World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:10, 2018

Analysis of Citation Rate and Data Reuse for Openly Accessible Biodiversity Datasets on Global Biodiversity Information Facility

Authors : Nushrat Khan, Mike Thelwall, Kayvan Kousha

Abstract : Making research data openly accessible has been mandated by most funders over the last 5 years as it promotes reproducibility in science and reduces duplication of effort to collect the same data. There are evidence that articles that publicly share research data have higher citation rates in biological and social sciences. However, how and whether shared data is being reused is not always intuitive as such information is not easily accessible from the majority of research data repositories. This study aims to understand the practice of data citation and how data is being reused over the years focusing on biodiversity since research data is frequently reused in this field. Metadata of 38,878 datasets including citation counts were collected through the Global Biodiversity Information Facility (GBIF) API for this purpose. GBIF was used as a data source since it provides citation count for datasets, not a commonly available feature for most repositories. Analysis of dataset types, citation counts, creation and update time of datasets suggests that citation rate varies for different types of datasets, where occurrence datasets that have more granular information have higher citation rates than checklist and metadata-only datasets. Another finding is that biodiversity datasets on GBIF are frequently updated, which is unique to this field. Majority of the datasets from the earliest year of 2007 were updated after 11 years, with no dataset that was not updated since creation. For each year between 2007 and 2017, we compared the correlations between update time and citation rate of four different types of datasets. While recent datasets do not show any correlations, 3 to 4 years old datasets show weak correlation where datasets that were updated more recently received high citations. The results are suggestive that it takes several years to cumulate citations for research datasets. However, this investigation found that when searched on Google Scholar or Scopus databases for the same datasets, the number of citations is often not the same as GBIF. Hence future aim is to further explore the citation count system adopted by GBIF to evaluate its reliability and whether it can be applicable to other fields of studies as well.

Keywords : data citation, data reuse, research data sharing, webometrics
Conference Title : ICCSIB 2018 : International Conference on Cybermetrics, Scientometrics, Informetrics and Bibliometrics
Conference Location : Barcelona, Spain
Conference Dates : October 29-30, 2018