Measuring the Impact of Biodiversity Datasets: Data Reuse, Citations and Altmetrics

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

Despite growing evidence of open biodiversity data reuse by scientists, information about how data is reused and cited is rarely 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) and content analyses of articles citing GBIF data. Results from quantitative and content analyses suggest that even though the number of studies making use of openly available biodiversity data has been increasing steadily, best practice for data citation is not yet common. It is encouraging, however, that an increasing number of recent articles (16 out of 23 in 2019) in biodiversity cite datasets in a standard way. A content analysis of a random sample of unique citing articles (n=100) found various types of background (n=18) and foreground (n=81) reuse cases for GBIF data, ranging from combining with other data sources to create species distribution modelling to software testing. This demonstrates some unique research opportunities created by open data. Among the citing articles, 27% mentioned the dataset in references and 13% in data access statements in addition to the methods section. Citation practice was inconsistent especially when a large number of subsets (12~50) were used. Even though many GBIF dataset records had altmetric scores, most posts only mentioned the articles linked to those datasets. Among the altmetric mentions of datasets, blogs can be the most informative, even though rare, and most tweets and Facebook posts were for promotional purposes.

Keywords

Open biodiversity data, altmetrics, data reuse, citation practice, citation analysis

Article Highlights

- Open data in biodiversity create unique research opportunities and are frequently reused for background and foreground research.
- Usage of multiple subsets complicates data citation in biodiversity. An alternative citation attribution method for GBIF is recommended.
- Blogs can be most informative of altmetric mentions even though rare. Tweets and Facebook posts are mostly promotional.

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 medicine, forensics, and evolutionary genetics (Anagnostou, Capocasa, Milia, & Bisol, 2013). Hence, open research data initiatives have been growing at different rates within different communities. However, publishing research data as first-class research outputs 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 because sharing and reusing globally collected research data is 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, in a survey of 370 international biodiversity science researchers, most (84%) agreed that "sharing article-related data is a basic responsibility" (Huang et. al., 2012, p. 401). The 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 for datasets on a large scale in an automated way.

Researchers have long recognized the need to provide attribution for dataset reuse. Ingwersen and Chavan (2011) suggested a Data Usage Index (DUI), an indicator based on search events and dataset download instances to demonstrate the impact of data creators and publishers. However, the use of persistent identifiers for datasets was not common at that time. At present, all datasets indexed on GBIF are provided with a DOI and when a combined data "subset" is downloaded from GBIF based on a search query (i.e., a collection of separate datasets), it is provided with its own DOI and accession date to cite. GBIF has also developed a semi-automated system to assign citations to the datasets included in subsets reused and cited by research articles.

Citing subsets complicates developing a standard model to estimate disparate and fractional contributions. 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." It is crucial that the original datasets are recognized in the right manner and their citations can be indexed by relevant systems, such as Google Dataset Search. 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 can be a major incentive for sharing (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). The current version of DCI indexes wider data repositories, such as Figshare, however. 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.

Bishop and Kuula-Luumi (2017) investigated data reuse cases for the UK Data Service (UKDS) and found that 64% of datasets were used for learning, followed by 15% for research purposes and 13% for teaching. This information was only available when UKDS required user registration for data download purposes and prior to 2013, none of the data collections were open data. When datasets are openly accessible and no information about usage purpose is requested from the users, it is difficult to track such societal impact. Altmetric sources could be useful in finding such use cases for datasets. Konkiel (2013) calls for using altmetrics to track various types of engagement that different stakeholders can have with a single dataset, such as discussions, formal references, and recommendations. It is not known, however, whether altmetric scores currently reflect such impact for datasets. Peters, Kraker, Lex, Gumpenberger and Gorraiz (2016) explored any relationship between citations and altmetric scores for research datasets in three different platforms and found that few cited research data had altmetrics, although it had increased in recent years. Importantly, no studies have published content analyses of altmetric mentions to understand whether the scores should be considered without context. This is particularly difficult because the content of each altmetric source must be accessed individually for such analyses.

This study looks beyond the numbers of altmetric sources and citation counts to explore how biodiversity data is reused and cited by the researchers and whether altmetric sources can be relied on to capture the impact of research data beyond research. 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 rates?
2. How quickly do dataset citations accrue? Has the number of articles citing GBIF datasets changed over recent years?
3. How do articles listed as citing datasets on GBIF reuse them, if at all?
4. Does the citation count on GBIF result from coherent citation practices? How does the simultaneous use of many subsets impact citation practice?
5. Do altmetric scores for GBIF datasets correlate with citation counts? Are altmetric scores informative about the impacts of open biodiversity data?

Methods

This research applies an exploratory method to study the citation and reuse practices of biodiversity datasets and assess the content of altmetrics sources that mention those 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 (Khan & Thelwall, 2019a). Further information on data reuse context in those articles was then collected to understand the reuse cases of open biodiversity data. Quantitative analysis was conducted for the altmetric scores collected for GBIF datasets and samples from four altmetrics sources were then used for content analysis (Khan & Thelwall, 2019b).

Data Collection

a) Data from the GBIF API

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 with a random number generator for a content analysis of articles that cited datasets. About 44% (437) of the 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 they could not be directly retrieved through the API.

The total number of unique citing articles in the random collection was 102 as some articles cited many datasets. The full text of two articles could not be accessed. However, one of those two articles had the associated dataset listed in the references. So, in total 100 articles were used for the content analysis to explore data reuse cases, but 101 articles for citation location count including the article with dataset citation in the references. The publication year, publishing journal, citation location, and contextual information of data reuse were collected for each one.

Since the data collection for citing articles was completed in 2018, 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.

b) Data from Altmetric Explorer

Data from Altmetric Explorer was collected for 43,971 GBIF dataset DOIs on April 21, 2019. Four altmetric sources – blogs, Twitter, Facebook and Wikipedia - were selected for this study as these can contain the contextual information necessary to understand whether and how a dataset was useful. Thus, the Altmetric Explorer data was split into four subsets - one for each altmetric source, where each record had received one or more mentions. A random sample of 100 dataset records from the blog subset was created for the phase 1 content analysis using a random number generator. Blogs were chosen since these gave the most detailed contextual information. For every dataset record in the blog sample, citation counts were also collected with Google Dataset Search, where available, to understand its coverage and compare citation counts between GBIF and Google. Based on the findings of phase 1, we focused on Occurrence dataset mentions in phase 2 and performed content analyses on a random sample of Occurrence datasets from each source. This gave four samples for phase 2.

Data Analyses

Preliminary explorations identified four types of datasets available on GBIF (GBIF, www.gbif.org/dataset-classes). *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. *Occurrence* datasets provide information about the location of individual organisms in time and space. *Sampling Event* datasets contain more granular information than Occurrence datasets, often containing abundant information to assess community composition for broader taxonomic groups. *Metadata-only* datasets describe undigitized resources in natural history and other collections.

After de-duplicating 169 records, 43,802 datasets were used for analysis. Citation counts (as reported by the GBIF API) were analyzed 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 research questions 3 and 4. A 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.

To explore research question 5, we performed content analyses and correlation tests for citation counts and altmetric scores. For the content analysis, mentions were examined to understand why these datasets were mentioned on the social web, what users talk about when discussing datasets on social media and whether altmetric mentions demonstrate dataset impact. At first the data was collected and analyzed for a random blog sample of 100 records. Based on the findings, the altmetrics dataset was then filtered for Occurrence dataset mentions only. The data collection method mentioned above were then repeated for the Occurrence subsets. In total, five random samples with more than one altmetric mentions were analyzed: 1) Blog sample for all types of datasets, 2) a. Blog sample, b. Tweeter sample, c. Facebook sample, and d. Wikipedia sample for Occurrence datasets.

Peters, Kraker, Lex, Gumpenberger and Gorraiz (2016) studied correlations between citation counts gathered from the Thomson Reuters Data Citation Index (DCI) and altmetric scores from PlumX, ImpactStory, and Altmetric.com. Their study found no correlation between the number of citations and the overall altmetric scores and observed that some research data can have high altmetric scores even though not cited. We also examined correlations between citation counts from GBIF and their altmetric scores to understand whether they produce similar results.

Results

Dataset quality and citation rate

Occurrence datasets are the most frequently cited, presumably because they offer direct evidence of the occurrence of a species (or other taxon) at a particular place on a specified date (Table 1).

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

Type	*Datasets*	*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 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 (Figure 1).

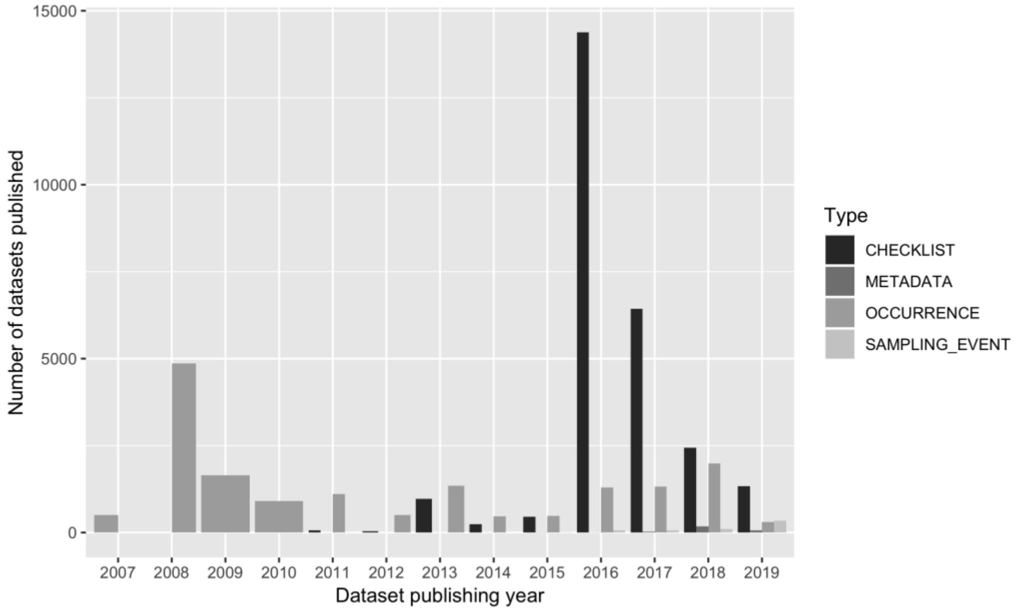

Figure 1. Distribution of different types of datasets published between 2007-2019

Citation growth rate and distribution of citing articles over the years

This study examines Occurrence datasets only since these are the type of datasets frequently reused and cited by articles. Figure 2 demonstrates a relatively consistent growth in posting Occurrence datasets. The mean number of citations 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 dataset citations.

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 downloads=2610) had much lower download counts than Occurrence datasets (n=343, average downloads=5211) in general.

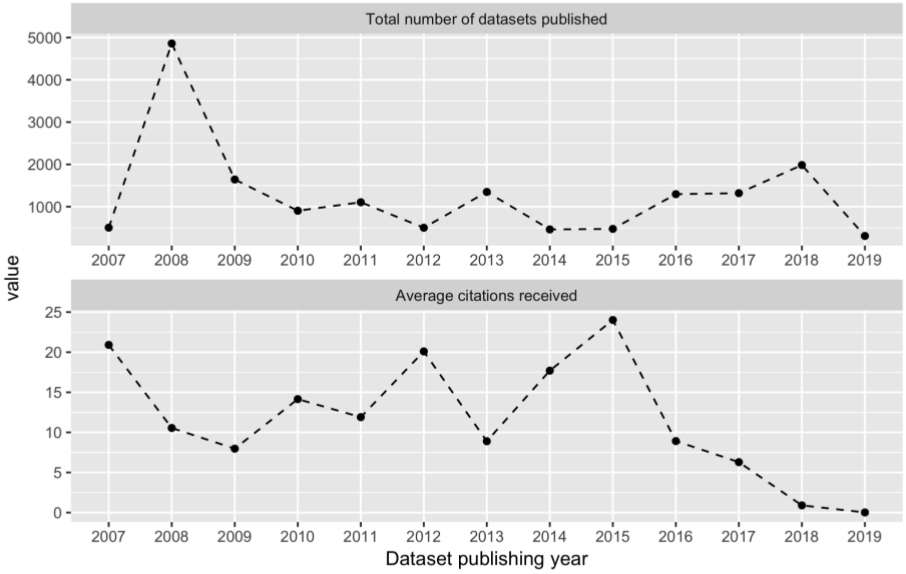

Figure 2. Number of occurrence datasets published, and average number of citations received

To date, 642 articles have been listed by GBIF as citing 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.

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

Article Publication Year	*Articles*	*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

Type of data use and reuse cases

Excluding the two articles for which full text could not be accessed, we analyzed the full texts of 100 citing articles to identify whether these articles used data from GBIF and why (Khan & Thelwall, 2019b). Based on the type of usage, we categorized them as foreground and background data (Wallis, Rolando, & Borgman, 2013). Foreground data are those needed to answer the particular research questions posed in a study and background data are the type of contextual information needed to establish research questions but not to answer the research questions.

GBIF was not mentioned as a data source anywhere in one article published in 2017. Among the remaining 99 articles, most used GBIF as a source of data to answer part of their research question. Two coders categorized the articles as ‘foreground’ or ‘background’ and identified in total 81 foreground use cases and 18 background use cases based on the information provided in those articles, especially the methods sections (kappa value for interrater reliability = 0.61, equating to “substantial” agreement).

Common foreground use cases are following: using occurrence data to create species distribution model, combining GBIF data with other databases to answer research questions, to analyze and observe the change in practice in collecting biodiversity data, investigate sampling and taxonomic bias, using the GBIF backbone taxonomy to solve taxonomic problems (synonyms), creating a species temperature index (STI) and a conservation network. We also marked data papers as foreground uses since the datasets produced are the basis of those articles.

Background use cases of biodiversity data include the following: using test datasets for software or tools testing purposes, explorations to establish the need for research in that area, simulation models, data mining, creating a baseline model, and comparing with previous records of a species' prior occurrence. Such varied use cases for biodiversity datasets demonstrate the importance of open data in this field and that it creates the opportunity to study biodiversity and related fields in many ways.

Citation practice for biodiversity datasets

A content analysis of 101 unique articles was conducted to understand citation practices in biodiversity articles citing GBIF datasets (Figure 3).

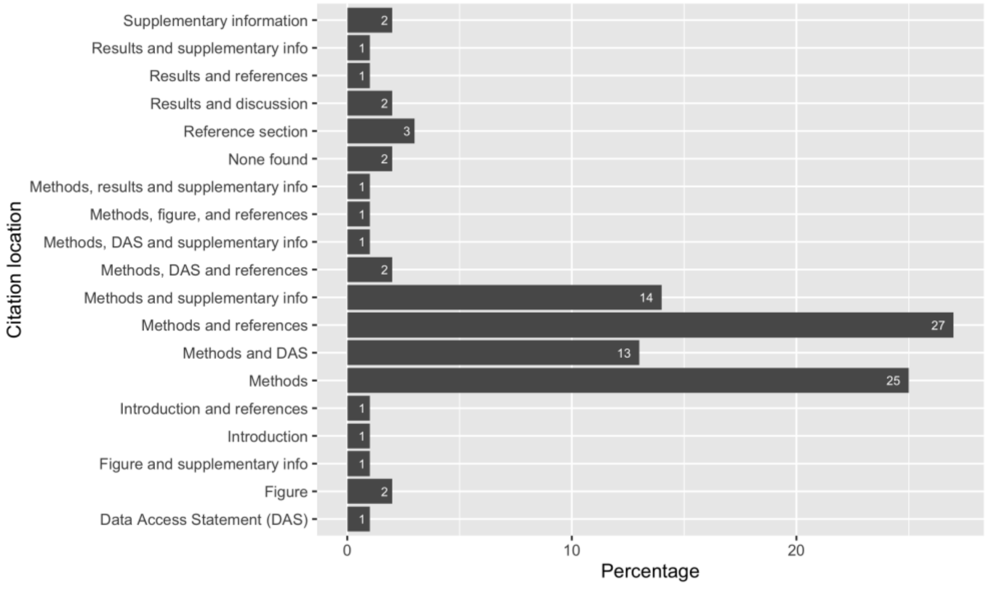

Figure 3. Citation location in randomly selected articles

Citations to GBIF datasets could not be located in two cases. For the remaining 99 datasets, 27% of the articles mentioned the dataset in their reference lists and 13% in data access statements in addition to the methods section, which is considered to be the standard citation practice. However, 25% 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 (14%). 59 articles in the sample were published in 2018 and 23 in 2019, and recent articles are increasingly adopting a standard method of citing data. 16 out of 23 articles published in 2019 had cited the datasets in references and data availability sections besides methods, which is encouraging.

Most (53%) articles listed one GBIF subset, but some cited many (9% cited at least 50 subsets). When comparing the number of subsets listed within the articles and on GBIF records of the corresponding articles, the number of subsets did not match for 5% of the articles. For example, "Species and river specific effects of river fragmentation on European anadromous fish species" (DOI: 10.1002/rra.3386) cited one GBIF subset but the record on GBIF for that article lists 16 subsets. Non-standard citation methods were especially employed by articles that used large numbers of datasets (12~50), perhaps making it difficult to include them all in the reference section.

Do altmetrics reflect data impact?

A small percentage of dataset records received a non-zero altmetric score. We found that a total of 2,111 (4.8%) datasets were mentioned in blogs, 7,271 (16.6%) were mentioned in tweets, 3,459 (7.9%) were mentioned on Facebook posts and 1,913 (4.4%) were mentioned on Wikipedia.

<u>Content Analysis - Phase 1</u>

Random sample 1 – Blog mentions (All datasets)

Blogs were used as the first sample for content analysis as blog posts do not have limited word counts, can be combined with different media, and bloggers who write about scientific topics are often domain experts (Shema, Bar-Ilan & Thelwall, 2014). Science blogging is used for explaining science to the general public and thus can bridge the gap between research and other parts of the society (Bornmann, 2015).

In the random sample of 100 GBIF dataset records that were mentioned by one or more blogs, 98 were Checklist datasets and only 2 were Occurrence datasets. The altmetric mentions were for articles associated with datasets and almost all the blogs mentioned discovery of a new species. Perhaps this is because Checklist datasets catalogue named organisms or taxa and the articles provide the most in-depth information for learning about these newly discovered species. The most frequently appearing blog was Species New to Science (n=52) and Earthling Nature and its Portuguese version, Natureza Terráquea (n=18) (Table 3). Others included blogs by citizen scientists, academic blogs that list faculty publications and blogs from news outlets. While learning about new discoveries in biodiversity is important, these only mentioned two dataset records in our sample, which led to phase 2.

Table 3. Top 5 blogs mentioning GBIF datasets

Blog name	Dataset mentions
Species New to Science	52
Natureza Terráquea	11
Earthling Nature	7
DNA Barcoding	6
Pensoft Blog	5

Google Dataset Search found no results for 43 records and, among the 57 records that were found, the citation counts varied greatly. The most striking differences were found for the dataset "Artportalen (Swedish Species Observation System)" (DOI: 10.15468/kllkyl) where GBIF listed 109 citations at the time of data collection and Google listed only 3 citations. Another example is for the checklist dataset, "Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny" (DOI: 10.1046/j.1439-0469.2001.00155.x), for which Google Dataset Search listed 28 citations and GBIF did not list any. Even though Google Dataset Search is still in beta form, the differences evidence that the current semi-automated system of applying citations to datasets may not be consistent and not open to general indexing systems at present.

Content Analysis - Phase 2

Checklist datasets received a majority of altmetric mentions (Table 4). This is not surprising as phase 1 showed that most of these datasets linked to the associated articles and new discoveries are likely to be mentioned on social media. Very few of the Occurrence datasets received altmetric mentions, with Twitter mentions being the most common and the rest having below 100 mentions. A random sample of 100 was used for Twitter and all of the results for blogs, Facebook and Wikipedia.

Table 4. Distribution of altmetric mentions for different types of datasets

	Occurrence	*Checklist*	*Sampling-event*	*Metadata-only*
Blogs	11	2,099	0	1
Twitter	403	6,700	128	40
Facebook	74	3,378	5	2
Wikipedia	5	1,908	1	0

Random sample 2a – Blog mentions (Occurrence datasets)

In total 11 datasets had received one or more blog mentions, which includes the two that were part of the random sample in phase 1. One post from the Teaching Biology blog was deleted, so 10 blog posts were analysed. Two were from the GBIF Blog and two were from the iPhylo blog, and the rest of the blog sources were different. All the blogs were written in English except for one blog in Dutch.

Even though rare, the blog posts provide useful insight on the usability and impact of datasets. All the blogged datasets received one or more citations (highest 284), as listed on GBIF, but the Google Dataset Search gave very different results again.

While the blog posts are often from the data publishers or creators, these can be useful sources of information for understanding unique use cases of open biodiversity data (Table 5). For example, the blog post "App combines computer vision and crowdsourcing to explore Earth's biodiversity, one photo at a time" talks about improving computerized species detection using research-grade observation data published on iNaturalist. Similarly, the blog post "Estimating changes in seasonal site occupancy using opportunistic observations" explains about a study that used the dataset to introduce a novel dynamic occupancy model that attempts to cope with known sources of bias including lack of absence data and variation in sampling effort. Other blog posts point out why the datasets are important, and they can be used for research and other purposes in the future.

Random sample 2b - Tweeter mentions (Occurrence datasets)

For this random sample of 100 Twitter mentions, we collected data on who tweeted and the content of the tweet except one missing tweet. Most (79) tweets were from the data publisher or GBIF and the rest were from data creator (9), domain experts (6), data management/ research data specialists (2), journal publishers (2), and a museum (1). 70 of these were tweets and retweets about publishing new data, four were on expansion and updates of existing datasets, and three on data paper publishing. A few tweets were less generic. For example, one tweet links to a news article by Chicago Tribune on the importance of data published on GBIF, another one indicates this dataset should be a starting point of many discussions, and another reply expresses thanks for sharing the data.

Table 5. Content from blog mentions of Occurrence datasets

Dataset title	*Blog title*	*Blog content*
Global compendium of Aedes albopictus occurrence (10.15468/7apj8n)	GBIF and impact: CrossRef, FundRef, and Altmetric	Blog post on understanding the impact of GBIF data
iNaturalist Research-grade Observations (10.15468/ab3s5x)	App combines computer vision and crowdsourcing to explore Earth's biodiversity, one photo at a time	Blog post on how research-grade observation data published on iNaturalist and GBIF is improving computerized species detection
EOD - eBird Observation Dataset (10.15468/aomfnb)	eBird 2017: Year in Review	Blog post on eBird data acquisition in 2017 with link to its dataset on GBIF.
International Barcode of Life project (iBOL) (10.15468/inygc6)	iBOL DNA barcodes in GBIF	Blog post by the dataset author on expansion of his dataset to include 2.7 million barcodes.
Artportalen (Swedish Species Observation System) (10.15468/kllkyl)	Estimating changes in seasonal site occupancy using opportunistic observations	Blog listed a study that used data from this dataset to introduce a novel dynamic occupancy model that attempts to cope with known sources of bias including lack of absence data and variation in sampling effort.
published Chenopodium vulvaria observations (10.15468/oyorvb)	Rejuvenating Centuries' Old Botany with Phytogeography	Blog post on biogeography using historical data on plant distribution and with a link to the associated geo-reference deposited on GBIF.
CABI Africa Invasive and Alien Species data (10.15468/pkgevu)	Largest Invasive Alien Plant dataset is now published online!	Blog by the data publisher on the context and impact of this large dataset on invasive series.
Xeno-canto - Bird sounds from around the world (10.15468/qv0ksn)	Vogelgeluiden van tienduizend vogelsoorten online beschikbaar	Blog post with the contextual and content description of the dataset made available on GBIF by the Netherlands Biodiversity Information Facility.
Occurrences of the invasive plant species Heracleum sosnowskyi Manden in the Komi Republic territory (10.15468/zo2svq)	Tracking the invasion of Sosnowsky's hogweed in the Komi Republic	Blog post and data paper on the collection and use of the associated dataset.
DNA barcoding the fishes of Lizard Island (Great Barrier Reef) (10.3897/bdj.5.e12409)	Tuesday reads	Blog post on a study that conducted short expedition to collect DNA barcodes of the fishes in Lizard Island.

Random sample 2c – Facebook mentions (Occurrence datasets)

There were 74 mentions of Occurrence datasets on Facebook and all were used for content analysis. Similar to Twitter, most Facebook posts were promotional and contained news on publishing new datasets, data papers, and updating existing datasets. Table 6 shows the distribution of content creators, where the Biodiversity Information System of Colombia (SiB Colombia) is the most active promoter.

Table 6. Distribution of content creators for Facebook mentions

Content creator	Number of posts
Biodiversity Information System of Colombia (SiB Colombia)	41
GBIF	17
Data journal publisher	7
Community	5
Journal publisher	2
Data creator	1

Even though Facebook posts were for promotional purposes, the posts often contained information on the usefulness of the dataset. Here is an example post from SiB – "The Selva Association published a set of #OpenData with more than 4,000 biological records from the Serranía del Darién, a key region for the movement of different animals between Central and South America. The objective of the project in which the registries were carried out was to determine the importance that this #biodiversity can have as a source of income and to formulate an ecotourism plan to ensure greater knowledge and better preservation of all the species in this area. These are some of the animals that were registered. Freely consult all the data: http://doi.org/10.15472/7okxxe."

Another post from GBIF sent out call for application for 2018 FGVCx Fungi Classification Challenge that used the Danish Mycological Society, fungal records database on GBIF for training and validating images. This type of use case shows creative ways of encouraging open data use outside of academic research.

Random sample 2d – Wikipedia mentions (Occurrence datasets)

Wikipedia mentioned only five datasets. Each received an altmetric score of 1 on Wikipedia. One of the datasets (DOI: 10.3897/bdj.5.e11794) is linked to the Biodiversity Data Journal and the author of the Wikipedia article is the first author of that paper. A second dataset (DOI: 10.3897/zookeys.73.840) is linked to ZooKeys journal and referred to the ZooKeys article for species information. In the Wikipedia article on plant Rheum lhasaense, the dataset (DOI:10.15468/o3pvnh) was used to describe its distribution, the article on Sirgenstein Cave used the dataset (DOI: 10.1594/pangaea.64558) to describe how species was organized and the article on Colpomenia sinuosa used the dataset (DOI: 10.5519/0002965) simply to refer to a synonym. This shows that the use of datasets for Wikipedia articles are similar to its use for academic articles but rare.

Correlation tests

We performed correlations tests by year and type of datasets. Few datasets had any altmetric mentions until 2015, ranging between 4 to 17 for 2007-2014. In the following years the number of datasets with altmetric scores are as follows – 101 in 2015, 3,136 in 2016, 3,547 in 2017, 1,082 in 2018, and 1,460 in 2019. From the observations above, the number of Checklist datasets published rocketed in 2016 (Figure 1) and most of the altmetric mentions were about Checklist datasets (Table 4). This explains the rapid growth of altmetric mentions from 2015 to 2016. Due to the lower number of mentions in the previous years, correlation tests were conducted for the years 2016 to 2018; excluding 2019 due to low number of datasets (n=7) with any journal article citations to perform correlation tests.

The total number of datasets that received any altmetric mentions for different types of datasets are: 8,773 Checklist datasets, 457 Occurrence datasets, 40 Metadata-only datasets, and 128 Sampling-event datasets. Below are the results for each type of platform, in terms of years (Table 7) and dataset types (Table 8), excluding Metadata-only datasets. Results show no strong correlations between citations and altmetric mentions when compared for different years. Correlation tests for different type of datasets show moderate to strong correlation between number of Tweets and citations for Occurrence and Sampling-event datasets respectively.

Table 7. Correlations between citations and altmetric scores between 2016-19

	Blogs	*Twitter*	*Facebook*	*Wikipedia*
2016	Non-significant (n=419)	Weak (n=2,410, r = 0.28, p=0.000)	Non-significant (n=543)	Non-significant (n=653)
2017	Weak negative (n=844, r = -0.16, p=0.000)	Non-significant (n=2,610)	Non-significant (n=2,084)	Weak negative (n=900, r= -0.153, p=0.000)
2018	Weak negative (n=285, r = -0.143, p=0.016)	Weak negative (n=865, r = -0.128, p=0.000)	Weak (n=514, r =0.1, p=0.027)	Non-significant (n=222)

Table 8. Correlations between citations and altmetric scores for different types of datasets

	Blogs	*Twitter*	*Facebook*	*Wikipedia*
Checklist	Weak negative (n=2,099, r= -0.24, p=0.000)	Non-significant (n=6,700)	Non-significant (n=3,378)	Non-significant (n=1,907)
Occurrence	Moderate, non-significant (n=11, r= 0.3, p= 0.369)	Moderate (n=403, r= 0.459, p=0.000)	Weak (n=74, r=0.19, p=0.104)	Not enough data
Sampling-event	No mentions	Strong (n = 128, r= 0.629, p=0.000)	Moderate, non-significant (n=5, r= 0.592, p=0.293)	No mentions

Are we attributing citations in the correct way?

The current semi-automatic approach of GBIF assigns citations to all original datasets when any subsets downloaded from GBIF are cited by a research article. While theoretically this is the right approach, it does not consider the factor that after downloading the subsets researchers often curate the data to fit their purpose. For example, one of the data reuse articles in our sample explored the usefulness of Digital Accessible Knowledge (DAK) in biodiversity for terrestrial mammals distributed across the Iberian Peninsula and found that out of 796,283 retrieved records, 616,141 records were unfit for their use due to quality issues (Escribano, Galicia, & Ariño, 2019). Another study had to heavily filter 294,704,442 occurrences downloaded from GBIF, resulting in the deletion of more than half of the occurrences. Even though this is unavoidable when using open data for specific research purposes, assigning citations for the datasets that contained deleted data would result into erroneous citation counts. To explore this issue further, we looked into a dataset that contains a single record of marine mammals (Marine mammals of the Seaflower Biosphere Reserve, DOI: 10.15472/uzo3mq, GBIF with the following UUID: 6f2b8f8d-4e29-40b8-a022-e3a0e642c89e). The only observation the dataset contains is of Stenella attenuata (Gray, 1846) from Colombia. However, when we checked one of the four articles citing this dataset, we found the article – "The shrews (Cryptotis) of Colombia: What do we know about them?" DOI: 10.12933/therya-19-760. The article used the GBIF subset 10.15468/dl.hjv2ad that was derived by searching for "Colombia" and 5,552,450 occurrences were downloaded. However, shrews are not marine animals and live in forest and cultivated areas (https://en.wikipedia.org/wiki/Colombian_small-eared_shrew). Therefore, the observation in the original dataset was not used in the article and listing it as a citing article would be misleading.

Recommendation

To avoid the issue with attributing citations mentioned above, we propose that GBIF allows depositing the cleaned datasets used for research purposes separately and assign DOI to those subsets. After registering those datasets into GBIF, they can apply the same semi-automated approach to apply citations to the specific datasets from which occurrences or geo-references were used. Currently, the subsets downloaded from GBIF are assigned a generic "GBIF Occurrence Download" title, which is not informative in terms of understanding the content of that subset or differentiating between multiple subsets when citing those subsets. Our proposed change will have the following benefits: 1. Assign citations to the correct datasets only, 2. Inform other users about the dataset content by providing meaningful title, and 3. Allow other users to learn about various use cases of GBIF datasets when they explore the list of datasets linked to citing articles, which can lead to generation of newer ideas of research by identifying trends and gaps.

Limitations

The content analyses results presented here are from relatively small samples of citing research articles and altmetric sources. This is due to multiple reasons: 1. The list of citing articles and altmetric contents in our samples had to be manually curated by accessing individual datasets on GBIF and individual mentions on Altmetrics Explorer since bulk download is not currently possible, 2. Citing articles were collected from different journals with different access restrictions, which limits downloading all full texts at once, and 3. Due to different citation practices and reuse cases, manual content analysis was conducted to capture the information that would be needed to develop automated methods in the future. The process is time consuming, however. Easy access to the citing literature and contents of altmetric sources would open up the opportunity to develop a systematic approach using text mining methods that would reduce speed up or eliminate manual analyses.

Discussion

This study explores data citation and reuse practices 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. Types of data reuse cases are diverse and indicate that open biodiversity data supports creative research in the field of biodiversity and beyond. For example, by creating species distribution models with existing data, researchers can identify scopes of new research, determine any changes in biodiversity in a particular area, and compare their findings against a baseline model. This demonstrates the impact of open data and researchers, data managers, and policy makers should identify how this type of knowledge flow can be encouraged in other fields and for different data formats as well.

Giving proper attribution is important to recognize the efforts of the data creators and publishers. 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. This study identified that not all downloaded data are reused as data cleaning almost always takes place before using data for a specific research and recommends an alternative approach for GBIF to attribute citations to datasets to avoid erroneous mass citations.

This study also investigated whether altmetric scores are informative about the impacts of open biodiversity data. The correlation test results of citation counts and altmetric mentions, and content analysis of sample 1 show that Checklist datasets tend to receive high altmetric score due to their link to new findings or discovery. However, when researchers search for data on GBIF they probably find more instances of Occurrence datasets and less Checklist datasets. Therefore, citation counts and altmetric mentions are not correlated or inversely related in these cases. Occurrence datasets showed moderate correlations for Twitter and blog posts, and a weak correlation for Facebook posts. Even though most of these social media posts are from the data publishers and data creators, perhaps promotion leads to more reuse of such open datasets or popular datasets are promoted more frequently. However, tweets were less rich in content than Facebook and blog posts and did not provide any insight on data reuse cases. This is probably because of the previous character count limitation on Twitter and tweeters focusing more on promotion. Among the four platforms compared, blog posts can be informative for those who are keen to learn about usability and use cases of open data. Such blog posts should be encouraged more to advocate creative solutions using open research data and capture societal impact.

Declarations

Funding

This study was funded by the University of Wolverhampton.

Conflicts of interest/Competing interests

Not applicable

Availability of data and materials

All data created during this study are available on Figshare at https://doi.org/10.6084/m9.figshare.8181098.v1 (Khan & Thelwall, 2019a) and https://doi.org/10.6084/m9.figshare.11357693 (Khan & Thelwall, 2019b).

Code availability

Not applicable.

Acknowledgements

The study is an extended version of the paper "Data Citation and Reuse Practice in Biodiversity – Challenges of Adopting a Standard Citation Model", which was presented at the 2019 ISSI Conference in Rome, Italy.

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.

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting qualitative data reuse: A decade on. *Sage Open*, *7*(1), 2158244016685136.

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

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, *103*(3), 1123-1144.

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.

Escribano, N., Galicia, D., & Ariño, A. H. (2019). Completeness of Digital Accessible Knowledge (DAK) about terrestrial mammals in the Iberian Peninsula. *PloS one*, *14*(3), e0213542.

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. (2019a). Dataset supporting "Data Citation and Reuse Practice in Biodiversity". FigShare. Dataset. 10.6084/m9.figshare.8181098.v1

Khan, N., & Thelwall, M. (2019b). Dataset supporting "Measuring the Impact of Biodiversity Datasets: Data Reuse, Citations and Altmetrics". FigShare. Dataset. 10.6084/m9.figshare.11357693

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.

Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. *Bulletin of the American Society for Information Science and Technology*, *39*(6), 27-32.

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.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723-744.

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.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology*, *65*(5), 1018–1027.

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*.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PloS One*, *8*(7), e67332.