Review of: "Retrieving SSH Journals Citation Information from three datasets (COCI, META and ERIH-PLUS) - Workflow v1"

Maddalena Ghiotto[1]

1 University of Bologna

Potential competing interests: No potential competing interests to declare.

1. Overview

The purpose of the research is clearly stated, and the goals of the research are solid, as they contribute to fill in an important lack of knowledge on the mapping and evaluation of data present in existing resources. For this reason, the research is overall valid and represents a meaningful contribution to field-specific science.

Nonetheless, the methodology at its current state, despite having a clear, reasoned skeleton, appears to be very general and too abstract, omitting significant information on data manipulation that compromise the reproducibility of the workflow and makes it difficult to evaluate its actual feasibility.

2. Design and Technical quality

In the abstract it is correctly described that the workflow requires python programming language. It is likely that there are more requirements to fulfil to be able to reproduce the methodology. For example, it could be useful to consider whether there are requirements for the correct reading of data, considering formats and volume.

The lack of detail about the structure of input and outcome data of every step of the workflow and about what functions need to be run in each step (or how are they designed) makes it difficult to understand how the different steps of the methodology are supposed to be performed. This undermines the verifiability of validity of the results and represents a major issue to the quality of the workflow.

In particular, the source datasets are declared and cited, which is positive since it enables access to sources verification. Their structure and how different data will be connected is declared and the visual map is very intuitive, but there are no data examples beside column names.
This could be problematic for non-experts in the field. For example, the authors intend to connect datasets by means of unique identifiers. But this information is implicit and may be misunderstood by those who are not accustomed to these particular kind of publication data. At the current definition of methodological detail, if someone ignores the fact that a single venue might have multiple identifiers, it could be likely for them to incur in errors during data manipulation.

As a solution, to make the methodology better understandable and reproducible by future scientists, it could be helpful to:

- Clearly state what tools will be used in each part of the research and how can they be accessed.
- Add real examples and more thorough description of how the data are supposed to look at the input and at the output of every step, as well as specifying the algorithmic processes that perform data manipulation. Either by providing runnable code examples, specifying function implemented and used or designing flow diagrams.

Moreover, to ensure quality and enhance reliability of research outcome, the methodology would benefit if the authors could specify how they foresee to test the correctness of the process and the results.

3. Data treatment

As far as publication of outcomes and fair treatment of data are concerned, the authors did not state how they envision to present their results and ensure FAIR data treatment.
If such guidelines are adopted, it would be good practice to include in the workflow how the authors intended to ensure findability, availability, interoperability, and reusability of their data. To instruct future scientists on how to handle data in the same way.

On the contrary, if such procedures are not implemented, explaining the reasons underlying this choice could help enhance transparency towards the public and preserve trustworthiness.

4. Conclusions and further questions

In conclusion, if improved in its specificities, the workflow has a great potential of reusability, considering that protocols.io workflows are publicly accessible and runnable. Moreover, sub-steps of the workflow may be useful for other research conducted on OpenCitations data.

Here are some further questions to help the authors improve their methodology:

1. How are source datasets analysed? What are the main aspects one should focus on when analysing these datasets?
2. How are the data cleaned? What is the relevant information you mentioned?
3. How does your merged dataset look like?
4. What python functions will you need for the completion of each task?
5. How is the design of your analysis operations? How will they be able to retrieve the specific data you need?
6. What could be an efficient way to present the results?