Review of: "Finding citations for PubMed: A large-scale comparison between five open access data sources"

Silvio Peroni[1]

1 University of Bologna

Potential competing interests: The author(s) declared that no potential competing interests exist.

Metadata

Title: Finding citations for PubMed: A large-scale comparison between five open access data sources

Author: Zhentao Liang, Jin Mao, Kun Lu, Gang Li

Submitted to: Scientometrics

Review

In this article, the authors described a study that compares five data sources that provide their data freely concerning the citations for the literature included in PubMed. It is indeed a fascinating study highlighting the coverage of existing freely available citation data sources for a specific macro-area. Moreover, it allows one to monitor the status of the free availability of these data and their effectiveness in bibliometrics studies compared with proprietary data sources, i.e. Scopus and Web of Science.

While I do not have significant comments for the study, I think that essential aspects of the work should be revised. Therefore, I will discuss these aspects as follows. Before commenting on these aspects, though, a full disclosure: I am one of the Directors of OpenCitations and, as such, I am responsible for one of the data sources used by the authors in their experiment, i.e. COCI.

Open Access vs Open Data

Open Access is a term that is associated with traditional publications (i.e. articles). When we talk about data, as the authors do in this paper, the correct term to use would be Open Data. Thus, it would be good to rephrase it as "... five open data sources" in the title and the paper. However, for being more specific with the particular domain in consideration, I would suggest using something along the lines of

- open bibliographic data sources
- open citation indexes
- open bibliographic databases
- etc.

The meaning of "open" in Open Access / Open Data

The term "open" applied to publications (i.e. Open Access) and data (i.e. Open Data) identifies a set of principles that go beyond the pure free availability to these objects - e.g. see https://oaspa.org/information-resources/frequently-asked-questions/#FAQ1. Indeed Open Data does not only mean that you can freely access such data, but that you can also "use, modify, and share [them] for any purpose (subject, at most, to requirements that preserve provenance and openness)" (see https://opendefinition.org/). This is today the intended definition of Open Data.

Thus, according to this definition, some of the "open access data sources" mentioned by the authors are not "open" at all, only freely accessible. I think that this distinction should be made explicitly in the paper. Indeed, I would value a lot a comparison between (A) proprietary services (Scopus, Web of Science), (B) open citation indexes (COCI, NIH-OCC, MAG), and (C) freely available data sources (Dimensions, S2ORC). Both (B) and (C) enable reproducibility, but only (B) are "open" as meant by the community – indeed, the non-commercial clause in the license of Dimensions and S2ORC is enough for not considering them "open" in terms of the definition above (see https://opendefinition.org/licenses/nonconformant/). Thus, I suggest that the authors revise the text to consider these three categories A, B, and C, listed above.

Citing the data

There are several references (see the section "Collecting and matching citations") to the data that have been used in the study. Some of them have been taken from complete dumps. Some of these dumps are research objects per se and are identified with persistent identifiers (e.g. DOI). As such, they should be added to the reference list and cited. In particular, both COCI and NIH-OCC releases are available on Figshare and can be cited using the related DOIs. Also, dumps of MAG are available on the Internet Archive (see https://archive.org/search.php?query=microsoft%20academic%20graph). However, it seems the data used by the authors are not those included there.

About MAG discontinuation

Microsoft has recently announced (https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/) the discontinuation of MAG. It would be good to know how this may affect the reproducibility of the study done by the authors – since they used the data available via an Azure installation, as far as I know. Would it be possible to publish somewhere (e.g. Zenodo) all the DOI-to-DOI citation links retrieved in MAG, considering that the license of MAG data should allow the authors of doing so? That would be great to enable the experiment's reproducibility in the long term, at least for what concerns MAG (and COCI and NIH-OCC, which have been already archived online).

More citations from MAG

The authors wrote that they extracted only DOI-to-DOI citations from MAG. However, MAG contains also non-DOI-identified articles. Some of them may, in principle, have a PMID and, thus, could be included in the PubMed dataset. Thus, why the authors did not consider the possibility of matching some metadata (title, first author, year of publication, etc.) of MAG articles with no DOIs against PubMed to consider also them in the analysis? Section 3 of a recent Visser et al.'s article published on QSS [1] also provides a strategy for doing it that could be adopted.

New release of COCI

It has not been advertised yet on the website (it will happen on Monday, 2 August 2021). However, the new release of COCI was published on Figshare a few days ago [2]. It contains more than 1.09 billion citations (including Elsevier ones). I know that it would take time, but I would love to see updated figures in the article that include this new release (even in comparison with the version of COCI considered currently in the article).

Final remarks

I do believe that this paper deserves to be published in Scientometrics, considering the importance of the topic addressed. However, I believe that all the aspects above should be addressed carefully before accepting it.

References

1. ^*Martijn Visser, Nees Jan van Eck, Ludo Waltman. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. doi:10.1162/qss_a_00112.*
2. ^*OpenCitations. (2021). COCI CSV dataset of all the citation data - July 2021 release. Figshare.*