Review of: "Investigating DOIs' Classes of Errors"

Arianna Moretti[1]

1 University of Bologna

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

Peer Review of Investigating DOIs' Classes of Errors Protocol

Metadata and Reviewer Presentation.

Reviewer Report 21 April 2021 – dx.doi.org/10.17504/protocols.io.bt65nrg6.

Arianna Moretti, Digital Humanities and Digital Knowledge student at University of Bologna, Classical Philology and Italian Studies Department, Bologna, Italy.
Arianna is currently attending the Open Science course under Professor Silvio Peroni, with whom she is also working at her thesis in OpenCitations[1] project.
She declares she wants her identity to be public for this review.

Introduction to the Review

This is a review work on the second version of *Investigating DOIs' classes of errors* [2]: a protocol by Ricarda Boente, Arcangelo Massari, Cristian Santini and Deniz Tural. The reviewed protocol is aimed at illustrating step by step the structure and the main functions of a software which is going to be developed to identify classes of errors in invalid referenced DOIs in citational data from Crossref [3], which is the source of OpenCitations COCI Index [4]. Accordingly, the software is also supposed to correct the classified DOI names and to store their modified version, together with their error class, in an output CSV file.

The protocol was developed as a part of Open Science course project (2020/2021) under Professor Silvio Peroni. The second version of *Investigating DOIs' classes of errors* was uploaded on protocols.io on the 13th of April by Arcangelo Massari.

Overall Impression

The protocol comes with a brief but still clear and concise abstract, declaring the purpose and listing the four classes of errors that are going to be used for the previously invalid DOIs classification. Since among these four classes also “previously invalid DOIs become valid” and “other type errors” are included, no analysed invalid DOI is supposed to be excluded from the classification task.

Since the first step, the procedure appears to be unambiguously stated, so to be understandable not only for the target users, but also for a broader public. The source Zenodo [5] CSV file is provided in the first step of the protocol, and an explanation about its structure is succinctly provided by the authors.

The second step is aimed at managing the first class of errors, i.e.: identifiers that were invalid at the time of the creation of the original CSV file but that became valid in the meantime. Also in this step, the natural-language explanation appears clear enough to imagine the structure of the software to be developed. Further, the DOI Proxy URL is provided [6].

The third step is devoted to the presentation of the *Error analysis*. In particular, the authors state that they are going to limit their activity to database mapping errors (related to a data-entry error), demonstrating the soundness of this choice by providing mentions of the referenced works used to define their classification (Buchanan, 2006 and Xu et al., 2019).

The data cleaning process of the factually invalid DOIs is presented in the fourth step, which offers a very well-organized workflow throughout a subdivision in subsections. Furthermore, it is remarkable that the authors also declare the intention to use regex to implement pattern matchings for both identifying corruptions at the beginning and cleaning corruptions at the end of the DOI. This aspect represents an added value since it allows the target users – with at least a minimum expertise in the field – to figure out the prospected structure of the code.

However, even if the overall process is explained plainly and results easy to follow, there is a consistent lack in the specification of the programming language which the author plan to use for the software development. Consequently, a potential user is also denied the possibility to know the needed version of the programming language to execute the workflow and reproduce the experiment.

Method and Analysis

Methodology-focused Aspects

The reviewed protocol addresses specific research questions – nominally the identification of classes of error and reparation of invalid DOIs – which are plainly declared in the abstract, and all the structured process appears coherent in the task achievement.

However, the conclusive output material is presented only in part. In fact, at the step 4.4, the authors mention an “output CSV file” where they store “each cited DOI” in its corrected form, and “a value corresponding to the error class to which it belongs”. Yet, in the fifth step it is stated that they will “provide the number of [fixed] DOI names” and “the number of DOI names for each class of errors”, but there is no mention to the statistical tool or visualisation they plan to adopt.

Overall, it is easy to notice that the protocol is technically sound, and that the procedure presented is very likely to lead to a relevant outcome. Further, since the input data are already provided in the file format that is supposed to be managed by the software, once that the code will be made available by the authors it will be quite easy to test the quality of the output material. In addition to that, the accuracy of the authors in paying explicit attention to the quality of the output is explicated in several passages of the protocol. A clear example is provided in the fifth step, in which the corrected DOI names are newly checked through the DOI Proxy, before the output is returned.

Moving on to some weaker aspects, it is possible to observe that there is no mention about where the data underlying the findings will be published once the study will be completed. However, it is likely that this information is stated in the Data Management Plan; in this case, this lack could be easily compensated by linking the DMP in the protocol itself.

As far as the methodology feasibility is concerned, some extra details could be provided; above all – as already mentioned – the programming language used.

Further, the choice of using CSV, which is an open standardised format, is a remarkable point with respect to openness; however, for completeness' sake, it would be advisable to make mention of the source of the input CSV file, i.e.: who is the provider of the input data.

Impact

The reviewed protocol meets a need of a specific target subset of the scientific community and it opens new paths in the Open Science field by making the reconstruction of corrupted citational data possible. The achievement of this aim is not only of primary importance in Bibliometric Studies, but results in concrete advantages in any kind of research activity, facilitating the process of publications' identifiers resolution.

However, the task was defined and assigned in the context of the University course of Open Science 2020-2021; consequently, the adequate development of such a protocol is first supposed to fulfil a request of the target scientific community.

Furthermore, the authors provide added value by sharing information on the data distribution among the four classes of error, allowing supplementary studies and evaluations on the extracted data. This latter choice seems to be oriented to popularize the outcomes of the underlying research activity, in an informative fashion which could lend to some widely understandable data visualization.

No explicit mention to other resources performing either similar or complementary activities was made. However, at the third step, both Buchanan (Buchanan, 2006) and Xu (Xu et al., 2019) are explicitly cited as references in the definition of classes of error.

Availability

Since the resource was uploaded on protocols.io [7], it is easily consultable and it comes with a consistent number of metadata, among which a DOI URI [8] and a canonical citation associated with the protocol [9]. In the same Metadata section, a CCBY licence specification is provided [10]. Further, all the authors of the protocol are identified by their names and ORCID codes.

Despite the overall completeness of the metadata section, no sustainability plan was directly specified for the medium and long-term maintenance of the resource. However, the protocol is declared to be in development and optimization phase, and it is likely that such information will be provided in a further update of the protocol development or just made available throughout a link to the Data Management Plan.

Reusability

As a new resource, there is no evidence of potential usage by a wider community for this protocol. Nevertheless, since it addresses a widespread issue in a smart, accurate but still quite general fashion, it is possible to imagine that it could be modified in some specific steps and extended so to be reused also in the management of other kind of identifiers besides DOIs.

However, at the current state of the resource development, the potential for reusability should still be improved by providing

a more detailed documentation and some tutorials for potential users, in addition to a clear explanation of how users are expected to use the data and the software to be developed.

As regards the limits of this protocol, authors clearly state the possibility that a given DOI stays invalid also after the correction attempt throughout the cleaning process. Though, there is no clear mention of whether further actions are planned to be performed for still invalid DOIs after this step (i.e., Are they simply excluded from the count of fixed DOIs? Is the information about their persistent invalidity going to be stored in the output CSV file, maybe adding a third column – besides modified DOI name and class of error – declaring the newly checked status of the DOI name?).

Design and technical Quality

With respect to an exemplar protocol, e.g. : *A methodology for gathering and annotating the raw-data/characteristics of the documents citing a retracted article V.1* ***[11]***, the design of the reviewed protocol seems to follow resource-specific best practices in its overall very clear and logically sounded structure, which not only results adequate for the resolution of the given task, but also adds informative content about numbers of fixed identifiers and their distributions over the error classes. Nevertheless, in view of further developments of the resource, the protocol might be extended with the inclusion of informative images, detailed usage examples and some more verbose explanations, so to facilitate users' comprehension of the global functioning of the procedure.

Writing and Presentation

As concerns the formal aspects, the protocol is presented in a plainly intelligible fashion, written in standard English, and generally expressed with grammatical accuracy.

Some improvable details:

1. Only for the sake of consistency, the titles of the protocol steps should be uniformed according to a standard: either nominal or verbal. In fact, in the current version some of them are expressed in nominal form, while others in gerund verbal form.
2. Some of the protocol steps descriptions end with a punctuation mark, other do not. Also in this case, it would be suggestable to make this aspect uniform.
3. According to PLOS ONE guidelines [12], the title of the protocol should contain the word "Protocol".

Conclusions and Future Works

Overall, *Investigating DOIs' classes of errors* protocol is clearly sounded and addresses well-defined research questions, breaking new ground in the scientific community in general and providing the possibility to contribute to Open Science field in particular, by making available to OpenCitations project some previously unusable data.

The protocol is already structured in a linear way and in a highly comprehensible fashion; however, here it is provided in summary a list of all the suggestions for the authors, in view of further improvements:

1. If possible, specify the programming language the software is going to be implemented with.
2. Make mention of tools and visualization that are going to be adopted for the presentation of the output data related to the number of fixed DOI names and their distribution over classes.

3. Consider the possibility of adding an additional column in the output CSV file in order to store the information about the current state (“still invalid” / “valid”) of the corrected DOI names.
4. Consider the possibility of making explicit mention of the source of the input data (i.e.: Crossref data processed in Open Citations software for the inclusion in COCI index).
5. Consider the possibility of making explicit mention of the relation to the OpenCitations project, and to COCI Index in particular.
6. Consider the possibility of including the Data Management Plant in the protocol, also simply by linking it.
7. As soon as possible, try to include some tutorials or more verbose explanations in order to show the process functioning also to unexperienced users.
8. Under a merely formal point of view, try to stick to one single naming convention, either the nominal or the verbal one, for the titles of the protocol steps, and add punctuation marks at the end of each step.
9. To be compliant with PLOS ONE guidelines, add the word “Protocol” in the title of the protocol.

In conclusion, the reviewed protocol is of very good quality and it already demonstrates a high potential, even if it is still in its development phase. For this reason, it would be necessary to make further reviews of the same document in some weeks, in order to keep track of its progresses and achievements.

References

- 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
- 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
- Ross-Hellauer, T. (2017). What is open peer review? A systematic review. F1000Research, 6, 588. https://doi.org/10.12688/f1000research.11369.2

- https://direct.mit.edu/qss/pages/submission-guidelines
- https://iswc2021.semanticweb.org/resources-track
- https://github.com/open-sci/2020-2021
- https://journals.plos.org/
- http://www.semantic-web-journal.net/reviewers

[1] https://opencitations.net/

[2] dx.doi.org/10.17504/protocols.io.bt65nrg6

[3] https://www.crossref.org/

[4] https://github.com/opencitations/index/tree/master/coci

[5] https://zenodo.org/

[6] https://www.doi.org/factsheets/DOIProxy.html

[7] https://www.protocols.io/

[8] dx.doi.org/10.17504/protocols.io.bt65nrg6

[9] 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

[11] dx.doi.org/10.17504/protocols.io.bdc4i2yw

[12] https://journals.plos.org/plosone/s/reviewer-guidelines#loc-study-protocols