Review of: "IRIS Software Protocol v1"

Matteo Guenci[1]

1 University of Bologna

Potential competing interests: No potential competing interests to declare.

In this peer review, thoughtful feedback is offered on the workflow aimed at exploring the role of IRIS within the academic literature landscape. Constructive thoughts and suggestions are provided on the clarity, coherence, and potential enhancements of the methodology, particularly focusing on the integration of diverse datasets, the interpretation of results, and the alignment with the overarching research question. The goal is to contribute to the refinement and development of this promising project, highlighting areas for improvement.

In the initial sections of this protocol, a reference is made to: "ODS_L1_IR_ITEM_DESCRIPTION.csv": the string of the authors and other related metadata of publications.

While reviewing this, I found myself uncertain about the meaning of "the string of the authors." It's possible to infer that it may refer to the textual representation of authors' names, including given names, family names, or similar details. However, a clearer explanation would greatly benefit the reader's understanding.

Moving forward, concerning the section detailing the conversion of CSV files into DataFrames —iris_id_df, iris_master_df, and iris_relation_df— there arises a question: Are only these three DataFrames involved, or are additional files also included? The uncertainty arises from the fact that it is only stated: “The CSV files are converted into DataFrames.” Clarifying this aspect would provide a clearer understanding for the reader.

Later on, in the subsequent section that is about the same three DataFrames, it is stated that a merging operation is applied to the mentioned DataFrames. This makes me think that the operation is applied on the three mentioned DataFrames only, but being the fact that it's not clearly stated, it makes me wonder if this process exclusively involves them or encompasses others as well.

The section discussing the use of the OpenCitations Meta Sparql endpoint to obtain CSV files through queries lacks specificity regarding the nature of the queries employed, rendering it difficult to evaluate the resulting files. This and the following phases of the workflow appear to be in an early developmental stage, making it challenging to assess their progress and viability.

The primary concern regards the clarity of how the importance and relevance of IRIS in academic literature will be evaluated, as well as the lack of an explicit connection to the research question throughout the workflow.

In other words, what's not clear from reading the workflow is how the working group responsible for producing this software will assess the importance and relevance of IRIS in the context of academic literature. Although pragmatically,

the workflow seems well conceived, what's missing is a general connection to the research question in the explanation of the results that are obtained through the operations applied, which are well described.

For now, it's clear that the working group has a solid understanding of what needs to be done and how to do it, but the reader remains disconnected as it's not precisely explained in the individual steps of the process what the results mean, how they are weighted, assessed, and used.

Since the project is in an embryonic phase, many things will surely become clearer as the work progresses, but precisely because this workflow expresses a complex and interesting work, it would be useful to be able to follow its developments step by step, staying at the same level as the team handling it, or at least being able to follow their reasoning in order to offer more fruitful advice if needed.

Regarding reproducibility, there don't seem to be any particular issues since the requirements are well expressed at the beginning of the work and seem to be within reach for everyone. The websites from which the working group gathers its data are all well exposed and accessible. The individual responses to the subsections of the research question are well imagined, but as mentioned before, due to the embryonic nature of the work, they are clearly not very informative. The section (yet to be developed) regarding data visualization will be very useful to the reader.

In conclusion, this seems to be a promising workflow for a very interesting project that can potentially reveal important insights and become even more promising with well-spread clarity, mostly regarding the exposition and the subsequent interpretation of the results obtained from the manipulation of the data.