Review of: "Classes of errors in DOI names"

Sara Coppini

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

Definitions

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

Peer Review
Defined by Jeffrey Beck et al.

Open Peer Review (OPR), Peer Review

0. Preliminary note

This Data Management Plan is part of a project of the Open Science course (a.y. 2020/2021) of the University of Bologna, held by professor Silvio Peroni. The research that arises in this course is focused on the analysis of incorrect DOIs and their classes of errors.

General metadata about the Data Management Plan is provided (founder, grant, organization involved, and researchers' names and DOIs). This resource describes two datasets planned to be created by the authors, the code for data analysis and the output dataset of the research, which descriptions will be both taken into account in this review. A short explanation of the datasets is helpful to contextualize the content of the DMP, although some more information about their role and their genesis within the project carried out by the scholars could help the reader to have a more precise preliminary idea about the datasets.

The nature and the content of the dataset are clearly stated in the preliminary section of each dataset description part, which allows the reader to better orientate himself in the rest of the report. In the latter's first section, both datasets are provided relevant information about the purpose of the data collection or generation, the type of data involved in the project, its format, origin, expected size, and also data utility.

1. Assessment of existing data

As regards the dataset "Classes of errors in DOI names: code", in the second part of the Dataset Description main section, some questions about data reuse are answered. It is claimed that existing data is reused to combine it with other data and to develop services. Which data will be reused is explicitly declared, as is the fact that there are no copyright issues or similar issues.

For the second dataset involved, "Classes of errors in DOI names: output dataset", questions about data reuse are also

answered. We know which data were reused, but not where the dataset resides or other related information, which should still be retrieved since the source of existing data is declared.

2. Information on new data

The information on data to be produced appears to be realistic and according to the research and methodology proposed in the main report provided for the Data Management Plan, both for the coding dataset and the output dataset. Starting from the short general descriptions provided for the datasets, it is plausible that those are all data planned to be generated from the research, even if it is not clearly asserted in the DMP introduction, so there is no factual evidence.

For both datasets descriptions, the purpose and the modality of data collection are noted, as well as the formats in which the data will be analyzed and stored. The type and origin of data are reported as well, providing a clear overview of both datasets, from the collection to the final definition of data involved in the project.

There is no clear indication of how data will be documented in the strict sense of the term, but questions about metadata, naming conventions, identifiers, and similar are answered (negatively).

More attention is wished for these aspects concerning making data findable and interoperable, for instance, greater consideration to the compilation and characterization of metadata and its vocabulary, which should be effective and appropriate.

3. Quality assurance of data

About procedures for quality assurance of collected and produced data, it is specified that for both datasets the template used is Horizon 2020; but there is no other information about methods for data validation or standards applied during data collection and data entry.

As regards codes of research practice adhered to, it is clear from the outset that the entire project fits into the FAIR and Open Data domain, so special attention is paid to these criteria in the description of both project datasets. As mentioned, there are some gaps regarding the interoperability and findability of the two datasets.

Finally, for both datasets, it is explicitly affirmed that there are no documented procedures for the quality assurance of the data.

4. Backup and security of data

Topics as backup and security of data are the main subjects of sections 3.2 and 5 for both datasets descriptions. The data storage platform was chosen is stated to be secure with backup and recovery, but the procedures for backup are not clearly specified, for example, those of the institutions involved in the project or those of the platforms on which the datasets will be loaded. A policy is not specified, for instance to specify the backup frequency or copies. If possible, it would be better to specify more information about the storage and safeguarding of data by the tools used in the project. However, for both datasets is considered a version control system.

5. Expected difficulties in data sharing

Some aspects of data reuse and sharing are addressed transparently and directly for both datasets: since there are neither ethical nor legal issues that can impact sharing the data, it is affirmed that all of the data will be openly accessible and the modality and platforms devoted to access to data are distinctly specified. Those platforms are declared to be secure for the data and do not require special methods or tools to access the data, avoiding possible issues in data access. Data is thought to be available for reuse immediately, but no support was thought for data reuse.

Perhaps it could be verified on the guidelines of the platforms used for data sharing if there are exceptional cases in which the sharing of data could be limited, and in the case addressing the issue if you want different data management than that guaranteed by the tool.

No further issues could be addressed since no personal and disclose information is involved in the project and the authors stated that there are not any ethical or legal issues that can have an impact on data sharing.

6. Copyright/intellectual property right

License is specified for each dataset, and it is an internationally recognized license in both cases; but reasons of the choice could be specified for clarity, especially if it depends on the license of source data or the combination of various sources of data. However, since they are CC licenses in both cases, there should not be legal or accessibility problems in the data reuse step.

7. Responsibilities

Regarding responsibilities, data management responsibilities have not been allocated to named individuals, it is just affirmed that the whole research team is responsible for the data management. When possible, it is more correct to specify the names and identifiers of the members and their respective roles and tasks in data processing (not only in the management process).

8. Preparation of data for sharing and archiving

Regarding sharing and archiving tasks or issues, there is no evidence that data will be well documented during research to provide high-quality contextual information and/or structured metadata for secondary users. In the first part of the description of the dataset, other resources should be linked or referenced to is useful for documenting the method of data collection, origin, circumstances, processing, and analysis of data.

As mentioned, the lacking of suitable metadata appears to be a downside also for data sharing and archiving, thus it is doubly recommended to improve this part of the data preparation.

9. Other relevant topics

9.1 Reusability

The project in which this resource was created belongs to a domain linked with a wide community, thus reusability is a quite relevant point to take into account since the reuse of data collected and generated for this project and documented in this resource is plausible and desirable.
As for now, the resource is easy to use and reuse, but there is also potential for improving and extensibility to meet future requirements, also because it is an ongoing project.
It is also specified to whom the datasets might be useful (data utility).

9.2 Design & Technical quality

The design of the resource follows resource-specific best practices since it is structured according to the Argos template for data management plans. Out of the total number of questions in the model that would be appropriate to answer, the resource answers most of them. However, in cases where the answer is no, the reasons could also be given.
Furthermore, the resource is compliant with FAIR principles both for the datasets it describes and for itself, since it is available in both human- and machine-readable formats (JSON and PDF). However, for datasets description is available only within the resource and not in terms of VoID/DCAT/DublinCore.

9.3 Availability

The resource is published at a persistent URI, which is a DOI. A canonical citation is associated with the resource. The resource is publicly available, accessible, and findable since it is registered in Zenodo, which is a widely used generic repository.
It is not explicitly specified a sustainability plan for the resource, but it is evident that is frequently updated since there are two different versions of the resource for now.
Finally, the resource also adopts open standards, which added to all the other aspects mentioned above makes it unassailable from the point of view of availability.

References

1. Boente, Ricarda, Massari, Arcangelo, Santini, Cristian, & Tural, Deniz. (2021). Classes of errors in DOI names. 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.