papers.
- In all these fields, fostering open access (without distinguishing between gold and green) is always a better research impact maximization strategy than relying on strictly paywalled papers.
- Having a green copy of a paper is the most impactful research communication strategy overall and the best strategy in 19 fields out of 22.
- Green is nearly always more effective than relying strictly on gold (20 out of 22 fields).
- Gold is the best strategy in biology and biomedical research and very close to green in clinical medicine (likely a reflection of the NIH and Wellcome Trust OA mandates).
- Gold has the least impact in six fields.

Table 1 Research impact of paywalled (not OA) versus open access (OA) papers

Field	Paper (*n*)	Reference (*n*)	not OA (*arc*)	OA (*arc*)	gold (*arc*)	green (*arc*)
TOTAL	3,350,910	34,865,430	0.81	1.23	1.06	1.28
Agriculture, Fisheries & Forestry	138,025	804,386	0.85	1.18	0.73	1.35
Biology	151,424	1,882,514	0.74	1.17	1.33	1.18
Biomedical Research	291,325	5,581,332	0.80	1.14	1.16	1.09
Built Environment & Design	16,648	84,825	0.83	1.28	0.79	1.35
Chemistry	317,930	2,432,155	0.90	1.24	0.65	1.34
Clinical Medicine	823,924	9,323,440	0.81	1.28	1.25	1.28
Communication & Textual Studies	28,178	37,152	0.78	1.93	0.81	2.16
Earth & Environmental Sciences	117,429	1,332,707	0.82	1.16	0.82	1.20
Economics & Business	66,037	607,155	0.65	1.25	0.67	1.27
Enabling & Strategic Technologies	250,651	2,404,079	0.89	1.18	0.79	1.30
Engineering	193,856	1,029,715	0.85	1.25	0.86	1.36
General Arts, Humanities & Social Sciences	3,932	11,757	0.65	1.69	0.99	1.65
General Science & Technology	31,793	1,906,904	0.93	1.10	0.84	1.20
Historical Studies	25,468	50,016	0.80	1.58	0.68	1.91
Information & Communication Technologies	97,786	582,010	0.72	1.23	0.98	1.27
Mathematics & Statistics	107,426	558,567	0.78	1.14	1.12	1.22
Philosophy & Theology	17,117	28,107	0.70	1.74	0.76	1.98
Physics & Astronomy	424,091	3,954,894	0.75	1.27	0.92	1.34
Psychology & Cognitive Sciences	70,022	1,026,674	0.69	1.23	1.15	1.19
Public Health & Health Services	85,703	804,085	0.83	1.17	1.00	1.23
Social Sciences	86,513	421,516	0.69	1.49	0.89	1.63
Visual & Performing Arts	5,632	1,440	0.83	2.19	1.17	2.69

Source: Computed by Science-Metrix and 1science using OAIndx and the Web of Science

Known limits and potential effects on findings

These data examine the impact of scholarly and research papers on the research community. This measure is therefore only an indication of the impact of these papers on further research, not of broader impact. Although one can surmise that open access papers may have a substantially greater economic and social impact than strictly paywalled papers, one cannot infer this from the present data. Furthermore, these data examine only communication strategies based on the use of scholarly and research journals; other media (e.g., books, conference proceedings) may have a greater impact on the research community, such as books in the humanities.

OAIndx comprises as many papers outside the WoS as it does within the WoS. It is presently impossible to know the precise effect of studying only half of the open access papers available. The "hidden" part is presently difficult to study as it isn't covered in the two mainstream citation databases. The WoS and Scopus both have similar selection strategies centred on large, established, highly cited journals, mostly in English and using a Latin character set—whereas OAIndx aims to be as comprehensive as possible, provided papers are published in quality-controlled scholarly journals.

Currently, the 1science harvesting technology has a recall of about 75%, meaning about 25% of false negative OA papers have been categorized here as paywalled, although they are actually available in open access. Considering that OA papers generally have higher impact, one can conclude that if all true positives (articles with an OA version) were removed from the paywalled category, the impact of paywalled papers would be even lower in several fields. Please note that all the findings are based on aggregate values and can in no way be used to predict the performance of individual papers.

Discussion

These data strongly corroborate the hypothesis that, regardless of the reason why, open access papers have a citation advantage over papers that are not found to have an openly accessible version. That said, there are strong data (based on a distribution comprising 17,453,126 papers) going against the hypothesis of an advantage accrued to an earlier availability of papers in OA. As shown in Figure 1 (and in all similar OA availability curves), a generalized latency is created by embargoes and by researchers being busy doing other things and therefore not posting papers promptly.

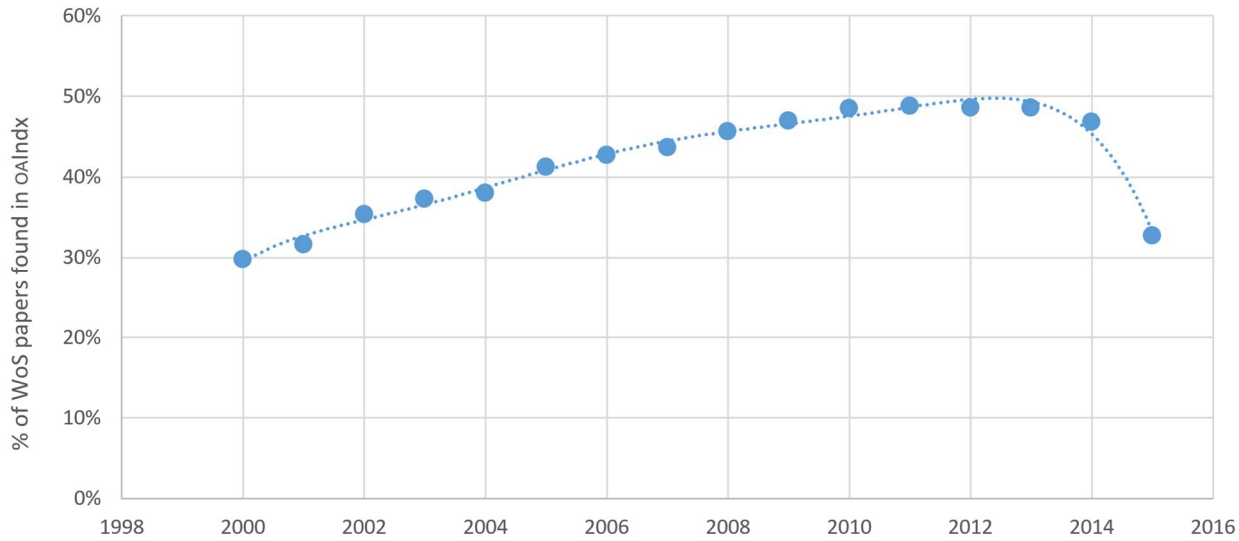

Figure 1 **Number of OA papers per publication year in OAIndx**
Source: Computed by Science-Metrix and 1science using OAIndx and the Web of Science

Availability curves for open access show that the percentage and the overall number of OA papers are actually always lower for the most recent years (see also Archambault, 2014). This evidence provides a strong argument against the "head start" hypothesis: many papers that eventually find their way into open access are actually not available in OA until one or two years after publication. Hence, on average, non-OA papers have a head start, and ascribing a citation advantage to OA that would be due to a longer citation window seems to be founded on beliefs, small sample sizes, a narrow study concentrating on fields where green OA does indeed have a head start (which is certainly the case in physics due to the well-established habit of using arXiv), inadequate protocols, or any combination of these factors.

The conjecture that is left open by this note is that of a selection bias. Although it would be surprising if such a large difference in the citedness of OA papers (ARC = 1.23) versus non-OA papers (ARC=0.81)—which means that on average gratis OA papers are 52% more highly cited than strictly paywalled papers—could be largely ascribed to a selection bias, none of the data presented here can refute that belief. In fact, if there is such a thing as a selection bias, the long citation window used in the present study and the selection of publication years (2007–2009) when OA wasn't as mainstream as today, could amplify that bias. A forthcoming study will address self-selection with a sufficiently large study size, using varied and large citation sources, and proper controls.

Conclusion

This short communication confirms, yet again, that a citation advantage exists for OA papers and that this is especially true for those made available through green routes. This note refutes the idea that open access papers have a head start vis-à-vis strictly paywalled papers. While a head start is likely present in physics and perhaps also in some other fields, the data shown here and in Archambault (2014)—the two studies on OA that surpass the million-papers study size—demonstrate that OA papers suffer from a lag between publication and availability in OA, which is the effect of embargoes and other phenomena. Thus, OA papers certainly do not benefit from a head start. Despite their late start and the ensuing shorter citation window, OA papers have an indisputable citation advantage.

Acknowledgments

The authors would like to thank Ivy Anderson, Kristin Antelman, Stevan Harnad, Peter Suber and Mat Willmott for their comments on an earlier version of this note. Any limitations and inaccuracies are strictly ours.

References

Archambault, E. et al. (2014). *Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013*. Produced by Science-Metrix for the European Commission. http://science-metrix.com/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

Hicks, D. (2016-07-22). *OACA – the open access citation advantage*. http://sciencemetrics.org/oaca-open-access-citation-advantage/

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*. 411(6837): 521. http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Suber, P. (2008). Gratis and Libre Open Access, *SPARC Open Access Newsletter*, August 2008 issue, http://sparcopen.org/our-work/gratis-and-libre-open-access/