Journal of Applied Ecology

BRITISH ECOLOGICAL SOCIETY

Journal of Applied Ecology 2017, **54,** 612–617 doi: 10.1111/1365-2664.12767

COMMENTARY

Is citizen science an open science in the case of biodiversity observations?

Quentin Groom[1]*, Lauren Weatherdon[2] and Ilse R. Geijzendorffer[3]

[1]*Botanic Garden Meise, Nieuwelaan 38, 1860 Meise, Belgium;* [2]*UNEP World Conservation Monitoring Centre (UNEP-WCMC), 219 Huntingdon Road, Cambridge CB3 0DL, UK; and* [3]*Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), Aix-Marseille Université, UMR CNRS IRD Avignon Université, Technopôle Arbois-Méditerranée, Bât. Villemin – BP 80, F-13545 Aixen-Provence cedex 04, France*

Summary

1. There is a high demand for biodiversity observation data to inform conservation and environmental policy, and citizen scientists generate the vast majority of terrestrial biodiversity observations. As this work is voluntary, many people assume that these data are openly available for use in conservation and scientific research.

2. Here, the openness of biodiversity observation data that are contributed to the Global Biodiversity Information Facility is examined by the data provider. Contrary to what many people assume, data sets from volunteers are among the most restrictive in how they can be used.

3. *Policy implications*. The assumption that voluntary data collection leads to data sharing does not recognize the wishes and motivations of those who collect data, nor does it respect the crucial contributions of these data to long-term monitoring of biodiversity trends. To improve data openness, citizen scientists should be recognized in ways that correspond with their motivations. Furthermore, organizations that manage these data should make their data sharing policies open and explicit.

Key-words: biodiversity, citizen science, data licensing, data mobilization, data sharing, GBIF, open data, reproducibility, sustainability, volunteers

Introduction

Citizen science in biodiversity research covers a wide variety of volunteer activities, from the collection of casual observations through to conducting detailed species monitoring (Wiggins & Crowston 2011). The skills of citizen scientists range from general members of the public with little scientific experience through to expert amateurs and retired professionals. The EuMon project has calculated that volunteers outnumber professionals 18 to 1 in species monitoring in Europe (EuMon 2015). The voluntary aspect of the time invested by citizen scientists is generally interpreted as being motivated primarily by its contribution to society and that society should profit from this effort through openly accessible data. For example, in the European Union's *Digital Agenda for Science*, citizen science is listed as a subcategory of open science (European Commission 2015), implying that citizen science data are open, permissively licensed and available to all.

However, the motivations of citizen scientists are diverse and include a general interest in a specific species or question, involvement in a community with similar interests, recognition for personal achievements, learning new skills and contributing to environmental activism (Bell *et al*. 2008; Rotman *et al*. 2012; Tulloch *et al*. 2013). Given the diversity of their motivations, data sharing could have many potential advantages for citizen scientists, such as ensuring the persistence of their data, safeguarding their scientific legacy and increasing the visibility and impact of their observations through use in others' research. Likewise, citizen scientists can benefit from the openness of others' data for their own projects.

Access to citizen scientists' data is not only essential for science, but also for continual monitoring and environmental impact assessment. For example, continental and global policy instruments, such as the Convention on Biological Diversity, have a pressing need for biodiversity

*Correspondence author. E-mail: quentin.groom@plantentuinmeise.be

data to fulfil their reporting requirements (Geijzendorffer *et al*. 2015). Open access to biodiversity data promotes their use, encourages novel applications and supports reproducible science (Arzberger *et al*. 2004; Tenopir *et al*. 2011; Thessen & Patterson 2011). This is reflected in the increasing openness of governmental and taxpayer-funded environmental data (e.g. http://www.data.gov/; http://data.gov.uk/; http://data.gov.be/). Likewise, scientists are becoming more open with their data and publications (Piwowar 2011). Some scientific journals, such as the Journal of Applied Ecology and other British Ecological Society Journals mandate that data used within research articles, are deposited in public digital repositories. Funding agencies are also actively requesting open access to research data; for example, the European Framework Programme for Research and Innovation, Horizon 2020, is conducting a pilot on open access to research data.

In this paper, we examine the openness of biodiversity observation data in relation to the sources of these data to identify the relative openness of citizen science data.

Materials and methods

The Global Biodiversity Information Facility (GBIF) is a long-standing, influential global resource on biodiversity distribution information with a pan-taxonomic approach. It is funded by governments of participating countries to provide open information exchange on biodiversity. As such, it is a highly suitable data set for this analysis. Observation data from GBIF are only one of a diverse range of data types collected by citizen science projects. However, unlike many other data sets, the data available in GBIF are accessible with comparatively clear licensing (Table 1).

Table 1. An explanation of some widely used data sharing licences from Creative Commons and Open Data Commons

CC0 and ODC-PDDL	Under the Creative Commons zero dedication and the Open Data Commons Public Domain Dedication and Licence, the copyright holder releases the work to the public domain waiving their rights under copyright law. This does not change the conventions of scientific citation
CC-BY and ODC-By	Under the Creative Commons Attribution Licence and the Open Data Commons Attribution Licence users can use the copyright material; however, they wish, even for commercial purposes, as long as they provide attribution
CC-BY-NC:	Under the Creative Commons Attribution Non-commercial licence, users can use the copyright material for non-commercial purposes, as long as they provide attribution
CC-BY-SA:	Under the Creative Commons Attribution Share Alike licence, users can use the copyright material however they wish, even for commercial purposes, as long as they provide attribution to the originator and the derivatives are shared under the same licence

CC0, CC-BY and CC-BY-NC are the licence options recommended by GBIF. Strictly speaking, Open Data Commons licences are more suitable for data licensing, but Creative Commons are nevertheless widely used.

Data set metadata were extracted from GBIF using R (version 3.2.0) on 19 January 2016 using the 'RGBIF' package (version 0.9.0) (Chamberlain *et al*. 2015). Where the legal right to use the data was explicitly mentioned, it was represented by either a short rights statement in the metadata or a link to a longer licence document. The 'rights' statements or URL to a licence was extracted for all occurrences and survey data sets with one or more observations. A total of 12 458 data sets were extracted, but only 11% of these data sets included an explicit data-usage-rights statement in the data set metadata. The licensing information can be found in three places in the 'RGBIF' output: once in the data set metadata and twice in the occurrence record, wherein the licensing is noted in both the 'rights' and 'accessRights' fields. The 'rights' field is a deprecated term that preceded the 'accessRights' term, originating from the Darwin Core standard. Darwin Core is used in GBIF to define fields in the data base and as a data exchange format (Wieczorek *et al*. 2012). When a rights statement was missing from the data set metadata, the occurrence-level rights information was obtained from the first record of each data set. It is assumed that the rights within a data set are uniform and that the first record is representative of all records in that set. Rights statements for a further 0·25% of data sets were obtained this way.

Licensed data sets use standard licences such as a Creative Commons or an Open Data Licence, while the remainder use bespoke licence statements of various sorts. To simplify the interpretation of the different licences, the intention of the licence holder was interpreted and simplified into seven categories. Each of these categories was then given a data openness score from zero to three, from the least to the most open, respectively. Details of these categories are given in Table 2.

The sources of the data sets were classified into thirteen types based on the data set names and provider names (Table 3). These types were chosen by reviewing the word frequency in the data providers' titles. For example, common words in data provider titles included 'University', 'Museum', 'Institute', 'Research' and 'Herbarium'. If the data provider's

Table 2. The interpreted data usage rights from GBIF data sets, their data openness score and the number of data sets in each class

Interpreted intention of the licence	Data openness score	Number of data sets
Requires permission to use	0	72
Non-commercial usage, with attribution and share alike	1	38
With attribution and share alike	1	1
Non-commercial usage	1	16
Non-commercial usage, with attribution	1	322
With attribution	2	492
Public domain	3	546
Not specified or ambiguous		10 971

name could not be used to assign a type or was ambiguous, the data set description, domain name and organization's website were used to guide attribution. The majority of data sets are described in English; for all others, *Google Translate* was used to interpret the titles and descriptions. Finally, the names of the data sets were reviewed to ensure they had been classified correctly. For example, some citizen science and scientific society data sets are submitted to GBIF by data centres of various types and can be recognized from their data set name. If these could be identified, they were then reclassified. It is acknowledged that many institutions fall within multiple types – for example, some museums will also be research institutions and *vice versa* – but each organization was assumed to belong to their self-identified type and that the data set's description took precedence over the provider's name. It is also assumed that all data sets consist of one type of data. We used a non-parametric Mann–Whitney U-test to determine the significance of differences between data openness scores, which was considered a rigorous test on these categorical and bounded scores.

Results

Our assessment showed that citizen science data sets comprise 10% of data sets on GBIF, but account for 60% of all observations. The largest data set by far is from eBird (Cornell Lab of Ornithology 2015), which is a citizen science data set that contains over 200 million observations.

When comparing the data openness scores of GBIF data sets with the data source types (Fig. 1), the citizen science projects ranked low on the openness of their data, although the vast majority do not include a licence statement (mean data openness scores 1·67, $n = 33$), whereas institutions such as museums, educational institutions and research institutes ranked higher (mean data openness scores 2·13, $n = 335$; 2·04, $n = 324$ and 2·01, $n = 219$, respectively). Commercial organizations ranked the most open (mean data openness score 2·82, $n = 11$). Of course,

Table 3. An overview of the different types of data provider defined for this study, with the number of data sets they provided with and without a specific licence statement and the total number of observations and an example of a data set and data provider

Type	Description	With licence	Without licence	Records (millions)	Example
Commercial Organizations	Organizations whose primary focus is the generation of wealth	11	3	0·05	'Barrow Island Ants' from Chevron Australia
Governmental Institutions	State-funded organizations primarily providing services to and from the state	23	496	5·38	'Observaciones de Especies Silvestres' from Ministerio del Medio Ambiente de Chile
Foundation	Charitable NGO organizations	24	23	5·26	'Aves Bosque Seco Chicamocha' from Fundación Natura Colombia
Museums	Collections of preserved biodiversity specimens, mainly animal	334	726	72·30	'NBM Unionoids' from New Brunswick Museum
Societies	Associations of individuals, often containing a proportion of amateurs	39	52	5·58	'British Lichen Society - BLS Lichen Data base: Scotland' from UK National Biodiversity Network
Educational Institutions	Universities, colleges and schools	324	304	13·54	'Insect Occurrence Data from MIZA' from Universidad Central de Venezuela
Research Institutions	Organizations with a specific focus on biodiversity research	219	405	29·27	'Alterra (NL) - Entomofauna inventory in peat swamps' from Alterra, Wageningen UR
Botanical gardens or herbaria	Collections of living and/or preserved plant specimens	41	50	8·56	'Hortus Botanicus Sollerensis Herbarium' from Sóller Botanical Garden Foundation
Information Facilities or data centres	Facilities specifically created for data aggregation and curation	368	293	37·21	'2011 Breeding Bird Survey of Taiwan' from Taiwan Biodiversity Information Facility
Parks Authorities or Nature Reserves	Organizations responsible for the conservation and management of national parks and wildlife reserves	26	13	0·44	'BioGIS – Hamaarag' from Israel Nature and Parks Authority
Citizen science Projects	Projects specifically using volunteers to collect observation biodiversity data	33	1231	218·89	'Garden Bird Surveys' from Canberra Ornithologists Group
Data publishers	Organizations devoted to online publication and preservation of scientific data	7	6938	<0·001	'Planktic foraminifera counts at CTD station PS55/030' published by Pangaea

Provider types in bold are considered to contain observations largely from volunteers

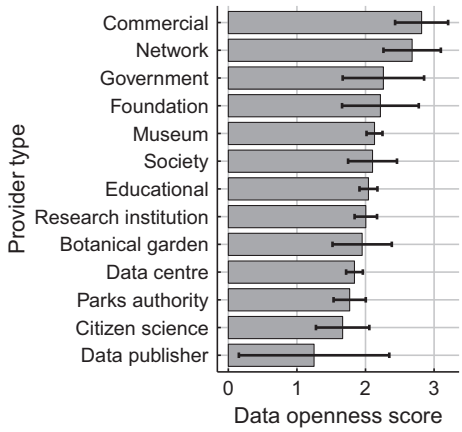

Fig. 1. The average data openness score of data sets on GBIF separated by the organization type of the data set provider. Only data sets with an explicit expression of data usage rights have been included. A data set with a score of zero is not usable without express permission of the owner; a data set with a score of one does not permit commercial use, requires acknowledgement and may have other restrictions; a score of two only requires acknowledgement; a score of 3 is given to data sets which are completely open. Error bars show the 95% confidence interval using a t distribution.

some data from citizen science projects were entered into GBIF using other organizations as intermediaries and are therefore hidden from our view. This is particularly true of data centres, which also score poorly (1·80, $n = 368$) and of societies (2·10, $n = 39$). A comparison of the scores of the three data set types together ('citizen science', 'societies' and 'data centres') demonstrated lower scores for these predominantly volunteer-provided data sets (mean 1·83, $n = 440$), than for all other data sets together (mean 2·10, n = 1048) (Mann–Whitney U-test, W = 194680, $P < 0·01$).

Discussion

The results confirm the important contribution of citizen scientists to biodiversity research. However, contrary to expectations, biodiversity data sets on GBIF derived from citizen science projects were often associated with more restrictive licences than other data types, and frequently restricted the data use by commercial organizations. Data centres, which distribute citizen science data as an intermediary, also receive low scores. Even though these data are collected voluntarily, the circumstances under which these data are managed and distributed seem to result in more restrictive data sharing. Scientific societies scored better on their accessibility. As these societies often have a largely voluntary membership, this raises questions on why their openness differs from citizen science projects. This category does, however, contribute fewer observations, forming only 2% of those contributed by citizen science and data centres. Surprisingly, commercial organizations scored highest on open access to biodiversity data. However, as the provisioning of biodiversity data is not the core business of these organizations, they form only a small fraction of providers and observations.

HETEROGENEOUS LICENSING

Within the GBIF data sets, there are a wide variety of licences. For example, 26% of licensed data sets restrict commercial usage, presumably to avoid undermining potential revenue sources for the data provider. However, they may be unaware that this stipulation also prevents not-for-profit research that they may assume is permitted (Hagedorn *et al.* 2011). This limitation is also true for the 88% of GBIF data sets that lack licensing information. Although it is perhaps assumed by some users that no licence information implies that the data can be used openly, this is not the case (Groom *et al.* 2015). Academic users can probably risk using these data, but the potential risk is much higher for commercial users. Policymakers should be aware that this makes it difficult to outsource the reporting of biodiversity targets that require these data. The heterogeneity of licensing is yet another obstacle that users need to resolve. The GBIF acknowledges these problems of data licensing and is transitioning to a simpler obligatory system that offers only three licensing options (Table 1) (Desmet & Aelterman 2013; GBIF 2015).

DATA USE BARRIERS

In addition to licensing issues, there are also an unknown number of organizations and individuals who hold data but do not share these openly. For instance, commercial companies may not want to share data that might be used against them (e.g. urban development projects can be delayed or blocked by the presence of protected species). Furthermore, we are aware from personal experience that many publicly available citizen science data sets are obfuscated by reducing their spatial or temporal resolution. For example, volunteers provide observations with a precise grid reference and date, but the data providers only supply summary data to GBIF, combining all observations for a year and grid cell into one record. In personal communications with several European GBIF nodes, they acknowledged that much of their country's data was obfuscated before these were provided to GBIF. An example is the National Biodiversity Network in the UK that provides most of the UK data available in GBIF, most of which comes from volunteer observers. At least 50% of these observations have their coordinates obfuscated at the request of the data providers (NBN Trust 2015).

The fact that some biodiversity observation data are either restricted, obfuscated or inaccessible may have several probable causes. For instance, data holders may use a conservation-based argument, such as protecting

locations of species vulnerable to persecution or exploitation. Data holders may also withhold data at the request of a landowner or because they did not receive legal permission to access the area where the observations were made. Among professional scientists, funding shortages and institutional support for data openness are important reasons for not sharing data (Tenopir *et al.* 2011; Fecher, Friesike & Hebing 2015).

THE MANDATE FOR SHARING

These reasons might also inhibit citizen scientists' data sharing, but there are additional reasons. For example, the mandate for decisions on sharing citizen science data is not held by the citizens themselves, but with intermediary organizations such as data distribution centres or citizen science organizations. Multiple reasons can cause these organizations to be unwilling to share data. For instance, data can be used as leverage to fund their activities, or to obtain acknowledgement of their contributions, particularly by being included as authors on publications. With the difficulty of finding sufficient funding, these are understandable reasons for withholding data, even though they considerably reduce the value of the data and can act contrary to the missions of these organizations. Indeed, commodification is a serious area of conflict between amateurs, their managing organizations and data aggregators (Ellis & Waterton 2005). Funding agencies should recognize that sustainability is fundamental to the reliable provisioning of good quality biological observations to long-term monitoring. For instance, some scientific societies have an enviable record in longevity. The Botanical Society of Britain and Ireland and the Audubon Society were established in 1856 and 1896, respectively, among other examples. The value of these sustainable models should not be underestimated.

MOVING FORWARD

There are some inspiring examples of advancements that can be applied. One crucial step forward is for organizations managing citizen science data to implement explicit data management policies with standard licences to prevent misconceptions regarding data sharing. A good example is Wikipedia, which has clear policies and no restrictions on commercial usage, but requires acknowledgement and 'share alike'. In the field of biodiversity observations, iNaturalist.org allows users to select from a range of Creative Commons licences, including releasing their data into the public domain.

Volunteers tend to have a particularly local conservation-based focus, whereas professionals may concentrate on international issues (Turnhout & Boonman-Berson 2011). This can lead to a lack of understanding between professional organizations and amateur societies, which have different goals and perspectives (Ellis & Waterton 2005). For example, volunteers are perhaps less likely to be motivated by citation in academic journals, but welcome acknowledgements that are visible to their local peer group. Some of the most productive volunteer observers collect biodiversity data for their own projects, such as producing regional floras or breeding bird atlases. Sharing their data becomes an interesting option for these citizen scientists if they can be assured that their own projects are not affected. Data users should support the activities of citizen scientists and societies through acknowledgement of their contributions in a way that matters for the citizen scientists.

This study illustrates that although citizen scientists contribute important biodiversity data to GBIF, the openness of their data ranks low among the different data providers studied. Several methods to stimulate data sharing are feasible. The assumption that voluntary data collection leads to data sharing does not do justice to those who collect data, nor does it acknowledge the contributions of these data to long-term monitoring of biodiversity trends. To improve data openness, citizen scientists should be encouraged in ways that correspond with their motivations. A first step would be for data distribution centres to delineate clearer licensing approaches and to thereby to enable citizen scientists to select appropriate levels of data accessibility.

Acknowledgements

The authors would like to thanks Graham French, Andre Heughebaert, Donald Hobern, Cees Hof, Rachel Stroud, Dagmar Triebel and Ben Wheeler from the GBIF nodes and the National Biodiversity Network, UK. We acknowledge the support of EU BON (http://www.eubon.eu), funded by the EU Framework Programme under grant agreement No. 308454.

Data accessibility

Data are available from Zenodo (Groom 2016).

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, **3**, 135–152.

Bell, S., Marzano, M., Cent, J., Kobierska, H., Podjed, D., Vandzinskaite, D. *et al.* (2008) What counts? Volunteers and their organisations in the recording and monitoring of biodiversity. *Biodiversity and Conservation*, **17**, 3443–3454.

Chamberlain, S., Ram, K., Barve, V. & Mcglinn, D. (2015) *Package 'rgbif': Interface to the Global 'Biodiversity' Information Facility 'API'*. http://cran.r-project.org/web/packages/rgbif/rgbif.pdf (accessed 25 January 2016).

Cornell Lab of Ornithology (2015) EOD - eBird Observation Dataset, 2015-11-05. http://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e (accessed 25 January 2016). doi: 10.15468/aomfnb.

Desmet, P. & Aelterman, B. (2013) Interpreting licenses of GBIF registered data. https://github.com/Datafable/gbif-data-licenses (accessed 25 January 2016).

Ellis, R. & Waterton, C. (2005) Caught between the cartographic and the ethnographic imagination: the whereabouts of amateurs, professionals, and nature in knowing biodiversity. *Environment and Planning D: Society and Space*, **23**, 673–693.

EuMon (2015) Professionals and volunteers involved in species monitoring http://eumon.ckff.si (accessed 31 January 2016).

European Commission (2015) Open Science. https://ec.europa.eu/digital-agenda/en/open-science (accessed 5 August 2016).
Fecher, B., Friesike, S. & Hebing, M. (2015) What drives academic data sharing? *PLoS ONE*, **10**, e0118053.
Geijzendorffer, I.R., Regan, E.C., Pereira, H.M., Brotons, L., Brummitt, N., Gavish, Y. *et al.* (2015) Bridging the gap between biodiversity data and policy reporting needs: an essential biodiversity variables perspective. *Journal of Applied Ecology*, doi: 10.1111/1365-2664.12417.
Global Biodiversity Information Facility (2015) Terms of use. http://www.gbif.org/terms/licences (accessed 31 January 2016).
Groom, Q. (2016) Data licences and organization type of contributors to the Global Biodiversity Information Facility as of 19 January 2016. *Zenodo*, doi:10.5281/zenodo.45363.
Groom, Q.J., Desmet, P., Vanderhoeven, S. & Adriaens, T. (2015) The importance of open data for invasive alien species research, policy and management. *Management of Biological Invasions*, **6**, 119–125.
Hagedorn, G., Mietchen, D., Morris, R., Agosti, D., Penev, L., Berendsohn, W. & Hobern, D. (2011) Creative Commons licenses and the non-commercial condition: Implications for the reuse of biodiversity information. *ZooKeys*, **150**, 127–149.
NBN Trust (2015) NBN Gateway Data Assessment. http://www.nbn.org.uk/nbn_wide/media/Documents/NBN%20Trust%20papers/TTE15-02-P12-NBN-Gateway-Data-Holding-Certificate-June-2015.pdf (accessed 31 January 2016).
Piwowar, H.A. (2011) Who Shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*, **6**, e18657.
Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D. & Jacobs, D. (2012) Dynamic changes in motivation in collaborative citizen-science projects. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (eds S. Poltrock & C. Simone), pp. 217–226. ACM, New York, NY, USA.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. & Frame, M. (2011) Data sharing by scientists: practices and perceptions. *PLoS ONE*, **6**, e21101.
Thessen, A.E. & Patterson, D.J. (2011) Data issues in the life sciences. *ZooKeys*, **150**, 15–51.
Tulloch, A.I., Possingham, H.P., Joseph, L.N., Szabo, J. & Martin, T.G. (2013) Realising the full potential of citizen science monitoring programs. *Biological Conservation*, **165**, 128–138.
Turnhout, E. & Boonman-Berson, S. (2011) Databases, scaling practices, and the globalization of biodiversity. *Ecology and Society*, **16**, 35.
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE*, **7**, e29715.
Wiggins, A. & Crowston, K. (2011) From conservation to crowdsourcing: a typology of citizen science. *System sciences (HICSS), 2011 44th Hawaii international conference*, pp. 1–10. IEEE.

Received 4 April 2016; accepted 3 August 2016
Handling Editor: Marc Cadotte