Variable Bibliographic Database Access Could Limit Reproducibility

TAD DALLAS, ALYSSA-LOIS GEHMAN, AND MAXWELL J. FARRELL

Bibliographic databases provide access to scientific literature through targeted queries. The most common uses of these services, aside from accessing scientific literature for personal use, are to find relevant citations for formal surveys of scientific literature, such as systematic reviews or meta-analysis, or to estimate the number of publications on a certain topic as a measure of sampling effort. Bibliographic search tools vary in the level of access to the scientific literature they allow. For instance, Google Scholar is a bibliographic search engine which allows users to find (but not necessarily access) scientific literature for no charge, whereas other services, such as Web of Science, are subscription based, allowing access to full texts of academic works at costs that can exceed $100,000 annually for large universities (Goodman 2005). One of the most commonly used bibliographic databases, Clarivate Analytics–produced Web of Science, offers tailored subscriptions to their citation indexing service. This flexibility allows subscriptions and resulting access to be tailored to the needs of researchers at the institution (Goodwin 2014). However, there are issues created by this differential access, which we discuss further below.

Variation in access to bibliographic databases

Although bibliographic databases provide access to a wide variety of scientific literature, the results obtained from user queries can vary from university to university—dependent on the level of access provided by the subscription—suggesting that some existing meta-analyses may not be reproducible and that the conclusions of these studies may depend on database access. This creates variation in institutional access to scientific literature through at least two distinct pathways.

First, institutional access to the Web of Science differs in terms of which databases may be accessed. These databases can be general (e.g., Science Citation Index Expanded) or geographically focused (e.g., Korea Citation Index). Web of Science has what they refer to as the "Core Collection," which is a set of ten general citation databases bundled together (Goodwin 2014). However, another 14 citation databases (Thomson Reuters 2011) are available to institutions, increasing the potential variability in institutional access.

Second, institutional access to the Web of Science differs in terms of *depth*, or the temporal coverage from the present year into the past. For example, an institution with a depth of 50 years would only have access to articles published after 1968. Although this time window could include all the relevant literature, depth restriction could also lead to the exclusion of foundational work. This can be especially troublesome for species-specific queries, such as species descriptions based on early zoological records that could fall outside the depth period.

Why is this a problem?

Variation among research institutions in bibliographic database access—as a function of either number of accessible citation databases or temporal depth—creates variation in search results in terms of the number and identity of citations. This becomes an issue when researchers attempt to reproduce studies with different bibliographic database access. Given the rather large annual price tag on bibliographic database services like the Web of Science or Scopus, smaller or more resource-limited institutions may not have access to the resources necessary to develop or reproduce some published analyses. Unlike the paywall around accessing articles in full text, the variation in citation identity and number created by differential database access is masked; researchers can be unaware of the citations they are missing.

Meta-analyses and systematic reviews are approaches that allow for the synthesis of multiple lines of evidence to gauge the level of support for a well-studied but poorly synthesized phenomenon (Moher 2015). When conducting formal literature reviews, authors use bibliographic databases to search for published literature using specific terms related to the phenomenon of interest. Therefore, variation in search terms, inclusion criteria, and access to bibliographic databases has the potential to strongly affect the conclusions drawn from these studies (Higgins 2003) independent of established meta-analytical protocols such as PRISMA (Moher 2009).

Apart from formal literature reviews, researchers commonly use the number of citations from bibliographic queries as surrogates for research or sampling effort. For example, citation counts from bibliographic databases are often used in cross-species comparative analyses as a measure of sampling effort per species (Lindenfors 2007,

Downloaded from https://academic.oup.com/bioscience/advance-article-abstract/doi/10.1093/biosci/biy074/5050469
by University of Durham user
on 11 July 2018

Olival 2017). Institutional variation in citation database access and depth can strongly influence these citation counts and therefore has the potential to bias studies using citation count as a measure of research effort.

What can researchers do to address the issue?

We recognize that the Web of Science is a valuable resource for scientific research, and it is not our suggestion that scientists abandon its use. However, scientists should be cognizant of bibliographic database limitations and how variation in database access may influence the reproducibility and overall findings of their research efforts. To ensure the reproducibility of scientific analyses relying on bibliographic databases, it is imperative that database subscription details be reported. For meta-analyses, citations for academic works included in the analysis should be provided. If researchers continue to use citation counts as a proxy for sampling completeness, care should be taken to make sure the results are insensitive to database access, to develop different measures of sampling completeness, or to use accessible data sources that allow reproducible citation count estimation. For institutions with limited bibliographic database access, Google Scholar is a free alternative to find citations (Harzing 2008), although there is no guarantee that each citation is from a peer-reviewed article.

Ideally, analytical code produced by researchers could be built upon as scientific literature accumulated. That is, the consistency of analytical results could change over time with accumulating scientific studies. An approach that allowed for the easy addition of citations—or a programmatic way to access changing citation counts—would provide a dynamic assessment of support for a given idea in the case of meta-analyses, and a clear way to examine how incorporating potential publication or sampling biases influences overall results in the case of quantifying sampling effort using citation counts.

Acknowledgments

The Macroecology of Infectious Disease Research Coordination Network (funded by National Science Foundation–National Institutes of Health–US Department of Agriculture no. DEB 131223) provided useful discussions and support for this work. MJF was supported by a Natural Sciences and Engineering Research Council Vanier Canada Graduate Scholarship.

Funding statement

Funding for this study was provided by NSF grant number DEB 131223.

References cited

Goodman D, Deis L. 2005. Web of Science (2004 version) and Scopus. Charleston Advisor 6: 5–21.

Goodwin C. 2014. Web of Science. Charleston Advisor 16: 55–61.

Harzing AWK, Van der Wal R. 2008. Google Scholar as a new source for citation analysis. Ethics in Science and Environmental Politics 8: 61–73.

Higgins J, et al. 2003. Measuring inconsistency in meta-analyses. British Medical Journal 327: 557–560.

Lindenfors P, Nunn CL, Jones KE, Cunningham AA, Sechrest W, Gittleman JL. 2007. Parasite species richness in carnivores: Effects of host body mass, latitude, geographical range and population density. Global Ecology and Biogeography 16: 496–509.

Moher D, et al. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLOS Med 6 (art. e1000097).

Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews 4 (art. 1). (8 June 2018; *https://doi.org/10.1186/2046-4053-4-1*)

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. 2017. Host and viral traits predict zoonotic spillover from mammals. Nature 546: 646–650. (8 June 2018; *www.nature.com/doifinder/10.1038/nature22975*)

Thomson Reuters T. 2011. Web of Science: The definitive resource for global research. Thomson Reuters. (8 June 2018; *http://wokinfo.com/media/pdf/WoSFS_08_7050.pdf*)

Tad Dallas is a postdoctoral researcher at the Centre for Ecological Change at the University of Helsinki, in Finland. Alyssa-Lois Gehman is a National Science Foundation postdoctoral research fellow at the University of British Columbia, in Vancouver, Canada. Max Farrell is graduate student at McGill University, in Montreal, Quebec, Canada.

doi:10.1093/biosci/biy074

Downloaded from https://academic.oup.com/bioscience/advance-article-abstract/doi/10.1093/biosci/biy074/5050469
by University of Durham user
on 11 July 2018