This document is a rebuttal letter to the reviews received on the Data Management Plan:
Cioffi, Alessia, Coppini, Sara, Moretti, Arianna, & Shahidzadeh Asadi, Nooshin. (2021). Investigating Missing Citations in COCI. Zenodo.

The reviews considered in this document are:

1) Cristian Santini. (2021). Review of: "Investigating Missing Citations in COCI (1.0)". Qeios. doi:10.32388/DIA06O.

2) Ricarda Boente. (2021). Review of: "DMP Investigating Missing Citations in COCI (zenodo.4671487)". Qeios. doi:10.32388/T0UF3H.

Answers and comments provided to the suggestions and notes of the reviews are written by the very same authors of the Data Management Plan: Nooshin Shahidzadeh, Alessia Cioffi, Arianna Moretti and Sara Coppini.

Cristian Santini. (2021). Review of: "Investigating Missing Citations in COCI (1.0)". Qeios. doi:10.32388/DIA06O.

In Part 2 [...] no proper identification and description of reusable data is given in this section. [...] We would suggest being more specific about the provenance of this [source] data and its accessibility, and also to describe it in the proper section. [...] there's no specific description of the license through which collected data was accessed and, therefore, few concerns may arise with respect to the sensitiveness of this data. [...] we suggest being more specific about the assessment of existing data, by inserting information to identify the data re-used and the licenses that regulate its accessibility in Part 2.

We see that many concerns arose about the reuse of existing data and their license, which led us to clarify this part accurately. In the first place, for the dataset called "Missing Citations in COCI", (now renamed as "Missing Citations in COCI: Publishers Analytics Result") we specified the source of reused data, where the dataset resides and how we are using these existing data. In the second place, for the dataset called "Code for Missing Citations Analysis in COCI" we corrected the answers to questions in section 2, specifying that we did not reuse any preexisting code for this code dataset, so we ought not to provide information about the reuse of any data.

Then, we specified the license that regulates its accessibility both in section 2, where information about reused data is provided, and in section 3, where we specify the licenses for both datasets, justifying them also in the light of the license of the existing data that were reused to create the output data of this project.

In section 3.4 the team states that there is no documented procedure for quality assurance of data. This is a potential drawback in a project that concerns the increase of data quality in the Open Science arena.

In the resources produced within this project, for "quality assurance" - as stated by Economic and Social Research Council, UKRI. (2019). Data management plan: Guidance for peer reviewers[1] - we either refer to:

a. Methods for data validation or standards applied during data collection and data entry;
b. Codes of research practice adhered to;
c. Transcription templates used.

In this sense, we did not document procedures for quality assurance in data collection since, theoretically, the existing data that we reused were produced in an Open Science and FAIR environment.

Furthermore, all the data we use and elaborate in our workflow are provided by referenced and trustable resources:

[1] https://esrc.ukri.org/files/about-us/policies-and-standards/data-management-plan-guidance-for-per-reviewers/

a. OpenCitations' COCI for the input dataset (i.e.: invalid_dois.csv);
b. Crossref API services for doi prefixes for publishers' identification;
c. doi.org API service to check DOIs validity.

Anyway, as a general assumption, we do not think that any more accurate quality assurance procedures are to be declared, both because of the nature of our research questions and since - as stated above - we structured the whole workflow in an Open Access and FAIR environment.

[...] the team is not consistent in stating if they will support data reuse or not for both datasets (see differences in section 3.4 of dataset descriptions in DMP) and it is not clear if they will adopt two different strategies for supporting reuse for the two datasets.

[...] In the DMP it is not clearly detailed how the team will ensure data reuse after their research project finishes. More specifically, it is not easy to establish if they will keep the data accessible from the original repositories (e.g. Zenodo and Github), if they will collaborate with external institutions, such as archives or universities, or if they will apply other strategies of data sharing. The reasons for this unclearness are still the inconsistencies and the lack of details in the way the authors describe the strategies for the two datasets (see Part 4 [Allocation of resources] in both dataset descriptions).

First of all, thank you for this constructive criticism because we care a lot about FAIR principles and reusability is one of them. For this reason, we have decided to take this point very seriously. The changes made to the Data Management Plan are as follows:

- We specified that both datasets will be kept accessible (as they already are) from the original repositories in which they have been stored since the beginning of the project;
- For what concerns other data sharing strategies, we plan to publish our output data, at least as a sample, on a website where we will summarise the structure of our workflow, present our working environment and context, state the principles behind our project development, and show some visualizations on our output material, in order to make our results more intelligible for people who may get interested in our outcomes, even without being experts in the field.
- We addressed precisely and adequately questions in Section 4, "Allocation of resources", since they are directly related to reusability, interoperability and accessibility; as well as roles in data management and ensuring data reuse after the end of the project. In particular, we specified the team members' names in response to the question on who will handle the data for both datasets, since we will all be doing them equally. Then, we also specified suitable strategies for ensuring data reuse for both datasets, such as relying on persistent infrastructures as Github to store our datasets.

[...] few details are provided with respect to the strategies adopted to assure persistency of data. The authors declare that the repositories through which data will be made available will be kept secure with backups and recovery processes; however, no information with respect to the modality and frequency of these procedures is given.
Both Zenodo and Github provide different ways and processes to backup data, so we have specified in the new version of the Data Management Plan that the platforms we have chosen to store the datasets are adequate also from this point of view.

[...] a recurrent drawback in the DMP is the inconsistency of information provided for the two datasets. While the authors state that they will use GitHub to make accessible the CSV file, they state that the source code will be made available only with Zenodo.
We decided, at the end of the review of the Data Management Plan, to compare the information provided for each dataset, checking that there were no inconsistencies and whether in some cases it was better to opt for methodological and procedural uniformity in the data management of the datasets.

[...] However, the main issues of the Data Management Plan consist in the lack of consistency and detail of the information that they provide about the strategies to make data persistent, sufficiently documented and secure. For example, in Part 3 of the source code dataset description, they provide no information with respect to the use of version control systems or possible metadata and documentation describing the software. Or in Part 4 of the description of the CSV data related to the missing citations, where no detailed information is provided of whether they will ensure data allocation on sustainable and reliable platforms.
GitHub and Zenodo - i.e., the platforms we have chosen to store and share our input and output datasets - are persistent, reliable and secure. They allow you to accurately document resources and their developments, such as new versions and updates. Evidently, not all of this information I can be clear to who does not know these two tools. For this reason, where possible, we have specified these characteristics in order to allow any reader to fully understand the quality and reliability of these infrastructures.

[...] we suggest to use a version control system for code sharing.
We admit that it was not clarified that for both datasets the platforms on which we make them available already have version control systems themselves, if by this term we refer generically to a system that records changes to a file or set of files over time so that you can recall specific versions later. Indeed, we have chosen GitHub as the main repository where to store both of our datasets, which provides a version control system. Zenodo also provides a version control system for the published resource, and allows you to view versions older than the current one of the same file.

Nonetheless, it would have been appropriate, especially for their software, to insert in their plan a specific documentation to give to the users technical information about the tool used in the research.

Additional documentation regarding technical information about the tool used in the research goes beyond the template used to write this Data Management Plan. However, the point of the observation is clear, and in fact we tried to address it to the extent which was allowed by the framework we worked in. To this end, the lack of technical details on the infrastructures used during the project was compensated by the attempt to specify, in every point of the DMP where it was possible and appropriate, the information concerning these technical details.

Ricarda Boente. (2021). Review of: "DMP Investigating Missing Citations in COCI (zenodo.4671487)". Qeios. doi:10.32388/T0UF3H.

One formal issue can be noticed in the mentioning of the authors. In the pdf printed version of the DMP, one author remains without a last name, being called "Alessia null".

The reviewers correctly mentioned the absence of the surname of a team member in the DMP data. This was a third-party technical problem that we fixed by stemming from ORCID and finding its root.

Concerning the assessment of existing data, the source of the CSV file that will be used could be made explicit in point 2 (Reusable data). By providing a link to the openly accessible source, the reader of the data management plan could be informed easily and effectively about what type of data is used in the research.

We all agreed on the validity of this point of the reviewer, and for this reason we welcomed the suggestion to link the data we use as primary source of our research. In fact, since we declare that the data is openly accessible, and indeed it truly is, it is logical to provide a link leading to it. We included this in our new version of the DMP.

By dividing the data management plan in two datasets, all data generated by the research seems to be covered. The dataset "Missing Citations in COCI" could be renamed, though, in order to make the purpose of this dataset clearer to the reader. Also, in 1.2 there is a typo "exiting" instead of "existing" source. In the second dataset, the introduction specifies well the expected outcome format. However, it would be recommendable to make this point also clear in 1.3, where there is only a very generic sentence at the moment. Also in this section, there are a lot of data formats mentioned (text file, etc.) without being explained on how or why they will be generated.

A typo in the DMP has been correctly noticed by the reviewing team and has been fixed since. We were reminded by the reviewers that our explanation for the output files formats in the second database was not satisfactory, and for this reason we rewrote the sentence used in point 1.3 of the second dataset to better explain which formats we are outputting and why. It has also been mentioned that the name of our second dataset was not adequate and clear enough, which is the reason why we changed it to "Missing Citations in COCI: Publishers Analytics Result".

Like this, it also seems realistic that the data management plan will be followed. In point 4.2, the responsibilities could maybe be clearer divided, as it might be easier to have one data manager than sharing the responsibility. In a group of four people though, sharing can maybe be more realistic, so a specification of the responsibilities is more an optional improvement in this case in my opinion.

The reviewers make an optional suggestion to elect only one team member as the data manager, albeit taking in account that it might be more compatible with the group dynamics to keep this responsibility a shared one. After discussing this point at length, we decided to move on with our previous decision at this point as this DMP belongs to a very tightly meshed group project in which the main goal is for all members to learn every possible point. In fact, as it is the first Open Science project for all the members of the group, we decided to share responsibilities and too keep the roles exchangeable, so that all of us could participate in equal manner to all the different phases and layers of the workflow.

References

1. Santini, Cristian. (2021). Review of: "Investigating Missing Citations in COCI (1.0)". Qeios. doi:10.32388/DIA06O.
2. Boente, Ricarda. (2021). Review of: "DMP Investigating Missing Citations in COCI (zenodo.4671487)". Qeios. doi:10.32388/T0UF3H.
3. Cioffi, Alessia, Coppini, Sara, Moretti, Arianna, & Shahidzadeh Asadi, Nooshin. (2021). Investigating Missing Citations in COCI. Zenodo.
4. Economic and Social Research Council, UKRI. (2019). Data management plan: Guidance for peer reviewers.