Data Citation and Mahalo Button: Collecting and Sharing Dataset Usage in DIAS

*Asanobu Kitamoto[1], Yoko Nakahara[2], Toshiyuki Shimizu[3], Hiroyuki Shimai[2], Masatoshi Yoshikawa[2]

1. National Institute of Informatics, 2. Kyoto University, 3. Kyushu University

Data citation is an essential component for appreciating the contribution of data creators. We discuss two challenges of data citation, namely collecting and sharing dataset usage, based on our experiences in the DIAS (Data Integration and Analysis System) Project (https://diasjp.net/).

The first challenge, collecting dataset usage, is about extracting mentions about the dataset from publications such as papers and other documents and uniquely identifying them from a list of target datasets. The recommended practice in the open science movement is to specify the dataset's DOI in the paper's reference section. This practice leads to easier extraction and identification because we know where and how to analyze the mention. To check if the current practice follows the recommended pattern, we collected the mention of the datasets provided by DIAS and categorized them by data citation patterns.

The target of the analysis is 357 DIAS datasets, among which 146 have data DOIs. We used those DOIs and other search keywords on Google Scholar to find the usage of DIAS datasets in scholarly publications. We also collected dataset usage from user reports requested by the dataset's terms of use. Finally, we categorized the mention of DIAS datasets by data citation patterns, namely where (the reference section, the acknowledgment section, the designated section, the body section, and others) and how (with the data DOI or without the data DOI).

The result suggests several data citation patterns. First, in the reference section, the data DOI was frequently used, indicating that the authors know how to cite the dataset or follow a recommended citation format. Second, the acknowledgment section is still used frequently without the data DOI, partly because data creators still recommend it. Third, in the body or the designated section, the dataset is still mentioned as the URL or the dataset name but not as the data DOI. This result suggests that the mention of the dataset has a considerable variation, which resulted in the difficulty of automating the collection and identification of dataset usage.

To solve this problem, we proposed the "Mahalo Button" (https://mahalo.ex.nii.ac.jp/) to help data curators and users share dataset usage. Mahalo Button is installed on the landing page of the dataset. Data curators and users can register a dataset usage using the DOI (Digital Object Identifier) of research publications derived from the dataset, with a "thank you message" describing the detail of dataset usage. The registered information is then shared via the button on the landing page, which helps potential dataset users learn how to use the dataset for their research. As the previous paragraph suggests, the automatic analysis of dataset usage from publications is still challenging due to the variety of data citation patterns. The approach of the Mahalo Button can complement the automatic approach with the help of data curators and users and allow us to share dataset-derived publications that DOI-based automated citation services might fail to extract.

The Mahalo Button is already active on the dataset landing page of DIAS, and several other data repositories have already started using this button. In the initial step, data curators need to help collect dataset usage, but in the next step, data users will play an essential role in the voluntary collection of dataset usage as a response to the dataset's terms of use and as an expression of appreciation to data creators. As the Mahalo Button grows as a visible sharing point for dataset usage, it visualizes the contribution of data creators and eventually promotes the release of open data on a data repository.

Keywords: Open Science, Data Citation, Mahalo Button, Dataset Usage, DIAS, DOI