Data Citation and Reuse Practice in Biodiversity – Challenges of Adopting a Standard Citation Model

Nushrat Khan[1], Mike Thelwall[2] and Kayvan Kousha[3]

[1]n.j.khan@wlv.ac.uk, [2]m.thelwall@wlv.ac.uk, [3]k.kousha@wlv.ac.uk
University of Wolverhampton, Wulfruna St, Wolverhampton, WV1 1LY (United Kingdom)

Abstract

Openly available research data promotes reproducibility in science and results in higher citation rates for articles published with data in biological and social sciences. Even though biodiversity is one of the fields where data is frequently reused, information about how data is reused and cited is not often openly accessible from research data repositories. This study explores data citation and reuse practices in biodiversity by using openly available metadata for 43,802 datasets indexed in the Global Biodiversity Information Facility (GBIF). Quantitative analysis of dataset types and citation counts suggests that the number of studies making use of openly available biodiversity data has been increasing in a steady manner. Citation rates vary for different types of datasets based on the quality of data, and similarly to articles, it takes 2-3 years to accrue most citations for datasets. Content analysis of a random sample of unique citing articles (n=101) for 437 cited datasets in a random sample of 1000 datasets suggests that best practice for data citation is yet to be established. 26.7% of articles are mentioned the dataset in references, 12.9% are mentioned in data access statements in addition to the methods section, and only 2% are mentioned in all three sections, which is important for automatic extraction of citation information. Citation practice was inconsistent especially when a large number of subsets (12~50) were used. This calls for adoption of a standard citation model for this field to provide proper attribution when using subsets of data.

Introduction

Reproducible science is of major importance to the scientific community and the datasets reported in research articles are rich source for this. Data sharing practices seem to be more common in some fields, such as medical, forensic, and evolutionary genetics (Anagnostou, Capocasa, Milia, & Bisol, 2013). Hence, open research data initiatives have been growing quickly within different communities as data sharing aids reproducibility and reduces the chances of researchers generating data that has already been collected. However, publishing research data as first-class research output opens the door to more complex questions for researchers and policy makers –from how to define a dataset to establishing best practices of citing datasets in a specific field (Borgman, 2012; Kratz & Strasser, 2014; Starr et al., 2015; Silvello, 2018).

This study focuses on biodiversity datasets because it seems that sharing and reusing of globally collected research data are more common in this field, with primary data uses being ecological studies, taxonomic works, and phylogenetic analyses (Magurran et al., 2010; Troudet et al., 2018). For instance, a survey of 370 international researchers in biodiversity sciences indicated that most (84%) agreed that "sharing article-related data is a basic responsibility". (Huang et. al., 2012, p. 401). Global Biodiversity Information Facility (GBIF, www.gbif.org/) was used as a data source because this group has been working towards developing data publishing standards for biodiversity from an early stage (Moritz et al., 2011) and the platform holds large number of diverse datasets from different countries. Furthermore, it supports an application programming interface (API) to collect citation counts to datasets on a large scale in an automated way.

Researchers have recognized the need to provide attribution to datasets long ago. Ingwersen and Chavan (2011) suggested using Data Usage Index (DUI), an indicator based on search events and dataset download instances to demonstrate the impact of data creator and publishers. However, use of persistent identifiers for dataset was not a common practice at that time. At present, most data reuse in biodiversity uses sets of multiple datasets and these are provided with a DOI and accession date when downloaded to cite those subsets using their in-house style. GBIF has developed a semi-automated system to assign citations to the main datasets that were included in the subsets re-used and cited by a research article.

Citing subsets complicates developing a standard model to provide the most useful information to the readers and users. As indicated by Kratz and Strasser (2014, p. 6), "…to reproduce an analysis performed on a subset of a larger dataset, the reader needs to know exactly what subset was used (e.g., a limited range of dates, only the adult subjects, wind speed but not direction). Datasets vary so widely in structure that there may not be a good general solution for describing subsets." The recently developed Dataset Search system of Google can detect the subsets mentioned indexed by DataCite. However, the original datasets need to be recognized and should be indexed by general indexing systems as well.

Citation information is not captured by most data publishing platforms due to difficulties with automating the process, caused by a lack of standards in citation styles. This makes GBIF an interesting source of information to study current data citation and reuse practices in this field and poses questions about what should be the best citation practice to make them machine-readable and how to develop a standard citation model.

Background

Citing datasets as professional reward for sharing has been mentioned by researchers in biodiversity and other fields to be a major incentive for making data openly available (Piwowar, 2011; Edmundus et al., 2012; Enke et al., 2012; Kim & Zhang, 2015; Kratz & Strasser, 2015; Sayogo & Pardo, 2013). The number of publications using GBIF data and citing GBIF has rapidly increased since 2007 (Costello et al., 2013). However, few datasets are cited in a standard format in biodiversity and the citation style is often determined by the editors for their journal (Costello et al., 2013). This is similar to life sciences data in Dryad, where the number of articles citing data in works cited section was only 8% as of 2014 (Mayo, Vision & Hull, 2016).

Previous studies have used the WoS Data Citation Index (DCI) to analyze data citation practices (Robinson-García, Jiménez-Contreras & Torres-Salinas, 2016; Park & Wolfram, 2017). However, there is evidence that DCI is relatively biased towards hard sciences and as of 2016, four repositories represented around 75% of the database (Robinson-García, Jiménez-Contreras & Torres-Salinas, 2016), although the current version of DCI indexes wider data repositories such as Figshare. Nevertheless, citation information available for each dataset on GBIF is not captured by DCI. This is an important omission, given the importance of this repository for biodiversity research and its relatively mature architecture.

The following research questions address the lack of knowledge about citation practices in GBIF – 1) Does the type of dataset or quality of information available affect citation rate? 2) How quickly do dataset citations accrue? Has the number of articles citing GBIF datasets changed over the past years? 3) Does the citation count on GBIF result from coherent citation practices? How does the use of large number of subsets impact citation practice?

Methods

This research applies an exploratory method to study the citation and reuse practice of biodiversity datasets. Quantitative analysis was used for the GBIF metadata and then content analysis was used for each unique citing article to collect information on citation location and data reuse context (Khan & Thelwall, 2019).

Data Collection

Metadata from 38,878 datasets was initially collected through the GBIF API in May 2018. The metadata fields retrieved included the dataset key, publishing organization key, dataset DOI, dataset type, title, description, language, homepage URL, citation, citation count, creation date, and last modification date.

A random sample of 1,000 datasets was then selected for a content analysis of articles that cited datasets. About 44% (437) of datasets in the sample had at least one citing article. Between October 2018 and March 2019, a random citing article and its associated metadata was manually collected for each of the 437 datasets for full-text analysis. Download counts were also manually collected since that could not be directly retrieved through the API.

The total number of unique citing articles in the random collection was 102 as some articles appeared repeatedly for a majority of datasets. One article could not be accessed, so 101 articles were used for content analysis. The publication year, publishing journal, citation location, and contextual information of data reuse were collected for each one.

Since the data collection of citing articles was completed in 2019, an updated dataset with 43,971 datasets and a list of all citing articles for them was collected on April 6, 2019. This dataset was used to explore the distribution of all unique citing articles over publishing years.

Data Analyses

Preliminary exploration identified four types of datasets available on GBIF (GBIF, www.gbif.org/dataset-classes) – 1) Checklist datasets provide a catalogue or list of named organisms or taxa and can be used as a rapid summary or baseline inventory of taxa in a given context, 2) Occurrence datasets provide information about the location of individual organisms in time and space, 3) Sampling Event datasets contain more granular information than Occurrence datasets, often comprising of abundant information to assess community composition for broader taxonomic groups, and 4) Metadata-only datasets describe undigitized resources like those in natural history and other collections.

After de-duplicating 169 records, 43,802 datasets were used for analysis. Citation counts were analysed for all types of dataset to explore the first research question. The creation dates for each dataset were processed and average citations were calculated for the years between 2007 and 2019 for Occurrence datasets to explore how long it takes to accrue dataset citations. The list of all citing articles was de-duplicated to identify all unique articles and was used to explore the distribution over each publishing year.

A content analysis for 101 unique citing articles was conducted for a random sample of 1000 datasets for exploring the research question 4. The Spearman correlation between download and citation counts was calculated for the 437 cited datasets to help assess whether they reflect a similar type of impact.

Results

Average citations for each dataset type suggest that, Occurrence datasets are most frequently cited since they offer direct evidence of the occurrence of a species (or other taxon) at a particular place on a specified date.

Table 1. Type and number of datasets published between 2007-19 and average citations

Type	*Number*	*Percentage (%)*	*Citations per dataset*
Occurrence	16,712	93.2%	9.82
Checklist	26,216	6.4%	0.43
Metadata-only	286	0.0%	0.06
Sampling Event	588	0.4%	1.32

Prior to 2011, Occurrence datasets were the only type of datasets made available on GBIF except for two Sampling Event datasets that were published in 2007. Despite of the evidence of a higher number of citations received by Occurrence datasets, there was a rapid increase in publishing Checklist datasets in 2016 and it is unclear why.

This study focuses on Occurrence datasets only since these are the type of datasets frequently reused and cited by articles. Figure 1 demonstrates a relatively consistent growth of Occurrence datasets. Mean citation received per occurrence dataset was 9.82, with the highest of 24.02 for occurrence datasets published in 2015 and a lowest of 0.9 for 2018. The drop in average citations per paper after 2015, indicates that, as for articles, it takes 2-3 years to accrue most citations for datasets.

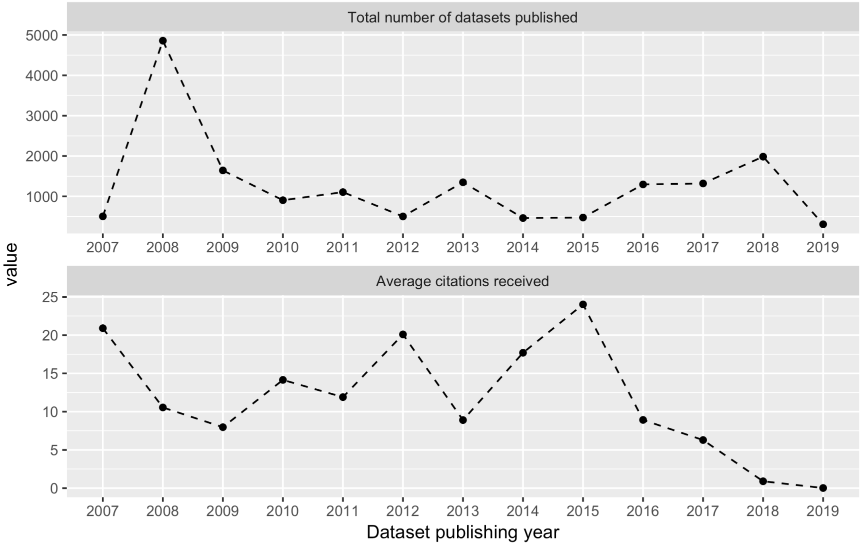

Figure 1. Number of occurrence datasets published, and average number of citations received

A correlation test was conducted for download and citation counts for the random sample of 437 cited datasets, finding a very strong positive correlation (rho = 0.787, p=0.000). Thus download counts and citation counts suggest a similar kind of impact. Because of this, early download counts might be a good indicator of longer term citation counts. Similar to the citation count findings above, Checklist datasets (n=92, average download=2610.38) had much lower download counts than Occurrence datasets (n=343, average download=5210.92) in general.

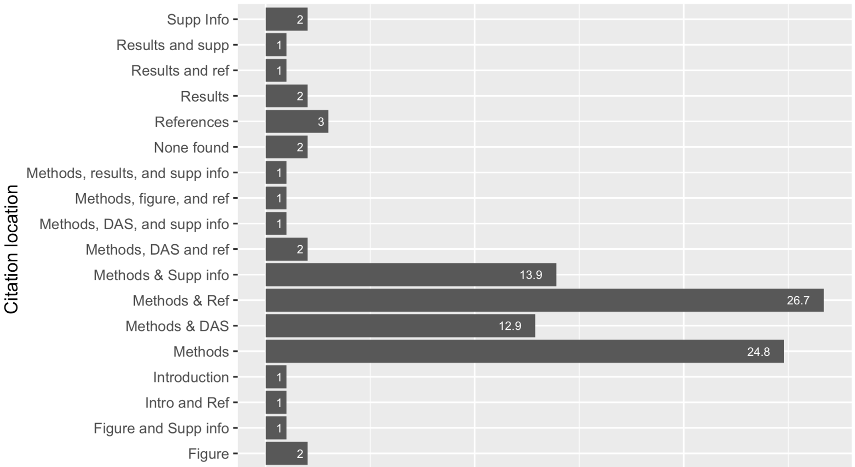

Figure 2. Citation location in randomly selected articles

A content analysis of 101 unique articles was conducted to understand citation practices in biodiversity articles citing GBIF datasets (Figure 2). Citation for GBIF dataset could not be located in two datasets. For the remaining datasets, 26.7% of the articles mentioned the dataset in their reference lists and 12.9% in data access statements in addition to the methods section, which is considered to be the standard citation practice. However, 24.8% mentioned the datasets in the methods section only within the text, which is difficult to find with indexing systems. Mentions in methods and supplementary material sections were also common (13.9%).

Most (52.5%) articles listed one subset, but some cited many (8.6% cited 50 subsets) where the number of subsets did not match for 4.9%. The non-standard citation method was used especially by articles that used large number of datasets (12~50), perhaps making it difficult to include them all in the reference section. Some articles appeared repeatedly for a majority of datasets with top 5 appearing as a citing article for more than 10,000 datasets. Those are usually the studies that used large number of records, ranging from 200 to 600 million occurrence data.

Table 2. Publication year of all citing articles mentioned on GBIF

Article Publishing Year	*Number*	*Percentage (%)*
2013	4	0.6
2014	5	0.8
2015	23	3.6
2016	70	10.9
2017	178	27.7
2018	260	40.5
2019	102	15.9

To date, 642 articles have been listed as citing article by GBIF for a total of 43,802 datasets. From the data in Table 2, it is obvious that data reuse in this field (at least from this source) has been increasing since 2013 as the number of citing articles has been growing consistently. The growth indicates the importance of openly available biodiversity data for researchers.

Discussion

This study explores data citation and reuse practice in biodiversity. It found evidence that openly available biodiversity data on GBIF is frequently reused by researchers and that the number of articles reusing and citing data retrieved from GBIF has been increasing steadily.

Citing data in references or data access statements is becoming more common but citation practices remain inconsistent across different journals. Articles using many data subsets pose extra challenges for citing in an appropriate manner. Publishing a data paper for the articles using many subsets and citing the paper itself could be a solution to this issue (Chavan & Penev, 2011). However, a refined and standard model should be adopted to address this problem when a data paper is not available. The model should also define a better way to inform the users regarding the reused subset rather than using only "GBIF Occurrence Download" as the reference title.

References

Anagnostou, P., Capocasa, M., Milia, N., & Bisol, G. D. (2013). Research data sharing: Lessons from forensic genetics. Forensic Science International: Genetics, 7(6), e117-e119.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, *12*(15), S2.

Costello, M. J., & Wieczorek, J. (2014). Best practice for biodiversity data management and publication. *Biological Conservation*, *173*, 68-73.

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, *28*(8), 454-461.

Edmunds, S. C., Pollard, T. J., Hole, B., & Basford, A. T. (2012). Adventures in data citation: sorghum genome data exemplifies the new gold standard. *BMC research notes*, *5*(1), 223.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. *Ecological Informatics*, *11*, 25-33.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters, 5(5), 399-406.

Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, *12*(15), S3.

Khan, N., & Thelwall, M. (2019). Dataset supporting "Data Citation and Reuse Practice in Biodiversity". figshare. Dataset. 10.6084/m9.figshare.8181098.v1

Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library & Information Science Research*, *37*(3), 189-200.

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, *3*.

Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific data*, *2*.

Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS One*, *10*(2), e0117619.

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in ecology & evolution*, *25*(10), 574-582.

Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *International Journal of Digital Curation*, *11*(1), 150-155.

Moritz, T., Krishnan, S., Roberts, D., Ingwersen, P., Agosti, D., Penev, L., Cockerill, M., & Chavan, V. (2011). Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics*, *12*(15), S1.

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, *111*(1), 443-461.

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PloS one*, *6*(7), e18657.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19-S31.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6-20.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R., Duerr, R., Haak, L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J., Lin, J., Nielsen, L., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M. and Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1, p.e1.

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964-2975.

Troudet, J., Vignes-Lebbe, R., Grandcolas, P., & Legendre, F. (2018). The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it?. *Systematic biology*.