Review of: "Classes of errors in DOI names (DMP)"

Nooshin Shahidzadeh Asadi

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

Assessment of Existing Data

The DMP cites a dataset for citations to invalid DOI-identified entities for use in both their datasets. The cited dataset is published on Zenodo under the CC0 licence, hence there are no copyright complexities when reusing data. This dataset has been generated when adding Crossref data to COCI and regenerating it would be difficult and useless, if not impossible. So reusing this dataset is a smart choice by the project collaborators.

Information on New Data

This project is set out to create two sets of new data: one for software code and one for the output achieved by the software. For the code dataset, the only information given is that the output will be a Python code with its size in the KB range. As for the output dataset, the exact structure and formatting of the CSV file they mention is not expressed and the description is generally vague. A positive point is the good estimates they make of the size of their data based on information available to them at this point.

It is expressly mentioned that they will not have support for data reuse which can be a negative point on its own, but has an unbalance with the mention of full and public availability of the data. More specific data on the point of how the data will be available to the public and

Quality Assurance of Data

Hardly any quality assurance data is present in the DMP for this project. For both their dataset, the research group declares no plans to add quality-related metadata (or any metadata at all), and the only standard they state to be following is a standard for formats (no detail given for the output dataset). Naming conventions, version numberings, etc. are not included in the DMP which can be a weak point for the project.

Backup and Security of Data

The DMP contains information about dealing with sensitive data. The plan is to store them on secure, managed storage for limited time. This is mainly adequate, but more specificity on the frequency of backups and their type would be more agreeable. The main storage for the data is declared to be secure with backup and recovery, with no special tools or methods needed to access the data. A version control system (GitHub) is mentioned for both datasets.

Expected Difficulties in Data Sharing

A main drawback of this DMP is its lack of any data about data sharing and support for reuse. They merely mention that all their data will be openly accessible, but offer no additional information on how this availability will take place and, moreover, state in their DMP that they have no plans to make this data interoperable or provide data reuse support. The licenses used are Open Data Commons Public Domain Dedication and Licence and Creative Commons Non-Commercial, more information on the way both these are planned to be used could be useful and enlightening.

Copyright/intellectual property right

The project does not use or create any data that is protected by proprietary copyright, and as mentioned before the used data is under CC0. Their plan is to mainly use CC licenses which would create no additional legal problems in data reuse stage, but more information on where they would use which license is lacking.

Responsibilities

The data management responsibilities are planned to be share across all team members. This is a logical choice for a project of this size, but more data on the main focus points of each member is advised.

Preparation of Data for Sharing and Archiving

There are no plans in this DMP to add metadata to either of their datasets, which is a major drawback of this plan. This would impair many archiving and sharing purposes and, for this reason, it is strictly advised to add searchable metadata to the project and the DMP.