Review of: "Investigating DOIs' classes of errors"

Alessia Cioffi

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

Definitions

Peer Review
Defined by Jeffrey Beck et al.

Open Peer Review (OPR)
Defined by Tony Ross-Hellauer

Peer Review, Open Peer Review (OPR)

About the reviewer

Alessia Cioffi, Digital Humanities and Digital Knowledge, student at the University of Bologna, Bologna, Italy. She is currently following a course on Open Science held by Professor Peroni.

Introduction

In this work it is analyzed the second version of the protocol ‘Investigating DOIs' classes of errors’ created by Boente R., Massari A., Santini C., Tural D. The project of which this protocol is part is run into the Open Science course held by Prof. Silvio Peroni at the University of Bologna in 2020/2021. The main objective of the team is to investigate the DOIs present in citational data obtained from Crossref and get useful information in order to create an automated way to correct them. Therefore the ‘Investigating DOIs' classes of errors’ project has two main stated steps:

1. individuate and classify the classes of errors to which the DOIs are subject to;
2. try to fix as many errors as possible, according to the classes of errors.

The protocol, which has the aim of fixing in a standard way the process that the authors adopted to carry out the above mentioned tasks, has been realized with the platform Protocols.io and is available at this link.
N.B. This review is made in a moment in which the protocol is not yet finished but it is in an intermediate step while the project is still in development. Therefore it wants to be a way through which to provide its authors a feedback on the positive aspects of the work and some suggestions on how the work can be further improved in this particular step of the work.

Overall impression

The reviewed protocol is well-organized from a structural point of view. All the steps planned by the authors are clearly stated and structured in a logical sequence which also an external and non expert reader would understand. Two things which are particularly appreciable and useful are:

1. the fact that the authors added to the protocol the data they are using (i.e. the input CSV file), showing in this way their intention to provide the protocol with the files containing the necessary data required to carry out, and therefore replicate, the task;
2. the fact that they added the bibliography they used in order to produce their project. This is another relevant information for openness, which allows the reader to find out the same information they obtained.

Moreover, it is clear that the manuscript provides a valid rationale for the ongoing study, with clearly identified and justified research questions.

However the protocol is also subject to some weaknesses.

First of all there are some steps and aspects of the project which are not clearly expressed, and therefore would not be of immediate comprehension for an external person. These two aspects are:

1. An incongruence between different parts of the protocol: in the introduction all the classes of errors are briefly listes. Among these classes there is also the set of previously invalid DOIs now become valid. But in the main steps of the protocol (1-5) the fact that also this one is a class of errors for the DOIs is not explicitly repeated and therefore not so clear. I would suggest to repeat also in the main steps of the protocol, maybe in a more extensive way than in the introduction, which are the four classes of errors identified, the moment in which they are identified and the way in which classified.
2. There are some procedural steps which are not clearly stated from a processing point of view: how do you concretely keep track, in the first passage ‘Checking the DOI names' invalidity’ of the DOI status valid/invalid? And in the fourth step (data cleaning process) what happens to the DOIs that in the end may not have been modified by the cleaning procedure? Are they left out in this case or in some way they are counted in some statistics?

A second issue is that even if, rightly, the authors have posted the link to the CSV file they will use for their work, it is missing the provenance of these data. Having more information about the source can be useful to keep track of the file during time.

The third, and last, issue is the fact that the programming language adopted and its relative version are not clearly stated in the protocol. This aspect is crucial since it is an essential requirement for reproducibility.

Analysis

Impact

This protocol breaks new ground since it defines in a critical way all the steps of a project which aims to identify and classify for the first time a series of DOIs errors. Solving this problem means providing to Open Citations a wider ground for its citational data. The current project can, indeed, be a good solution for this issue. Another useful aspect, carried out by the protocol, is that of providing the final results of the data which the researchers were able to achieve. This is essential in order to provide effective reviews on the whole work.

Another relevant aspect is that the contents of this protocol are of interest for the entire scientific community, also for some

researchers which may be non-experts in the computer science field. Therefore, there are some computational aspects of the protocol which may be not immediate for this part of the researchers and that, if improved, would give open access to all. For instance, the specification of the programming language used and a reference to resources which can give information about that (e.g. the 'Getting started section' of w3schools or another open platform which allows to have good basis), could be a good starting point to make the resource more accessible and available also for non-experts.

Reuse

From the perspective of the reuse of the resource, the team has done a good job, even if there are some aspects which can be further improved. Since the resource has been structured in a generic way, it has the advantage of being reusable also in other circumstances. Indeed apart from being sufficiently resolutive for the task at hand, or at least supposed to be so since the work is not yet finished, it seems that it can be applied also to other contexts. For instance, the extensive use of the regular expressions as a way for cleaning the DOIs is a system which could be adapted also to other contexts in which there are wrong identifiers that are wanted to be cleaned automatically. Also, there is a potential for extensibility to meet future requirements, but we can't be sure at the moment since the code is not yet available. Another good point is that the protocol adopts open standards. For instance it provides as input file a CSV, a standardized file format.

An issue is that, in the current circumstances, it is not easy to reuse or replicate the resource. Apart from the already mentioned flaws, there other two objections:

1. the resource does not include a clear explanation of how others are expected to use the data and and the software;
2. it has not yet been presented a sustainability plan;
3. it has not been specified where the resources will be made available once the project will be finished and in which format.

Finally, the resource description does not provide explicit specifications of what the resource can and what it cannot do. For instance, it is not explicitly said but it comes out that not all the IDs are solved and there are no further specifications about the role of these non correct DOIs in the project.

Metadata

For what regards the metadata the work is complete. All the relevant metadata required to define and find the resource on the web are present. Indeed since the protocol has been created through Protocol.io it has the following features, either manually selected or automatically attributed by the platform:

1. it creates of a persistent identifier of the resource (URI);
2. it provides a complete citation to the resource that can be easily used by others when citing this protocol;
3. It provides a CC-BY license.

The protocol is also publicly available and findable, which is a relevant aspect with respect to the fact that it is part of an Open Science project.

To conclude, all the authors are provided with an ORCID which allows them and their protocol to be easily searchable on the web and adds another element of openness.

Testing the protocol

At this very moment it is not possible to test the whole process of the protocol. Indeed even if the steps are presented clearly and they are sound with respect to the research questions, there are some missing elements which do not allow a complete testing. The first two steps can be partially carried out because there have been added the relevant resources and literature which allow to replicate almost all the passages. The last two steps, instead, lack the relevant specification which would allow others to test it (e.g. the Regular expressions are only mentioned, they have not been specified yet, and there is no specification about the format in which the final output data will be saved).

Conclusions

The structure of the protocol is clear and well-articulated. The manuscript is presented in an intelligible fashion and written in standard English. The overall work done is good and appreciable. The protocol addresses in a structured way all the steps which allow the team to retireve the invalid DOis classes errors and clean them. There are some flaws that can be easily solved, since they are mainly linked to the lack of useful information rather than being conceptual errors.

A suggestion for further improvements about the presentation part would be that of adding the part of the code necessary for the researcher to replicate the processes and some images for the parts that are not code. Another useful aspect could be that of improving the information in the abstract, or creating a 'before starting' section in which the authors should provide a wider explanation about the context in which the project have been addressed (e.g. the operating system, the programming language, where to get information about it and how to get started to it).

Once the protocol will be finished, it shall be revised, with a special attention to the changes that will be made and to the newly added elements.

References

Ivan Heibi, Silvio Peroni 2020. A methodology for gathering and annotating the raw-data/characteristics of the documents citing a retracted article. protocols.io https://dx.doi.org/10.17504/protocols.io.bdc4i2yw

Ricarda Boente, Deniz Tural, Cristian Santini, Arcangelo Massari 2021. Investigating DOIs' classes of errors. protocols.io https://dx.doi.org/10.17504/protocols.io.bt65nrg6
Version created by Arcangelo Massari

Peroni, S. (2021). Citations to invalid DOI-identified entities obtained from processing DOI-to-DOI citations to add in COCI (1.0). Zenodo.