Cui Bono? Cumulative Advantage in Open Access Publishing

David Pride, Matteo Cancellieri and Petr Knoth

The Knowledge Media Institute, The Open University, Milton Keynes, UK.
{david.pride, matteo.cancellieri, petr.knoth}@open.ac.uk

Abstract. This study examines the differences in production and consumption of Open Access (OA) literature across institutional prestige variables and examines who is gaining the most benefit from the adoption of current OA publishing practices. In this approach we define production as the publication of OA literature (as a proportion of all research literature produced) by an entity (author, institution, country, continent). We define consumption as evidence of using OA literature as measured by citations to OA literature. Using data points for over 24,000 institutions we examine the role of institutional prestige in the Open Access landscape. Overall, we find medium to strong correlations between OA production and OA consumption. We find that higher ranked institutes are both greater producers and consumers of Open Access literature. Importantly, we find a stronger correlation for higher ranked institutions compared to lower ranked ones when using ranking data from the Times Higher Education (THE) World University Rankings. This indicates that it is the higher ranked and more prosperous institutes that are best placed to benefit from current Open Science and Open Access publishing structures.

1 Introduction

We approach our initial study with the use of two paradigms, production and consumption. In this approach we define production as the publication of OA literature (as a proportion of all research literature produced) by an entity (author, institution, country, continent). We define consumption as evidence of using OA literature by an entity as measured by citation to OA literature.

Using the production and consumption framework, it is possible to measure production and consumption in multiple ways. In this study, we focus on measuring the OA Production Rate, i.e. the proportion of all papers produced by an entity that are OA. While OA production is somewhat straightforward to measure, there are multiple ways in which OA Consumption could be measured. For instance, one option would be to measure the proportion of OA paper downloads by an entity. However, such data are not currently publicly available. As a result, we estimate the OA Consumption Rate as the proportion of OA references an entity (authors, institution, country, etc.) cited in the research papers this entity produced (as a proportion of total references).

Terminology	Description
Production	The publication of research papers by an entity (continent, country, institute)
OA Production	Research papers produced by an entity. We use Unpaywall data to distinguish between OA and non-OA literature.
OA Production Rate	The proportion of OA research papers produced by an entity.
Consumption	The use of a research paper by an entity as measured by citations in that entity's publications.
OA Consumption	The use of an OA research paper as measured by citations in that entity's publications.
OA Consumption Rate	The proportion of OA research papers used by an entity. In our work, we use as evidence of use the act of citing OA research literature in manuscripts produced by an entity.

Table 1. Terminology used in this study

The subsequent analyses of OA consumption take as a basic assumption that one can only cite what one has read. We understand this is a somewhat imperfect assumption due to two potential confounding factors: (1) that people may indeed often cite articles that they have in fact not read, and (2) "shadow library" websites (most prominently Sci-Hub). As for the former point, we acknowledge that it has been shown that authors sometimes cite research that they have not read (Ball 2002; Bornmann and Daniel 2008). Given our quantitative methods, we are unable to take account of this factor here. We hence treat this as a limitation of our study.

2 Related work

Recent work by [1] investigated the production of OA literature around the globe based on institutions present in the THE rankings. They found that, in 2017, the 100 top-ranked universities made 80–90% of their research publications available as Open Access. In 2017, [2] undertook a comparison of institutional performance using data from the Leiden Ranking and found that research performance differences among universities mainly stem from size, disciplinary orientation and country location. The authors state that this result underlines, yet again, that larger universities systematically over-perform in citation rankings. However, the exact cause remains under-researched [3]. Regarding citation behaviours, most studies conclude that OA articles receive more citations than articles that are behind paywalls ([4], [5], [6]).

3 Data Sources

Data regarding institutions, authors and articles for these experiments come from the Microsoft Academic Graph (MAG) dataset [7] which as of June 2021 contains

260,423,032 papers. We use the university ranking data from THE World University Rankings and from the Leiden Rankings [8] which we derive our performance / prestige metrics when undertaking comparisons of individual institutions. We used the Unpaywall API to ascertain the OA status of each paper in the dataset.

4 Methodology

In this study we investigate the level of production and consumption of OA literature at an institutional level. We first examine the levels of OA production and consumption (measured as a proportion of all production). We then correlate this at the institutional level, and also measure this using THE ranking and Leiden Ranking data.

We used MAG data to collate all papers with complete metadata to the publishing institution. This methodology identifies 219m paper / institute pairs, representing 84% of the total MAG corpus. We then collated metadata and all known citations for all papers where the institution and author data were complete. From the complete MAG data, we were able to collate identifiers for 219m papers by 44m authors from 24,000 individual institutions (all figures are close approximations). We then use the Unpaywall API[1] to ascertain the OA status of each paper.

5 Results

5.1 Institutional ranking and OA consumption

This study was undertaken to determine whether there was a link between the prestige of an institution, using a range of different ranking methodologies, and the levels of OA consumption at these institutes.

When using ranking data from the THE rankings, we find a statistically significant difference in the amount of OA content cited by differently ranked institutions. Institutions ranked in the top third on average cite 13% more open access content than those in the bottom third. (Figure 1).

Figure 2 uses the same dataset of papers, authors, institutions and citations but uses the ranking taken from the Leiden Rankings. The Leiden Ranking are based on bibliometric data. The THE rankings use a proprietary ranking system, the exact calculations for which are not publicly available. It is therefore an interesting result that we only observe a difference in citation rates when using the THE data. The results obtained using the THE ranking data would suggest that lower ranked institutes tend to cite a smaller percentage of OA research papers than their higher ranked counterparts which seems counter-intuitive.

[1] https://unpaywall.org/products/api

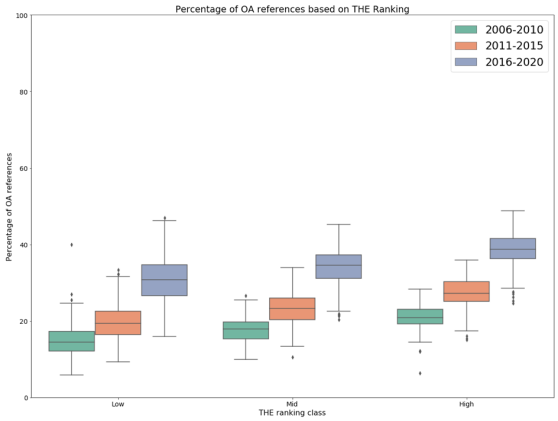

Fig. 1. Percentage of OA references over time (2006-2020) by THE Ranking

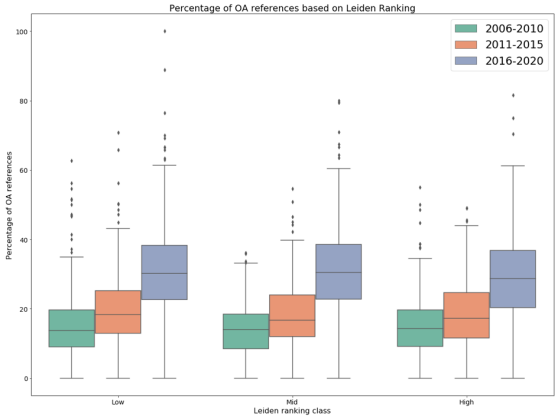

Fig. 2. Percentage of OA references over time (2006-2020) by Leiden Ranking

5.2 Correlation between production / consumption and institutional ranking

We examined the correlation between the production and consumption of OA literature for institutions in the THE World University Rankings using Pearson's Correlation Coefficient.
Figure 3 shows the correlation between OA production and OA consumption at the institutional level for two different time periods.

Each dot is a single institution and covers all institutes in THE rankings. The dots are coloured and sized according to the institutions' ranking. It can been seen from these results that during both time periods, there was a far stronger correlation between the production and consumption of OA literature

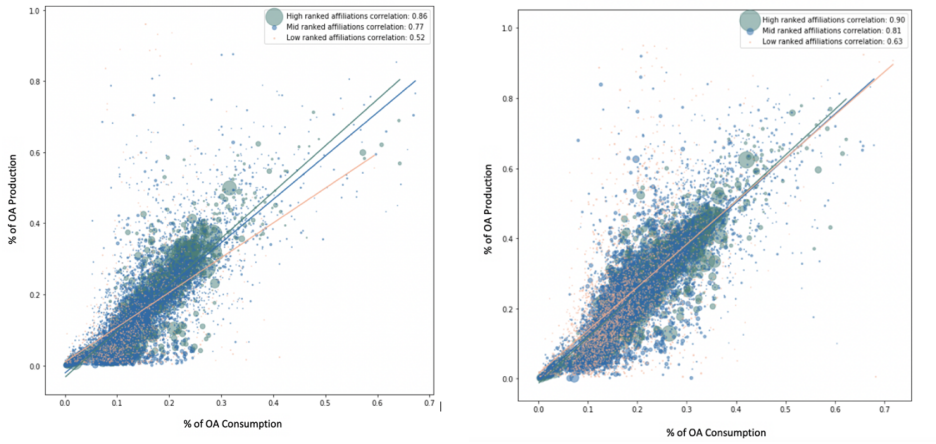

Fig. 3. Correlation of OA production and OA Consumption based on THE ranking. Left: 2011-2015, Right: 2016-2020

for institutes ranked in the top third of institutes using the THE rankings. For the period 2011-215, the highest ranked institutes showed a very strong correlation of *r=0.86* whereas for the lowest ranked institutes this figure was significantly weaker at *r=0.52*. For the second period, 2016-2020, the correlation for all ranks increased and the gap closed slightly; for high ranked institutes *r=0.90* and for the lowest ranked, *r=0.63*.

This change was largely driven by lower ranked institutions increasing rates of production of OA literature. Overall, however, the lower ranked institutions both produce and consume less OA when measured using the THE World University Rankings. There are several reasons why this may be the case. Higher ranked institutions were early adopters in building OA infrastructure and potentially realised its benefits earlier than the lower ranked institutions. The size, wealth or location of the institution in question are all potential confounding factors here and these differences remain to be investigated in future work.

6 Conclusion

The narrative that has formed throughout this study is that the production and consumption of OA literature is highly correlated at the institutional level. We observe the more highly ranked institutions, when using THE rankings, are both greater producers and greater consumers of OA than lower-ranked institutions. One explanation for this phenomenon might be that higher ranked institutions had the resources to invest in OA, that they became the first movers, advocates and adopters of OA, and that their strategy is being followed by the lower ranked institutions. A recent study by Siler et al. (2018) showed that, for the field of Global Health, lower-ranked institutions are more likely to publish in

closed outlets. Their rationale here is that this is due to the cost of Article Processing Charges (APCs) levied by the publishers. This study shows some early indications that it is the higher ranked, better funded institutes which are best placed to capitalise on the Open Access movement.

7 Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 824612 related to the Observing and Negating Matthew Effects in Responsible Research and Innovation Transition (ON-MERRIT) project.

References

1. Neylon C, Hosking R, Montgomery L, Wilson KS, Ozaygen A, Brookes-Kenworthy C, et al. Evaluating the impact of open access policies on research institutions. eLife. 2020.
2. Frenken K, Heimeriks GJ, Hoekman J. What drives university research performance? An analysis using the CWTS Leiden Ranking data. Journal of informetrics. 2017;11(3):859-72.
3. Bornmann L, Mutz R, Daniel HD. Multilevel-statistical reformulation of citation-based university rankings: The Leiden ranking 2011/2012. Journal of the American Society for Information Science and Technology. 2013;64(8):1649-58.
4. Holmberg K, Hedman J, Bowman TD, Didegah F, Laakso M. Do articles in open access journals have more frequent altmetric activity than articles in subscription-based journals? An investigation of the research output of Finnish universities. Scientometrics. 2020;122(1):645-59.
5. Hajjem C, Harnad S, Gingras Y. Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. arXiv preprint cs/0606079. 2006.
6. Kousha K, Abdoli M. The citation impact of Open Access agricultural research: A comparison between OA and non-OA publications. Online Information Review. 2010.
7. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJ, et al. An overview of microsoft academic service (mas) and applications. In: Proceedings of the 24th international conference on world wide web; 2015. p. 243-6.
8. Waltman L, Calero-Medina C, Kosten J, Noyons EC, Tijssen RJ, van Eck NJ, et al. The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. Journal of the American society for information science and technology. 2012;63(12):2419-32.