Review of: "Investigating Missing Citations in COCI (1.0)"

Cristian Santini

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

Introduction

In this review we analyse the Data Management Plan (DMP)[1] realized by Alessia Cioffi, Sara Coppini, Nooshin Shahidzadeh Asadi and Arianna Moretti for their research entitled “Investigating Missing Citations in COCI”. The DMP was realized in order to formally recognize and set the data management as part of a research workflow which follows the *FAIR* principles (Findability, Accessibility, Interoperability and Reusability). Therefore, in our review, we will assess if their Data Management Plan suits the aims of maximizing the availability and reusability of the data generated and collected by their research project.

Context

Before looking in depth at the DMP, we would like to describe a little bit the context and the aims of their research, in order to simplify the lecture to unaware readers. The research “Investigating Missing Citations in COCI” by Cioffi, Coppini, Shahidzadeh and Moretti is a research project currently running in the context of the “Open Science” course held by Prof. Silvio Peroni at the University of Bologna in 2020/2021. The aim of this research is to infer from invalid DOIs present in citations data collected from Crossref[2] the specific publishers which provide incorrect metadata and the publishers to which invalid DOIs point to. As outcome of their research, the authors expect to realize two kinds of data: source code for data analysis and a dataset presenting the results of their inquiry.

Assessment of existing data

In our review, we first focus on the information that the authors provide for identifying possible existing data related to their research question or secondary sources which can be reused in their work. In *Part 2* of the *Dataset Description* of their DMP, both for the source code and for the research outcomes, the authors state that they will re-use existing data; however, no proper identification and description of reusable data is given in this section. The only information provided for secondary data is given in sections *1.2* and *1.3* of the dataset description for their source code files: indeed, the research team states that they will collect data from an already existing dataset to combine and analyse data and that this dataset is a CSV file. We would suggest being more specific about the provenance of this data and its accessibility, and also to describe it in the proper section.

Information on new data

One of the positive aspects of the DMP provided by Cioffi, Coppini, Shahidzadeh and Moretti regards the way they describe the data which will be generated from their research. According to what is stated in the DMP, the team will propose two types of data: a source code written in Python and a CSV file containing, for each publisher retrieved from the Crossref API, the number of invalid DOIs provided and the number of invalid citations pointing to it from other publishers. Both the datasets are expected to be delivered under open access licenses; thus, fully reusable. They state that the data generated will be useful for people involved in the OpenCitation's COCI project for research and policy-making. In addition, the authors plan to make data accessible in a open platform developed following *FAIR* principles, Zenodo, and, therefore, they expect it to release metadata describing the datasets using standardized vocabularies.

Quality assurance

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 addition, 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.

Backup and security of data

Furtherly, 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. Moreover, 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 suggest to use a version control system for code sharing.

Expected difficulties in data sharing

The authors deal with the potential difficulties of data sharing by using open formats (like Python and CSV files) which allow for interoperability and can be processed by supported open-source tools. Moreover, the datasets generated will be made accessible through licenses that allow maximum re-uses, such as the ISC license for the source code files and *Open Data Commons Public Domain Dedication and Licence 1.0* for the CSV file. Finally, data will be provided through free and open platforms like Zenodo and Github. 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.

Copyright/intellectual property right

As the DMP states, the research carried by Cioffi, Coppini, Shahidzadeh and Moretti will not raise specific copyright issues, since they will not process from it any sensitive data. However, 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.

Responsibilities

The Data Management Plan is a source also to establish data management responsibilities. As the researchers declare in the DMP, the four team members are equally involved in the process of data management. We found this choice suitable for a small computational project in which there is no division of roles in the research workflow.

Preparation for data sharing

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 esabilish 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* in both dataset descriptions).

Conclusion

The Data Management Plan provided by Cioffi, Coppini, Shahidzadeh and Moretti for their research “Investigating missing citations in COCI” seems a valid first draft of a just-started research project which aims to follow the *FAIR* principles in the management of data. The authors express their concerns in maximizing the reuse of data generated by using standardized and interoperable formats and by publishing their datasets with public licenses and on open platforms such as Github and Zenodo. 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. Finally, 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 also recommend using *Part 7* to specify in detail other mentioned data management strategies, like those concerning backup and data quality or those concerning the support for data re-use.

References

1. ^*Cioffi, A., Coppini, S., Shahidzadeh Asadi, N., & Moretti, A. (2021). Investigating Missing Citations in COCI (1.0). Zenodo.*
2. ^*Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo.*