Review of: "Investigating Invalid DOIs in COCI"

Arcangelo Massari[1]

1 University of Bologna

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

The premises. About the study's rationale and impact

The protocol entitled *Investigating Invalid DOIs in COCI* (http://doi.org/10.17504/protocols.io.bt5xnq7n) describes a procedure aimed at identifying, given a list of valid citing DOIs and invalid cited DOIs, which publishers have published incorrect metadata, and which publishers the invalid citations refer to. Also, in the event of previously invalid DOIs becoming valid, the procedure illustrates how to get the number of these now valid citations.

The research question appears clear and justified, but there seems to be a discrepancy with the answers obtained: the findings, in fact, are not limited to reporting the names of the publishers, but compose some sort of ranking of the most deserving publishers, based on how many correct metadata were sent, thus answering a question that was not originally asked. Therefore, it is advisable either to change the question or to change the answer, in order to make it clear to the reader what to expect from the research work.

Focusing back on the premises, in the abstract it is said that the COCI REST API will be used to obtain information about the publishers, while the actual protocol uses exclusively the Crossref API for this purpose. Therefore, I suggest modifying this small imprecision, certainly justified by the evolution of the research work over time.

Furthermore, the reasearch does not offer any bibliographic references about previous works on this problem, and there is no mention of what further contribution is given by the protocol. It would be interesting to understand how the process that led to this specific methodology was elaborated to fill the gaps of the pre-existing research.

In any case, I believe that the research in question is of enormous relevance in the field of Scientometrics, since a huge amount of DOI names is not verified and reported with various kinds of errors - as evidenced, for example, in *Errors in DOI indexing by bibliometric databases* (Franceschini et al., 2015). Furthermore, the work is of particular interest to the promoter of the research itself, that is COCI, which could thus be enriched with all the cited DOI names previously excluded because they were not valid.

The methodology. About the protocol's technical soundness

Delving into the protocol's technical aspects, the data structure to be obtained in output was clearly and completely described, that is a JSON file with three keys: two for the "responsible_publishers" and "receiving_publishers", both containing a dictionary, and another for "total_number_of_corrected_dois", containing a number. Although there is not yet

a result, it is already possible to imagine it by reading the paragraph "Creating the output JSON file" of the protocol. However, the workflow turns out to be just partially technically valid, as it seems to lead to significant results regarding the validation of DOI names and the recovery of publishers in the case of a valid prefix, while it appears incomplete in the case of an invalid prefix. The procedure described consists in searching Crossref for the references of the citing DOI and using the relative metadata to conduct a bibliographic query on Crossref itself. However, it is not clear how this can lead to the correction of the original wrong DOI, for the following reasons:

- The references reported by Crossref for each DOI are many: how is it possible to understand which one corresponds to the wrong DOI?
- Assuming that the correct reference has been identified, by querying Crossref with the corresponding metadata, all relevant works are returned: how is it possible to understand which is the right one?
- Assuming that the correct query output has been identified, who ensures that the DOI reported there is the correct one? Besides, the wrong starting DOI name came from Crossref too.
- Moreover, the unstructured field is often not sufficient to return meaningful results; other metadata, such as the author, is necessary. It is also essential to do a heuristic check of the result a posteriori. For further information on how to conduct a heuristic analysis based on the match between two metadata dictionaries, see chapter 3.2 and the Appendix of *Large-scale comparison of bibliographic data. sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic* (Visser, Martijn et al., 2005), available at the following address: https://arxiv.org/abs/2005.10732.

In summary, correcting a DOI is a complex problem. It is so complex that a good solution could be to delegate the addressing of this question to a separate protocol, such as the one proposed in *Investigating DOIs' classes of errors* (doi: 10.17504/protocols.io.bt65nrg6).

The reproducibility. About the input and output

With the exception of the point described above, the methodology has been described clearly enough to be reproduced and, although the protocol is in development and there is no code or output yet, it is already possible to understand the rationale of the underlying algorithm. Furthermore, the input data has already been made public via the link to the corresponding repository on Zenodo (Peroni, 2021), to the benefit of reproducibility. As for the output, however, it has not been indicated how the results obtained will be supported, that is through which statistical system or visualization means.

Conclusions

To conclude, the protocol has been well described and, although under development, it already offers a clear overview of the method to be adopted. However, it is necessary to correct some small errors of inconsistency between the premises and the conclusions. Furthermore, adding a chapter on literature review and another on results' presentation is vital, as is delegating to an external protocol the solution of a problem that goes beyond the research questions specified in the premises.

References

1. Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. Scientometrics, 102(3), 2181–2186. https://doi.org/10.1007/s11192-014-1503-4.
2. Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo. https://doi.org/10.5281/ZENODO.4625300.
3. Visser, M., van Eck, N.J., Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. arXiv. https://arxiv.org/abs/2005.10732v2.